KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

Essays on Knowledge Access and Cumulative Innovation



Dissertation presented to obtain the degree of Doctor in Business Economics

by

Thomas Schaper

Number 764 2021

KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

Essays on Knowledge Access and Cumulative Innovation

Dissertation presented to obtain the degree of Doctor in Business Economics

by

Thomas Schaper

Number 764 2021

Daar de proefschriften in de reeks van de Faculteit Economie en Bedrijfswetenschappen het persoonlijk werk zijn van hun auteurs, zijn alleen deze laatsten daarvoor verantwoordelijk.

Since the dissertations in the series published by the Faculty of Economics and Business are the personal work of their authors, only the latter bear full responsibility.

Committee

Supervisors:

Prof. Dr. Samuel Arts (promotor)	KU Leuven
Prof. Dr. Reinhilde Veugelers (co-promotor)	KU Leuven

Doctoral committee:

Prof. Dr. Dirk Czarnitzki	KU Leuven
Prof. Dr. Hanna Hottenrott	Technical University of Munich
Prof. Dr. Stijn Kelchtermans	KU Leuven
Prof. Dr. Martin Watzinger	University of Muenster

Chair:

Prof. Dr. Vincent Molly	KU Leuven

Acknowledgements

To my PhD advisors, first and foremost. For their close guidance, invaluable expertise and endurance throughout this journey. To Sam Arts, for walking into my office in late spring of 2016, saying 'Hi, I heard you are interested in doing a PhD – want to grab a coffee?' and briefly after that, together with Reinhilde, deciding to hire me on a PhD track at KU Leuven as your first PhD student ever. For your meticulous comments and feedback, often available in near real time. For pushing me to be critical and always go the extra-mile, no matter whether it came to data work, writing papers or (successful) grant applications. And for your continuous advice and encouragement you gave me over the past 5 years. To Reinhilde Veugelers, for your support and for being a great mentor. For sharing with me your contagious curiosity for science, your rigorous scrutiny and your immense expertise. For always, and without exception, taking the time to have extensive meetings and discussions with me on my research, despite your absurdly busy agenda. Realizing how much you care about this was truly inspiring. I am only starting to reckon how much I may have learned from you by noticing that a lot of my sentences nowadays start with 'You know, Reinhilde always says...'. Thank you Sam, thank you Reinhilde – it was a challenge and a great honour being your student.

To my exceptional doctoral committee. For their constructive feedback that helped improving my papers greatly, and for being particularly accommodating in this process. To Dirk Czarnitzki, for inviting a random 'management guy' from Germany to come over and visit the renowned MSI department at KU Leuven years ago. For providing me, like many others of FEB's new doctoral cadets, with my first armament for econometric street-credibility (those were the days!). For many

great talks after hours. And for looking after me, when things weren't so great. You are a very good person, Dirk. To Hanna Hottenrott, for your excellent comments, your openness and academic positivism that you conveyed to me in the final stages of this PhD. Also for your praise and problem solving skills. I am looking much forward to be working together in the coming years. To Stijn Kelchtermans, for taking me to the SEI Consortium as my 'foster advisor'. For always being there for me with precious advise, profound suggestions and ideas – accompanied with your great sense of humor. To Martin Watzinger, for all the great input, and for being extremely approachable. And for sharing many valuable insights from your own work, which I cite so often. To Vincent Molly, for your flexibility and willingness to become chairman of my committee on a certainly very short notice.

To the Research Foundation – Flanders and to the KU Leuven, for investing so much money in my human capital.

To Mattia Nardotto, for agreeing to start a joint paper with me at a crucial moment of my trajectory, which became one of the chapters of this thesis. And for writing letters for me. Working with you taught me a lot about research and academic career development from a new angle. Thank you, Mattia. To René Belderbos, Bart Leten, Jo Seldeslachts, Bart Van Looy and Bruno Cassiman, for always having open ears and doors, and appreciated advice for me over the years. To Michael Stephan, for giving me my first job in academia (ever, actually) and for being a great boss.

To my dear friends I met at MSI. To Dennis J. Verhoeven, my brother from a different mother[1] and MSI's chief spillover-amplifier. For sharing thousands of ideas, an office, a (slightly too loudly) outspoken passion for patents, and calling for hours just to catch up when we were far apart. For making a great sales force. To Adrián 'MC' Kovács, a genuinely gracious mind and perhaps the sharpest knife around. For being a steady companion in countless and adventurous late nights at the office, incl. the Great Battle of the Bats of 2018, and remembering all about the 90s as vividly as I do. For helping me out on plenty of occasions. To Jeroen Mahieu, for being a notoriously altruistic, sociable and true-hearted amigo. And my much younger 'twin brother', making strangers greet me walking around the

[1] Although coincidence has it that our mothers were – suspiciously – born on the same day.

cafeteria. To Nuria Boot, for opening doors for me over the past years,[2] and for reliably reminding me about dinner time on a daily basis. For always telling me not to freak out - I wish I had a fraction of your admirable serenity. To Jesse Wursten, for happily optimizing my code, gladly validating my econometrics, and being an utterly agreeable and smart co-author. Thank you guys. Knowing that I could always count on you was of unwavering support in these years.

To my fabulous former office mates. To Michela Bergamini, Manuel Gigena, Federico De Michiel, Linde Colen, Daniela Silvestri and Francesco Biancalani. For many laughters, coffee breaks, chats about life and, of course, the intricacies of patent data. For filling in for me in many supervisions. For your warmth and joviality.

To all my other colleagues and friends from the (extended) MSI family. Especially to Maarten Rabijns, Cem Ermagan, Marcelina Grabowska„ Felix Bracht, Steven Vanhaverbeke, Céline Aerts, Kristof Van Criekingen, Wytse Joosten, Maikel Pellens, Nima Moshgbar, Leonard Treuren, Antonio Della Malva, Charro Moens-Seynaeve, Kieran Dobson, Mathias Boënne, Nazareno Braito, Florence Benoit and Jeroen Van den bosch. For countless great times and memories, both in and outside the office. To all from far and near occasionally gathering at MSI's debating society (and sometimes glee-club) in Standonckstraat, Leuven. To Isabelle Theys, for solving all problems always. To Hendrik Meder, for being a great matchmaker.

To everybody at FEB's Campus Carolus Antwerp. Especially to my research group colleague Jianan Hou, to Elien Voermans, Liliana Kowalczyk, Ingrid Van Melkebeke and Carlos Ocampo y Vilas, for being very kind and helpful people.

To my friends from Leuven outside of MSI. To Arina Gorbatyuk, Quynh Hoang and José Mendoza. For being great, exhilarating off-work company.

To my friends and former colleagues from times at Uni Marburg. To Nico Schmid, Roman Kersting, Tobias Engler, Steffi Lorenz, Christel Dehlinger, Christian Rötz, Kai Blume, Michael Schulz, Johannes Beyenbach, Marianne Hock, Peter Süß, Aylish Dietrich and Martin Schneider.

Some very personal words. To my partner Stella Pickert, the best human I have ever met. For your love, your joy, your patience, and your shelter. I can never thank you enough for this. To my mother, a strong woman with a heart of gold. For your unconditional love and for raising me to who I am today. Here's to you, Mama.

[2] Btw., I still have your card.

To my beloved grandmother, O. For being my dearest conversation partner for so many decades and teaching me most everything I know about life. And for sending me to university. To the best brother in the world, Stefan. For standing by my side and never deserting me since I could crawl. To my two fathers, who could not live to see this day. For kite flying and teaching me how to fix a bike. I miss you. I hope I would have made you proud.

To Monika and Chris Nelson, and to Stefan and Doris Delp. For caring for me for my entire life and being there for me as my family. To Milan Gulyás, for being a loyal friend. To Stella's parents, for welcoming me into their family with open arms and making it so easy for me to be their son-in-law.

Lastly, to the wonderful people of Leuven, for being so adorable and welcoming. Flanders will always feel like home to me.

Thank you all.

Thomas Schaper
Leuven, October 2021

For my family.
To Stella.

Contents

Acknowledgements i

Table of Contents vii

List of Figures xii

List of Tables xvii

General Introduction 1

Motivation 1
Contributions 5
Outline of the Dissertation 8

1 Not Like the Others: Frontier Scientists and High-Impact Technology 15

1.1 Introduction 15
1.2 Prior Literature & Theoretical Considerations 20
1.2.1 Science as an input for technology 20
1.2.2 The importance of scientific human capital for links between science and technology 22
1.2.3 Frontier authors and their use of science in inventions . . . 25
1.3 Research Setting & Data 27
1.3.1 Frontier author links in biomedical technology 27
1.3.2 Data sources 28
1.3.3 Measures 29
1.3.4 Sample & summary statistics 33

1.3.5 Sample selection considerations . . . 37
1.4 Econometric Analysis . . . 39
1.4.1 Frontier author patents and technological impact . . . 39
1.4.2 Frontier author patents and the use of frontier SNPRs . . . 47
1.5 Conclusion . . . 59
Appendices . . . 69
1.A Appendix A . . . 69
1.B Appendix B . . . 76
1.C Appendix C . . . 86
1.D Appendix D . . . 93
1.E Appendix E . . . 95

2 Online Repositories, Search Costs and Cumulative Innovation 99
2.1 Introduction . . . 99
2.2 Background . . . 105
2.2.1 External search and prior art search costs . . . 105
2.2.2 The International AIDS Patent Database project . . . 107
2.3 Data . . . 111
2.3.1 Information about AIDS DB patents . . . 111
2.3.2 Patent universe data . . . 112
2.3.3 Scientific publication universe data . . . 112
2.3.4 Final database . . . 113
2.4 Effects on Cumulative Patent Citations . . . 116
2.4.1 Empirical strategy . . . 116
2.4.2 Within AIDS DB Comparison: Effects on Patents without vs. with Front Page References to HIV/AIDS . . . 124
2.4.3 Timing of effects . . . 130
2.4.4 Mechanism: Differential effects for patents with intrinsically higher search costs . . . 132
2.4.5 Second order effects on citations to referenced scientific publications . . . 136
2.5 Changes in the Quality and Reach of Knowledge Spillovers . . . 138
2.5.1 Citation-level estimation model . . . 138

2.5.2 Quality of knowledge flows . . . 141
2.5.3 Geographic and social distance of spillovers . . . 143
2.6 Conclusion . . . 147
Appendices . . . 155
2.A Appendix tables . . . 155
2.B Appendix figures . . . 166
2.C Examples of AIDS DB patents with vs. without front-page reference to HIV/AIDS . . . 171
2.D External matched control group results . . . 177

3 The Impact of Broadband Internet on Invention: Evidence from the U.K. 185
3.1 Introduction . . . 185
3.2 The UK Broadband Roll-Out and the Search for Technical Prior Art 191
3.2.1 The UK broadband market . . . 191
3.2.2 The search for technical prior art online . . . 194
3.3 The Data . . . 196
3.3.1 Broadband market data. . . . 197
3.3.2 Data on patenting activity . . . 197
3.3.3 Census data . . . 200
3.3.4 Descriptive statistics . . . 200
3.4 Empirical Analysis & Results . . . 203
3.4.1 Identification strategy . . . 203
3.4.2 Inventive activity by LEs . . . 208
3.4.3 Historical placebo and concerns about instrument exogeneity 212
3.4.4 Mechanisms . . . 215
3.5 Conclusion . . . 220
Appendices . . . 227
3.A Appendix tables . . . 227
3.B Appendix figures . . . 241

General Conclusion 243
Main Findings . . . 244

Implications 247
Limitations and Avenues for Future Research 250

List of Figures

1.1 Author Selection into Industry Patenting 38
1.2 Patent Impact and Frontier Author Incidence 45
1.A1 Author Selection into Industry Patenting - within Authors Comparison 75
1.B1 Time Trends in Author Patents Impact Premium 84
1.B2 Time Trends in Author Patents Impact Premium - Hits 85

2.1 AIDS DB access page, fall 1996 108
2.2 Geographic Dispersion of AIDS DB Inventors 114
2.3 Frequencies of AIDS DB Patents over Technology Fields and Geographic Origin 115
2.4 Group Means *within AIDS DB* Comparison Yearly Patent Citations 121
2.5 Yearly Effect for Patents without HIV/AIDS Front Page Reference 131
2.B1 AIDSDB Link on USPTO Website, Fall 1996 166
2.B2 AIDSDB Access Page II, Fall 1996 166
2.B3 AIDSDB Access Page III, Fall 1996 167
2.B4 AIDSDB Access Page IV, Patent View, Fall 1996 167
2.B5 Spike in Daily U.S. Patent Filings prior to TRIPS Reforms Enactment on June 8th 1995 168
2.B6 Group Means *within AIDS DB* Comparison Yearly Patent Citations, Matched on Pre-Trends 169
2.B7 Yearly Effect for Patents without HIV/AIDS Front Page Reference, Matched on Pre-Trends 170
2.C1 Example 1- Lippa et al. vs. Eckstein et al. (1989) 172
2.C2 Example 2- Rubinstein vs. Al-Sioufi (1990) 174

2.C3 Example 3- Osther et al. vs. Co et al. (1996) 175
2.D1 Group Means Comparison Yearly Patent Citations 179
2.D2 Yearly Effect on Patent Citations 183

3.1 Share of LEs with Broadband Access by Quarter 193
3.2 Geolocation of Inventors across Great Britain 1975–2010 199
3.3 Cumulative of *within NUTS3* ADSL Activation Lag across Great Britain 204
3.4 Areal Variation in Rainfall and ADSL Activation 205
3.5 Mismatch between MSOA and LE Catchment Areas 206
3.B1 NUTS3 Regions (Counties) of Great Britain (N=167) 241
3.B2 Census 2001 MSOAs Across Great Britain (N=3,374) 242

List of Tables

1.1 Summary Statistics 34
1.2 Frontier Author Patents and Technological Impact 42
1.3 Frontier SNPR Patents and Technological Impact 49
1.4 Frontier SNPR Timing and Technological Impact 50
1.5 Frontier Author Patents and Frontier SNPR Probability 52
1.6 Frontier Author Patents, SNPRs and Technological Impact 55
1.7 Frontier Author Patents and Technological Impact - Closeness & Timing to Frontier SNPRs 57
1.A1 Summary Statistics 72
1.A2 Author-Level Summary Statistics 73
1.A3 Author Selection into Industry Patenting 74
1.B1 Frontier Author Patents and Technological Impact - Robustness . . 76
1.B2 Frontier Author Patents and Technological Impact - Log(cits) OLS 77
1.B3 Frontier Author Patents and Technological Impact - Country-level Residuals 78
1.B4 Frontier Author Patents and Technological Impact - Top5% Paper Authors 79
1.B5 Frontier Author Patents and Technological Impact - Alternative Frontier Timing 80
1.B6 Frontier Author Patents and Technological Impact - Splits by First Authors 81
1.B7 Frontier Author Patents and Technological Impact - Bottom vs. Top-Impact 82
1.B8 Frontier Author Patents and Technological Impact - Field Splits . . 83
1.C1 Frontier SNPR Patents and Technological Impact 89

1.C2 Frontier SNPR Patents and Technological Impact – Robustness . . 90
1.C3 Frontier SNPR Patents and Technological Impact – Top5% Cited Papers 91
1.C4 Frontier SNPR Patents and Technological Impact – Timing of SNPRs 92
1.D1 Frontier Author Patents and Frontier SNPR Probability – Robustness 93
1.D2 Frontier Author Patents and Frontier SNPR Probability – Top5% Cited Papers 94
1.E1 Frontier Author Patents, SNPRs and Technological Impact – Robustness 95
1.E2 Frontier Author Patents, SNPRs and Technological Impact – Robustness 96
1.E3 Frontier Author Patents, SNPRs and Technological Impact – Top5% Cited Papers 97
1.E4 Frontier Author Patents, SNPRs and Technological Impact – Ex-ante SNPRs 98

2.1 Summary Statistics, Patents with vs. without Front Page Reference to HIV/AIDS 120
2.2 Effect for Patents without HIV/AIDS Front Page Reference, within AIDS DB 126
2.3 Effect for Patents without HIV/AIDS Front Page Reference, excluding Top-Citing Examiners 128
2.4 Differential Effects for Patents with Intrinsically Higher Search Costs, within AIDS DB 133
2.5 Second Order Effect on SNPRs in Patents without HIV/AIDS Front Page Reference 137
2.6 Effects on Quality of Generated Spillovers 142
2.7 Effects on the Reach of Generated Spillovers 145
2.A1 Responsiveness to AIDS DB Deposit of Citation Rates from Patents within Individual Examiners 155
2.A2 Top-20 HIV/AIDS Inventors vs. Authors, by the End of 1996 . . . 156
2.A3 Effect for Patents without HIV/AIDS Front Page Reference, within AIDS DB, PPML Estimates 157

2.A4 Effect for Patents without HIV/AIDS Front Page Reference, within AIDS DB, Patents Observed in all Sample Years 158
2.A5 Effect for Patents without HIV/AIDS Front Page Reference, within AIDS DB, Matched on Pre-Trends 159
2.A6 Effect for Patents without HIV/AIDS Front Page Reference, within AIDS DB, Weighted Citations 160
2.A7 Differential Effects for Patents Based on Recency of Grant, within AIDS DB 161
2.A8 Summary Statistics, Quality and Reach of Spillovers Pre-AIDS DB 162
2.A9 Effects on the Geographic Reach of Generated Spillovers 163
2.A10 Effects on the Social Distance (SD) Reach of Generated Spillovers, Author-Inventor Graph 164
2.A11 Effects on the Social Distance (SD) Reach of Generated Spillovers, Inventor Graph 165
2.D1 Summary Statistics, until Year of AIDS DB Deposit 178
2.D2 Main Effect on Patent Citations, Matched Sample 181

3.1 Patent level summary statistics 201
3.2 Node (LE) level yearly summary statistics 202
3.3 Yearly Patents per Node (Log), Rainfall Sample 209
3.4 Yearly Patents per Node-MSOA (Log), Demographics Sample . . . 210
3.5 Yearly Citation-Weighted Patents per Node (Log). Rainfall and Demographics Samples 211
3.6 Yearly *Public* Patents per Node (Log). Rainfall and Demographics Samples 212
3.7 Historical Placebo Test for Yearly Patents per Node (Log). Rainfall and Demographics Samples 213
3.8 Probability of Citing a Same Country Patent 217
3.9 Citation Lag in Years, Foreign Cited Patents Only 218
3.10 Probability of *Examiner* Referencing a Same Country Patent 218
3.11 Citation Lag of *Examiner Added Citations* in Years, Foreign Cited Patents Only 219

3.A1 Yearly Patents per Node (Log), Rainfall Sample – Pre-Broadband Active LEs . . . 227
3.A2 Yearly Patents per Node (Log), Rainfall Sample – Pre-Broadband Established Firms . . . 228
3.A3 Yearly Patents per Node (Log), Rainfall Sample – Pre-Broadband Active LEs . . . 228
3.A4 Yearly Citation Weighted Patents per Node (Log), Rainfall Sample – Pre-Broadband established firms . . . 229
3.A5 Citation Lag in Years, Domestic Cited Patents Only . . . 229
3.A6 Applicant-Added Citations per Patent (Log) . . . 230
3.A7 Probability of Citing a Same Country Patent – Pre-Broadband Established Firms . . . 230
3.A8 Probability of Examiner Adding a Same Country Patent Reference – Pre-Broadband Established Firms . . . 230
3.A9 Firm Size Splits (above Median) for Same Country Citations . . . 231
3.A10 Probability of Reliance on Science [01] . . . 231
3.A11 Probability of Reliance on Science [01] – Pre-Broadbrand Established Firms . . . 231
3.A12 Number of Inventors (Log) per Patent . . . 232
3.A13 Number of Inventors (Log) per Patent – Pre-Broadband Established Firms . . . 232
3.A14 Probability of New Inventor Entry [01] . . . 233
3.A15 Probability of New Inventor Entry [01] – Pre-Broadband Established Firms . . . 233
3.A16 Firm Size Splits for New Inventor Entry [01] . . . 233
3.A17 Probability of Geographic Co-Location of Inventor Teams (Same Country) . . . 234
3.A18 Probability of Geographic Co-Location of Inventor Teams (Same Country) – Pre-Broadband Established Firms . . . 234
3.A19 Probability of New Inventor Collaboration . . . 235
3.A20 Probability of New Inventor Collaboration – Pre-Broadband Established Firms . . . 235
3.A21 Probability of Applicant Citing a US Patent Document . . . 236

3.A22 Probability of Applicant Citing a *Foreign* US Patent Document . . 236
3.A23 Probability of *Examiner Added* Citation to US Patent Document . . 236
3.A24 Probability of *Examiner Added* Citation to *Foreign* US Patent Document 237
3.A25 Probability of Applicant Citing a US Patent Document – Pre-Broadband Established Firms Only 237
3.A26 Probability of Applicant Citing a *Foreign* US Patent Document – Pre-Broadband Established Firms Only 238
3.A27 Probability of Applicant Citing an EP Patent Document 238
3.A28 Probability of Applicant Citing a *Foreign* EP Patent Document . . . 238
3.A29 Probability of *Examiner Added* Citation to EP Patent Document . . 239
3.A30 Probability of *Examiner Added* Citation to *Foreign* EP Patent Document 239
3.A31 Probability of Applicant Citing an EP Patent Document – Pre-Broadband Established Firms Only 240
3.A32 Probability of Applicant Citing a *Foreign* EP Patent Document – Pre-Broadband Established Firms Only 240

General Introduction

Motivation

Access to existing knowledge is essential for technical change and societal progress. In the production of knowledge, prior advances are the foundation of new ideas fostering welfare and economic growth. This notion of *cumulativeness* is most popularized by the allegorical expression of [being able to see further by] 'standing on the shoulders of giants', attributed originally to Isaac Newton in a letter to fellow scientist Robert Hooke in 1675. In modern economists' thought, the stock of knowledge constitutes a public good in societies, and its accessibility is central to macroeconomic theory models of endogenous growth, such as Romer (1986), Romer (1990), Aghion and Howitt (1990), or Weitzman (1998), but can be found back already in the works of Marx (1867), Schumpeter (1942), or Machlup (1962). The importance of access to knowledge is historically documented by the case of Gutenberg's invention of the printing press in 1436, which led to an unprecedented diffusion of the human knowledge stock due to cost efficient codified replicability with major and long-lasting influence on economic prosperity in Europe, as shown by Dittmar (2011).

Underlying the reasoning of macrotheorists for the relevance of access to existing knowledge is grounded on its elemental significance for innovation, as a driving engine of growth (Jones, 1995; Mokyr, 2005). The list of historical examples of important scientific discoveries, technological inventions and entirely new businesses prominently relying on significant knowledge inputs from prior art is very long: Einstein's theory of general relativity, Edison's light bulb, DNA amplification tech-

niques, or Google's page rank algorithm. All of these led to innovation linked to increases in aggregate productivity and had major impacts on consumers and firms in product markets. A key objective of both scientific publication as well as the patent system is, indeed, to provide access to the public good through the disclosure and dissemination of knowledge in order to generate spillovers and, thereby, increasing social returns to research and development in the economy.

The value of knowledge for innovation, however, is not constant. It differs strongly across types of knowledge inputs, depending on their usefulness, and is sensitive to the timing of access. Based on a natural experiment provided by the disruption of scholarly communication during World War I, Iaria, Schwarz and Waldinger (2018) have shown compelling causal evidence for the importance particularly of access to frontier knowledge for the productivity of scientists and the relevance of their work for new technologies. Accordingly, a key challenge for both R&D policy and private firms remains to promote and acquire effective access to new insights at the frontier of knowledge that can be translated into on-going scientific inquiries and the development of new technical solutions.

However, here public policy objectives of maximizing social returns from the non-rival nature of knowledge collide with private actors' incentives of excludability and towards internalization of knowledge spillovers. The industrial organization literature traditionally considers such spillovers as free-floating externalities that can be captured by industry rivals, creating negative incentives to conduct R&D, and inducing firms to form alliances and cooperative networks to prevent such spillouts (e.g., Katz, 1986; d'Aspremont and Jacquemin, 1988). Taking the opposite view, the idea that knowledge spillovers can be freely captured is abandoned in the absorptive capacity literature (starting with Cohen and Levinthal, 1989), further initiating a scholarly debate in the economics of innovation evolving around effective mechanisms of external knowledge acquisition (e.g. Cassiman and Veugelers, 2006), finding strong resonance also in the management discipline. These debates are further supported and motivated by an ample body of empirical work showing strong clustering of knowledge spillovers along geographic (e.g., Jaffe, Trajtenberg and Henderson, 1993), political (e.g., Singh and Marx, 2013), public/ private domain (e.g., Jaffe, 1989; Belenzon and Schankerman, 2013) and collaborative network (e.g., Singh, 2005) boundaries.

Beyond the industrial organization and corporate strategy perspective, the question of knowledge access and innovation is a particularly relevant one also for the effective design and implementation of access-enhancing institutions. A recent and growing body of work has been dedicated to the empirical investigation of this question, especially focusing on efficiency and the elasticity of access costs to prior art on innovation. Furman and Stern (2011) show an amplification of cumulative impact of individual scientific discoveries following their inclusion into specialized biological resource centres. Similarly, Thompson and Hanley (2018) find increases in citation rates to scientific articles referenced in Wikipedia. Evaluating variation in physical access costs to codified knowledge, Furman, Nagler and Watzinger (2018) and Berkes and Nencka (2020) find significant increases in inventive activity in surrounding areas following the establishment of patent libraries and, respectively, open public libraries in the U.S.. A closely related stream of literature has provided evidence for a significant elasticity of access costs to knowledge inputs following variation in property rights, evaluating the effects of open access mandates (Bryan and Ozcan, 2016) and copyrights (Biasi and Moser, 2018) on the diffusion of scientific knowledge in science and its subsequent use in inventions, and more specifically the costs of patent IP protection on cumulative innovation (Williams, 2013; Galasso and Schankerman, 2015; Murray et al., 2016). Several recent studies have specifically investigated the importance of timing of disclosure of relevant knowledge following variation provided by the American Inventor Protection Act (Hegde, Herkenhoff and Zhu, 2020; Baruffaldi and Simeth, 2020; Lück et al., 2020), showing significant positive effects on rate of invention and the de-duplication of efforts.

Lastly, an important role in the context of knowledge access is often ascribed in the popular debate to the advent of modern information technology, in particular the internet. While scholars in other fields of the economic discipline – such as inequality, education or development economics – have investigated questions on the impact of internet access on a multiplicity of economic outcomes, the innovation literature has remained comparatively more silent in this regard. Studying the effects of BitNet adoption, a pre-internet information system connecting academic institutions, on university scientists, Agrawal and Goldfarb (2008) and Ding et al. (2010) find increases in research collaboration and access to more distant, in par-

ticular top-tier knowledge. Forman and van Zeebroeck (2012) and Forman and van Zeebroeck (2019) show similar intra-firm effects in technology development following firm-level ICT adoption. Based on a survey of German firms, Bertschek, Cerquera and Klein (2013) find that the arrival of broadband internet was associated with a higher number of new products and processes introduced by those firms gaining access.

The foregoing outline of existing scholarly work in this area highlights knowledge access as a multifaceted complex, whose determinants, effective modes and outcomes on cumulative innovation are not yet fully explicated. In particular in the modern information age, where knowledge appears as omni-available good which is replicable at no costs, the external validity of historical case studies to inform nowadays decision makers may be reduced. On the other hand, this implies that our understanding of access privilege to knowledge inputs in innovation, as well as retrieval costs and, in particular, the role of attention of researchers to distant knowledge, new insights and opportunities may need to be revised and more closely examined.

The aim of this dissertation is to extend prior work within these streams of research and contribute to our understanding of the relationship between knowledge access and cumulative innovation, in particular with regards to the role of:

1. *Human capital* and individual scientific capability for effective access to the frontier of knowledge in industrial technology;

2. *Search costs*, and how they can be effectively lowed in order to attract the attention of researchers in topic-specific inventive search;

3. *Information technology*, in particular broadband internet, and its impact on inventive activity through lowering access costs to more distant knowledge inputs.

The following section outlines the specific contributions to the literature made by the individual chapters of the dissertation.

Contributions

This cumulative dissertation consists of three, non-sequential chapters, each of which entails an original, empirical essay addressing a specific research question in the area of applied economics on knowledge access and innovation. Theoretically, all three chapters build on theories and conceptual underpinnings from the economics of science and innovation, and to some extent strategy, industrial organization, public economics and information systems. Methodologically, all three chapters are based on large scale econometric analyses and are rooted empirically in the universes of scientific publications and patent records. In some cases, these are complemented by additional data (areal census, meteorological, and proprietary ICT data) as well as web-scraping and disambiguation (probabilitic matching) techniques.

Chapter 1 is embedded in a vast literature in the economics of innovation investigating how fundamental science connects and contributes value to technological progress (in the tradition of Nelson, 1962; Griliches, 1986; Mansfield, 1995; Cohen, Nelson and Walsh, 2002; Fleming and Sorenson, 2004). Recent work by Ahmadpoor and Jones (2017) has provided evidence for the extent and the ramifications of this connect. We contribute to this debate by shedding new light into the mechanism through which science is effectively accessed in technology, especially with regards to hit inventions pushing technological progress and entering the far-right tail of the value distribution. Thereby, we focus in particular on the role of inventors simultaneously involved in the production of frontier knowledge, i.e. recent breakthroughs in science. The role of scientists-inventors has been advanced as critical for enabling science-technology linkages by prior literature (e.g., Cockburn and Henderson, 1998; Gittelman and Kogut, 2003; Breschi and Catalini, 2010), in particular that of collaboration with pivotal scientists in their fields (e.g., Zucker, Darby and Armstrong, 2002). However, prior research has not been able to identify *which types* of scientists-inventors create value in technology through *which mechanisms*. We argue that inventors involved in recent top-science are able to generate a technological impact premium through their better access and use of frontier knowledge in their inventions. Our findings provide consistent evidence for this pattern in author-inventor links and tracing scientific references in biomedical

patents. Our main contribution emphasizes the importance of the frontier nature of scientific involvement in enabling the relation between early access to frontier scientific advance and inventive performance, in particular with regards to the generation of technological breakthrough. Our paper, further, is the first of its kind to provide a comprehensive empirical link between all inventors and authors of scientific publications in the universe of biomedical research (and their outcomes in both science and technology), building on recent advances in library and data sciences, allowing to overcome the significant limitation of relatively small samples utilized by prior quantitative studies in the literature.

Chapter 2 addresses the question of effective design of access-enhancing institutions in facilitating the dissemination of relevant prior art and the realization of knowledge spillovers across dispersed communities of researchers. This chapter, thereby, aligns into a recent body of work investigating the importance of such institutions for cumulative innovation, following Furman and Stern (2011), Furman, Nagler and Watzinger (2018), Thompson and Hanley (2018), and Berkes and Nencka (2020). I contribute to this stream of literature by informing the debate about how topic-specific repositories can effectively facilitate access through a reduction in inventive search costs, allowing to draw the attention of researchers to distant and otherwise more hidden advances on a targeted subject matter in line with a specific policy objective. For this, I exploit the historical case of the first, disease-specific online repository of electronic patent documents targeting inventions with applications to HIV/AIDS, launched in the mid-1990s, and I empirically evaluate differential changes in cumulative citation rates following database deposit for patents whose link to the disease complex was previously non-obvious to be detected in standard bibliographic search by inventors. In line with predictions of a higher reduction in search costs, I find strong support for the effectiveness of the online repository as access-enhancing institution, evidenced by higher marginal citation impact of these more hidden patents, originating more often from more distant geographical areas and previously disconnected researcher communities and benefiting in particular knowledge access of inventor in private firms, embedded in more clustered institutional contexts. The findings of this chapter contribute to the literature in particular with regards to the effectiveness of topic-targeted databases on lowering search costs, which is a question of high currentness, and allow to in-

form how in the presence of information overflow such repositories may be useful to direct the scope and search of inventions towards socially desirable outcomes. Furthermore, the chapter contributes by developing an empirical methodology to assess variation in search costs, which can be replicated in a straightforward way in other contexts and subsequent studies.

Chapter 3 departs from the empirical demarcation of biomedical research, in which Chapters 1 and 2 are rooted, and moves the scope of investigation of the dissertation to the evaluation of role of modern information technology for knowledge access and its repercussions on areal inventive activity on the comprehensive scale of a national economy. As such, this chapter is embedded in a broader stream of research studying the effects of ICT on productivity (e.g., Czernich et al., 2011; Dittmar, 2011), and in particular on innovation(e.g. Agrawal and Goldfarb, 2008; Ding et al., 2010; Forman and van Zeebroeck, 2012, 2019). Specifically, we evaluate the impact of the arrival of ADSL broadband technology on invention in the U.K. in the early 2000s, and especially its effect on access in knowledge sourcing, by exploiting geographical variation in activation status due to delays in market entry of internet service providers across narrow catchment areas of local exchange points within regions. We contribute to the literature by showing a large and significant real effect of broadband access on patenting activity in activated areas, which we are further able to link to enhanced access to more geographically distant knowledge, evidenced by changes in the composition and recency of prior art references in newly filed patents. To the best of our knowledge, this paper provides the first, national economy-wide empirical evidence for the impact of broadband internet on cumulative innovation, which may be informative to current debates on broadband expansion ongoing in several developed and developing countries. We further contribute methodologically by developing new approaches for causal identification of the effect of broadband roll-outs using instrumental variables techniques based on areal mismatch between local exchange catchment areas and Census-based demographic zones, expanding further the set of instruments developed in the literature (e.g, by Gavazza, Nardotto and Valletti, 2019).

The next section provides an overview and short summary of each chapter.

Outline of the Dissertation

Chapter 1 – *Not Like the Others: Frontier Scientists and High-Impact Technology*

A vast body of literature in the economics of innovation recognizes the role of new scientific ideas to private sector innovation. Empirical studies over the past decades have amply documented a large and growing share of technology to be directly drawing on inputs from science. However, economists have also acknowledged that the links between scientific and technological advance may not be obvious, direct or immediate. Taking into account the highly skewed nature of both scientific progress and technology performance, it becomes of particular interest to know how recent scientific breakthrough - frontier knowledge - connects to high-impact technology, and through which channels it can be effectively accessed.

Relying on private sector biomedical patents from 1980-2009 and research articles indexed in PubMed, we study the role of scientists active at the knowledge frontier as brokers between new scientific findings and technological applications. We find that technological inventions developed by scientific authors - especially by those scientists who were recently involved in top-scientific research (frontier authors) - are more impactful and more likely to be a hit patent. This effect is in part driven by better access to frontier knowledge, especially knowledge originating in their close proximity. These findings foster the pivotal role of frontier researchers for inventive progress.

JEL: O30, O31, O34, O36.

Keywords: Industry-science links, inventor-authors, frontier science, (top) technological impact

Chapter 2 – *Online Repositories, Search Costs and Cumulative Innovation*

Efficient access to existing knowledge is essential to technical advance, given that scientific production is shaped by sequential and complementary efforts. Especially, *new* ideas often require visibility and the free flow of information and simultaneous

experimentation in order to prevail. Recent empirical work in the economics of science and innovation has started to evaluate the importance of access-enhancing institutions for cumulative innovation, yet little is known about the mechanisms and effective design of such institutions in order to shape intertemporal and spatial knowledge spillovers.

In this chapter, I investigate the cumulative technological impact of the CNIDR AIDS Database, the first, disease-targeted, online repository of electronic patent documents, launched in 1994. Tracing references from subsequent patents, I find that the marginal impact of the repository was largest (+30%) among patents for which the established disease-link was previously non-obvious to detect through standard bibliographic search, in line with predictions of stronger reduction of search costs. Further findings suggest that increased visibility and attention to more 'hidden' prior art particularly benefited private sector HIV researchers, and was reflected in enhanced diffusion of technological knowledge across scientific community and geographic boundaries.

JEL: I23, O31, O32, O33.

Keywords: Knowledge diffusion, search costs, information systems, patents

Chapter 3 – *The Impact of Broadband Internet on Invention: Evidence from the U.K.*

Prior knowledge is a crucial input for the creation of new technologies. This cumulativeness constitutes a fundamental underpinning of economic progress. Technical knowledge embedded in patent documents represents the most important source of codified information for inventors, next to scientific publications and tacit knowledge in-flows sourced from interpersonal networks. While the patent system provides private incentives to incur the cost of innovation, the social value of granting patents is contingent by the extent to which the publication of patent documents triggers knowledge spillovers. Information technologies provide superior means of storing and disseminating such codified information. Yet there is only limited evidence on the empirical elasticity of ICT induced reduction in access costs to information on cumulative innovation.

In this chapter, we investigate the impact of the arrival of the first generation

of broadband internet (ADSL) on knowledge sourcing and resulting inventive activity. Exploiting geographical variation in ADSL activation in the U.K. in the early 2000s, we find that broadband access increased the number of years filed patents significantly by between 10% to 55%, which was accompanied by a real effect of increased cumulative impact of new patents. Further evidence reveals that this increase was arguably driven by better access to geographically distant knowledge, and particularly more recent foreign knowledge, as documented by an increase in the share of foreign patent citations, and a 2-3 year decline in the average age of these foreign cited patents.

JEL: O31, O32, O33, O52.

Keywords: Cumulative innovation, information technology, patents

References

Aghion, Philippe, and Peter Howitt. 1990. "A model of growth through creative destruction."

Agrawal, Ajay, and Avi Goldfarb. 2008. "Restructuring research: Communication costs and the democratization of university innovation." *American Economic Review*, 98(4): 1578–90.

Ahmadpoor, Mohammad, and Benjamin F Jones. 2017. "The dual frontier: Patented inventions and prior scientific advance." *Science*, 357(6351): 583–587.

Baruffaldi, Stefano H, and Markus Simeth. 2020. "Patents and knowledge diffusion: The effect of early disclosure." *Research Policy*, 49(4): 103927.

Belenzon, Sharon, and Mark Schankerman. 2013. "Spreading the word: Geography, policy, and knowledge spillovers." *Review of Economics and Statistics*, 95(3): 884–903.

Berkes, Enrico, and Peter Nencka. 2020. "Knowledge access: The effects of Carnegie libraries on innovation."

Bertschek, Irene, Daniel Cerquera, and Gordon J Klein. 2013. "More bits–more bucks? Measuring the impact of broadband internet on firm performance." *Information Economics and Policy*, 25(3): 190–203.

Biasi, Barbara, and Petra Moser. 2018. "Effects of Copyrights on Science - Evidence from the US Book Republication Program." National Bureau of Economic Research Working Paper 24255.

Breschi, Stefano, and Christian Catalini. 2010. "Tracing the links between science and technology: An exploratory analysis of scientists' and inventors' networks." *Research Policy*, 39(1): 14–26.

Bryan, Kevin A, and Yasin Ozcan. 2016. "The impact of open access mandates on invention." *forthcoming Review of Economics and Statistics*.

Cassiman, Bruno, and Reinhilde Veugelers. 2006. "In search of complementarity in innovation strategy: Internal R&D and external knowledge acquisition." *Management science*, 52(1): 68–82.

Cockburn, Iain M, and Rebecca M Henderson. 1998. "Absorptive capacity, coauthoring behavior, and the organization of research in drug discovery." *The journal of industrial economics*, 46(2): 157–182.

Cohen, Wesley M, and Daniel A Levinthal. 1989. "Innovation and learning: the two faces of R & D." *The economic journal*, 99(397): 569–596.

Cohen, Wesley M, Richard R Nelson, and John P Walsh. 2002. "Links and impacts: the influence of public research on industrial R&D." *Management science*, 48(1): 1–23.

Czernich, Nina, Oliver Falck, Tobias Kretschmer, and Ludger Woessmann. 2011. "Broadband infrastructure and economic growth." *The Economic Journal*, 121(552): 505–532.

d'Aspremont, Claude, and Alexis Jacquemin. 1988. "Cooperative and noncooperative R & D in duopoly with spillovers." *The American Economic Review*, 78(5): 1133–1137.

Ding, Waverly W, Sharon G Levin, Paula E Stephan, and Anne E Winkler. 2010. "The impact of information technology on academic scientists' productivity and collaboration patterns." *Management Science*, 56(9): 1439–1461.

Dittmar, Jeremiah E. 2011. "Information technology and economic change: the impact of the printing press." *The Quarterly Journal of Economics*, 126(3): 1133–1172.

Fleming, Lee, and Olav Sorenson. 2004. "Science as a map in technological search." *Strategic management journal*, 25(8-9): 909–928.

Forman, Chris, and Nicolas van Zeebroeck. 2012. "From wires to partners: How the Internet has fostered R&D collaborations within firms." *Management science*, 58(8): 1549–1568.

Forman, Chris, and Nicolas van Zeebroeck. 2019. "Digital technology adoption and knowledge flows within firms: Can the Internet overcome geographic and technological distance?" *Research policy*, 48(8): 103697.

Furman, Jeffrey L, and Scott Stern. 2011. "Climbing atop the shoulders of giants: The impact of institutions on cumulative research." *American Economic Review*, 101(5): 1933–63.

Furman, Jeffrey L, Markus Nagler, and Martin Watzinger. 2018. "Disclosure and subsequent innovation: Evidence from the patent depository library program." National Bureau of Economic Research.

Galasso, Alberto, and Mark Schankerman. 2015. "Patents and cumulative innovation: Causal evidence from the courts." *The Quarterly Journal of Economics*, 130(1): 317–369.

Gavazza, Alessandro, Mattia Nardotto, and Tommaso Valletti. 2019. "Internet and politics: Evidence from UK local elections and local government policies." *The Review of Economic Studies*, 86(5): 2092–2135.

Gittelman, Michelle, and Bruce Kogut. 2003. "Does good science lead to valuable knowledge? Biotechnology firms and the evolutionary logic of citation patterns." *Management Science*, 49(4): 366–382.

Griliches, Zvi. 1986. "Productivity, R&D, and the Basic Research at the Firm Level in the 1970's." *American Economic Review*, 76(1): 141–54.

Hegde, Deepak, Kyle Herkenhoff, and Chenqi Zhu. 2020. "Patent disclosure and innovation." *Available at SSRN 3158031*.

Iaria, Alessandro, Carlo Schwarz, and Fabian Waldinger. 2018. "Frontier Knowledge and Scientific Production: Evidence from the Collapse of International Science*." *The Quarterly Journal of Economics*, 133(2): 927–991.

Jaffe, Adam B. 1989. "Real effects of academic research." *The American economic review*, 957–970.

Jaffe, Adam B, Manuel Trajtenberg, and Rebecca Henderson. 1993. "Geographic localization of knowledge spillovers as evidenced by patent citations." *The Quarterly Journal of Economics*, 108(3): 577–598.

Jones, Charles I. 1995. "R&D-based models of economic growth." *Journal of Political Economy*, 103(4): 759–784.

Katz, Michael L. 1986. "An analysis of cooperative research and development." *The RAND Journal of Economics*, 527–543.

Lück, Sonja, Benjamin Balsmeier, Florian Seliger, and Lee Fleming. 2020. "Early disclosure of invention and reduced duplication: An empirical test." *Management Science*, 66(6): 2677–2685.

Machlup, Fritz. 1962. *The production and distribution of knowledge in the United States.* Vol. 278, Princeton university press.

Mansfield, Edwin. 1995. "Academic research underlying industrial innovations: sources, characteristics, and financing." *The review of Economics and Statistics*, 55–65.

Marx, Karl. 1867. "Das kapital. Buch I. English edition: Marx, K.(1976)." *Capital. Volume*, 1.

Mokyr, Joel. 2005. "Long-term economic growth and the history of technology." In *Handbook of economic growth*. Vol. 1, 1113–1180. Elsevier.

Murray, Fiona, Philippe Aghion, Mathias Dewatripont, Julian Kolev, and Scott Stern. 2016. "Of mice and academics: Examining the effect of openness on innovation." *American Economic Journal: Economic Policy*, 8(1): 212–52.

Nelson, Richard. 1962. "The link between science and invention: The case of the transistor. The rate and direction of inventive activity."

Romer, Paul M. 1986. "Increasing returns and long-run growth." *Journal of political economy*, 94(5): 1002–1037.

Romer, Paul M. 1990. "Endogenous technological change." *Journal of political Economy*, 98(5, Part 2): S71–S102.

Schumpeter, Joseph Alois. 1942. *Socialism, capitalism and democracy.* Harper and Brothers.

Singh, Jasjit. 2005. "Collaborative networks as determinants of knowledge diffusion patterns." *Management science*, 51(5): 756–770.

Singh, Jasjit, and Matt Marx. 2013. "Geographic constraints on knowledge spillovers: Political borders vs. spatial proximity." *Management Science*, 59(9): 2056–2078.

Thompson, Neil C., and Douglas Hanley. 2018. "Science is Shaped by Wikipedia: Evidence From a Randomized Control Trial." *Available at SSRN 3039505*.

Weitzman, Martin L. 1998. "Recombinant growth." *The Quarterly Journal of Economics*, 113(2): 331–360.

Williams, Heidi L. 2013. "Intellectual Property Rights and Innovation: Evidence from the Human Genome." *Journal of Political Economy*, 121(1): 1–27.

Zucker, Lynne G, Michael R Darby, and Jeff S Armstrong. 2002. "Commercializing knowledge: University science, knowledge capture, and firm performance in biotechnology." *Management science*, 48(1): 138–153.

Chapter 1

Not Like the Others: Frontier Scientists and High-Impact Technology[1]

1.1 Introduction

The economics of innovation literature has since long recognized the value contribution of new scientific ideas to private sector invention and their importance for technological progress (Nelson, 1962; Griliches, 1986; Jaffe, 1989; Mansfield, 1995; Cohen, Nelson and Walsh, 2002; Fleming and Sorenson, 2004). A vast body of empirical work has demonstrated the large and increasing share of technologies to be reliant on inputs from science (Narin, Hamilton and Olivastro, 1997; Ahmadpoor and Jones, 2017). However, not all scientific knowledge may be relevant for industrial applications, nor may all of these depend on inputs from science (Stephan, 1996).

[1]This chapter is based on joint work with Sam Arts and Reinhilde Veugelers. I thank the members of my doctoral committee, Dennis Verhoeven and participants in seminars and presentations at KU Leuven and the 2018 CISS PhD Summer School for many useful comments and suggestions. Financial support from the Research Foundation - Flanders (FWO) is gratefully acknowledged under grant number 11B5918N.

And furthermore, the links between science and technology may not be obvious, direct or immediate (Iaria, Schwarz and Waldinger, 2018). What type of science and through which mechanisms science is translated into which new technologies is only starting to be examined (e.g. Veugelers and Wang, 2019; Watzinger and Schnitzer, 2019). Taking into account the highly skewed nature of both scientific progress and technology performance, concentrated in a few breakthroughs (Kuhn, 1962; Scherer and Harhoff, 2000; Wang, Song and Barabási, 2013), of particular interest is to know whether new scientific breakthroughs – frontier science – are more important for technological impact (Hicks et al., 2000; Poege et al., 2019), more specifically for generating inventive hits, and if so, how frontier science can be effectively accessed and translated into valuable technology.

Building further on the literature which identifies the critical role of inventor-authors (Gittelman and Kogut, 2003; Bonaccorsi and Thoma, 2007; Breschi and Catalini, 2010; Arts and Veugelers, 2020), we argue that inventions with not merely scientific authors involved, but particularly with inventors involved in the creation of recent breakthroughs in science – whom we label as "frontier authors" – will be more likely to have technology impact, even more so to become technology hits. An inherent advantage that we ascribe to such frontier authors lies in their enhanced (i.e. better and faster) access to the latest breakthroughs in science: They are able to utilize frontier knowledge early-on for technology development, while the general inventive workforce remains widely excluded due to high complexity, tacitness and uncertainty associated with new scientific ideas (Zucker, Darby and Armstrong, 2002). Considering the importance of access to frontier science for superior and outstanding technology performance (Poege et al., 2019; Baruffaldi and Poege, 2020), the presence of "frontier authors" on inventive teams will more likely generate high-impact technologies, particularly inventive hits. Furthermore, access gains to frontier science may be particularly strong but do not need to be limited to the use of their own frontier scientific work: Due to active participation, this advantage may also extend to being able to more effectively source and apply others' new frontier ideas and findings, as the network and absorptive capacity literature suggests (Rosenberg, 1990; Audretsch and Stephan, 1996; Cockburn and Henderson, 1998; Fabrizio, 2009).

Relying on all private sector biomedical patents filed at the United States Patents

and Trademarks Office (USPTO) from 1980–2009 and publications from PubMed, we empirically investigate the contribution of frontier authors to technical advance. The biomedical sector provides a particularly interesting setting for our analysis, given the strong reliance of private sector innovation on fundamental research (Henderson, Pisano and Orsenigo, 1999; Li, Azoulay and Sampat, 2017), as well as its vibrant interactions between firms and scientists (Cockburn and Henderson, 1998; Zucker, Darby and Brewer, 1998).

We use the biomedical setting to investigate the interrelations between frontier authors, frontier science inputs and their technological outcomes. For this, we first provide estimates comparing average citation and top-citations impact of patents without and with authors among their inventors and further distinguish among the latter by scientific capability, i.e. between patents of frontier authors and those of authors with other science backgrounds (non-frontier authors). In search for mechanisms possibly explaining technology performance differences across frontier, non-frontier and non-author patents, we subsequently investigate changes across these patents regarding the type and timing of references to scientific knowledge as prior art (SNPRs) – in particular references to frontier knowledge, which we classify as recent access links to new scientific breakthroughs. Finally, we explore whether and to which extent differences in access and proximity to frontier SNPRs allow to capture the high-impact premium of frontier author patents.

We present a series of main findings as stylized facts. First, among similar technologies, firms and inventor teams, patents with frontier authors generate substantially more technological impact compared to non-author patents - +51% citations on average – and, in particular, much higher excess probability (+69%) of technological breakthrough (hit patents) The relative impact premia of patents with non-frontier authors are significantly smaller (less than half in size), though still above non-author patent levels. Frontier author effects decrease strongly when using lower benchmarks for assessing the quality levels of scientific output, consistent with the idea that the frontier of knowledge in science is closely connected to technology and that leading scientists develop the most valuable inventions. This is further strengthened by our finding that impact premia of frontier author patents are contingent on the recency of involvement of authors in top-research – in line with the view of a rapidly moving knowledge frontier – and that they are par-

ticularly strong for patents with authors involved as first authors, i.e. who had a substantial role, in frontier science.

Second, exploring patent references to scientific articles, we find that patents with frontier authors are generally much more science reliant than non-author, but also non-frontier author patents. In particular, they are more than twice as likely than the latter to establish frontier science links (frontier SNPRs), i.e. patent references to a recent top-paper, which we additionally show to be directly linked to hit patents and higher average technological performance in comparison to other science-based and non science-based inventions (in line with prior findings from Ahmadpoor and Jones, 2017 and Poege et al., 2019). We further find that patents with frontier authors are particularly likely to reference frontier science articles, more strongly those originating in close proximity to their own scientific work - namely self- and network citations. In contrast, we find only low levels of frontier science reliance in patents with non-frontier authors, which are mostly not significantly different, sometimes even slightly lower, than those of non author patents. This is consistent with the idea of frontier author involvement allowing for access and, thus, increasing absorptive capacity Cohen and Levinthal (1989) towards the frontier of knowledge for technology development.

Finally, we find that the patents with frontier authors that make such frontier SNPRs generate the highest impact premium and are most likely to yield hit technologies. This holds not only compared to other inventions made by frontier authors, but also relative to patents with non-frontier authors and non-authors making references to frontier science. In particular with regards to breakthrough (hit) technology, differences frontier authors and other inventor types seem to stem to a large extent from their differences in use of frontier SNPRs. We interpret this as evidence suggestive of better use of frontier knowledge by frontier authors in technology, consistent with the idea of superior scientific capability and human capital of frontier authors that extends to the sphere of technology. We further find that excess impact related to the use of frontier science in frontier author patents is largest for cases of references to articles in the close proximity of the authors. This lends support to the idea of increased absorptive capacity through active participation in frontier science and highlights the important role of these authors as brokers from the frontier of knowledge to industrial applications. Nevertheless,

our results also suggest that differences in use of science measured through SNPRs do not fully explain impact differentials across patents of frontier author and other inventor categories, especially higher average citation rates. Further, we do not find important differences based on priority in access to frontier science, regardless of author or inventor type.

Our findings contribute, primarily, to the literature on returns from science involvement of inventors in industrial technology (Gittelman and Kogut, 2003; Bonaccorsi and Thoma, 2007; Gruber, Harhoff and Hoisl, 2013). Here, by identifying the importance of the frontier nature in the author-inventor link, we contribute evidence for significant heterogeneity in this relationship: First, by showing that the recency of involvement of authors in top-scientific performance is a crucial predictor for both impact and access to frontier knowledge in technology. And second, with our finding that frontier scientists are not merely "bridge-builders" between science and technology (Breschi and Catalini, 2010), but that their enhanced access to new scientific ideas is, in particular and directly, linked to technological breakthroughs. The observation of this 'dual-excellence' also adds to the literature highlighting the pivotal role of top-scientists for knowledge production, in particular in biomedical research (Zucker and Darby, 1996; Cockburn and Henderson, 1998; Azoulay, Graff Zivin and Wang, 2010; Agrawal, McHale and Oettl, 2017), which has been strongly centered around human capital externalities.

Our paper also contributes to the literature on corporate participation in research to increase downstream absorptive capacity in firm R&D (Rosenberg, 1990; Gambardella, 1992; Arora, Belenzon and Sheer, 2021), which has emphasized motives and contingencies behind firm-science linkages. Here, our findings establish frontier authors in firm development as a crucial element for capturing technological value out of scientific progress. Thereby, we complement prior literature showing recency effects of science participation on technology search outcomes (Fabrizio, 2009), and contribute the insight that this is particularly important, i.e. necessary and impactful, for top-science, at the frontier and in the tail of the quality distribution. Our findings also speak specifically to prior work on channels of early access for firms to scientific advance (Baruffaldi and Poege, 2020; Arora, Belenzon and Dionisi, 2021). We add to this recent stream by highlighting that active involvement of specifically authors at the frontier provides effective access to recent

top-science for generating superior technological impact, but conversely also that the usefulness of author frontier knowledge is fleeting and discounted over time.

Our results, more broadly, also contribute to the general literature on links between science and industrial technology (Mansfield, 1995; Cohen, Nelson and Walsh, 2002), in particular to prior work regarding the relevance of top-science and frontier knowledge for new technologies (Hicks et al., 2000; Iaria, Schwarz and Waldinger, 2018; Poege et al., 2019). Specifically, here we contribute by showing, first, that the link between research excellence and technological performance persists at the individual-level and, second, that actively involved authors make better translations of (the same) frontier science into technological applications, evidenced by superior impact and higher breakthrough rates. The latter finding adds in particular also to the body of work focusing on scientific non-patent references in patents (Narin, Hamilton and Olivastro, 1997; Callaert et al., 2006; Callaert, Pellens and Van Looy, 2014; Ahmadpoor and Jones, 2017), by emphasizing that the individual backgrounds of the inventors (in this case a frontier author) matter – beyond institutional, technological and temporal contexts- for the impact of scientific references made.

The following section elaborates in more detail the relevant state-of-the-art and proceeds to develop the specific research questions that we investigate in the paper.

1.2 Prior Literature & Theoretical Considerations

1.2.1 Science as an input for technology

Technical progress and scientific advance are long-known to be closely coupled. In the logic of the 'linear model' of innovation, science provides the understanding of basic principles which serve as a foundation for technical solutions (Bush, 1945; Godin, 2006; Balconi, Brusoni and Orsenigo, 2010). Theoretical models of economic growth emphasize the importance of scientific inputs for technical change (Romer, 1986, 1990; Jones, 1995; Weitzman, 1998). Empirical studies relying on evidence from surveys (Mansfield, 1995; Cohen, Nelson and Walsh, 2002; Giuri et al., 2007) and patent records (Narin, Hamilton and Olivastro, 1997; Ahmadpoor and

Jones, 2017) confirm that an increasingly large share of new technology is, either directly or indirectly, reliant on scientific knowledge. While in itself detached from practical application, basic research fuels innovation by generating a public good with large applied utility (Toole, 2012; Azoulay et al., 2019). Not only does science have repercussions on the creation of new technologies, but inventions drawing on scientific insight are also found to be more impactful and valuable among their kind: Using patent forward citations as an indicator for technological value, Fleming and Sorenson (2004) show that science-reliant patents have significantly more impact on follow-up invention, fostering the idea of science as a map conducive to more useful combinations in technological search. Besides the social value created for technology, scientific inputs have been found to also yield inventions with superior private returns (Griliches, 1979; Mansfield, 1991; Watzinger and Schnitzer, 2019), which constitutes an essential condition for the use and engagement of private firms with basic research (Rosenberg, 1990; Gambardella, 1992; Arora, Belenzon and Sheer, 2021).

However, not all scientific knowledge is equally useful for technology, and not all inventions depend on science access in the same way (Stephan, 1996). Both scientific and technological progress are often depicted as frontiers being advanced by a few breakthroughs, i.e. highly impactful, and often radical, new discoveries or hit inventions (Schumpeter, 1942; Kuhn, 1962; Dosi, 1982). Concordant with this view, impact in their respective realms is found to be highly concentrated on a small number of distinguished research articles and patents (Scherer and Harhoff, 2000; Wang, Song and Barabási, 2013). This immediately prompts the thought of a non-uniform link, and direct connection between excellence in science and top-performance in technology. A recent empirical literature has started to investigate this relationship more closely: Iaria, Schwarz and Waldinger (2018), in a natural experiment, provide causal evidence that access to the frontier knowledge leads to more scientific articles with relevance for technology. In their analysis of patent references to scientific articles, Hicks et al. (2000) find that the top-percentile of most cited scientific papers exhibits a multiple times higher likelihood of receiving patent citations. Further building on this, Poege et al. (2019) show on a comprehensive scale that a small share of high-quality scientific articles disproportionately accounts for technological value in patents. Moreover, Ahmadpoor and Jones (2017) show

that often scientific breakthroughs yield 'double-hits' and become highly impactful also in technology.

The ways in which such high-impact connections between specific types of science, on one hand, and top-performing technology, on the other, are established are, however, not obvious and have remained so far unanswered by the literature. We argue that individual human capital, specifically the simultaneous involvement of inventors in scientific production, plays a crucial role for access to science with a higher value potential for technology. We further argue that individual participation will be, in particular, essential for enabling the use of science at the knowledge frontier, i.e. to recent scientific breakthroughs which are especially difficult to externally absorb, for invention and that this specific connection creates high-impact potential in technology, which is superior to other modes of access to frontier science.

1.2.2 The importance of scientific human capital for links between science and technology

A long literature on innovation, starting from Arrow (1962), highlights the institutional and intellectual divide that exists between open, predominantly public, science and proprietary industrial technology, rooted in the conflicting logics of non-rivalry and the imperative appropriation of rents and the internalization of market spillovers. This becomes manifest, particularly, through the existence of largely disconnected communities of scientists and inventors, especially across the academe-corporate boundary. How to bridge this divide has been focus of a vast empirical literature.[2] A central role of knowledge brokers between such "small world" communities has been ascribed by prior work to individuals embedded simultaneously in both the science and technology realms, so-called "inventor-authors" (Balconi, Breschi and Lissoni, 2004; Breschi and Catalini, 2010).

[2] This question is central to streams of literature studying mechanisms of knowledge transfer, e.g., through academic commercialization (Mowery et al., 2001; Agrawal and Henderson, 2002; Debackere and Veugelers, 2005; Toole and Czarnitzki, 2007), geographic proximity (Jaffe, 1989; Belenzon and Schankerman, 2013; Bikard and Marx, 2020), access-enhancing institutions (Furman and Stern, 2011; Berkes and Nencka, 2020) and open access intellectual property rights (Murray and Stern, 2007; Williams, 2013; Bryan and Ozcan, 2016).

The idea of 'connectedness' is also prominently found in the absorptive capacity literature focussing on the outcomes of firm collaboration with academic scientists as means of access to scientific progress (Cockburn and Henderson, 1998; Fabrizio, 2009). Especially in the presence of radical and dynamic change of the knowledge structures of industries, access to the scientific frontier is found to be closely tied to the human capital externalities of pivotal, high-quality scientists ("stars") - as observed by several studies investigating the birth of the U.S. biotechnology industry (Zucker and Darby, 1996; Zucker, Darby and Brewer, 1998). Star-scientists embed a significant amount of tacit insight which appears crucial to the effective realization of technological opportunity originating from frontier knowledge.[3] Collaboration with top-scientists on matters of basic research is often described as effective, though indirect, mechanism to absorb and integrate these ideas into technological applications with superior commercial value (Zucker, Darby and Armstrong, 2002; Colen et al., 2014).

Relatively few papers in the literature have, however, investigated the direct technological outcomes of scientists in invention. From a perspective of human capital accumulation, Bianchi and Giorcelli (2020) find that switching from a technical to a science-based higher education was a strong predictor for future patenting among Italian graduates. In their seminal analysis of biotechnology firms and the value of their patents, Gittelman and Kogut (2003) show that patents with scientific authors among their inventors receive substantially higher forward citation rates compared to non-author inventor patents. Similar patterns are found in the field of nanotechnology by Bonaccorsi and Thoma (2007). In more recent work, Arts and Veugelers (2020) document how among inventors in industry holding PhDs, those showing more pronounced scientific motivations (i.e. a stronger individual taste for science) develop more novel and more impactful patents.

A significant gap in the literature, however, persists regarding our understanding of heterogeneity in this relationship, especially of the nature and scientific origins of the author-technology impact premium: Which type of authors develop the most valuable, impactful patents? And, in particular, technological breakthroughs?

[3] The human capital externalities of star scientists have been subject to several empirical investigations and a broadly related literature regarding, e.g., peer effects in scientific production Azoulay, Graff Zivin and Wang, 2010; Oettl, 2012 or recruiting outcomes (e.g. Agrawal, McHale and Oettl (2017)).

While the absorptive capacity literature has long recognized the existence of a disparity in outcomes of scientific collaboration with authors for firm performance – skewed towards highly pivotal scientists –, the body of prior work on direct technological outcomes of authors on patents, has so far failed to identify such contingencies. However, the link between author human capital and technology performance does not appear to be straightforward: It is non-obvious that excellence in the understanding and discovery of fundamental would directly translate into better use of the knowledge in technology applications, in particular not for hit technologies. In fact, others have argued that it may well require different minds and skills within an organization to translate ideas from science into high-impact technology (e.g. Allen et al., 1977).

We aim to contribute to this literature by developing two testable predictions: First, we argue that there exists significant heterogeneity in technological outcomes of authors conditional on their scientific human capital. Specifically, the involvement of inventors in top-scientific production, at the frontier of knowledge, will generate opportunity for new technological solutions stemming from understanding of new fundamental principles that will translate into inventions with superior value, relative to those of authors previously involved in science not at the frontier and to inventors without scientific background. Further, analogously to what has been observed for the link between science and technology, we postulate that there is a concordance between excellence in performance in one and the other sphere. Hence, we expect authors of breakthrough science not only to create more average value in technology, but that they will be particularly more likely to generate hit technologies.

Second, we expect the opportunity that involvement in frontier science provides for human capital of authors in technology development to be impersistent and decaying with time. Indeed, as the frontier of knowledge is rapidly moving (e.g. Mukherjee et al., 2017), following the obsolescence of their insight, also the impact advantage of top-authors in technology is likely to be discounted as time since top-scientific engagement passes. The positive link between recency of scientific involvement and technological outcomes has been highlighted by prior literature (e.g. Fabrizio, 2009; Gruber, Harhoff and Hoisl, 2013).[4] However, these prior studies

[4] Specifically, we argue that the marginal return of the use of (former) frontier knowledge diminishes

do not differentiate between the quality of recent scientific performance. Given the uneven utility of scientific knowledge for technological, however, we expect this decay to be relatively stronger for top-science involvement, grounded especially in the initially higher potential of frontier science to fuel breakthrough technology. Conversely, we expect that frontier authors, i.e. those involved in recent top-science, would contribute more value to invention than those author-inventors with top-science involvement dating back further, and especially that they would more likely generate technological hits.

1.2.3 Frontier authors and their use of science in inventions

In order to understand the contribution of frontier authors to technological progress, it is vital to identify the sources of their inventive performance premium. While scientific insight, in general, may impact technological search capabilities in a broader, more original and perhaps indirect way, which has been highlighted by prior literature (Bonaccorsi and Thoma, 2007; Gruber, Harhoff and Hoisl, 2013), as discussed above, the specific advantage of frontier authors among scientist-inventors is likely to be rooted in the enhanced access to the rapidly moving frontier of knowledge. This access advantage may be particularly conducive to the use of frontier science as prior art in the type of inventive solutions that frontier authors make.

Recent empirical studies have contributed evidence on the importance of timing of links to new scientific ideas for technology outcomes. Based on scientific references in patents, Poege et al. (2019) find that the correlation between recency and higher impact of technologies is strongest in the tail of the science-quality distribution, while it becomes marginal for patent links to lower quality science. Concomitantly, Arora, Belenzon and Dionisi (2021) show that first mover advantages in making references to scientific articles are associated with particularly higher private value of firm patents, especially for scientific articles with greater

with increasing temporal distance as this knowledge becomes a more commonly integrated component of new technologies and subsequent, more recent scientific ideas start to bear higher impact potential.Gruber, Harhoff and Hoisl (2013) argue that the decreasing relevance of their scientific insight makes inventors with prior science experience shift more towards the application of in-depth technological knowledge over their career, and find a concurrent decrease in recombinant breadth of their technologies.

utility for industrial applications. Both the studies of Baruffaldi and Poege (2020) and Arora, Belenzon and Dionisi (2021) find that corporate participation in science, engagement with the scientific community, and collaboration with scientists are effective channels of early access to scientific ideas. However, an open question remains whether and which types of scientists generate a technology premium for industrial applications, as well as to whether and through which types of links to new science this premium can be explained.

We argue that the advantage of frontier authors in the development of higher-impact and, in particular, hit technologies originates in large part from their superior access to frontier science and the active use of this knowledge in their inventions. To illustrate this better, it is useful to consider the very particular nature of frontier knowledge: Infant scientific breakthroughs are typically characterized by a high degree of complexity and tacitness, given that the understanding of their working principles is usually concentrated among a few pivotal individuals (Zucker, Darby and Armstrong, 2002). Often, the application of such knowledge initially entails high levels of uncertainty and requires multiple experimentation in order to prevail (Murray et al., 2016). This combination of implicitness and high-risk is likely to exclude the broader inventive workforce from effective early access to frontier science. Thereby, the individual connect between knowledge input and technological application becomes particularly important for frontier science, as this knowledge is particularly challenging to externally absorb. Following this argument, frontier authors will enjoy an advantage of both better capability as well as timing to use frontier knowledge. This will manifest itself in inventions of higher technological impact, but especially higher technological hit rates, given the direct link between successful access to top-science and breakthrough potential.

This advantage is likely to be contingent on the degree and the recency of involvement. In line with the idea of 'connectedness' (Audretsch and Stephan, 1996; Cockburn and Henderson, 1998), we expect the frontier author premium to be concentrated, but not limited, to links to frontier science originating in close proximity to the scientific work of the authors, i.e. stronger for self-reliance or scientific peer network links. Following higher closeness through involvement, we further expect frontier authors to establish links to frontier science particularly early on, often already before the scientific knowledge becomes generally available and, by

this, likely enjoy first mover advantages.

In summary of this section, following the above outlined argumentation, we will contribute to the literature by empirically investigating whether frontier authors are more likely to generate high-impact technology, in particular technological breakthrough, and subsequently tracing the sources of such performance premia at the invention-level. Thereby, we will be assessing, first, whether their patents systematically differ in the reliance on frontier knowledge, in particular through proximity and early access, and second, evaluating whether and to which extent such access advantage can help to explain their superior technological performance.

In the next section, we outline our research setting as well as our main empirical approach to investigate our research question.

1.3 Research Setting & Data

1.3.1 Frontier author links in biomedical technology

The scope of our empirical analysis is invention in the biomedical sector. This sector is an important one to study given its large public support for and private participation in basic research (Gambardella, 1992; Henderson and Cockburn, 1996; Li, Azoulay and Sampat, 2017), making questions of effective channels of scientific knowledge translation into industrial applications highly relevant for policy. Moreover, it is particularly interesting in the context of our paper given the strong connectedness between science and technology and the vibrant interactions between fundamental and applied researchers that many prior studies have highlighted (e.g. Cockburn and Henderson, 1998; Murray, 2002).

Furthermore, the literature has often emphasized the key role of 'star scientists' and their importance for industry to access the frontier of knowledge in biomedical science (Zucker, Darby and Armstrong, 2002; Colen et al., 2014). This concentration of relevant frontier knowledge in pivotal individuals makes questions of effective early access particularly interesting, considering that medical breakthroughs often bare the potential for double-hits (Ahmadpoor and Jones, 2017), i.e. to be directly and almost simultaneously linked to top-impact technologies. For our quest of understanding the exact conditions under which this process takes place, the

biomedical sector provides an ideal empirical setting.

1.3.2 Data sources

We analyze the relationship between frontier science access of inventors and technological impact in a sample of all biomedical utility patents granted by the U.S. Patent and Trademark Office (USPTO) between 1980 to 2009. We define biomedical inventions based on 3-digit USPC patent classes assigned to 'Drugs&Medical' category in the classification of Hall, Jaffe and Trajtenberg (2001).[5] Our main source for patent data is PATSTAT 2018 spring edition. We restrict our sample only to patents assigned exclusively to private corporations based on the PATSTAT PSN sector allocation, excluding collaborative patents between industry and universities or public research institutes.[6] We link this sample to the USPTO disambiguated inventor database (Li et al., 2014) and retain only those patents for which we have full inventor information.[7] Our estimation sample comprises 239,180 unique patented inventions developed by 161,624 unique inventors.

For all patents in this sample, we retrieve inventors who appear also as authors of scientific publications in PubMed until 2009.[8] For this purpose, we establish a unique link between inventors in the Li et al. (2014) database and authors in the *Author-ity* 2009 data (Smalheiser and Torvik, 2009; Torvik and Smalheiser, 2009;

[5] We use the 2006 version of the NBER classification. We retrieve information on USPC classes of patents from the USPTO historical files (Marco et al., 2015) and from Fierro (2014). In our main sample, we include all patents listing at least one biomedical class. However, all results are robust for the sub-set based on biomedical primary class only.

[6] Industry-university/public co-patents account for < 2% of all patents assigned to private corporations. An important limitation of patent data is that, while we observe the sectoral allocation of patent assignees, we cannot consistently identify the affiliation of inventors at the time of patenting. A significant residual share of patents in our sample may be inventions resulting from informal or contractual collaborations of private firms with individual, academic or public sector researchers.

[7] Around 10% of U.S. patents are missing in the Li et al. (2014) data, due to remaining ambiguous or not listed inventor names.

[8] PubMed is a comprehensive online bibliographic index database for all, but not exclusively, biomedical literature, maintained by the National Centre of Biotechnology Information (NCBI) at the U.S. National Library of Medicine (NLM) (see: pubmed.ncbi.nlm.nih.gov). It currently covers citations to 32 million biomedical publications worldwide. For a discussion of coverage and generalizability of PubMed for links to science by patents in our sample, see Appendix 3.B. Affiliations in PubMed refer to the time of publication and are only available for first listed authors.

Torvik, 2018). We restrict our PubMed sample to only scientific articles journals in the Science Citation Index (SCI) and link them to scientific fields in Clarivate Web of Science.[9] Patents in our estimation sample comprise 59,455 unique inventor-authors.

We retrieve scientific non-patent references (SNPRs) to all articles indexed in PubMed/ Web of Science SCI published until 2013, using data from the *PatCi* disambiguation (Agarwal et al., 2014) and Marx and Fuegi (2020).[10] We further source scientific article-to-article citations from Microsoft Academic Graph (MAG), in order to determine the medical content area and scientific impact of all publications in our data. For more details on data processing and different disambiguation approaches, see Appendix 3.B. Below we describe how these data sources are used to construct the key measures used in the analysis of the paper.

1.3.3 Measures

Patent impact

We use patent forward citations to quantify the impact of inventions in our sample, that is references received from subsequent generations of patents. Forward citations are a commonly used and established indicator in the literature to assess the quality and economic value of patents (Trajtenberg, 1990; Harhoff et al., 1999). Importantly, they capture social technological value, given they reflect how useful an invention is as input for subsequent patents in the overall economy. For each patent, we count the total number of citing patent families (DOCDB) over a moving window of 10-years since application. In line with prior literature (e.g. Trajtenberg, 1990; Ahmadpoor and Jones, 2017), we identify "hit patents" as those inventions entering the far-right tail of the measure, which we define as entering the upper 5th percentile in the citation distribution of their primary patent class-year cohort.

[9] These are a population of 6.7 million unique authors listed on 11.5 million scientific journal articles, and cover 58% of all publications indexed in PubMed for the same period.

[10] References to Web Of Science SCI indexed articles account for 87.6% of all SNPRs to PubMed. We use the data from Agarwal et al. (2014) as our main SNPR disambiguation due to higher coverage of references to the PubMed universe (see Appendix 3.B), which are the relevant science universe for our study. However, we replenish our data with information on examiner-added SNPRs from Marx and Fuegi (2020) for patents in which this information is available (granted after Jan 2001).

Frontier authors

To identify among inventors in our patent sample authors involved at the frontier of science, we exploit the richness our data entails regarding heterogeneity of inventors' footprints in fundamental research. We define the frontier of scientific knowledge in two dimensions: 1) *top-quality*, i.e. reaching the highest standard of scientific excellence, and 2) *recency* of discovery, respectively publication. For our default "frontier" measure, we consider scientific articles to meet the highest quality criterion if they are accepted for publication in one of the eight most prestigious journals of the biomedical discipline.[11] These journals, which we denote as "top-general" journals, usually span a broad range of topics and require submissions to surpass the highest peer review standards before being accepted for publication. Articles in these journals cover a large share of the most impactful scientific publications across fields. The use of this type of indicator for top-scientific performance has the advantage of being unaffected by future commercial success of patents associated with a scientific underpinnings, either through an author- or article reference link, and thereby avoids potential measurement issues arising from simultaneity in our relationship of interest.

However, obviously not all publications in top-general are of equally high quality. Therefore, to account more directly for the scientific excellence of the specific papers authored by an inventor, we estimate robustness of results using an alternative indicator based on 10-year scientific citations counts post publication, and consider as top-science all articles that enter the upper 5th percentile of the this distribution of their Web of Science field and year. In this case, to mitigate concerns of simultaneity in measurement, we exclude all patents from the estimation that we identify as "patent-paper-pairs" Murray, 2002; Murray and Stern, 2007; Magerman, Van Looy and Debackere, 2015.[12]

[11] These journals are: The New England Journal of Medicine, Lancet, JAMA, Nature, Science, Cell, Nature Medicine, British Medical Journal. Formally, we select these based on: 1) Being in the top-category of journals in a biomedical Web of Science field category according to the ECOOM FEB journal ranking (2021), 2) covering a broad range of fields and topics, i.e. being general-interest, and 3) scoring above the 90th percentile of average journal impact factor (JIF) among journals in the overall top-category of biomedical journals, across fields. We exclude review journals from this ranking.

[12] We detect "patent-paper-pairs" as patents that are linked to a scientific article through a non-patent reference which is i) published recently around or after the patent application (enabling disclosure),

We consider scientific knowledge to be *recent* if it was published within the last 3 years prior to the application of a focal patent. This cut-off is based on a plausible time frame of obsolescence of knowledge (e. g. Mukherjee et al. (2017); Fleming (2001)) and the distributions of most recent scientific involvement of inventors with author background on all patents in our biomedical sample. The 3-year cut-off implies a relatively strict definition of recency of scientific involvement, given that the patent development process – so the actual inventive search we are considering – usually precedes the filing date of a patent by 1–2 years. For robustness, however, we further investigate persistence and decay of effects when applying cut-offs for recency of 5 and 10 years prior to the patent application.[13]

Based on these dimension, we classify a patent as *frontier author patent* if at least one of the inventors listed on the patent has authored a top-general journal scientific publication in the 3 years or less preceding the patent application year.[14] Similarly, we also take into account if a patent involves non-frontier authors (either non-top science authors, or non-recent top-science authors). In case a patent combines multiple author types, we assign it to the exclusive category with the highest scientific quality, followed by the most recent involvement.[15]

Frontier SNPRs

We quantify and trace the actual use of scientific ideas in technical application by considering the scientific references listed in the patents as prior art. SNPRs are an imperfect measure of scientific underpinnings, and have been critically assessed in

ii) for which the number of inventors is smaller or equal to the number of authors on the publication, and iii) with whom there is a 100% overlap between inventors listed on the patent and authors on the publication, similar to the approach of Thompson, Ziedonis and Mowery (2018).

[13] The 3-year cut-off further accommodates the specificity of our empirical setting (80% of author patents in biomedical have a most recently active author in the past 3 years, only 7% a frontier (recent + top-publication) author. However, all results in the paper are fully qualitatively robust, though quantitatively reduced, when using alternative recency cut-offs (see further, section 1.4.1).

[14] We consider the priority filing date as application date of patents. In case of continuations or divisional filings, which are very common in biomedical (30%), we consider the application that directly lead to the ultimate patent as first filing (typically the U.S. application filing date).

[15] That is, we will assign a patent with a frontier author and a top-general journal author of more than 3 years before the patent application to the "frontier author" category, while we will assign a patent with a non-recent top author and a recent non-top author to the category "top, > 3 years".

the literature (Meyer, 2000; Cassiman, Veugelers and Zuniga, 2008). While they may not be able to capture all ways in which scientific knowledge contributes to a new application, they are usually a strong proxy indicator for a genuine and important relation between the cited scientific discovery as prior art for the development of the citing technology (Tijssen, 2001; Roach and Cohen, 2013; Callaert, Pellens and Van Looy, 2014).

Using our link to PubMed, we differentiate among types of SNPRs and classify a patent as *frontier SNPR patent* if at least one of its scientific references links to an article in a top-general scientific journal and has been only recently published, i.e. in the 3 years or less leading up to the patent application. Again, we account for non-frontier SNPR links in patents as well, by applying the same hierarchy of exclusive categories as for author-type measures. We build similar indicators for SNPRs to articles entering the far right tale of their field and year scientific citations distributions.

Furthermore, we determine for each frontier and non-frontier SNPR patent its closest social distance to any of the authors listed, and classify SNPRs as *close* if they refer to a publication of the same author (self-citation) or to the direct network of an author, i.e. a publication of a prior co-author not involving the focal author.

In addition, we use the time distance between PubMed publication year[16] of frontier (and other) scientific articles and the filing dates of citing patents to distinguish different moments regarding the speed and first use of SNPRs: First, we re-conceptualize *first to cite* links as being the first (among subsequently relying patents) to establish a link to a scientific publication in technology, in line with Arora, Belenzon and Dionisi (2021). Second, we detect *ex-ante* links to scientific publications, i.e. references to on-going scientific inquiries in articles that were not yet published, or just published, by the time of the priority patent application.[17] Third, we combine these categories to identify differentials among frontier links distinguishing further into first-mover and up-front connections to frontier science.

[16] PubMed collects the year of publication for each indexed scientific article, i.e. the year in which the article has first been published in a peer-reviewed scientific journal. Exact dates of publication are not systematically reported. The publication year does not necessarily correspond to the year the article was first indexed in MEDLINE or the PubMed Central database.

[17] In our data, these references are almost never inserted by examiners, but usually supplied by the applicants during the examination process.

1.3.4 Sample & summary statistics

Based on our above discussed measures, only 7% of patents rank in the category of frontier author patents, i.e. including at least one inventor with a top-general journal publication in the last 3 years. Overall, 17% of patents are developed by inventors with prior top-general journal achievements. For the definition of frontier authors based on top 5% most highly cited publications, the patent shares are similar, though slightly larger (13% frontier author patents). The strong connect between science and technology in biomedical, however, is resembled by the fact that almost 60% of patents in our sample have at least one inventor with prior scientific involvement (see Table 1.A1, in Appendix 3.B).[18]

Table 1.1 summarizes the main measures utilized in the paper by different types of inventor author-status of patents in the sample. The average biomedical patent in our data receives 25 citations over a 10 year window. This reflects, on one hand, the strong patent intensity of biomedical R&D, evidenced by high rates of incoming citations (Hall, Jaffe and Trajtenberg, 2001), and on the other hand, a heavily skewed impact distribution, considering that the median patent receives less than half of this amount of citations (see Table 1.A1, in Appendix 3.B). The non-parametric comparison of citation means reveals that patents with frontier authors – with an average of 28.75 follow-on citations (column 2) – generate significantly higher impact in subsequent technologies than non-frontier author patents (25.51 citations, column 3) and patents without scientific authors among their inventors (23.32 citations, column 4 of Table 1.1). A similar pattern emerges with regards to the frequencies of hit patents: About 9% of frontier author inventions enter the upper 5th percentile of their year-class impact distribution. Medical patents with non-frontier authors (7%) and no authors (6%) are still more often hits in a patent class compared to the population average, however their breakthrough rates are significantly lower than those of patented inventions by frontier authors.

The nexus between scientific research and industrial technology also becomes evident through patent-paper trails. As many as 40% of patents in our sample in-

[18] Considering all prior publications in PubMed, rather only those indexed in SCI journals, does not affect the share of author-patents significantly. In this case, 63% of patents show an inventor-author link.

Table 1.1. Summary Statistics

VARIABLES	All	Frontier author	Non-frontier author	No author
Patent forward citations	24.84	28.75	25.51	23.32
Top5% hit patent	0.07	0.09	0.07	0.06
Scientific reference (SNPR)	0.40	0.78	0.51	0.20
Frontier SNPR	0.06	0.27	0.06	0.02
Citation lag SNPR (years)	3.70	1.99	3.63	5.08
Citation lag frontier SNPR	1.23	1.09	1.31	1.23
Patent forward cites if SNPR == 1	27.52	29.10	26.62	29.52
Patent forward cites if frontier SNPR == 1	30.59	33.89	28.86	30.30
Top5% hit patent if SNPR == 1	0.08	0.09	0.08	0.08
Top5% hit patent if frontier SNPR == 1	0.12	0.13	0.12	0.11
Number of SNPRs	5.03	17.78	6.22	1.37
Number of inventors	2.99	4.00	3.47	2.20
Applicant firm patent stock	403.09	298.74	420.31	397.95

Notes: The table reports patent-level group means. Our sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980–2009. Forward patent citations are measured over a 10-year window (DOCDB-level). "Frontier authors" are inventors that authored a paper in a "top-general" journal in the prior 3 years. "Top-general" journals in biomedical sciences are New England Journal of Medicine, Lancet, JAMA, Nature, Science, Cell, Nature Medicine, British Medical Journal. "SNPR" (scientific reference) indicates at least one patent reference to a biomedical scientific article indexed in PubMed/ Web of Science SCI. "Frontier SNPRs" are references to papers in "top-general" journals published in the past 3 years before the patent application date. Inventor-author disambiguation is based on Li et al., 2014; Smalheiser and Torvik, 2009 and Torvik (2018). DB links established by the authors. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015). Scientific patent references to articles in PubMed/ SCI are sourced from Agarwal et al. (2014) and Marx and Fuegi (2020). Science citations are sourced from Microsoft Academic Graph.

clude at least one reference to a journal article in PubMed/ Web of Science SCI. This, again, resembles the importance and strong direct link of biomedical technology to scientific research, as observed by prior literature (e.g.Narin, Hamilton and Olivastro, 1997; Ahmadpoor and Jones, 2017). Science-reliance is by far most common for patents with frontier authors (almost 80% of patents), followed with some gap by inventions of non-frontier authors (50% including SNPRs), while patents of inventors without scientific involvement reference research articles at

significantly below average rates (20% only).

This pattern is confirmed for references to frontier science: These occur much less frequently – in only 6% of patents at least one – and frontier author patents show these much more often (almost one third of patents this group). Here, differences in frequency of reference are even much more pronounced with regards to non-frontier and no author patents. Important differences in SNPRs can also be observed regarding the speed, i.e. the time lags, at which citations to scientific publications are incoming across patents in the different groups of inventor-author types: While patents with frontier authors reference the average scientific publication in patent less than two years after its publication, non-author patents seem to access this knowledge only three years later. The gap citation lags between the frontier author and non-frontier author patents categories is still considerable, with new science taking more than one and a-half years longer to be picked up in references of patents in the latter group.[19] Among patents making references to frontier science, differences in citation lags are significantly narrowed across the three groups: While patents with frontier authors are still fastest to pick up frontier knowledge, with an average citation lag of 1.09 years, differences to other patents making such links, though statistically significant, appear economically almost negligible (less than three months).[20]

The third panel in Table 1.1 compares impact for patents making SNPRs, in general, and SNPRs to frontier science. As expected, and in line with prior literature, patents referencing scientific publications receive higher citation rates than the sample average (27.52 follow-up citations in 10 years; row 7, column 1). However, differences in impact from science use are not apparently significant across patents with frontier authors and non-author patents. Science-citing patents with authors who are not at the scientific frontier even receive below average citation rates (26.62 citations).[21] A similar observation can be made for patents referencing fron-

[19] The average (shortest) citation lag to top-general journal publications from patents in our sample is 6.3 years, which indicates that our 3 year cut-off for frontier science links falls roughly around the upper 30th percentile of most recent links.

[20] In a t-test for equality of means with uneven variance, the difference in citation lags to frontier SNPRs between non-frontier and non-author patents is statistically significant only below the 10%-level.

[21] In our sample, the frontier author vs. non author patents difference is only significant when using the alternative definition of frontier authors based on recent top5% highly cited publications. Such

tier scientific articles, while the frontier author group shows a clearly pronounced citations premium. This provides non-parametric evidence suggesting that the returns from scientific involvement for the use of science in technology, in particular frontier science, are non-general and that such involvement may pay off only for top-research, at least when considering average impact of technologies. Finally, with regards to hit patents, differences in frequency across patents from different author-inventor groups are narrow and all three categories rank significantly above average, suggesting on one hand that frontier science use generates particular high breakthrough rates (in line with, e.g. Poege et al., 2019), but on the other hand that the connections under which such links are established may be more indirect and nuanced across broad categories of author-inventors.

The bottom panel of Table 1.1 reports means of patent-level characteristics across patents of different author-inventor types relating to volume of references, team size and firm context. The average patent in the sample includes five scientific references to PubMed/ Web of Science SCI publications. However there is, again, strong skewness in this measure: 75% of patents include three references or less, while the upper 5th percentile of patents cite more than 25 different articles (see Table 1.A1, in Appendix 3.B). Strikingly, patents with frontier authors among their inventors appear to be crowding in the tail of this distribution, making 18 SNPRs on average. The median frontier author patents still references as many as 6 scientific publications. Similar to what we observed for the occurrence of SNPRs for non-author patents, also the intensity of reliance on scientific publications for this group is quite low. A typical patent in our sample lists between two to three inventors, while it becomes evident that frontier authors tend to work in significantly larger inventor teams (four members on average), especially in comparison to patents without authors (inventor duos). The opposite pattern emerges when comparing prior patent stocks of applicant firms.[22] Frontier authors appear on patents of significantly smaller firms, considering the number of their, time-weighted, prior patent applications.

authors' patents receive 30.51 citations on average over 10 years, excluding patent-paper-pairs.

[22] We calculate firm patent stocks with a perpetual inventor method (Griliches, 1979; Hall, Jaffe and Trajtenberg, 2005), using a 15% annual depreciation rate. This allows us to discount firm R&D assets over time, despite not observing the yearly amount of expenditure.

At the author-level, scientific footprints appear quite heterogeneous; Generally, individual authors represented in our sample achieve above average performance in science (ranking above the 70th percentile in article citations in their respective scientific fields), have more years of experience in scientific publishing than in patenting, more prior patents and larger networks than non-author inventors, and have last published a journal article 2.17 years before the focal patent application. For frontier authors, these distributions are more skewed, although they don't differ in patenting activity from non-frontier authors at the median. More author-level descriptive statistics are reported in Table 1.A2 in Appendix 3.B.

1.3.5 Sample selection considerations

Focus on biomedical research

The scope of our analysis is limited to biomedical research. This limitation arises, partly, due to the demarcation of our empirical data base. We can think of two ways in which this possibly affects the generalizability and external validity of our estimates: First, it could be that the more immediate and direct applicability of scientific knowledge makes the within-individual link particularly important in this area, leading us to potentially overvalue the role of frontier authors in technology when attempting to extrapolate to other areas. However, prior studies find the relationship between direct science connectedness and technological impact, in particular between high-quality science and superior impact, to very stable and differing only slightly in magnitude across technological fields (Ahmadpoor and Jones, 2017; Poege et al., 2019). This corroborates the view that frontier scientific insight may provide strong benefits for more useful applications, also in other areas.

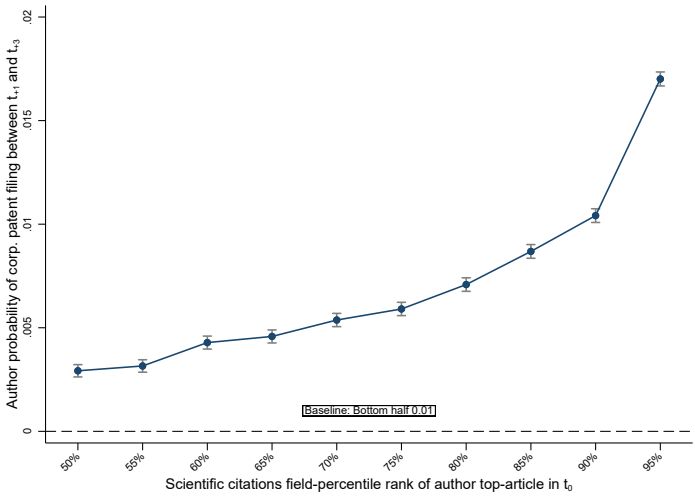
Second, one could also think about entrance into patenting as a positive selection mechanism. While in biomedical research, engaging in technology development appears a comparatively small step from fundamental science, in other fields, where this is less common, participation in corporate patenting may imply a significantly higher selection hurdle, involving substantial risk and higher opportunity costs. Accordingly, only scientists who can expect high gains and have a strong ability to translate their ideas into valuable applications, would enter commercial technology development. In this case, finding a positive relationship in our biomedical patents

would likely make us underestimate the importance of frontier authors – conditional on selection - for technological advance in other sectors.

Below we explore the relationship between performance in science and likelihood of entering into commercial patenting further for our underlying population of PubMed authors in the timeframe for our sample.

Author selection into patents sample

Figure 1.1. Author Selection into Industry Patenting



Notes: The figure plots parameter estimates from OLS regression of likelihood of entrance into corporate patenting on author top-article citations rank for the yearly cross-section of all authors on articles (N = 15,416,284) in PubMed/ Web of Science published between 1980 and 2005. The dependent variable measures the probability that an author appears as inventor on a U.S. patent assigned to a private corporation within 3 years following the publication year of the focal scientific article. The right-hand side/ x-axis indexes percentiles of the 10-year citation distribution of a WOS subject category field year cohort. The estimation model includes a full set of WOS-field category x year fixed effects. "Patent-paper pairs" are excluded. 95% confidence intervals are based on heteroscedasticity robust standard errors. Inventor-author disambiguation is based on Li et al., 2014; Smalheiser and Torvik, 2009 and Torvik (2018). DB links established by the authors. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015). Scientific patent references to articles in PubMed/ SCI are sourced from Agarwal et al. (2014) and Marx and Fuegi (2020). Science citations are sourced from Microsoft Academic Graph

In order to develop a better understanding regarding the selection mechanism

governing participation in corporate technology of medical scientists, we estimate the probability of entering our inventor sample for the population of all authors in PubMed/ Web of Science SCI publishing in a given year in the period 1980-2005 as a function of their highest ranked publication in that year in the citations distribution of a given field.

Figure 1.1 plots the estimated relationship, after absorbing field by year idiosyncratic variation. The likelihood of appearing as inventor on a firm patent in our sample within 3 years after publication to be 1% for authors of scientific articles in the bottom half of the 10-year citations distribution. However, this probability steeply increases for authors of higher impact papers moving towards the tail of the distribution. For authors of publications in the upper 10th percentile, the patenting probability is twice as high, for upper 5th percentile authors, it reaches +148% relative to the background rate.

This relationship is robust for author classification based on the ranking of journals, as well as when comparing expected participation rates for within-author publishing variation (see Table 1.A3 in Appendix 3.B). This shows that authors of recent top-science, i.e. frontier science as we define it, tend to be more likely to become inventors, suggesting a strong and direct connect between creation of knowledge at the frontier and technological opportunity.

1.4 Econometric Analysis

1.4.1 Frontier author patents and technological impact

To assess the inventive contribution of frontier authors to technological progress, relative to inventions of non-frontier authors and non-authors, we estimate the following regression for the patents in the sample:

$$Y_i = \alpha_0 + \beta_1 FrontierAuth_i + \sum_j \beta_j NonFrontierAuth_{j(i)} + \mathbf{X}_i'\gamma + \delta_{c(i)t(i)} + \eta_{s(i)d(i)} + \epsilon_i \tag{1.1}$$

Y_i captures technological impact through counts of forward citations over a 10-year period received by patent i, and alternatively as a binary indicator for the

likelihood of patent i to become a hit, i.e. to enter the upper 5th impact percentile of its primary 3-digit class and year cohort. $FrontAuth_i$ is a dummy variable which takes on the value 1 if a patent lists a frontier author as inventor, and 0 otherwise. The coefficient β_1 captures the difference in citation rates and top-impact probability for patents with frontier authors, relative to the background rate of non-author patents. We further include a set of exclusive category indicators accounting for non-frontier authors on patents, where $NonFrontierAuth_{j(i)} = 1$ if patent i includes non-frontier author type j – i.e. non-recent authors in top-journals, authors in top-field journals, authors in non-top journals - and 0 otherwise, whose impact differentials are captured by series of coefficients denoted as $\sum_j \beta_j$.

Our specification controls for a vector of patent level covariates $\boldsymbol{X}_i'$ that may affect both frontier author status and patent impact. In our baseline model, we adjust for the number of inventors listed on a patent, as this stochastically affects the likelihood of author occurrence and patents by larger teams tend to have systematically higher impact (Wuchty, Jones and Uzzi, 2007). We account for this by including a full set of fixed effects for the number of inventors, i.e. a series of dummy variables for each integer value taken on by inventor team size. This allows for non-parametric adjustment without having to impose functional form assumptions on the underlying distributions. To avoid unnecessarily restricting degrees of freedom given the long tail of the distribution, we replace team size values above the 95th percentile with their corresponding percentile ranks.[23]

We further include primary 3-digit class by application year fixed effects, in order to absorb permanent differences in citation rates and scientific involvement as well as time trends across technological fields. Finally, in our main model specification, we add a series of dummies for applicant firm patent stock decile rank interacted with a categorical variable capturing changes of decile rank (growing, shrinking, steady) of an applicant firm relative to its previous patenting year observation. These fixed effects aim to account for variation across institutional contexts and firm-level dynamics: For instance, frontier authors may be particularly likely to appear on patents of small start-ups or spin-offs (as summary statistics seem to suggest, see Table 1.1), which are characterize by high-growth rates and higher breakthrough likelihood (Schumpeter, 1942; Acs and Audretsch, 1990). On the other hand, large

[23] This generates a total of 11 equisized inventor team size category dummies.

pharmaceutical firms more often have in-house research departments, closer ties to external research departments and generate more valuable patents on average (Henderson and Cockburn, 1996). In order to estimate robustness of our baseline results, we further use several alternative specifications, including fixed effects for the number of authors, firm dummies, and parametric controls for author prior publications and experience effects in patenting, referring always to the maximum value across all authors on a patent.

In models for average technological impact, where the outcome measure is count distributed, we estimate equation (1.1) using Poisson pseudo-maximum likelihood estimation (PPML).[24] This Poisson estimator is highly efficient for regressions on count data and provides asymptotically correct standard errors with overdispersion, likely to be encountered in patent citations (Silva and Tenreyro, 2011). Moreover, it is equally consistent in the presence of fixed effects, but - due to its ability to include zero-values - avoids many issues associated with log-linearization or related data transformations under OLS (Silva and Tenreyro, 2006). For regressions of top-impact likelihood, we estimate equation (1.1) as linear probability models. We use robust variance-covariance matrices to account for heteroscedastic error terms.

Table 1.2 presents the estimation results for the relationship between technological impact of patents and frontier author status. Column 1 reports the baseline estimate for impact differentials of patents with authors relative to non-author patents. The point estimate of 0.193 indicates a 21.3% higher level of average forward citations for patents of inventors with prior scientific authorship (significant at the 1% level).[25] This observed author-patent premium is in line with findings in prior literature (Gittelman and Kogut, 2003; Bonaccorsi and Thoma, 2007). In Column 2, we then estimate heterogeneous effects on impact for patents of frontier authors and non-frontier authors, by splitting up author-status into sub-categories based on journal ranks and recency of authorship. The point estimates reveal an important disparity: Technological impact is magnified for patents with frontier

[24] We use the implementation of the PPML estimator with multiple levels of fixed effects provided by Correia, Guimaraes and Zylkin (2019).

[25] Coefficients from PPML estimation can be interpreted as semi-elasticities, identical to the log-linear model, i.e. as $\exp(\beta)$-1.

Table 1.2. Frontier Author Patents and Technological Impact

	PPML			OLS		
VARIABLES	(1) PatCit10	(2) PatCit10	(3) PatCit10	(4) Top5% hit	(5) Top5% hit	(6) Top5% hit
Author	0.193***			0.019***		
	(0.007)			(0.001)		
Frontier author		0.412***	0.429***		0.037***	0.038***
		(0.015)	(0.015)		(0.003)	(0.003)
Non-frontier author		0.176***			0.017***	
		(0.007)			(0.001)	
Author, top-general jrn, >3y			0.262***			0.024***
			(0.011)			(0.002)
Author, top field-jrn, last 3y			0.259***			0.025***
			(0.011)			(0.002)
Author, top field-jrn, >3y			0.184***			0.014***
			(0.012)			(0.002)
Author, no top-journal			0.106***			0.012***
			(0.009)			(0.001)
Constant	3.300***	3.293***	3.287***	0.055***	0.054***	0.054***
	(0.005)	(0.005)	(0.006)	(0.001)	(0.001)	(0.001)
Patent class × year	Yes	Yes	Yes	Yes	Yes	Yes
Number of inventors	Yes	Yes	Yes	Yes	Yes	Yes
Firm patent stock × Δ	Yes	Yes	Yes	Yes	Yes	Yes
Observations	237,567	237,585	237,585	237,572	237,590	237,590
Pseudo/ adjusted R^2	0.26	0.26	0.26	0.03	0.03	0.03

Notes: Each column reports parameter estimates from Poisson Pseudo Maximum Likelihood (columns (1)-(3)) and OLS (columns (4)-(6)) regressions of technological impact on inventor-author status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980-2009. The dependent variable measures 10-year window forward patent citations (DOCDB-level), in columns (1)-(3), and the probability that a patent enters the upper 5th percentile of the primary patent class-year-impact distribution, in columns (4)-(6). "Author" indicates that at least one inventor appears as author on a biomedical scientific publication indexed in PubMed/ Web of Science SCI prior to the patent application. "Frontier authors" are inventors that authored a paper in a "top-general" journal in the prior 3 years. "Top-general" journals in biomedical sciences are New England Journal of Medicine, Lancet, JAMA, Nature, Science, Cell, Nature Medicine, British Medical Journal. "Top-field" journals are all other journals in the top-category (score = 5) of journals in a biomedical WOS-field in the ECOOM FEB journals ranking (2021). Inventor-author disambiguation is based on Li et al., 2014; Smalheiser and Torvik, 2009 and Torvik (2018). DB links established by the authors. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015). Heteroscedasticity robust standard errors in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

authors, which receive +51% more follow-up citations than patents without scientific authors (reference group). This implies a 2-3 times larger impact premium compared to patents with only non-frontier authors among inventors (receiving respectively + 19% more citations than non-author patents).

In column 3, we further split up the category of non-frontier authors to investigate the differential mechanism of quality and timing of involvement: The coefficients imply a cascading order of magnitude of excess impact for author patents, which is highest for frontier author patents (+53%), followed with a large gap by patents with authors of more dated contributions to top-general journals and recent authors in top-field journals (both +30%, not significantly different from each other[26]), non-recent authors in top-field journals (+20%), and patents with authors in non-top journals (+11%). This allows for making two important observations: First, there seems to be significant additional value for technology from top-authorship over authorship in general, which becomes evident by the comparison between patent impact of authors in top- vs. non-top journals, but also in the distinction between general interest and field top-journals. Second, the value of recency of involvement seems to play out differently among top-general and top-field science, when comparing the increments in patent value between the categories of recent (frontier) and non-recent authorship in top-general journals (+76%) and in top-field journals (+50%), implying that frontier knowledge entails a critical recency dimension, which depreciates fast in terms of utility for generating a technology premium, significantly faster than for non-frontier knowledge.

In columns 4 to 6 of Table 1.2, we move from average value to investigating differences in the likelihood of generating hit technologies, i.e. patents that enter the upper 5% of the impact distribution of a technology class-year cohort, among frontier, non-frontier and non-author patents. Results imply a qualitatively similar relationship, however estimated effects are economically significantly larger than for average impact: The excess likelihood of becoming a technological hit for patents with frontier authors is about 69% higher compared to patents without scientific authors among inventors (baseline), which is more than twice the estimated increment for patent with only non-frontier authors (column 5, all co-

[26] Wald tests indicate all coefficient pairs to be significantly different from each other below the 1%-level.

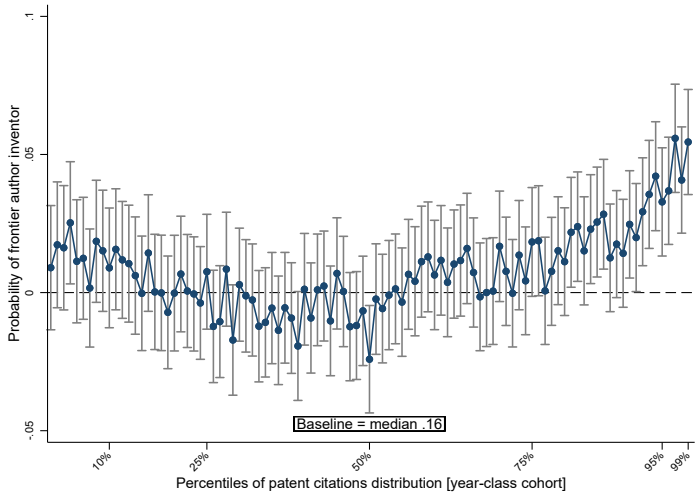
efficients significant at the 1%-level). This implies and important and crucial role of frontier scientists in the development of breakthrough technology. Noticeably, once controlling for contextual factors of breakthrough inventions, we find this adjusted excess rate of frontier author patents to be substantially larger compared to the non-parametric evidence reported in Table 1.1. The critical dimension of recency to frontier, and its relatively higher depreciation rate for technology value, is re-confirmed for hit patent rates in column 6.

We investigate robustness of these results in several ways. First, we estimate alternative specifications holding constant institutional context (firm fixed effects) and inventor/author level characteristics, such as the number of authors on a patent, their prior experience in patenting and publishing, as well as career tenure. Table 1.B1 in Appendix 1.B shows that our baseline estimates are fully robust and frontier author impact gains even more pronounced in these stricter specifications. We further show that effects on average citation impact are robust, yet likely understated in magnitude when using OLS log-linear estimation (Table 1.B2 in Appendix 1.B). Next, to mitigate concerns that the relationship between scientist involvement into private sector patenting might be affected by changes in national-level institutional regimes, such as Bayh-Dole type legislations (e.g. Mowery et al., 2001; Thompson, Ziedonis and Mowery, 2018), we additionally show robustness of results for within-country-year variation, while allowing regression residuals to be correlated within country-field pairs, in Table 1.B3 in Appendix 1.B. Finally, our findings are perfectly robust and very similar in sizes of effects when defining frontier authors based on recent 5% top-cited articles rather than journal outlets (Table 1.B4 in Appendix 1.B).

We further explore the importance of recency of involvement for frontier science, by shifting the cut-off used to delineate frontier authors to 5 and, respectively, 10 years prior to the patent application. Results reported in Table 1.B5 in Appendix 1.B show that the frontier author premium is robust for using alternative frontier timing, however decays in magnitude when shifting the frontier-recency cut-off more towards the past. We observe this pattern both for average patent citations and top-impact likelihood, consistent with the view of a rapidly moving frontier of knowledge and a depreciation of frontier insights as time passes since involvement.

In order to develop a better understanding of the dispersion of these effects,

Figure 1.2. Patent Impact and Frontier Author Incidence



Notes: The figure plots parameter estimates from OLS regression of likelihood of frontier author occurrence on percentile ranks in the patent citation distribution for primary class-year cohorts for the sample of all biomedical U. S. patents exclusively assigned to private firms (N = 237,585), granted between 1980–2009. The dependent variable measures the probability of a frontier author among the inventors listed on a patent. "Frontier authors" are inventors that authored a paper in a "top-general" journal in the past 3 years. "Top-general" journals in biomedical sciences are New England Journal of Medicine, Lancet, JAMA, Nature, Science, Cell, Nature Medicine, British Medical Journal. The regression model includes a full set of class x year, number of inventors, and firm patent stock decile fixed effects. 95% confidence intervals are based on heteroscedasticity robust standard errors. Inventor-author disambiguation is based on Li et al., 2014; Smalheiser and Torvik, 2009 and Torvik (2018). DB links established by the authors. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015).

we evaluate the incidence of frontier author patents across the technology impact distribution. Figure 1.2 shows that, relative to the median patent, the likelihood of frontier authors on the inventor team significantly increases for patents above the 75th percentile and displays another steep leap for patents in the top-regions above the 95th percentile, in line with the point estimate in column 5 of Table 1.2. However, we also find that frontier author patents are associated with a slightly above average probability of failure, measured as entering the bottom 25% of the impact distribution, as shown in Table 1.B7 in Appendix 1.B. This may imply that the, often cited, high risk-high gain nature associated with frontier science research

might be encountered also in technology endeavors of frontier authors.

Heterogeneity in frontier author impact

We further investigate heterogeneity of frontier author impacts in a number of ways. In Table 1.B6 in Appendix 1.B, we investigate whether frontier author impact in technology is likely contingent on the intensity of involvement in top-scientific research: We find that the premium of frontier authors is significantly stronger for patents of authors listed as first authors on frontier science publications, i.e. recent top-general journal articles, consistent with the view that frontier knowledge gains would be magnified for individuals who played a substantial role in top-scientific research. The value of substantial contribution (difference between coefficients for patents of first vs. non-first authors) for technological impact appears stronger for frontier science relative to the increment estimated for non-frontier science, and is especially pronounced for effects on technological breakthrough rates, i.e. the likelihood of becoming hit patents.

We explore heterogeneity of higher impact of frontier author patents over time and across technological fields. As shown in Figure 1.B1 in Appendix 1.B, the relatively higher level of citations to patents with frontier authors compared to non-frontier author and non-author patents is stable over time and persists for all years of our sample. However, with regards to top-impact likelihood, we observe a relative decline in author excess (both for frontier and non-frontier author patents) setting in towards the late 1990s and levelling out completely by 2002 (Figure 1.B2 in Appendix 1.B). The extent to which this apparent decline of the importance for frontier scientists for the creation of hit technologies in recent years might be related to the advent of information or meta-technologies, higher accessibility and connect between scientific methods and technology, or other factors, may provide an interesting question to be validated and explored by further research.

Lastly, we analyze heterogeneity of our estimates across fields within our sample, in order to inquire whether frontier author patents are particularly salient in certain technological areas.[27] Table 1.B8 in Appendix 1.B shows that the positive

[27] Patents within the category of 'Drugs&Medical' (Hall, Jaffe and Trajtenberg, 2001) are still rather heterogeneous, spanning from drug compounds (classes: 424, 435, 514), to biotechnology and genetically modified organisms (classes: 800, 930), to surgery and medical instruments (classes: 128, 600, 601, 602,

relationship between frontier author presence and superior impact persists across all involved subcategories (fields). The relatively highest impact of frontier author patents is observed in genetics / biotechnology, where the higher likelihood of top-impact of author patents is entirely concentrated on this group. This comes as not surprising considering the strong infancy state and frontier nature that this field was in during most of our sample period. Our results here reinforce prior findings on the pivotal role of top-scientists in the context of the genesis of this industry (Zucker, Darby and Brewer, 1998; Zucker, Darby and Armstrong, 2002). However, our estimates also indicate that frontier author patents are associated with robust superior impact also in fields like instruments, dentistry, optics and prothesis (subcategory: Other).

In the following section we aim to trace the sources of these frontier author impact advantages to differences in use of scientific inputs in their patents, in particular with regards to frontier science.

1.4.2 Frontier author patents and the use of frontier SNPRs

If patents of frontier authors generate more impact in technology, we expect this to be in part related to their access advantage and the better use of frontier knowledge in their patents. In order to investigate this particular mechanism as an explanation for frontier author impact, we rely on patent references to scientific prior art (SNPRs) and take advantage of the information these provide us with regard to the content and timing of scientific underpinnings of patented inventions. In order to evaluate whether the technology premium of frontier authors is related to differences in SNPRs made, we, first, investigate and confirm findings of prior literature whether references to frontier science, at different levels of recency and timing instances, are associated with superior impact of technology. We discuss findings from these estimation briefly in Section 1.4.2, and report more extensive results in Appendix 1.C of the paper. Second, we estimate regression models to inquire about the existence systematic differences relating to the use of and faster access to frontier SNPRs in patents with frontier authors among their inventors (Section 1.4.2). Third, we investigate to which extent the use of frontier SNPRs allows to explain

604, 606, 607) to a miscellaneous group including optics (351), dentistry (433) and prothesis (623).

the impact premia for average citation rates and hit likelihood of frontier author patents and whether their use of frontier knowledge, as captured in the SNPRs, differs in impact and hit probability generated from other users (Section 1.4.2).

Frontier SNPR links and technological impact

To investigate the relationship between SNPRs to frontier science in patents and technological impact, we estimate regression models of citation rates and hit likelihood on different types of SNPRs along the science quality and timing dimensions. A detailed discussion of our empirical specification can be found in Appendix 1.C of the paper.

Table 1.3 reports the main results from our estimation of the frontier SNPR patent impact relation. Our estimates confirm that science-referencing patents have superior impact and higher hit patent likelihood (colums 1 and 4). In particular, we estimate that these impact premia are substantially higher for patents making references to frontier-science publications, compared to patents with no SNPRs, but also to patents with SNPRs to less recent top-publications or articles in lower ranked journals or lower ranks of the scientific impact distributions (columns 2 and 5). In line with our prediction regarding the relation between frontier science and technological breakthrough, we observe largest coefficients on frontier SNPR dummies for regression models of hit patent likelihood: Patents relying on recent top-science have a more than +125% higher probability of entering the upper 5th percentile of the impact distribution, which is roughly threefold the excess probability of non-frontier SNPR patents (column 5). Estimates in columns 4 and 6 show that recency appears to matter significantly more as predictor for high-impact for patents with links to top-science, while it only has a marginal influence for patents relying exclusively on non-top SNPRs. This pattern is in line with prior findings of Poege et al. (2019).

However, the model coefficients of columns 4 and 6 further imply that – even among the top-ranked scientific outlets – there is persistent heterogeneity with regards to the importance of recency: The direct link between recent top-science and technological impact seems to be strongly concentrated on knowledge at the very frontier – in articles published in the top general interest journals -, while it

Table 1.3. Frontier SNPR Patents and Technological Impact

	PPML			OLS		
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	PatCit10	PatCit10	PatCit10	Top5% hit	Top5% hit	Top5% hit
SNPR	0.375***			0.033***		
	(0.008)			(0.001)		
Frontier SNPR		0.338***	0.403***		0.069***	0.070***
		(0.019)	(0.020)		(0.004)	(0.004)
Non-frontier SNPR		0.202***			0.020***	
		(0.012)			(0.002)	
SNPR, top-general jrn, >3y			0.284***			0.016***
			(0.018)			(0.003)
SNPR, top field-jrn, last 3y			0.294***			0.048***
			(0.018)			(0.003)
SNPR, top field-jrn, >3y			0.316***			0.016***
			(0.016)			(0.003)
SNPR, no top-journal			0.164***			0.018***
			(0.012)			(0.002)
Constant	3.254***	3.328***	3.307***	0.053***	0.055***	0.055***
	(0.004)	(0.006)	(0.006)	(0.001)	(0.001)	(0.001)
Patent class × year	Yes	Yes	Yes	Yes	Yes	Yes
Number of SNPRs	No	Yes	Yes	No	Yes	Yes
Firm patent stock × Δ	Yes	Yes	Yes	Yes	Yes	Yes
Observations	237,585	237,585	237,585	237,590	237,590	237,590
Pseudo/ adjusted R^2	0.26	0.27	0.27	0.03	0.03	0.03

Notes: Each column reports parameter estimates from Poisson Pseudo Maximum Likelihood (columns (1)-(3)) and OLS (columns (4)-(6)) regressions of technological impact on inventor-author status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980-2009. The dependent variable measures 10-year window forward patent citations (DOCDB-level), in columns (1)-(3), and the probability that a patent enters the upper 5th percentile of the primary patent class-year-impact distribution, in columns (4)-(6). "SNPR" (scientific reference) indicates at least one patent reference to a biomedical scientific article indexed in PubMed / Web of Science SCI. Examiner-given SNPRs are excluded from patents for which this information is available (granted after Jan 2001). "Frontier SNPRs" are references to papers in "top-general" journals published in 3 years or less before the patent application date. "Top-general" journals in biomedical sciences are New England Journal of Medicine, Lancet, JAMA, Nature, Science, Cell, Nature Medicine, British Medical Journal. "Top-field" journals are all other journals in the top-category (score = 5) of journals in a biomedical WOS-field in the ECOOM FEB journals ranking (2021). Scientific patent references to articles in PubMed/ SCI are sourced from Agarwal et al. (2014) and Marx and Fuegi (2020). Science citations are sourced from Microsoft Academic Graph. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015). Heteroscedasticity robust standard errors in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

rapidly diminishes for top-field publications and other scientific articles in lower ranked outlets. Here, our results provide evidence that the *frontier* nature of scientific knowledge underlying patents matters especially for high-impact effects. Extensive robustness of these findings when using alternative specifications and measurement is reported Appendix 1.C of the paper.

Table 1.4. Frontier SNPR Timing and Technological Impact

	PPML	OLS
	(1)	(2)
VARIABLES	PatCit10	Top5% hit
Frontier SNPR, first	0.313***	0.085***
	(0.028)	(0.006)
Frontier SNPR, not first	0.322***	0.071***
	(0.021)	(0.004)
Non-frontier SNPR, first	0.130***	0.040***
	(0.014)	(0.002)
Non-frontier SNPR, not first	0.222***	0.012***
	(0.012)	(0.002)
Constant	3.333***	0.054***
	(0.006)	(0.001)
Patent class × year	Yes	Yes
Number of SNPRs	Yes	Yes
Firm patent stock × Δ	Yes	Yes
Observations	237,585	237,590
Pseudo/ adjusted R^2	0.27	0.03

Notes: Columns report parameter estimates from Poisson Pseudo Maximum Likelihood (column (1)) and OLS (column (2)) regressions of technological impact on inventor-author status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980-2009. For details on variable measurement and data sources, see Table 1.3 and Section 1.3. Heteroscedasticity robust standard errors in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In Table 1.4 we explore heterogeneity of frontier SNPR impact for relative timing advantages, i.e. whether patents are the first to cite or followers in citing a given frontier article. Estimates in column 1 imply that first-mover advantages in frontier links do not translate to higher average citation impact. For non-frontier SNPRs priority is even inversely related to citation rates. However, we find that first-

time links affect technological hit likelihood significantly more (column 2). This is true for all types of SNPRs, but in particular also among frontier science referencing patents. First access to frontier science generates a +160% higher likelihood of hit patent relative to the background rate of no-SNPR patents. In Table 1.C4 in Appendix 1.C we show robustness of this pattern for ex-ante links to frontier science, i.e. early access references by patents already before the corresponding articles were published.

Frontier author patents and the likelihood of frontier SNPRs

Next, in this section we investigate whether patents with frontier authors differ from those without or non-frontier authors in the their references to science, in particular regarding the likelihood to link to frontier science and whether they are faster in establishing these links. For this purpose, we estimate linear probabilities for patents in the sample based equations of the following type:

$$SNPR_i = \alpha_0 + \beta_1 FrontierAuth_i + \sum_j \beta_j NonFrontierAuth_{j(i)} + X_i'\gamma + \delta_{c(i)t(i)} + \eta_{s(i)d(i)} + \epsilon_i \tag{1.2}$$

In its general form, the dependent variable of equation (1.2) is a binary indicator equal to 1 if a patent includes at least one scientific reference of the specific type, i.e. in case of frontier SNPRs to a top-general journal article published in 3 years or less before the patent application, and 0 otherwise. The predictors of interest are the same as in equation (1.1), namely a dummy variable for the presence of at least one frontier author among the inventors on a patent, and a series of dummies capturing effects for non-frontier author patents, with non-author patents constituting the reference category, providing the baseline probability for all estimation models.

Table 1.5 reports results from the estimation of equation (1.2) in sequential models. Coefficients in Column 1 confirm the finding in prior literature that patents of scientific authors are significantly more likely to make SNPRs (Gittelman and Kogut, 2003), but reveals that this additional likelihood is considerably higher for frontier author patents. Not only are these more likely to make such references at the extensive margin, but, conditional on citing science, frontier author patents also display a much higher number of SNPRs (column 2, estimated as PPML count

Table 1.5. Frontier Author Patents and Frontier SNPR Probability

	OLS	PPML		OLS	
	(1)	(2)	(3)	(4)	(5)
VARIABLES	SNPR [01]	# SNPRs	Frontier SNPR	Close frontier SNPR	First frontier SNPR
Frontier author	0.313***	0.853***	0.133***	0.088***	-0.012
	(0.004)	(0.021)	(0.005)	(0.005)	(0.014)
Non-frontier author	0.171***	0.371***	-0.005*		-0.027**
	(0.002)	(0.016)	(0.003)		(0.013)
Constant	0.291***	2.290***	0.131***	0.007***	0.286***
	(0.001)	(0.015)	(0.002)	(0.001)	(0.012)
with SNPR only		Yes	Yes	Yes	Yes
with frontier SNPR only				Yes	Yes
with author only				Yes	
Patent class × year	Yes	Yes	Yes	Yes	Yes
Number of inventors	Yes	Yes	Yes	Yes	Yes
Number of SNPRs	No	No	Yes	Yes	Yes
Firm patent stock × Δ	Yes	Yes	Yes	Yes	Yes
Observations	237,591	95,190	95,190	12,062	13,661
Pseudo/ adjusted R^2	0.30	0.20	0.19	0.10	0.09

Notes: Each column reports parameter estimates from OLS (columns (1), (3)-(5)) and Poisson Pseudo Maximum Likelihood (column (2)) regressions of scientific reference on inventor-author status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980-2009. The dependent variable (in columns (1), (3)-(5)) measures the probability that a patent makes a prior art reference (SNPR) to a scientific article indexed in PubMed/ Web of Science SCI. Examiner-given SNPRs are excluded from patents for which this information is available (granted after Jan 2001). "Frontier SNPRs" are references to papers in "top-general" journals published in 3 years or less before the patent application date. "Top-general" journals in biomedical sciences are New England Journal of Medicine, Lancet, JAMA, Nature, Science, Cell, Nature Medicine, British Medical Journal. "Frontier authors" are inventors that authored a paper in a "top-general" journal in the prior 3 years. "Close SNPRs" are self-citations or citations to the co-author network of at least one listed inventor. Inventor-author disambiguation is based on Li et al., 2014; Smalheiser and Torvik, 2009 and Torvik (2018). DB links established by the authors. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015). Scientific patent references to articles in PubMed/ SCI are sourced from Agarwal et al. (2014) and Marx and Fuegi (2020). Science citations are sourced from Microsoft Academic Graph. Heteroscedasticity robust standard errors in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

model). This underlines the importance of accounting for the number of references made, in order to be able to differentiate sheer volume effects from de-facto, higher probabilities of links.

As main insight in this section we find that, conditional on the same number of SNPRs, patents of frontier authors are significantly more likely to make frontier SNPRs (column 3, significant at the 1%-level). Such patents are approximately twice as likely to include references to articles at the knowledge frontier than patents without authors among their inventors. Especially noticeable, our estimates in column 3 also reveal that patents of authors who are not at the frontier of science do not significantly differ from patents without authors in their likelihood of frontier SNPR. The corresponding coefficient for non-frontier author patents is very small in size and significant only below the 10%-level.

In columns 4-5, we investigate differences in frontier SNPR likelihood conditional on closeness to own scientific work and relative timing of links. Among patents with authors making frontier SNPRs, the group with frontier authors is particularly more likely to rely on science closer to their own networks (column 4). While this is partly implied by construction (self-citations), it provides support the view of increased access through active participation. With regards to relative timing, however, we do not find frontier author patents to be more likely than patents without authors to make first-time references to frontier science, once we condition on frontier SNPRs (column 5). This implies that shorter lags to frontier science observed for frontier author patents (compare Table 1.1) seem to originate from a generally stronger reliance on frontier knowledge rather than from faster access per se. We find consistent patterns or ex-ante access links as well as for alternative specifications and measurement of frontier science, based on top5% highly cited publications, reported in Tables 1.D1 and 1.D2 in Appendix 1.D.

Frontier author patents, frontier SNPRs and high-impact technologies

Findings from the previous sections have established, first, that frontier SNPRs are associated with patents of higher impact, especially hit technologies, and, second, that patents with frontier authors are substantially more likely to make such frontier SNPRs, in particular those close to their own scientific work. At this point in our

investigation, we are interested to which extent these differentials in use of SNPRs may allow to explain the superior technological performance of patents of frontier authors (Section 1.4.1) and whether frontier patents differ in impact, which we attributed to the use of frontier science, also once we hold links to frontier science (and other scientific references) constant.

For this purpose, we combine author and SNPR parameters from our previous specifications into estimating equations taking the following augmented functional form:

$$Y_i = f(FrontierAuth_i, FrontierSNPR_i, NonFrontierAuth_i, NonFrontierSNPR_i, \boldsymbol{X}_i', e_i) \tag{1.3}$$

Similar to Sections 1.4.1 and 1.4.2, with the dependent variable Y_i we measure forward citation counts and top-impact likelihood of a focal patent. The right-hand side of specifications contains the predictors of interest in form of exclusive category binary indicators for patents with frontier, non-frontier and no authors, but we now combine these with the SNPR indicators and decompose frontier author status further by frontier SNPR, non-frontier SNPR and no SNPR patents. Similarly, we construct the same categories for non-frontier author and no author patents. This allows us to scrutinize whether higher frontier author impact reflects better use of frontier SNPRs or, eventually, a technology premium of frontier author patents which persists independent from the use of frontier SNPRs. In the main specifications, we include fixed effects for both the number of inventors and the number of SNPRs, next to class by year and firm patent stocks characteristics.

Table 1.6 reports the corresponding results. In column 1, estimates along the vector of coefficients implies that patents which are in the combined group of frontier author inventors and making frontier SNPRs capture the relatively highest average impact, across all groups relative to the baseline of no author patents not making SNPRs. Noticeably, patents with frontier SNPRs are those subcategories with the highest within-category impact premia also for the non-frontier author and no author patents categories. This further corroborates the view of frontier science exerting major influence on technological value. Moreover, when comparing across author-categories, we find that frontier author patents show the relatively largest impact effects, in each subcategory of frontier SNPR, non-frontier SNPR

Table 1.6. Frontier Author Patents, SNPRs and Technological Impact

	PPML	OLS
	(1)	(2)
VARIABLES	PatCit10	Top5% hit
Frontier author, frontier SNPR	0.537***	0.086***
	(0.029)	(0.006)
Frontier author, non-frontier SNPR	0.364***	0.031***
	(0.024)	(0.004)
Frontier author, no SNPR	0.423***	0.031***
	(0.029)	(0.004)
Non-frontier author, frontier SNPR	0.401***	0.078***
	(0.022)	(0.004)
Non-frontier author, non-frontier SNPR	0.289***	0.028***
	(0.014)	(0.002)
Non-frontier author, no SNPR	0.157***	0.017***
	(0.009)	(0.001)
No author, frontier SNPR	0.373***	0.065***
	(0.039)	(0.008)
No author, non-frontier SNPR	0.267***	0.030***
	(0.015)	(0.003)
Constant	3.249***	0.047***
	(0.008)	(0.001)
Patent class × year	Yes	Yes
Number of inventors	Yes	Yes
Number of SNPRs	Yes	Yes
Firm patent stock × Δ	Yes	Yes
Observations	237,585	237,590
Pseudo/ adjusted R^2	0.28	0.03

Notes: Each column reports parameter estimates from Poisson Pseudo Maximum Likelihood (column (1)) and OLS (column (2)) regressions of technological impact on inventor-author and scientific reference status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980-2009. For details on variable measurement and data sources, see Table 1.2, Table 1.3 and Section 1.3. Examiner-given SNPRs are excluded from patents for which this information is available (granted after Jan 2001). Heteroscedasticity robust standard errors in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

and no SNPR patents. In particular in the subcategories of no SNPR patents, their estimated excess impact is almost three times (+270%) higher compared to that of patents with only non-frontier authors.

This pattern is consistent for regressions of top5% hit patent likelihood, in reported in column 2. Here, the excess likelihood of frontier author patents with frontier SNPRs to become a hit is even multiple times larger relative to the background rate probability of non-author no SNPR patents (+183%), and generally frontier SNPR effects are particularly pronounced. However, the predictive surplus of frontier author patents without frontier SNPRs, relative to non-frontier author patents, appears significantly reduced with regards to technological breakthrough. This becomes particularly evident in the subcategory of patents with non-frontier SNPRs, where our estimates do not show significant differences across inventor-author type categories.

The nuanced differences across models of average impact and top-impact likelihood estimation highlight an interesting duality regarding the relationship between frontier author patents, frontier SNPRs and value in technology: While frontier SNPRs, and differences in the use of frontier SNPRs, do not fully explain the value contribution of frontier author patents on neither dimension, they appear to be much more important in predicting hit patents. This finding is in line with the idea that authors at the frontier of knowledge are particularly conducive to technological breakthrough given their better access and capability to translate this frontier knowledge into technological applications. Differences in scientific references made by these authors in patents seem to resemble this mechanism rather well.

In Table 1.7 we investigate impact heterogeneity for (frontier) author science links based on proximity and relative timing in references. Coefficient estimates in column 1 imply that a relatively higher excess impact share of frontier author patents can be attributed to those inventions relying on frontier science close to the authors own work, either through self-citations or references to their scientific co-author network. We estimate the impact premium of these patents to be more than 50% higher compared to patent of frontier authors relying on more distant frontier science as prior art. This differential is consistent when evaluating the probability of patents to become technological hits (column 3), whereas effects here are again

Table 1.7. Frontier Author Patents and Technological Impact – Closeness & Timing to Frontier SNPRs

	PPML	PPML	OLS	OLS
	(1)	(2)	(3)	(4)
VARIABLES	PatCit10	PatCit10	Top5% hit	Top5% hit
Frontier author, frontier SNPR, close	0.631***		0.106***	
	(0.038)		(0.009)	
Frontier author, frontier SNPR, not close	0.445***		0.069***	
	(0.036)		(0.007)	
Non-frontier author, frontier SNPR, close	0.362***		0.066***	
	(0.027)		(0.005)	
Non-frontier author, frontier SNPR, not close	0.438***		0.091***	
	(0.028)		(0.006)	
Frontier author, frontier SNPR, first		0.416***		0.088***
		(0.045)		(0.010)
Frontier author, frontier SNPR, not first		0.582***		0.085***
		(0.033)		(0.007)
Non-frontier author, frontier SNPR, first		0.445***		0.090***
		(0.034)		(0.007)
Non-frontier author, frontier SNPR, not first		0.384***		0.073***
		(0.024)		(0.005)
Author, non-frontier & no SNPR interactions	Yes	Yes	Yes	Yes
No author, SNPR interactions	Yes	Yes	Yes	Yes
Constant	3.249***	3.249***	0.047***	0.047***
	(0.008)	(0.008)	(0.001)	(0.001)
Patent class × year	Yes	Yes	Yes	Yes
Number of inventors	Yes	Yes	Yes	Yes
Number of SNPRs	Yes	Yes	Yes	Yes
Firm patent stock × Δ	Yes	Yes	Yes	Yes
Observations	237,585	237,585	237,590	237,590
Pseudo/ adjusted R^2	0.28	0.28	0.03	0.03

Notes: Each column reports parameter estimates from Poisson Pseudo Maximum Likelihood (column (1)) and OLS (column (2)) regressions of technological impact on inventor-author and scientific reference status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980-2009. The dependent variable measures 10-year window forward patent citations (DOCDB-level), in columns (1)-(2), and the probability that a patent enters the upper 5th percentile of the primary patent class-year-impact distribution, in columns (3)-(4). "Frontier authors" are inventors that authored a paper in a "top-general" journal in PubMed / Web of Science SCI in the prior 3 years. "Top-general" journals in biomedical sciences are New England Journal of Medicine, Lancet, JAMA, Nature, Science, Cell, Nature Medicine, British Medical Journal. "Frontier SNPRs" are references to papers in "top-general" journals published in 3 years or less before the patent application date. "Close SNPRs" are self-citations or citations to the co-author network of at least one listed inventor. Examiner-given SNPRs are excluded (for patents granted after Jan 2001). For details on data sources, see Table 1.2, Table 1.3 and Section 1.3 Heteroscedasticity robust standard errors in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

economically larger relative to the background rate of non-author patents.

This observation provides a potentially important insight into the contribution of frontier scientists to technological progress and the underlying mechanism: Frontier authors achieve highest impact in inventions where they directly access and use frontier knowledge related to their own capability and expertise. As implied by the estimates in column 1 and 3 in Table 1.7, their value contribution to technology is not significantly different when linking to more distant, though recent, top-science. But the difference they make seems to originate from the direct translation of their own insight and contribution to frontier science. This generates a performance premium which is unmatched by authors and other inventors more distant to that specific type of knowledge. This finding highlights the pivotal importance of active involvement as a governing access mechanism to breakthrough science and its immediate translation into technological applications with highly superior impact.

Furthermore, the model estimates in columns 2 and 4 of Table 1.7 suggest that the performance advantage of frontier author patents may be more related to better use, hence a superior scientific capability, rather than timing, i.e. faster access. The coefficients of frontier author patents in column 2 imply a larger average impact rate of follower patents in links to frontier science compared to first-mover links. With regards to predicting technological hits, there is no significant difference between these two categories of frontier author-science links. For non-frontier author patents, the opposite seems to be true and first-time links to frontier science are consistently associated with significantly higher average value and hit likelihood. This diverging pattern across patents in the two author categories of our data may reflect the higher variability in outcomes and higher risk of failure that is associated with first-mover links to frontier knowledge, which frontier authors make significantly more often due to their much higher reliance on frontier science. We find some, yet limited, evidence for higher hit probability of frontier author ex-ante access links to frontier science, compared to their ex-post links. However, this pattern is consistent for non-frontier author patents (see Table 1.E4 in Appendix 1.E.

Qualitative robustness of the results in this section is provided in Tables 1.E1 to 1.E3 in Appendix 1.E, including estimates using stricter specifications accounting for inventor/ author level characteristics and firm fixed effects, and sensitivity for

alternative measurement of top-science, defined as 5% most highly cited papers.

1.5 Conclusion

Understanding the mechanisms through which frontier science supports and contributes to technological performance is an important quest, which is of equal interest to scholars, decision makers and societies at large. Only recently, economists have started to examine and validate predictions of this connect empirically.

In this paper, we investigate the contribution of inventors simultaneously involved in the production of frontier scientific research for the advancement of technology, i.e. the creation of superior value invention and, in particular, hit patents. We explore whether, how and to which extend frontier scientists push technological performance by facilitating this linkage between frontier science and invention through more immediate access and better use of new top-scientific ideas in technological applications. Based on our empirical analysis of technological impact, inventor-author links and science references of all corporate biomedical U.S. patents between 1980 to 2009, we make several contributions to the literature: First, our findings imply that patents with authors participating at the scientific frontier generate considerably above-average impact in technology and are associated with a substantially higher probability of top-impact, relative to other author or non-author patents. This relationship is strongly dependent on the quality level of their scientific output and the recency of involvement, lending support to the frontier nature of their scientific engagement to drive this connection. Second, we find that patent referencing frontier science, i.e. recent top-scientific research – measured in citation links to articles in top-prestigious scientific journals or that enter the far-right tail of the science quality distribution from the previous 3 years – have significantly higher impact and technology hit likelihood compared to patents with less timely or lower quality references and to patents not relying on scientific prior art. Our results imply that early access to science, in particular first-mover access, is strong predictor for hit technologies, however, to a lesser extent for frontier than non-frontier science. This suggests that recency may play a more important role than priority in valuable access and use of frontier science.

Third, we find that such references to new frontier science are systematically

(much) more likely to occur in patents with frontier authors and that, in line with predictions, this access premium is particularly strong for articles closer to their own scientific output and network. However, frontier authors are not necessarily more likely to establish priority links to frontier science, beyond what is explained by their general stronger reliance on this knowledge in patents.

Fourth, our findings suggest that frontier authors might be particularly good in applying frontier science, as evidenced by impact differentials between frontier author patents with references to frontier science, other frontier author patents, and patents of other authors (and non-author patents) with references to frontier science. We find that the frontier author-science link in patents is a particularly strong predictor for hit patents, i.e. breakthrough technology. This is an important insight, given the non-obviousness of this direct link, as emphasized by prior literature (Allen et al., 1977; Gittelman and Kogut, 2003; Gruber, Harhoff and Hoisl, 2013). Further analyses suggest that frontier author links to frontier science that is in their close proximity – self- and network citations – are particularly likely to lead to such high-impact and hit technologies.

Our findings entail several normative implications. If the involvement of corporate inventors in fundamental science enables access and use of new scientific insights in industry technology which generate superior social value, this makes a strong case for the importance of industry participation in basic research. It also raises potential concerns of welfare losses arising due to the recently observed decline of basic science investment in corporate R&D (Arora, Belenzon and Patacconi, 2018), which might endanger the functioning of such fragile yet high-value generating connections effectively bridging public upstream research and downstream industrial development. This provides an important implication not only for policy but also for management and resource allocation decisions in R&D strategy. On the other hand, firms may seek other, promising channels to engage with the basic science community (as, e. g., suggested by the recent work of Baruffaldi and Poege (2020) and Arora, Belenzon and Dionisi (2021)), either through informal or formal collaboration. While this calls upon policy makers for an inclusive design of institutional regimes governing involvement of academic scientists in corporate technology, the question about which of these channels (corporate participation in basic research, academic commercialization, or hybrid forms) might be most

effective for knowledge transfer is an on-going and interesting avenue for future research.

Our findings also highlight, once more, the uneven distributions in scientific knowledge production and the importance of frontier research for economic prosperity, often emphasized in central theories and empirical contributions in the economics and philosophy of science. In this paper we provide evidence that frontier science and frontier technology have a direct and intimate connection, transcending the traditional distinctions between basic and applied research.

Considering that the immediate gains from this connection are usually proceeded by significant risk taking and uncertainty of outcomes, this calls for particular attention of policy makers and businesses to the specific needs of frontier research when it comes facilitating conditions and resource allocation.

Lastly, our empirical insights, obviously, reflect non-random matches and patterns of strong selectivity in the universe of data points that we study. Specific intervention or program evaluation studies are needed and may be an interesting avenue for future research, in order to causally validate some of the mechanisms that our broad-scale evidence implies.

References

Acs, Zoltan J, and David B Audretsch. 1990. *Innovation and small firms.* Mit Press.

Agarwal, Sneha, Miles Lincoln, Haoyan Cai, and Vetle I Torvik. 2014. "Patci—a tool for identifying scientific articles cited by patents." GSLIS Research Showcase.

Agrawal, Ajay, and Rebecca Henderson. 2002. "Putting patents in context: Exploring knowledge transfer from MIT." *Management science*, 48(1): 44–60.

Agrawal, Ajay, John McHale, and Alexander Oettl. 2017. "How stars matter: Recruiting and peer effects in evolutionary biology." *Research Policy*, 46(4): 853–867.

Ahmadpoor, Mohammad, and Benjamin F Jones. 2017. "The dual frontier: Patented inventions and prior scientific advance." *Science*, 357(6351): 583–587.

Allen, Thomas J, et al. 1977. "Managing the flow of technology: Technology transfer and the dissemination of technological information within the R&D organization." *MIT Press Books*, 1.

Arora, Ashish, Sharon Belenzon, and Andrea Patacconi. 2018. "The decline of science in corporate R&D." *Strategic Management Journal*, 39(1): 3–32.

Arora, Ashish, Sharon Belenzon, and Bernardo Dionisi. 2021. "First-mover Advantage and the Private Value of Public Science." National Bureau of Economic Research.

Arora, Ashish, Sharon Belenzon, and Lia Sheer. 2021. "Knowledge spillovers and corporate investment in scientific research." *American Economic Review*, 111(3): 871–98.

Arrow, Kenneth. 1962. "Economic Welfare and the Allocation of Resources for Invention (pp. 609-626)." *Universities-National Bureau Committee for Economic and Council (ed.), The Rate and Direction of Inventive Activity: Economic and Social Factors, Princeton.*

Arts, Sam, and Reinhilde Veugelers. 2020. "Taste for science, academic boundary spanning, and inventive performance of scientists and engineers in industry." *Industrial and Corporate Change*, 29(4): 917–933.

Audretsch, David B, and Paula E Stephan. 1996. "Company-scientist locational links: The case of biotechnology." *The American economic review*, 86(3): 641–652.

Azoulay, Pierre, Joshua S Graff Zivin, and Jialan Wang. 2010. "Superstar extinction." *The Quarterly Journal of Economics*, 125(2): 549–589.

Azoulay, Pierre, Joshua S Graff Zivin, Danielle Li, and Bhaven N Sampat. 2019. "Public R&D investments and private-sector patenting: evidence from NIH funding rules." *The Review of economic studies*, 86(1): 117–152.

Balconi, Margherita, Stefano Breschi, and Francesco Lissoni. 2004. "Networks of inventors and the role of academia: an exploration of Italian patent data." *Research policy*, 33(1): 127–145.

Balconi, Margherita, Stefano Brusoni, and Luigi Orsenigo. 2010. "In defence of the linear model: An essay." *Research policy*, 39(1): 1–13.

Baruffaldi, Stefano, and Felix Poege. 2020. "A Firm Scientific Community: Industry Participation and Knowledge Diffusion." 20-10.

Belenzon, Sharon, and Mark Schankerman. 2013. "Spreading the word: Geography, policy, and knowledge spillovers." *Review of Economics and Statistics*, 95(3): 884–903.

Berkes, Enrico, and Peter Nencka. 2020. "Knowledge access: The effects of Carnegie libraries on innovation."

Bianchi, Nicola, and Michela Giorcelli. 2020. "Scientific education and innovation: from technical diplomas to university STEM degrees." *Journal of the European Economic Association*, 18(5): 2608–2646.

Bikard, Michaël, and Matt Marx. 2020. "Bridging academia and industry: How geographic hubs connect university science and corporate technology." *Management Science*, 66(8): 3425–3443.

Bonaccorsi, Andrea, and Grid Thoma. 2007. "Institutional complementarity and inventive performance in nano science and technology." *Research policy*, 36(6): 813–831.

Breschi, Stefano, and Christian Catalini. 2010. "Tracing the links between science and technology: An exploratory analysis of scientists' and inventors' networks." *Research Policy*, 39(1): 14–26.

Bryan, Kevin A, and Yasin Ozcan. 2016. "The impact of open access mandates on invention." *The Review of Economics and Statistics*, 1–45.

Bush, Vanevar. 1945. "The endless frontier." *National Science Foundation–EUA. Washington*.

Callaert, Julie, Bart Van Looy, Arnold Verbeek, Koenraad Debackere, and Bart Thijs. 2006. "Traces of prior art: An analysis of non-patent references found in patent documents." *Scientometrics*, 69(1): 3–20.

Callaert, Julie, Maikel Pellens, and Bart Van Looy. 2014. "Sources of inspiration? Making sense of scientific references in patents." *Scientometrics*, 98(3): 1617–1629.

Cassiman, Bruno, Reinhilde Veugelers, and Pluvia Zuniga. 2008. "In search of performance effects of (in) direct industry science links." *Industrial and Corporate Change*, 17(4): 611–646.

Cockburn, Iain M, and Rebecca M Henderson. 1998. "Absorptive capacity, coauthoring behavior, and the organization of research in drug discovery." *The journal of industrial economics*, 46(2): 157–182.

Cohen, Wesley M, and Daniel A Levinthal. 1989. "Innovation and learning: the two faces of R & D." *The economic journal*, 99(397): 569–596.

Cohen, Wesley M, Richard R Nelson, and John P Walsh. 2002. "Links and impacts: the influence of public research on industrial R&D." *Management science*, 48(1): 1–23.

Colen, Relinde Maria, Rene Belderbos, Bart Leten, and Stijn Kelchtermans. 2014. "Reaching for the stars: the contingent performance effects of basic research collaboration." DRUID working paper, https://conference. druid. dk/acc_papers ….

Correia, Sergio, Paulo Guimaraes, and Thomas Zylkin. 2019. "PPMLHDFE: Stata module for Poisson pseudo-likelihood regression with multiple levels of fixed effects."

Debackere, Koenraad, and Reinhilde Veugelers. 2005. "The role of academic technology transfer organizations in improving industry science links." *Research policy*, 34(3): 321–342.

Dosi, Giovanni. 1982. "Technological paradigms and technological trajectories: a suggested interpretation of the determinants and directions of technical change." *Research policy*, 11(3): 147–162.

Fabrizio, Kira R. 2009. "Absorptive capacity and the search for innovation." *Research policy*, 38(2): 255–267.

Fierro, Gabe. 2014. "Processing USPTO patent data." *Coleman Fung Institute for Engineering Leadership*.

Fleming, Lee. 2001. "Recombinant uncertainty in technological search." *Management science*, 47(1): 117–132.

Fleming, Lee, and Olav Sorenson. 2004. "Science as a map in technological search." *Strategic management journal*, 25(8-9): 909–928.

Furman, Jeffrey L, and Scott Stern. 2011. "Climbing atop the shoulders of giants: The impact of institutions on cumulative research." *American Economic Review*, 101(5): 1933–63.

Gambardella, Alfonso. 1992. "Competitive advantages from in-house scientific research: The US pharmaceutical industry in the 1980s." *Research policy*, 21(5): 391–407.

Gittelman, Michelle, and Bruce Kogut. 2003. "Does good science lead to valuable knowledge? Biotechnology firms and the evolutionary logic of citation patterns." *Management Science*, 49(4): 366–382.

Giuri, Paola, Myriam Mariani, Stefano Brusoni, Gustavo Crespi, Dominique Francoz, Alfonso Gambardella, Walter Garcia-Fontes, Aldo Geuna, Raul Gonzales, Dietmar Harhoff, et al. 2007. "Inventors and invention processes in Europe: Results from the PatVal-EU survey." *Research policy*, 36(8): 1107–1127.

Godin, Benoît. 2006. "The linear model of innovation: The historical construction of an analytical framework." *Science, Technology, & Human Values*, 31(6): 639–667.

Griliches, Zvi. 1979. "Issues in assessing the contribution of research and development to productivity growth." *The bell journal of economics*, 92–116.

Griliches, Zvi. 1986. "Productivity, R&D, and the Basic Research at the Firm Level in the 1970's." *American Economic Review*, 76(1): 141–54.

Gruber, Marc, Dietmar Harhoff, and Karin Hoisl. 2013. "Knowledge recombination across technological boundaries: Scientists vs. engineers." *Management Science*, 59(4): 837–851.

Hall, Bronwyn H, Adam B Jaffe, and Manuel Trajtenberg. 2001. "The NBER patent citation data file: Lessons, insights and methodological tools."

Hall, Bronwyn H, Adam Jaffe, and Manuel Trajtenberg. 2005. "Market value and patent citations." *RAND Journal of economics*, 16–38.

Harhoff, Dietmar, Francis Narin, Frederic M Scherer, and Katrin Vopel. 1999. "Citation frequency and the value of patented inventions." *Review of Economics and statistics*, 81(3): 511–515.

Henderson, Rebecca, and Iain Cockburn. 1996. "Scale, scope, and spillovers: The determinants of research productivity in drug discovery." *The Rand journal of economics*, 32–59.

Henderson, Rebecca, Gary P Pisano, and Luigi Orsenigo. 1999. "The pharmaceutical industry and the revolution in molecular biology: interactions among scientific, institutional, and organizational change."

Hicks, Diana, Anthony Breitzman Sr, Kimberly Hamilton, and Francis Narin. 2000. "Research excellence and patented innovation." *Science and Public Policy*, 27(5): 310–320.

Iaria, Alessandro, Carlo Schwarz, and Fabian Waldinger. 2018. "Frontier knowledge and scientific production: evidence from the collapse of international science." *The Quarterly Journal of Economics*, 133(2): 927–991.

Jaffe, Adam B. 1989. "Real effects of academic research." *The American economic review*, 957–970.

Jones, Charles I. 1995. "R & D-based models of economic growth." *Journal of political Economy*, 103(4): 759–784.

Kuhn, Thomas Samuel. 1962. *The Structure of Scientific Revolutions.* Chicago: University of Chicago Press.

Li, Danielle, Pierre Azoulay, and Bhaven N Sampat. 2017. "The applied value of public investments in biomedical research." *Science*, 356(6333): 78–81.

Li, Guan-Cheng, Ronald Lai, Alexander D'Amour, David M Doolin, Ye Sun, Vetle I Torvik, Z Yu Amy, and Lee Fleming. 2014. "Disambiguation and co-authorship networks of the US patent inventor database (1975–2010)." *Research Policy*, 43(6): 941–955.

Magerman, Tom, Bart Van Looy, and Koenraad Debackere. 2015. "Does involvement in patenting jeopardize one's academic footprint? An analysis of patent-paper pairs in biotechnology." *Research Policy*, 44(9): 1702–1713.

Mansfield, Edwin. 1991. "Academic research and industrial innovation." *Research policy*, 20(1): 1–12.

Mansfield, Edwin. 1995. "Academic research underlying industrial innovations: sources, characteristics, and financing." *The review of Economics and Statistics*, 55–65.

Marco, Alan C, Michael Carley, Steven Jackson, and Amanda Myers. 2015. "The uspto historical patent data files: Two centuries of innovation."

Marx, Matt, and Aaron Fuegi. 2020. "Reliance on science: Worldwide front-page patent citations to scientific articles." *Strategic Management Journal*, 41(9): 1572–1594.

Meyer, Martin. 2000. "Does science push technology? Patents citing scientific literature." *Research policy*, 29(3): 409–434.

Mowery, David C, Richard R Nelson, Bhaven N Sampat, and Arvids A Ziedonis. 2001. "The growth of patenting and licensing by US universities: an assessment of the effects of the Bayh–Dole act of 1980." *Research policy*, 30(1): 99–119.

Mukherjee, Satyam, Daniel M Romero, Ben Jones, and Brian Uzzi. 2017. "The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: The hotspot." *Science advances*, 3(4): e1601315.

Murray, Fiona. 2002. "Innovation as co-evolution of scientific and technological networks: exploring tissue engineering." *Research policy*, 31(8-9): 1389–1403.

Murray, Fiona, and Scott Stern. 2007. "Do formal intellectual property rights hinder the free flow of scientific knowledge?: An empirical test of the anti-commons hypothesis." *Journal of Economic Behavior & Organization*, 63(4): 648–687.

Murray, Fiona, Philippe Aghion, Mathias Dewatripont, Julian Kolev, and Scott Stern. 2016. "Of mice and academics: Examining the effect of openness on innovation." *American Economic Journal: Economic Policy*, 8(1): 212–52.

Narin, Francis, Kimberly S Hamilton, and Dominic Olivastro. 1997. "The increasing linkage between US technology and public science." *Research policy*, 26(3): 317–330.

Nelson, Richard. 1962. "The link between science and invention: The case of the transistor. The rate and direction of inventive activity."

Oettl, Alexander. 2012. "Reconceptualizing stars: Scientist helpfulness and peer performance." *Management Science*, 58(6): 1122–1140.

Poege, Felix, Dietmar Harhoff, Fabian Gaessler, and Stefano Baruffaldi. 2019. "Science quality and the value of inventions." *Science advances*, 5(12): eaay7323.

Roach, Michael, and Wesley M Cohen. 2013. "Lens or prism? Patent citations as a measure of knowledge flows from public research." *Management Science*, 59(2): 504–525.

Romer, Paul M. 1986. "Increasing returns and long-run growth." *Journal of political economy*, 94(5): 1002–1037.

Romer, Paul M. 1990. "Endogenous technological change." *Journal of political Economy*, 98(5, Part 2): S71–S102.

Rosenberg, Nathan. 1990. "Why do firms do basic research (with their own money)?" *Research Policy*, 19(2): 165–174.

Scherer, Frederic M, and Dietmar Harhoff. 2000. "Technology policy for a world of skew-distributed outcomes." *Research policy*, 29(4-5): 559–566.

Schumpeter, Joseph Alois. 1942. *Socialism, capitalism and democracy.* Harper and Brothers.

Silva, JMC Santos, and Silvana Tenreyro. 2006. "The log of gravity." *The Review of Economics and statistics*, 88(4): 641–658.

Silva, JMC Santos, and Silvana Tenreyro. 2011. "Further simulation evidence on the performance of the Poisson pseudo-maximum likelihood estimator." *Economics Letters*, 112(2): 220–222.

Smalheiser, Neil R, and Vetle I Torvik. 2009. "Author name disambiguation." *Annual review of information science and technology*, 43(1): 1–43.

Stephan, Paula E. 1996. "The economics of science." *Journal of Economic literature*, 34(3): 1199–1235.

Thompson, Neil C, Arvids A Ziedonis, and David C Mowery. 2018. "University licensing and the flow of scientific knowledge." *Research Policy*, 47(6): 1060–1069.

Tijssen, Robert JW. 2001. "Global and domestic utilization of industrial relevant science: patent citation analysis of science–technology interactions and knowledge flows." *Research Policy*, 30(1): 35–54.

Toole, Andrew A. 2012. "The impact of public basic research on industrial innovation: Evidence from the pharmaceutical industry." *Research Policy*, 41(1): 1–12.

Toole, Andrew A, and Dirk Czarnitzki. 2007. "Biomedical academic entrepreneurship through the SBIR program." *Journal of Economic Behavior & Organization*, 63(4): 716–738.

Torvik, Vetle I. 2018. "Author-Linked data for Author-ity 2009." University of Illinois at Urbana-Champaign.

Torvik, Vetle I, and Neil R Smalheiser. 2009. "Author name disambiguation in MEDLINE." *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(3): 1–29.

Trajtenberg, Manuel. 1990. "A penny for your quotes: patent citations and the value of innovations." *The Rand journal of economics*, 172–187.

Veugelers, Reinhilde, and Jian Wang. 2019. "Scientific novelty and technological impact." *Research Policy*, 48(6): 1362–1372.

Wang, Dashun, Chaoming Song, and Albert-László Barabási. 2013. "Quantifying long-term scientific impact." *Science*, 342(6154): 127–132.

Watzinger, Martin, and Monika Schnitzer. 2019. "Standing on the shoulders of science." CEPR Discussion Paper No. DP13766.

Weitzman, Martin L. 1998. "Recombinant growth." *The Quarterly Journal of Economics*, 113(2): 331–360.

Williams, Heidi L. 2013. "Intellectual property rights and innovation: Evidence from the human genome." *Journal of Political Economy*, 121(1): 1–27.

Wuchty, Stefan, Benjamin F Jones, and Brian Uzzi. 2007. "The increasing dominance of teams in production of knowledge." *Science*, 316(5827): 1036–1039.

Zucker, Lynne G, and Michael R Darby. 1996. "Star scientists and institutional transformation: Patterns of invention and innovation in the formation of the biotechnology industry." *Proceedings of the National Academy of Sciences*, 93(23): 12709–12716.

Zucker, Lynne G, Michael R Darby, and Jeff S Armstrong. 2002. "Commercializing knowledge: University science, knowledge capture, and firm performance in biotechnology." *Management science*, 48(1): 138–153.

Zucker, Lynne, Michael Darby, and Marilynn Brewer. 1998. "Intellectual Human Capital and the Birth of US Biotechnology Enterprises." *American Economic Review*, 88(1): 290–306.

Appendices

1.A Appendix A

Details on author & SNPR disambiguation in patents

Inventor-author links. To identify inventors simultaneously (or previously) active in basic scientific discovery, we establish a link between all individuals listed as U.S. patent inventors (n = 2,665,709) between 1975 to 2011 in the Li et al. (2014) database and all author identities in the *Author-ity* 2009 disambiguated database of author names (n = 9,243,805) in PubMed up until 2009 Smalheiser and Torvik, 2009; Torvik and Smalheiser, 2009. For this, we use the author-linked data provided by Torvik (2018) and match each author-patent instance to the series of all unique inventor identifiers in the same patent in Li et al. (2014). This results into an interim data set, having author IDs (*au_id*) and author name strings on the left hand side and a vector of possible inventor IDs (*lower*) and their name strings on the right-hand side, which are connected to a joint patent number.

In a next step, we use standard within-patent probabilistic name matching based on Jaro-Winkler similarity and Levinstein Distance across all author-patent-inventor instances in the data to determine the unique name pairs with the highest match probabilities and define a series of thresholds above which we accept a match to be unambiguous.[28] In Level 1 of this procedure, we accept only unique matches of author-inventor pairs with a Jaro-Winkler similarity score above 80% based on full name strings. In Level 2, out of the remaining, we accept all those pairs that score high (>90%) on last name on both similarity measures, in case a match is unique above this threshold. In Level 3, finally, we accept those that yield unique matches on a threshold above 90% for Jaro-Winkler only, but slightly lower for Levinstein. In this way, we are ultimately able to match 295,694 out of the 309,395

[28] Jaro-Winkler similarity is a comparative measure yielding a score between 0 and 100 for the partial agreement between any two strings, taking into account the length of a string and also, in part, common human errors in spelling of names and number series. The Levinstein Distance is another commonly used similarity measure which calculates the number of string changes required from transforming one string into another. We use Jaro-Winkler as main algorithm, and Levinstein D to confirm accepted matches.

authors linked to a patent in the Torvik (2018) data (95,67%). Out of these 293,249 (99,08%) are uniquely disambiguated, 2,534 (0.86%) are matched to two Li et al. (2014) inventor IDs, 141 (0.05%) are linked to three inventor IDs, and 40 (0.01%) to four to maximum seven inventor IDs. We keep these very few cases of multiple matches in our sample, given their occurrence is unlikely to be correlated to any of the outcomes in our analysis. The remaining, very small, fraction of author-patent cases, for which we could not establish an automatic link are basically impossible to disambiguate even manually (usually very short or truncated name strings).

Next, we conduct a validation exercise of our matching approach. For this, we randomly sampled 200 candidates of inventor-author links (including null links, for which we do extensive web search) from patents in the sample, and manually compare inventor name, location, assignee, co-inventors, narrow topic of patent to the matched author name, affiliation, location, co-authors, field of their (in time) closest publication(s) in PubMed. We further compare career patents' assignees, location, co-inventors to career publications' affiliations, locations, co-authors for matched pairs to assess their temporal co-evolution and plausibility. We find that, based on this procedure, 93.3% of inventor-author links are correctly identified (95.4% of positive links); 4.5% are very likely false positive author-patent links in the Torvik (2018) data. The remaining 2.2% are false negatives from our probabilistic name matching approach. In our validation set, no link observations were false positives. SQL-code for our inventor-author match is available upon request.

Scientific references (SNPRs). We use two data sources to identify references to scientific articles in U.S. patents; First, the *PatCi* algorithm (Agarwal et al., 2014), which disambiguates patent references to all articles in PubMed until 2013. Second, we rely on the *PCS* dataset from Marx and Fuegi (2020), which provides references to all articles in Microsoft Academic Graph, including articles in PubMed.

We compare these two sources vis-a-vis with regards to their accuracy (precision and recall) for the patents in our sample: We extract two random samples of 100 SNPR links (including null links), one based on *PatCi*, one based on *PCS*. We then manually check whether the reference listed (and disambiguated) on the USPTO patent PDF document corresponds to the linked publication in PubMed online, to identify false positives. Second, we validate whether the other disambiguation

finds the same match, to identify conditional false negatives. Finally, we repeat this procedure for the second extracted random sample. We find that both datasets have a precision of 100% (no false positive links) in the validation set. However, while all PubMed SNPRs of *PCS* are correctly found in *PatCi*, *PCS* fails to detect about 22% of total references that are in *PatCi*. Given the importance of accurately identified references to PubMed for our paper, we therefore use *PatCi* as our main SNPR disambiguation. However, we use *PCS* to replenish information on examiner added references (available for patents granted after January 2001), which is not available in *PatCi*.

Finally, we evaluate what the restriction on only SNPR to PubMed implies for generalizability and coverage of total scientific references made by patents in our sample. For this purpose, we compare the overlap between total references to scientific publications in *PCS*/ Microsoft Academic Graph (MAG) (spanning all scientific disciplines) and references to PubMed indexed articles. This allows us to infer about how many links to non-biomedical science supporting the inventions in our sample we might miss, i.e. not observe in our data. We find this share of links to be very small and likely to be innocuous for internal validity of our estimates: In total, *PCS* identify 1,879,926 unique SNPRs from patents in our sample to MAG. In contrast, with *PatCi* we observe only 1,617,074 unique SNPRs for the same patents to PubMed (86% of total MAG links). While 14% of missed links might still seem relatively considerable, we find this share to be significantly reduced when considering only references to (peer-reviewed) scientific journals in Web of Science, indexed in the Journal Citations Report (JCR): Our PubMed/ WOS SCI SNPRs cover 94.9% of the total 1,492,540 references to scientific articles in WOS JCR that are made by patents in our sample, implying that only 5% of links go to scientific discoveries outside of what we defined as the relevant population for our paper. This reinforces our belief of our results to be representative for the universe of biomedical sciences.

Additional summary & exploratory statistics

Table 1.A1. Summary Statistics

VARIABLES	Obs	Mean	SD	Median	75th	90th	Min	Max
Patent forward citations (10y)	239,181	24.84	41.07	12	28	58	0	966
Top5% hit patent	239172	0.07	0.25	0	0	0	0	1
Number of inventors	239,181	2.99	2.12	2	4	6	1	32
Inventor-author	239,163	0.59	0.49	1	1	1	0	1
Number of inventor-authors	239,163	1.22	1.56	1	2	3	0	26
Frontier author (top-general jrn)	239,181	0.07	0.25	0	0	0	0	1
Frontier author (top5% paper)	239,181	0.13	0.33	0	0	1	0	1
Scientific reference (SNPR)	239,181	0.40	0.49	0	1	1	0	1
Number of SNPRs	239181	5.03	16.54	0	3	13	0	569
Frontier SNPR (top-general jrn)	239,181	0.06	0.24	0	0	0	0	1
Frontier SNPR (top5% paper)	239,181	0.10	0.30	0	0	0	0	1
Citation lag top-general SNPR (y)	34,051	6.26	6.94	4	8	14	-9	73
Citation lag top5% SNPR (y)	47,942	5.08	5.75	4	7	12	-17	63
Patent-paper pair	239,181	0.02	0.13	0	0	0	0	1
Inventor prior patents (max)	239,163	21.84	38.43	10	24	50	1	1,807
Inventor prior publications (max)	239,163	12.52	26.51	2	13	36	0	326
Applicant firm patent stock	239,180	403	1,014	57	374	1,025	0	31,559

Notes: Our sample contains all biomedical U.S. patents exclusively assigned to private firms, granted between 1980-2009. Forward patent citations are measured over a 10-year window (DOCDB-level). Top 5% hits are patents that enter the upper 5th percentile of the primary patent class-year-impact distribution. "Inventor-author" indicates that at least one inventor appears as author on a biomedical scientific publication indexed in PubMed/ Web of Science SCI prior to the patent application. "Frontier authors" are inventors that authored a paper in a "top-general" journal (or alternatively a top 5% paper) in the prior 3 years. "Top-general" journals in biomedical sciences are New England Journal of Medicine, Lancet, JAMA, Nature, Science, Cell, Nature Medicine, British Medical Journal. "Top5% papers" are scientific articles that enter the upper 5th percentile of the citation distribution of a WOS subject category field year cohort. "SNPR" (scientific reference) indicates at least one patent reference to a biomedical scientific article indexed in PubMed/ Web of Science SCI. "Frontier SNPRs" are references to papers in "top-general" journals (or alternatively top 5% papers) published in 3 years or less before the patent application date. "Patent-paper pairs" are patents citing a scientific paper with 100% inventor-author overlap and published in direct temporal proximity to the patent application. Inventor-author disambiguation is based on Li et al., 2014; Smalheiser and Torvik, 2009 and Torvik (2018). DB links established by the authors. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015). Scientific patent references to articles in PubMed/ SCI are sourced from Agarwal et al. (2014) and Marx and Fuegi (2020). Science citations are sourced from Microsoft Academic Graph.

Table 1.A2. Author-Level Summary Statistics

VARIABLES	Obs	Mean	SD	Median	75th	90th	Min	Max
Prior publications	291,203	14.47	24.65	6	15	36	1	326
Prior publications (all PubMed)	291,203	18.95	31.84	8	20	48	1	397
Years since first publication	291,203	11.25	8.87	9	16	24	0	88
Years since last publication	291,203	2.17	3.68	1	3	7	0	55
Prior co-authors	291,203	40.47	57.81	21	48	96	0	1,742
Top-article citations pct (10y)	291,203	70.91	30.41	83	97	100	1	100
Frontier author (top-general jrn)	291,203	0.08	0.27	0	0	0	0	1
Frontier author (top5% paper)	291,203	0.16	0.37	0	0	1	0	1
Prior patents	291,203	19.15	40.17	9	20	42	1	648
Years since first patent	291,203	7.92	6.92	6	12	18	0	103
Prior co-inventors	291,203	15.71	18.58	10	20	35	0	718
Number of unique authors	59.455							

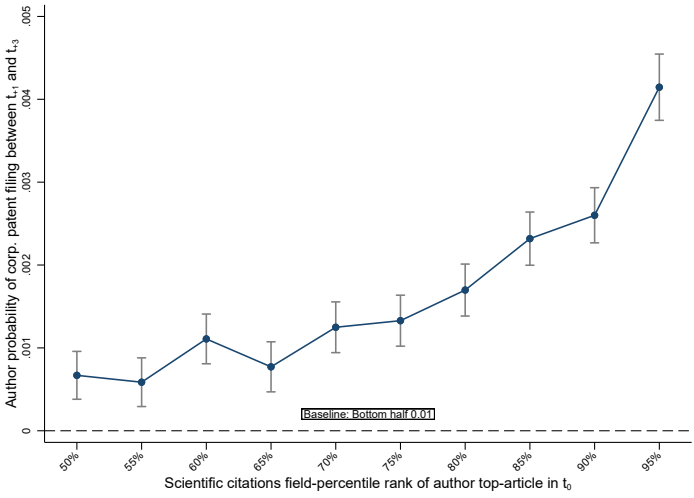
Notes: The table provides author-level statistics for all inventors appearing as author on a biomedical scientific publication indexed in PubMed/ Web of Science SCI prior to patent application for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980–2009. "Frontier authors" are inventors that authored a paper in a "top-general" journal (or alternatively a top 5% paper) in the prior 3 years. "Top-general" journals in biomedical sciences are New England Journal of Medicine, Lancet, JAMA, Nature, Science, Cell, Nature Medicine, British Medical Journal. "Top5% papers" are scientific articles that enter the upper 5th percentile of the citation distribution of a WOS subject category field year cohort. "SNPR" (scientific reference) indicates at least one patent reference to a biomedical scientific article indexed in PubMed/ Web of Science SCI. "Frontier SNPRs" are references to papers in "top-general" journals (or alternatively top 5% papers) published in 3 years or less before the patent application date. "Patent-paper pairs" are patents citing a scientific paper with 100% inventor-author overlap and published in direct temporal proximity to the patent application. Inventor-author disambiguation is based on Li et al., 2014; Smalheiser and Torvik, 2009 and Torvik (2018). DB links established by the authors. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015). Scientific patent references to articles in PubMed/ SCI are sourced from Agarwal et al. (2014) and Marx and Fuegi (2020). Science citations are sourced from Microsoft Academic Graph.

Table 1.A3. Author Selection into Industry Patenting

	OLS				
VARIABLES	(1) CorpPat3y	(2) CorpPat3y	(3) CorpPat3y	(4) CorpPat3y	(5) CorpPat3y
Frontier author (top-general jrn)	0.0096***	0.0016***			
	(0.0002)	(0.0002)			
Author in top field-jrn	0.0078***	0.0015***			
	(0.0001)	(0.0001)			
Frontier author (top5% paper)			0.0163***	0.0037***	0.0033***
			(0.0002)	(0.0002)	(0.0002)
Top10% paper author			0.0092***	0.0020***	0.0018***
			(0.0002)	(0.0002)	(0.0002)
Constant	0.0108***	0.0137***	0.0133***	0.0137***	0.0133***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
WOS field × year	Yes	No	Yes	No	No
Year only	No	Yes	No	Yes	Yes
Author	No	Yes	No	Yes	Yes
Patent-paper pairs excl.	No	No	No	No	Yes
Observations	15,416,284	131,00,464	15,416,284	13,100,464	13,100,464
Adjusted R^2	0.02	0.47	0.02	0.47	0.47

Notes: Each column reports parameter estimates from OLS regression of likelihood of entrance into corporate patenting on top-author status for the yearly cross-section of all authors on articles in PubMed/ Web of Science published between 1980 and 2005. The dependent variable measures the probability that an author appears as inventor on a U.S. patent assigned to a private corporation within 3 years following the publication year of the focal scientific article. "Top-general" journals in biomedical sciences are New England Journal of Medicine, Lancet, JAMA, Nature, Science, Cell, Nature Medicine, British Medical Journal. "Top-field" journals are all other journals in the top-category (score = 5) of journals in a biomedical WOS-field in the ECOOM FEB journals ranking (2021). "Top5% papers" are scientific articles that enter the upper 5th percentile of the citation distribution of a WOS subject category field year cohort. "Patent-paper pairs" are patents citing a scientific paper with 100% inventor-author overlap and published in direct temporal proximity to the patent application. Inventor-author disambiguation is based on Li et al., 2014; Smalheiser and Torvik, 2009 and Torvik (2018). DB links established by the authors. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015). Scientific patent references to articles in PubMed/ SCI are sourced from Agarwal et al. (2014) and Marx and Fuegi (2020). Science citations are sourced from Microsoft Academic Graph. Standard errors in parentheses are heteroscedasticity robust in columns (1)-(3) and clustered at the author-level in columns(2)-(4). Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure 1.A1. Author Selection into Industry Patenting – within Authors Comparison



Notes: The figure plots parameter estimates from OLS regression of likelihood of entrance into corporate patenting on author top-article citations rank for the yearly cross-section of all authors on articles (N = 13,100,464) in PubMed/ Web of Science published between 1980 and 2005. The dependent variable measures the probability that an author appears as inventor on a U. S. patent assigned to a private corporation within 3 years following the publication year of the focal scientific article. The right-hand side / x-axis indexes percentiles of the 10-year citation distribution of a WOS subject category field year cohort. The estimation model includes a full set of author and publication year fixed effects. "Patent-paper pairs" are excluded. 95% confidence intervals are based on author-level clustered standard errors. "Patent-paper pairs" are excluded. 95% confidence intervals are based on heteroscedasticity robust standard errors. Inventor-author disambiguation is based on Li et al., 2014; Smalheiser and Torvik, 2009 and Torvik (2018). DB links established by the authors. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015). Scientific patent references to articles in PubMed/ SCI are sourced from Agarwal et al. (2014) and Marx and Fuegi (2020). Science citations are sourced from Microsoft Academic Graph.

1.B Appendix B

Table 1.B1. Frontier Author Patents and Technological Impact - Robustness

	PPML			OLS		
VARIABLES	(1) PatCit10	(2) PatCit10	(3) PatCit10	(4) Top5% hit	(5) Top5% hit	(6) Top5% hit
Author	0.065***			0.008***		
	(0.008)			(0.001)		
Frontier author		0.057***	0.084***		0.005*	0.012***
		(0.014)	(0.017)		(0.003)	(0.003)
Author, top-general jrn, >3y			0.025*			0.008***
			(0.015)			(0.003)
Author, top field-jrn, last 3y			0.052***			0.009***
			(0.012)			(0.002)
Author, top field-jrn, >3y			-0.005			-0.002
			(0.013)			(0.002)
Log(inventor age)	Yes	No	No	Yes	No	No
Log(# prior patents)	Yes	Yes	Yes	Yes	Yes	Yes
Log(# prior publications)	No	Yes	Yes	No	Yes	Yes
Constant	3.435***	3.589***	3.582***	0.071***	0.095***	0.094***
	(0.010)	(0.015)	(0.015)	(0.002)	(0.003)	(0.003)
Patent class × year	Yes	Yes	Yes	Yes	Yes	Yes
Number of inventors	Yes	Yes	Yes	Yes	Yes	Yes
Number of authors	Yes	Yes	Yes	Yes	Yes	Yes
Applicant firm	Yes	Yes	Yes	Yes	Yes	Yes
Observations	223,320	134,205	134,205	223,395	134,245	134,245
Pseudo/ adjusted R^2	0.48	0.49	0.49	0.13	0.14	0.14

Notes: Each column reports parameter estimates from Poisson Pseudo Maximum Likelihood (columns (1)-(3)) and OLS (columns (4)-(6)) regressions of technological impact on inventor-author status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980-2009. The dependent variable measures 10-year window forward patent citations (DOCDB-level), in columns (1)-(3), and the probability that a patent enters the upper 5th percentile of the primary patent class-year-impact distribution, in columns (4)-(6). "Author" indicates that at least one inventor appears as author on a biomedical scientific publication indexed in PubMed/ Web of Science SCI prior to the patent application. "Frontier authors" are inventors that authored a paper in a "top-general" journal in the prior 3 years. "Top-general" journals in biomedical sciences are New England Journal of Medicine, Lancet, JAMA, Nature, Science, Cell, Nature Medicine, British Medical Journal. "Top-field" journals are all other journals in the top-category (score = 5) of journals in a biomedical WOS-field in the ECOOM FEB journals ranking (2021). Inventor-author disambiguation is based on Li et al., 2014; Smalheiser and Torvik, 2009 and Torvik (2018). DB links established by the authors. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015). Heteroscedasticity robust standard errors in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.B2. Frontier Author Patents and Technological Impact - Log(cits) OLS

	OLS		
VARIABLES	(1) Log(1+PatCit10)	(2) Log(1+PatCit10)	(3) Log(1+PatCit10)
Author	0.130***		
	(0.005)		
Frontier author		0.284***	0.294***
		(0.010)	(0.010)
Non-frontier author		0.117***	
		(0.005)	
Author, top-general jrn, >3y			0.179***
			(0.008)
Author, top field-jrn, last 3y			0.174***
			(0.007)
Author, top field-jrn, >3y			0.103***
			(0.008)
Author, no top-journal			0.069***
			(0.006)
Constant	2.491***	2.488***	2.484***
	(0.004)	(0.004)	(0.004)
Patent class × year	Yes	Yes	Yes
Number of inventors	Yes	Yes	Yes
Firm patent stock × Δ	Yes	Yes	Yes
Observations	237,573	237,591	237,591
Pseudo/ adjusted R^2	0.26	0.26	0.26

Notes: Each column reports parameter estimates from OLS regressions of technological impact on inventor-author status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980–2009. The dependent variable measures natural logarithm(1 + count of 10-year forward patent citations) (DOCDB-level). "Author" indicates that at least one inventor appears as author on a biomedical scientific publication indexed in PubMed/ Web of Science SCI prior to the patent application. "Frontier authors" are inventors that authored a paper in a "top-general" journal in the prior 3 years. "Top-general" journals in biomedical sciences are New England Journal of Medicine, Lancet, JAMA, Nature, Science, Cell, Nature Medicine, British Medical Journal. "Top-field" journals are all other journals in the top-category (score = 5) of journals in a biomedical WOS-field in the ECOOM FEB journals ranking (2021). Inventor-author disambiguation is based on Li et al., 2014; Smalheiser and Torvik, 2009 and Torvik (2018). DB links established by the authors. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015). Heteroscedasticity robust standard errors in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.B3. Frontier Author Patents and Technological Impact – Country-level Residuals

	PPML			OLS		
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	PatCit10	PatCit10	PatCit10	Top5% hit	Top5% hit	Top5% hit
Author	0.171***			0.014***		
	(0.016)			(0.001)		
Frontier author		0.333***	0.344***		0.023***	0.023***
		(0.044)	(0.046)		(0.004)	(0.004)
Non-frontier author		0.159***			0.013***	
		(0.017)			(0.002)	
Author, top-general jrn, >3y			0.189***			0.012***
			(0.018)			(0.002)
Author, top field-jrn, last 3y			0.228***			0.018***
			(0.026)			(0.003)
Author, top field-jrn, >3y			0.148***			0.008***
			(0.038)			(0.002)
Author, no top-journal			0.121***			0.013***
			(0.020)			(0.002)
Constant	3.329***	3.324***	3.319***	0.058***	0.058***	0.058***
	(0.010)	(0.010)	(0.010)	(0.001)	(0.001)	(0.001)
Patent class	Yes	Yes	Yes	Yes	Yes	Yes
Country × year	Yes	Yes	Yes	Yes	Yes	Yes
Number of inventors	Yes	Yes	Yes	Yes	Yes	Yes
Applicant firm	Yes	Yes	Yes	Yes	Yes	Yes
Observations	237,955	237,973	237,973	237,954	237,972	237,972
Pseudo/ adjusted R^2	0.28	0.28	0.28	0.03	0.03	0.03

Notes: Each column reports parameter estimates from Poisson Pseudo Maximum Likelihood (columns (1)-(3)) and OLS (columns (4)-(6)) regressions of technological impact on inventor-author status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980-2009. The dependent variable measures 10-year window forward patent citations (DOCDB-level), in columns (1)-(3), and the probability that a patent enters the upper 5th percentile of the primary patent class-year-impact distribution, in columns (4)-(6). "Author" indicates that at least one inventor appears as author on a biomedical scientific publication indexed in PubMed/ Web of Science SCI prior to the patent application. "Frontier authors" are inventors that authored a paper in a "top-general" journal in the prior 3 years. "Top-general" journals in biomedical sciences are New England Journal of Medicine, Lancet, JAMA, Nature, Science, Cell, Nature Medicine, British Medical Journal. "Top-field" journals are all other journals in the top-category (score = 5) of journals in a biomedical WOS-field in the ECOOM FEB journals ranking (2021). Inventor-author disambiguation is based on Li et al., 2014; Smalheiser and Torvik, 2009 and Torvik (2018). DB links established by the authors. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015). Standard errors in parentheses are clustered at the country × class level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.B4. Frontier Author Patents and Technological Impact – Top5% Paper Authors

	PPML			OLS		
VARIABLES	(1) PatCit10	(2) PatCit10	(3) PatCit10	(4) Top5% hit	(5) Top5% hit	(6) Top5% hit
Author	0.189***			0.019***		
	(0.007)			(0.001)		
Frontier author		0.391***	0.400***		0.036***	0.037***
		(0.012)	(0.012)		(0.002)	(0.002)
Non-frontier author		0.149***			0.016***	
		(0.008)			(0.001)	
Author, top5% paper, > 3y			0.288***			0.026***
			(0.013)			(0.002)
Author, no top5% paper			0.119***			0.014***
			(0.008)			(0.001)
Constant	3.304***	3.294***	3.291***	0.055***	0.054***	0.054***
	(0.005)	(0.006)	(0.006)	(0.001)	(0.001)	(0.001)
Patent class × year	Yes	Yes	Yes	Yes	Yes	Yes
Number of inventors	Yes	Yes	Yes	Yes	Yes	Yes
Firm patent stock × Δ	Yes	Yes	Yes	Yes	Yes	Yes
Patent-paper pairs excl.	Yes	Yes	Yes	Yes	Yes	Yes
Observations	233,402	233,420	233,420	233,407	233,425	233,425
Pseudo/ adjusted R^2	0.26	0.26	0.26	0.03	0.03	0.03

Notes: Each column reports parameter estimates from Poisson Pseudo Maximum Likelihood (columns (1)-(3)) and OLS (columns (4)-(6)) regressions of technological impact on inventor-author status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980-2009. The dependent variable measures 10-year window forward patent citations (DOCDB-level), in columns (1)-(3), and the probability that a patent enters the upper 5th percentile of the primary patent class-year-impact distribution, in columns (4)-(6). "Author" indicates that at least one inventor appears as author on a biomedical scientific publication indexed in PubMed/ Web of Science SCI prior to the patent application. "Frontier authors" are inventors that authored a "top5%" paper in the prior 3 years. "Top5% papers" are scientific articles that enter the upper 5th percentile of the citation distribution of a WOS subject category field year cohort. "Patent-paper pairs" are patents citing a scientific paper with 100% inventor-author overlap and published in direct temporal proximity to the patent application. Inventor-author disambiguation is based on Li et al., 2014; Smalheiser and Torvik, 2009 and Torvik (2018). DB links established by the authors. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015). Scientific patent references to articles in PubMed/ SCI are sourced from Agarwal et al. (2014) and Marx and Fuegi (2020). Science citations are sourced from Microsoft Academic Graph. Heteroscedasticity robust standard errors in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.B5. Frontier Author Patents and Technological Impact – Alternative Frontier Timing

Frontier-science cut-off	≤ 3y	≤ 5y	≤ 10y	≤ 3y	≤ 5y	≤ 10y
		PPML			OLS	
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	PatCit10	PatCit10	PatCit10	Top5% hit	Top5% hit	Top5% hit
Frontier author	0.412***	0.382***	0.358***	0.037***	0.035***	0.036***
	(0.015)	(0.014)	(0.012)	(0.003)	(0.002)	(0.002)
Non-frontier author	0.176***	0.173***	0.167***	0.017***	0.017***	0.016***
	(0.007)	(0.008)	(0.008)	(0.001)	(0.001)	(0.001)
Constant	3.293***	3.293***	3.292***	0.054***	0.054***	0.054***
	(0.005)	(0.005)	(0.005)	(0.001)	(0.001)	(0.001)
Patent class × year	Yes	Yes	Yes	Yes	Yes	Yes
Number of inventors	Yes	Yes	Yes	Yes	Yes	Yes
Firm patent stock × Δ	Yes	Yes	Yes	Yes	Yes	Yes
Observations	237,567	237,585	237,585	237,572	237,590	237,590
Pseudo/ adjusted R^2	0.26	0.26	0.26	0.03	0.03	0.03

Notes: Each column reports parameter estimates from Poisson Pseudo Maximum Likelihood (columns (1)-(3)) and OLS (columns (4)-(6)) regressions of technological impact on inventor-author status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980-2009. The dependent variable measures 10-year window forward patent citations (DOCDB-level), in columns (1)-(3), and the probability that a patent enters the upper 5th percentile of the primary patent class-year-impact distribution, in columns (4)-(6). "Author" indicates that at least one inventor appears as author on a biomedical scientific publication indexed in PubMed/ Web of Science SCI prior to the patent application. "Frontier authors" are inventors that authored a paper in a "top-general" journal in the prior 3, 5 or 10 years, depending on the defined cut-off. "Top-general" journals in biomedical sciences are New England Journal of Medicine, Lancet, JAMA, Nature, Science, Cell, Nature Medicine, British Medical Journal. "Top-field" journals are all other journals in the top-category (score = 5) of journals in a biomedical WOS-field in the ECOOM FEB journals ranking (2021). Inventor-author disambiguation is based on Li et al., 2014; Smalheiser and Torvik, 2009 and Torvik (2018). DB links established by the authors. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015). Heteroscedasticity robust standard errors in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.B6. Frontier Author Patents and Technological Impact – Splits by First Authors

	PPML			OLS		
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	PatCit10	PatCit10	PatCit10	Top5% hit	Top5% hit	Top5% hit
First author	0.213***			0.020***		
	(0.008)			(0.001)		
Author, never first author	0.144***			0.017***		
	(0.010)			(0.002)		
Frontier author, first		0.509***	0.484***		0.060***	0.058***
		(0.024)	(0.024)		(0.004)	(0.004)
Frontier author, not first		0.280***	0.238***		0.017***	0.015***
		(0.018)	(0.018)		(0.003)	(0.003)
Non-frontier author, first		0.189***			0.016***	
		(0.008)			(0.001)	
Non-frontier author, not first		0.153***			0.016***	
		(0.009)			(0.001)	
Auth, top-gen, >3y, first			0.227***			0.017***
			(0.015)			(0.002)
Auth, top-gen, >3y, not first			0.192***			0.018***
			(0.014)			(0.002)
Auth, top field, ≤3y, first			0.221***			0.021***
			(0.014)			(0.002)
Auth, top field, ≤3y, not first			0.163***			0.019***
			(0.014)			(0.002)
Auth, top field, >3y			0.146***			0.010***
			(0.012)			(0.002)
Author, no top-journal			0.073***			0.009***
			(0.009)			(0.001)
Constant	3.297***	3.295***	3.320***	0.054***	0.055***	0.057***
	(0.005)	(0.005)	(0.005)	(0.001)	(0.001)	(0.001)
Patent class × year	Yes	Yes	Yes	Yes	Yes	Yes
Number of inventors	Yes	Yes	Yes	Yes	Yes	Yes
Firm patent stock × Δ	Yes	Yes	Yes	Yes	Yes	Yes
Observations	237,567	237,585	237,585	237,572	237,590	237,590
Pseudo/ adjusted R^2	0.26	0.26	0.26	0.03	0.03	0.03

Notes: Each column reports parameter estimates from Poisson Pseudo Maximum Likelihood (columns (1)-(3)) and OLS (columns (4)-(6)) regressions of technological impact on inventor-author status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980–2009. For details on variable measurement and data sources, see Table 1.B1 and Section 1.3. Heteroscedasticity robust standard errors in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.B7. Frontier Author Patents and Technological Impact - Bottom vs. Top-Impact

	OLS				
	(1)	(2)	(3)	(4)	(5)
VARIABLES	Bottom25%	Top50%	Top25%	Top5%	Top1%
Frontier author	0.010***	0.049***	0.066***	0.037***	0.011***
	(0.004)	(0.005)	(0.004)	(0.003)	(0.001)
Non-frontier author	−0.001	0.017***	0.025***	0.017***	0.005***
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
Constant	0.792***	0.542***	0.277***	0.054***	0.011***
	(0.001)	(0.002)	(0.002)	(0.001)	(0.000)
Patent class × year	Yes	Yes	Yes	Yes	Yes
Number of inventors	Yes	Yes	Yes	Yes	Yes
Firm patent stock × Δ	Yes	Yes	Yes	Yes	Yes
Observations	237,590	237,590	237,590	237,590	237,590
Pseudo/ adjusted R^2	0.02	0.02	0.03	0.03	0.02

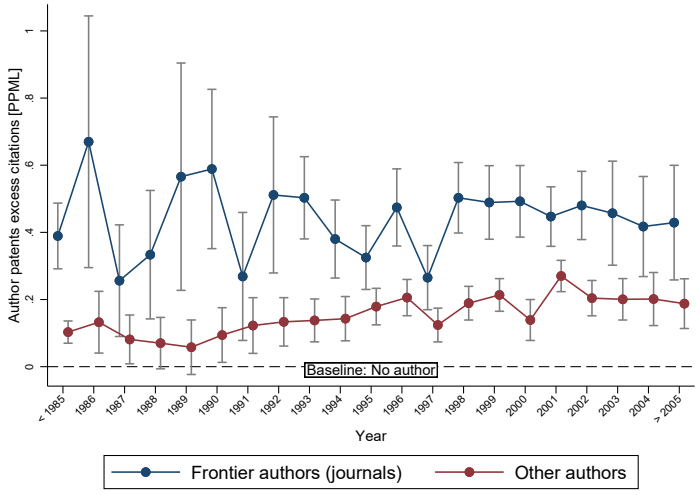
Notes: Each column reports parameter estimates from OLS regressions of technological impact on inventor-author status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980-2009. The dependent variable measures the probability that a patent enters the indicated percentile rank of it primary patent class-year impact distribution. "Author" indicates that at least one inventor appears as author on a biomedical scientific publication indexed in PubMed/ Web of Science SCI prior to the patent application. "Frontier authors" are inventors that authored a paper in a "top-general" journal in the prior 3 years. "Top-general" journals in biomedical sciences are New England Journal of Medicine, Lancet, JAMA, Nature, Science, Cell, Nature Medicine, British Medical Journal. "Top-field" journals are all other journals in the top-category (score = 5) of journals in a biomedical WOS-field in the ECOOM FEB journals ranking (2021). Inventor-author disambiguation is based on Li et al., 2014; Smalheiser and Torvik, 2009 and Torvik (2018). DB links established by the authors. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015). Heteroscedasticity robust standard errors in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.B8. Frontier Author Patents and Technological Impact – Field Splits

Technology fields	Drugs	Instr	Genetics	Other	Drugs	Instr	Genetics	Other
	PPML				OLS			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	PatCit10	PatCit10	PatCit10	PatCit10	Top5% hit	Top5% hit	Top5% hit	Top5% hit
Frontier author	0.369***	0.348***	0.544***	0.479***	0.038***	0.029***	0.051***	0.060**
	(0.018)	(0.036)	(0.107)	(0.079)	(0.003)	(0.008)	(0.017)	(0.024)
Non-frontier author	0.146***	0.184***	0.373***	0.172***	0.016***	0.015***	0.005	0.024***
	(0.010)	(0.011)	(0.098)	(0.025)	(0.002)	(0.002)	(0.009)	(0.005)
Constant	2.764***	3.657***	2.440***	3.853***	0.043***	0.056***	0.058***	0.049***
	(0.009)	(0.008)	(0.076)	(0.017)	(0.001)	(0.001)	(0.005)	(0.003)
Patent class × year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of inventors	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm patent stock × Δ	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	120,628	67,215	4,989	13,218	120,628	67,214	4,989	13,218
Pseudo/ adjusted R^2	0.11	0.17	0.09	0.35	0.01	0.01	0.01	0.01

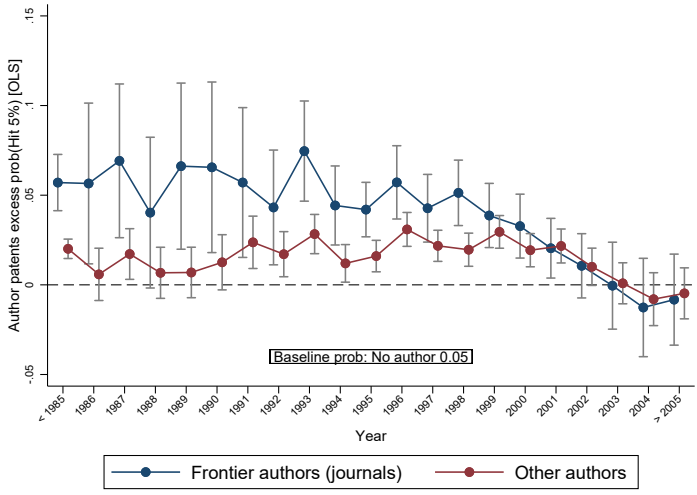
Notes: Each column reports parameter estimates from Poisson Pseudo Maximum Likelihood (columns (1)-(4)) and OLS (columns (5)-(8)) regressions of technological impact on inventor-author status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980-2009, across technological fields. Fields are based on subcategories in the classification of Hall, Jaffe and Trajtenberg (2001) (version 2006). The dependent variable measures 10-year window forward patent citations (DOCDB-level), in columns (1)-(3), and the probability that a patent enters the upper 5th percentile of the primary patent class-year-impact distribution, in columns (4)-(6). "Author" indicates that at least one inventor appears as author on a biomedical scientific publication indexed in PubMed/ Web of Science SCI prior to the patent application. "Frontier authors" are inventors that authored a paper in a "top-general" journal in the prior 3 years. "Top-general" journals in biomedical sciences are New England Journal of Medicine, Lancet, JAMA, Nature, Science, Cell, Nature Medicine, British Medical Journal. "Top-field" journals are all other journals in the top-category (score = 5) of journals in a biomedical WOS-field in the ECOOM FEB journals ranking (2021). Inventor-author disambiguation is based on Li et al., 2014; Smalheiser and Torvik, 2009 and Torvik (2018). DB links established by the authors. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015). Heteroscedasticity robust standard errors in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

FIGURE 1.B1. TIME TRENDS IN AUTHOR PATENTS IMPACT PREMIUM



Notes: The figure plots parameter estimates from Poisson Pseudo Maximum Likelihood regressions of technological impact on inventor-author status for the sample of all biomedical U.S. patents exclusively assigned to private firms, applied for between 1985-2005. The dependent variable measures 10-year window forward patent citations (DOCDB-level). "Frontier authors" are inventors that authored a paper in a "top-general" journal in the past 3 years. "Top-general" journals in biomedical sciences are New England Journal of Medicine, Lancet, JAMA, Nature, Science, Cell, Nature Medicine, British Medical Journal. The regression model includes a full set of class x year, number of inventors, and firm patent stock decile fixed effects. 95% confidence intervals are based on heteroscedasticity robust standard errors. Inventor-author disambiguation is based on Li et al., 2014; Smalheiser and Torvik, 2009 and Torvik (2018). DB links established by the authors. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015).

Figure 1.B2. Time Trends in Author Patents Impact Premium – Hits



Notes: The figure plots parameter estimates from OLS regressions of technological impact on inventor-author status for the sample of all biomedical U.S. patents exclusively assigned to private firms, applied for between 1985–2005. The dependent variable measures the probability that a patent enters the upper 5th percentile of the primary patent class-year-impact distribution. "Frontier authors" are inventors that authored a paper in a "top-general" journal in the past 3 years. "Top-general" journals in biomedical sciences are New England Journal of Medicine, Lancet, JAMA, Nature, Science, Cell, Nature Medicine, British Medical Journal. The regression model includes a full set of class x year, number of inventors, and firm patent stock decile fixed effects. 95% confidence intervals are based on heteroscedasticity robust standard errors. Inventor-author disambiguation is based on Li et al., 2014; Smalheiser and Torvik, 2009 and Torvik (2018). DB links established by the authors. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015).

1.C Appendix C

Frontier science links and high-impact patents

In this appendix section, we verify whether the use of frontier knowledge in patents is associated with higher impact and top-impact, as suggested by prior literature Ahmadpoor and Jones, 2017; Poege et al., 2019, irrespective of author status. We estimate the following regression equation for frontier science impact:

$$Y_i = \alpha_0 + \beta_1 FrontierSNPR_i + \sum_j \beta_j NonFrontierSNPR_{j(i)} + \mathbf{X}_i'\gamma + \delta_{c(i)t(i)} + \eta_{s(i)d(i)} + \epsilon_i \tag{1.4}$$

Analogously to Section 1.4.1, we measure technological impact outcomes by counts of forward citations and, alternatively, as likelihood of hit patent. Our main predictor of interest in this section is the presence of a scientific reference to a top-research article published in the past 3 years (or less) before the patent application, measured by the binary indicator $FrontierSNPR_i$. The coefficient for β_1 captures the influence of frontier reference status on the expectation of patent impact in our model equation 1.4. We account for the presence of non-frontier SNPRs through a series of dummy variables for different reference types j – non-recent top-journal articles, top-field journal articles, and non top-journal articles -, captured by $\sum_j \beta_j$. We additionally adjust for patent level covariates, primarily fixed effects for the number of SNPRs, and further patent characteristics in robustness analyses. To allow for approximately homogeneous fixed effects bins, we replace the count of SNPRs with their corresponding percentile ranks for patents above the 90th percentile of science references and group patents with less than two references into one dummy category.[29] Our estimation model also includes a full set of primary patent class by year fixed effects and applicant firm patent stock decile by decile rank change dummies.

Table 1.C1 below reports the econometric results from the estimation of equation 1.4, using PPML and OLS on the different outcome distributions. Patents with references to scientific articles have significantly higher impact (column 1) and higher top-performance likelihood (column 5), which is in line with findings in

[29] Our estimation models include a total of 23 SNPR count fixed effect categories.

prior literature (Fleming and Sorenson, 2004; Ahmadpoor and Jones, 2017). When comparing timing heterogeneity of references within science citing patents, we find recency of SNPRs not per se to be associated with higher impact; While patents relying on publications from the last 3 years prior to their application have a higher likelihood of being hit-technologies (column 6), they show slightly lower average citation rates than patents referencing only more dated scientific articles (column 2).

In column 3, we augment this distinction by a science quality dimension and estimate heterogeneous effects for references to frontier science (recent, top-general journals) and non-frontier science. We find patents linking to frontier articles to receive on average +40% more follow-up citations than patents without SNPRs. This excess premium is almost twice as large as the extra-citation rate of science-based patent only referencing non-frontier SNPRs (+22%). Moreover, we find the premium of frontier SNPR patents to be magnified with regards to their breakthrough likelihood (column 7 of Table 1.C1): The probability of patents referencing a recent, top-general journal articles to become a hit patent in the upper 5th percentile of the distribution is 125% higher than for patents without SNPRs. This excess probability is threefold the effect estimated for patents referencing only non-frontier SNPRs. In columns 4 and 8, we further find the value of recency to be non-uniform across the quality dimension, and by far strongest as a predictor for technological impact among patents making references to top-general interest journals, while the increment in citations and hit likelihood is significantly from citing more recent articles is significantly reduced for patents referencing top-field journal science and below.

These estimates confirm findings from prior literature showing that higher quality science is directly associated with superior impact in technology, especially hit technologies (Ahmadpoor and Jones, 2017; Poege et al., 2019). In particular, we confirm the observation of Poege et al. (2019) that recency seems to matter disproportionately more for links to top-science and exerts only a marginal role in technology referencing publications in the lower part of the science quality distribution.

In Table 1.C2 below, we show robustness of these results for stricter comparisons based on patent level characteristics, as well as for different definitions of recency to frontier and alternative measures for science quality: In Table 1.C3, we further

show that our estimates are fully robust and effect sizes even larger when measuring frontier science as recent links to articles in top upper tail of the scientific citations distribution, regardless of journal outlets.

We also investigate differences in patent impact on the dimension of relative timing of references to science, in particular to frontier science. Several papers have suggested a high relevance of early access links for the creation of technological value (Baruffaldi and Poege, 2020; Arora, Belenzon and Dionisi, 2021). Our results reported in Table 1.C4 show that links to on-going research – measured as ex-ante links from patents to scientific articles which are not yet published, or just published, at the time of patent application – are not associated with more average citation rates, but have particularly higher breakthrough probability, i.e. to become a hit patent in a given technology class-year cohort (columns 1 and 5). This pattern is persistent, yet less evidently pronounced for frontier science links (columns 2 and 6). In columns 3-4 and 7-8, we further investigate the role first-mover advantages in the use science, in particular frontier science. In line with Arora, Belenzon and Dionisi (2021), we find that first-mover links to science are not associated with higher social value, as measured by follow-up citations, also not for first links to frontier science (columns 3-4). However, estimates in columns 7-8 imply that first-to-use links are a significantly strong predictor for breakthrough invention, and appear on patents which are particularly more likely to be hits in their technological domain. This is an important insight, providing evidence for the pivotal role of first-access to science, in particular to frontier science, for pushing the boundaries of technological progress.

Table 1.C1. Frontier SNPR Patents and Technological Impact

	PPML				OLS			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	PatCit10	PatCit10	PatCit10	PatCit10	Top5% hit	Top5% hit	Top5% hit	Top5% hit
SNPR	0.375***				0.033***			
	(0.008)				(0.001)			
SNPR, recent (last 3y)		0.194***				0.046***		
		(0.013)				(0.002)		
SNPR, not recent (>3y)		0.208***				0.007***		
		(0.012)				(0.002)		
Frontier SNPR			0.338***	0.392***			0.069***	0.073***
			(0.019)	(0.020)			(0.004)	(0.004)
Non-frontier SNPR			0.202***				0.020***	
			(0.012)				(0.002)	
SNPR, top-gen jrn, >3y				0.274***				0.018***
				(0.018)				(0.003)
SNPR, top field-jrn, ≤3y				0.284***				0.051***
				(0.018)				(0.003)
SNPR, top field-jrn, >3y				0.284***				0.051***
				(0.018)				(0.003)
SNPR, no top-jrn, ≤3y				0.122***				0.028***
				(0.015)				(0.003)
SNPR, no top-jrn, >3y				0.189***				0.011***
				(0.014)				(0.002)
Constant	3.254***	3.338***	3.328***	3.311***	0.053***	0.053***	0.055***	0.054***
	(0.004)	(0.006)	(0.006)	(0.006)	(0.001)	(0.001)	(0.001)	(0.001)
Patent class × year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of SNPRs	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Firm patent stock × Δ	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	237,585	237,585	237,585	237,585	237,590	237,590	237,590	237,590
Pseudo/ adjusted R^2	0.26	0.27	0.27	0.27	0.03	0.03	0.03	0.03

Notes: Each column reports parameter estimates from Poisson Pseudo Maximum Likelihood (columns (1)-(4)) and OLS (columns (5)-(8)) regressions of technological impact on inventor-author status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980-2009. The dependent variable measures 10-year window forward patent citations (DOCDB-level), in columns (1)-(4), and the probability that a patent enters the upper 5th percentile of the primary patent class-year-impact distribution, in columns (5)-(8). "SNPR" (scientific reference) indicates at least one patent reference to a biomedical scientific article indexed in PubMed / Web of Science SCI. Examiner-given SNPRs are excluded from patents for which this information is available (granted after Jan 2001). "Frontier SNPRs" are references to papers in "top-general" journals published in 3 years or less before the patent application date. "Top-general" journals in biomedical sciences are New England Journal of Medicine, Lancet, JAMA, Nature, Science, Cell, Nature Medicine, British Medical Journal. "Top-field" journals are all other journals in the top-category (score = 5) of journals in a biomedical WOS-field in the ECOOM FEB journals ranking (2021). Scientific patent references to articles in PubMed/ SCI are sourced from Agarwal et al. (2014) and Marx and Fuegi (2020). Science citations are sourced from Microsoft Academic Graph. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015). Heteroscedasticity robust standard errors in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.C2. Frontier SNPR Patents and Technological Impact – Robustness

	PPML				OLS			
VARIABLES	(1) PatCit10	(2) PatCit10	(3) PatCit10	(4) PatCit10	(5) Top5% hit	(6) Top5% hit	(7) Top5% hit	(8) Top5% hit
SNPR	0.180***				0.016***			
	(0.008)				(0.001)			
SNPR, recent (last 3y)		0.111***				0.036***		
		(0.013)				(0.002)		
SNPR, not recent (>3y)		0.103***				-0.001		
		(0.012)				(0.002)		
Frontier SNPR			0.196***	0.233***			0.049***	0.052***
			(0.018)	(0.020)			(0.004)	(0.004)
Non-frontier SNPR			0.104***				0.011***	
			(0.011)				(0.002)	
SNPR, top-gen jrn, >3y				0.150***				0.009***
				(0.018)				(0.003)
SNPR, top field-jrn, ≤3y				0.173***				0.034***
				(0.017)				(0.004)
SNPR, top field-jrn, >3y				0.166***				0.003
				(0.015)				(0.003)
SNPR, no top-jrn, ≤3y				0.068***				0.023***
				(0.014)				(0.003)
SNPR, no top-jrn, >3y				0.080***				0.003
				(0.013)				(0.002)
Constant	3.533***	3.567***	3.562***	3.550***	0.060***	0.058***	0.060***	0.059***
	(0.004)	(0.006)	(0.006)	(0.006)	(0.001)	(0.001)	(0.001)	(0.001)
Patent class × year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of SNPRs	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Number of inventors	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Applicant firm	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	223,373	223,373	223,373	223,373	223,448	223,448	223,448	223,448
Pseudo/ adjusted R^2	0.48	0.48	0.48	0.48	0.13	0.13	0.13	0.13

Notes: Each column reports parameter estimates from Poisson Pseudo Maximum Likelihood (columns (1)-(4)) and OLS (columns (5)-(8)) regressions of technological impact on inventor-author status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980-2009. The dependent variable measures 10-year window forward patent citations (DOCDB-level), in columns (1)-(4), and the probability that a patent enters the upper 5th percentile of the primary patent class-year-impact distribution, in columns (5)-(8). "SNPR" (scientific reference) indicates at least one patent reference to a biomedical scientific article indexed in PubMed / Web of Science SCI. Examiner-given SNPRs are excluded from patents for which this information is available (granted after Jan 2001). "Frontier SNPRs" are references to papers in "top-general" journals published in 3 years or less before the patent application date. "Top-general" journals in biomedical sciences are New England Journal of Medicine, Lancet, JAMA, Nature, Science, Cell, Nature Medicine, British Medical Journal. "Top-field" journals are all other journals in the top-category (score = 5) of journals in a biomedical WOS-field in the ECOOM FEB journals ranking (2021). Scientific patent references to articles in PubMed/ SCI are sourced from Agarwal et al. (2014) and Marx and Fuegi (2020). Science citations are sourced from Microsoft Academic Graph. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015). Heteroscedasticity robust standard errors in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.C3. Frontier SNPR Patents and Technological Impact – Top5% Cited Papers

	PPML		OLS	
	(1)	(2)	(3)	(4)
VARIABLES	PatCit10	PatCit10	Top5% hit	Top5% hit
Frontier SNPR	0.312***	0.346***	0.070***	0.069***
	(0.016)	(0.018)	(0.003)	(0.003)
Non-frontier SNPR	0.201***		0.019***	
	(0.012)		(0.002)	
SNPR, top5% paper, >3y		0.273***		0.013***
		(0.017)		(0.003)
SNPR, no top5% paper, last 3y		0.167***		0.031***
		(0.015)		(0.003)
SNPR, no top5% paper, >3y		0.204***		0.013***
		(0.013)		(0.002)
Constant	3.326***	3.316***	0.054***	0.054***
	(0.006)	(0.006)	(0.001)	(0.001)
Patent class × year	Yes	Yes	Yes	Yes
Number of SNPRs	Yes	Yes	Yes	Yes
Firm patent stock × Δ	Yes	Yes	Yes	Yes
Patent-paper pairs excl.	Yes	Yes	Yes	Yes
Observations	233,420	233,420	233,425	233,425
Pseudo/ adjusted R^2	0.27	0.27	0.03	0.03

Notes: Each column reports parameter estimates from Poisson Pseudo Maximum Likelihood (columns (1)-(2)) and OLS (columns (3)-(4)) regressions of technological impact on inventor-author status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980-2009. The dependent variable measures 10-year window forward patent citations (DOCDB-level), in columns (1)-(2), and the probability that a patent enters the upper 5th percentile of the primary patent class-year-impact distribution, in columns (3)-(4). "SNPR" (scientific reference) indicates at least one patent reference to a biomedical scientific article indexed in PubMed / Web of Science SCI. Examiner-given SNPRs are excluded from patents for which this information is available (granted after Jan 2001). "Frontier SNPRs" are references to "top5% papers" published in 3 years or less before the patent application date. "Top5% papers" are scientific articles that enter the upper 5th percentile of the citation distribution of a WOS subject category field year cohort. "Patent-paper pairs" are patents citing a scientific paper with 100% inventor-author overlap and published in direct temporal proximity to the patent application. Scientific patent references to articles in PubMed/ SCI are sourced from Agarwal et al. (2014) and Marx and Fuegi (2020). Science citations are sourced from Microsoft Academic Graph. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015). Heteroscedasticity robust standard errors in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.C4. Frontier SNPR Patents and Technological Impact – Timing of SNPRs

	PPML				OLS			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	PatCit10	PatCit10	PatCit10	PatCit10	Top5% hit	Top5% hit	Top5% hit	Top5% hit
SNPR, ex-ante	0.173***				0.047***			
	(0.016)				(0.003)			
SNPR, ex-post	0.206***				0.018***			
	(0.012)				(0.002)			
Frontier SNPR, ex-ante		0.323***				0.079***		
		(0.028)				(0.006)		
Frontier SNPR, ex-post		0.332***				0.071***		
		(0.021)				(0.004)		
Non-frontier SNPR, ex-ante		0.161***				0.041***		
		(0.017)				(0.003)		
Non-frontier SNPR, ex-post		0.205***				0.018***		
		(0.012)				(0.002)		
SNPR, first to cite			0.131***				0.042***	
			(0.014)				(0.002)	
SNPR, not first			0.222***				0.013***	
			(0.012)				(0.002)	
Frontier SNPR, first				0.313***				0.085***
				(0.028)				(0.006)
Frontier SNPR, not first				0.322***				0.071***
				(0.021)				(0.004)
Non-frontier SNPR, first				0.130***				0.040***
				(0.014)				(0.002)
Non-frontier SNPR, not first				0.222***				0.012***
				(0.012)				(0.002)
Constant	3.339***	3.330***	3.343***	3.333***	0.056***	0.054***	0.056***	0.054***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.001)	(0.001)	(0.001)	(0.001)
Patent class × year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of SNPRs	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Firm patent stock × Δ	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	237,585	237,585	237,585	237,585	237,590	237,590	237,590	237,590
Pseudo/ adjusted R^2	0.26	0.27	0.27	0.27	0.03	0.03	0.03	0.03

Notes: Each column reports parameter estimates from Poisson Pseudo Maximum Likelihood (columns (1)-(4)) and OLS (columns (5)-(8)) regressions of technological impact on inventor-author status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980-2009. The dependent variable measures 10-year window forward patent citations (DOCDB-level), in columns (1)-(4), and the probability that a patent enters the upper 5th percentile of the primary patent class-year-impact distribution, in columns (5)-(8). "SNPR" (scientific reference) indicates at least one patent reference to a biomedical scientific article indexed in PubMed / Web of Science SCI. Examiner-given SNPRs are excluded from patents for which this information is available (granted after Jan 2001). "Ex-ante SNPRs" are references to articles non yet or *just* published at the time of the patent application. "Frontier SNPRs" are references to papers in "top-general" journals published in 3 years or less before the patent application date. "Top-general" journals in biomedical sciences are New England Journal of Medicine, Lancet, JAMA, Nature, Science, Cell, Nature Medicine, British Medical Journal. "Top-field" journals are all other journals in the top-category (score = 5) of journals in a biomedical WOS-field in the ECOOM FEB journals ranking (2021). Scientific patent references to articles in PubMed/ SCI are sourced from Agarwal et al. (2014) and Marx and Fuegi (2020). Science citations are sourced from Microsoft Academic Graph. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015). Heteroscedasticity robust standard errors in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

1.D Appendix D

Table 1.D1. Frontier Author Patents and Frontier SNPR Probability – Robustness

	OLS	PPML	OLS		
	(1)	(2)	(3)	(4)	(5)
VARIABLES	SNPR [01]	# SNPRs	Frontier SNPR	Close frontier SNPR	First frontier SNPR
Frontier author	0.028***	0.051***	0.115***	0.060***	0.017
	(0.004)	(0.017)	(0.005)	(0.006)	(0.012)
Log(# prior patents)	Yes	Yes	Yes	Yes	Yes
Log(# prior publications)	Yes	Yes	Yes	Yes	Yes
Constant	0.571***	2.804***	0.195***	-0.011	0.347***
	(0.005)	(0.027)	(0.005)	(0.008)	(0.019)
with SNPR only		Yes	Yes	Yes	Yes
with frontier SNPR only				Yes	Yes
with author only				Yes	
Patent class × year	Yes	Yes	Yes	Yes	Yes
Number of inventors	Yes	Yes	Yes	Yes	Yes
Number of authors	Yes	Yes	Yes	Yes	Yes
Number of SNPRs	No	No	Yes	Yes	Yes
Applicant firm	Yes	Yes	Yes	Yes	Yes
Observations	134,248	72,103	72,103	10,760	10,760
Pseudo/ adjusted R^2	0.37	0.48	0.27	0.19	0.15

Notes: Each column reports parameter estimates from OLS (columns (1), (3)-(5)) and Poisson Pseudo Maximum Likelihood (column (2)) regressions of scientific reference on inventor-author status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980-2009. The dependent variable (in columns (1), (3)-(5)) measures the probability that a patent makes a prior art reference (SNPR) to a scientific article indexed in PubMed/ Web of Science SCI. Examiner-given SNPRs are excluded from patents for which this information is available (granted after Jan 2001). "Frontier SNPRs" are references to papers in "top-general" journals published in 3 years or less before the patent application date. "Top-general" journals in biomedical sciences are New England Journal of Medicine, Lancet, JAMA, Nature, Science, Cell, Nature Medicine, British Medical Journal. "Frontier authors" are inventors that authored a paper in a "top-general" journal in the prior 3 years. "Close SNPRs" are self-citations or citations to the co-author network of at least one listed inventor. Inventor-author disambiguation is based on Li et al., 2014; Smalheiser and Torvik, 2009 and Torvik (2018). DB links established by the authors. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015). Scientific patent references to articles in PubMed/ SCI are sourced from Agarwal et al. (2014) and Marx and Fuegi (2020). Science citations are sourced from Microsoft Academic Graph. Heteroscedasticity robust standard errors in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.D2. Frontier Author Patents and Frontier SNPR Probability – Top5% Cited Papers

	OLS	PPML	OLS					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	SNPR [01]	# SNPRs	Frontier SNPR	Close front SNPR	1st front SNPR	Front5% SNPR	Cl front5% SNPR	1st front5% SNPR
Frontier author	0.288***	0.737***	0.062***	0.040***	-0.014	0.090***	0.063***	-0.029***
	(0.003)	(0.019)	(0.004)	(0.004)	(0.014)	(0.004)	(0.003)	(0.011)
Non-frontier author	0.163***	0.344***	-0.003		-0.026**	-0.016***		-0.040***
	(0.002)	(0.017)	(0.003)		(0.013)	(0.003)		(0.010)
Constant	0.290***	2.285***	0.133***	0.020***	0.286***	0.228***	0.009***	0.368***
	(0.001)	(0.015)	(0.002)	(0.002)	(0.012)	(0.003)	(0.001)	(0.009)
with SNPR only		Yes	Yes	Yes	Yes	Yes	Yes	
with frontier SNPR only				Yes	Yes		Yes	Yes
with author only				Yes			Yes	
Patent class × year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of inventors	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of SNPRs	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Firm patent stock × Δ	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Patent-paper pairs excl	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	237,591	95,190	95,190	12,062	13,661	95,190	19,522	22,508
Pseudo/ adjusted R^2	0.30	0.20	0.18	0.06	0.09	0.20	0.06	0.06

Notes: Each column reports parameter estimates from OLS (columns (1), (3)-(8) and Poisson Pseudo Maximum Likelihood (column (2)) regressions of scientific reference on inventor-author status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980-2009. The dependent variable (in columns (1), (3)-(8)) measures the probability that a patent makes a prior art refrence (SNPR) to a scientific article indexed in PubMed/ Web of Science SCI. Examiner-given SNPRs are excluded from patents for which this information is available (granted after Jan 2001). In columns (3)-(5), "Frontier SNPRs" are references to papers in "top-general" journals published in 3 years or less before the patent application date. "Top-general" journals in biomedical sciences are New England Journal of Medicine, Lancet, JAMA, Nature, Science, Cell, Nature Medicine, British Medical Journal. Alternatively, in columns (6)-(8), "Frontier SNPRs" are references to "top5% papers" published in 3 years or less before the patent application date. "Top5% papers" are scientific articles that enter the upper 5th percentile of the citation distribution of a WOS subject category field year cohort. "Close SNPRs" are self-citations or citations to the co-author network of at least one listed inventor. "Frontier authors" are inventors that authored themselves a "top5%" paper in the prior 3 years. "Patent-paper pairs" are patents citing a scientific paper with 100% inventor-author overlap and published in direct temporal proximity to the patent application. Inventor-author disambiguation is based on Li et al., 2014; Smalheiser and Torvik, 2009 and Torvik (2018). DB links established by the authors. Patent data is sourced from PATSTAT 2018 and Marco et al. (2015). Scientific patent references to articles in PubMed/ SCI are sourced from Agarwal et al. (2014) and Marx and Fuegi (2020). Science citations are sourced from Microsoft Academic Graph. Heteroscedasticity robust standard errors in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

1.E Appendix E

Table 1.E1. Frontier Author Patents, SNPRs and Technological Impact – Robustness

	PPML	OLS
	(1)	(2)
VARIABLES	PatCit10	Top5% hit
Frontier author, frontier SNPR	0.186***	0.044***
	(0.028)	(0.006)
Frontier author, non-frontier SNPR	0.090***	0.007*
	(0.023)	(0.004)
Frontier author, no SNPR	0.078***	0.005
	(0.025)	(0.005)
Non-frontier author, frontier SNPR	0.159***	0.045***
	(0.021)	(0.005)
Non-frontier author, non-frontier SNPR	0.065***	0.007***
	(0.013)	(0.002)
Log(# prior patents)	Yes	Yes
Log(# prior publications)	Yes	Yes
Constant	3.554***	0.087***
	(0.017)	(0.003)
Patent class × year	Yes	Yes
Number of inventors	Yes	Yes
Number of authors	Yes	Yes
Number of SNPRs	Yes	Yes
Applicant firm	Yes	Yes
Observations	134,205	134,245
Pseudo/ adjusted R^2	0.49	0.14

Notes: Each column reports parameter estimates from Poisson Pseudo Maximum Likelihood (column (1)) and OLS (column (2)) regressions of technological impact on inventor-author and scientific reference status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980-2009. For details on variable measurement and data sources, see Table 1.2, Table 1.3 and Section 1.3. Examiner-given SNPRs are excluded from patents for which this information is available (granted after Jan 2001). Heteroscedasticity robust standard errors in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.E2. Frontier Author Patents, SNPRs and Technological Impact – Robustness

	PPML	OLS
	(1)	(2)
VARIABLES	PatCit10	Top5% hit
Frontier author, frontier SNPR	0.854***	0.100***
	(0.026)	(0.005)
Frontier author, non-frontier SNPR	0.541***	0.040***
	(0.022)	(0.003)
Frontier author, no SNPR	0.415***	0.031***
	(0.029)	(0.004)
Non-frontier author, frontier SNPR	0.679***	0.091***
	(0.018)	(0.004)
Non-frontier author, non-frontier SNPR	0.432***	0.036***
	(0.010)	(0.002)
Non-frontier author, no SNPR	0.155***	0.017***
	(0.009)	(0.001)
No author, frontier SNPR	0.603***	0.076***
	(0.036)	(0.008)
No author, non-frontier SNPR	0.383***	0.036***
	(0.013)	(0.002)
Constant	3.173***	0.043***
	(0.006)	(0.001)
Patent class × year	Yes	Yes
Number of inventors	Yes	Yes
Number of SNPRs	No	No
Firm patent stock × Δ	Yes	Yes
Observations	237,585	237,590
Pseudo/ adjusted R^2	0.27	0.03

Notes: Each column reports parameter estimates from Poisson Pseudo Maximum Likelihood (column (1)) and OLS (column (2)) regressions of technological impact on inventor-author and scientific reference status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980-2009. For details on variable measurement and data sources, see Table 1.2, Table 1.3 and Section 1.3. Examiner-given SNPRs are excluded from patents for which this information is available (granted after Jan 2001). Heteroscedasticity robust standard errors in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.E3. Frontier Author Patents, SNPRs and Technological Impact – Top5% Cited Papers

	PPML	OLS
	(1)	(2)
VARIABLES	PatCit10	Top5% hit
Frontier author, frontier SNPR	0.479***	0.091***
	(0.023)	(0.005)
Frontier author, non-frontier SNPR	0.433***	0.032***
	(0.019)	(0.003)
Frontier author, no SNPR	0.403***	0.035***
	(0.018)	(0.003)
Non-frontier author, frontier SNPR	0.391***	0.076***
	(0.020)	(0.004)
Non-frontier author, non-frontier SNPR	0.268***	0.028***
	(0.014)	(0.002)
Non-frontier author, no SNPR	0.136***	0.016***
	(0.009)	(0.001)
No author, frontier SNPR	0.373***	0.071***
	(0.026)	(0.006)
No author, non-frontier SNPR	0.266***	0.029***
	(0.015)	(0.003)
Constant	3.243***	0.045***
	(0.008)	(0.001)
Patent class × year	Yes	Yes
Number of inventors	Yes	Yes
Number of SNPRs	Yes	Yes
Firm patent stock × Δ	Yes	Yes
Patent-paper pairs excl	Yes	Yes
Observations	233,420	233,425
Pseudo/ adjusted R^2	0.28	0.03

Notes: Each column reports parameter estimates from Poisson Pseudo Maximum Likelihood (column (1)) and OLS (column (2)) regressions of technological impact on inventor-author and scientific reference status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980–2009. "Frontier authors" are inventors that authored themselves a "top5%" paper in the prior 3 years. "Top5% papers" are scientific articles that enter the upper 5th percentile of the citation distribution of a WOS subject category field year cohort. For more details on variable measurement and data sources, see Table 1.2, Table 1.3 and Section 1.3. Examiner-given SNPRs are excluded from patents for which this information is available (granted after Jan 2001). Heteroscedasticity robust standard errors in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.E4. Frontier Author Patents, SNPRs and Technological Impact - Ex-ante SNPRs

	PPML	OLS
	(1)	(2)
VARIABLES	PatCit10	Top5% hit
Frontier author, frontier SNPR, ex-ante	0.516***	0.091***
	(0.043)	(0.009)
Frontier author, frontier SNPR, ex-post	0.546***	0.083***
	(0.034)	(0.007)
Non-frontier author, frontier SNPR, ex-ante	0.363***	0.080***
	(0.035)	(0.008)
Non-frontier author, frontier SNPR, ex-post	0.413***	0.077***
	(0.024)	(0.005)
Author, non-frontier & no SNPR interactions	Yes	Yes
No author, SNPR interactions	Yes	Yes
Constant	3.249***	0.047***
	(0.008)	(0.001)
Patent class × year	Yes	Yes
Number of inventors	Yes	Yes
Number of SNPRs	Yes	Yes
Firm patent stock × Δ	Yes	Yes
Observations	237,585	237,590
Pseudo/ adjusted R^2	0.28	0.03

Notes: Each column reports parameter estimates from Poisson Pseudo Maximum Likelihood (column (1)) and OLS (column (2)) regressions of technological impact on inventor-author and scientific reference status for the sample of all biomedical U.S. patents exclusively assigned to private firms, granted between 1980-2009. "Ex-ante SNPRs" are references to articles non yet or *just* published at the time of the patent application. For more details on variable measurement and data sources, see Table 1.2, Table 1.3 and Section 1.3. Examiner-given SNPRs are excluded from patents for which this information is available (granted after Jan 2001). Heteroscedasticity robust standard errors in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Chapter 2

Online Repositories, Search Costs and Cumulative Innovation[1]

2.1 Introduction

Scientific advance is shaped by sequential and complementary intellectual efforts. Efficient access to existing information is, therefore, essential for technical progress (Jones, 2003; Mokyr, 2005). This becomes particularly salient in times of acute intensification of disease-targeted research activities, as during pandemics or public health crises. Especially, *new* ideas necessitate the free flow of information and simultaneous experimentation in order to prevail (Rosenberg, 1976; Murray et al.,

[1] This chapter is based on a single-authored manuscript of the same title, available at SSNR. I thank the members of my doctoral committee, Adrián Kovács, Jeroen Mahieu, Maikel Pellens, Leonard Treuren, Dennis Verhoeven, Jesse Wursten and participants in seminars and presentations at KU Leuven, TU Eindhoven, VU Amsterdam, ZEW Mannheim, the 2021 Annual Meeting of the European Policy for Intellectual Property (EPIP) Association, DRUID 2021 and the 2020 SEI Consortium for many useful comments and suggestions. Financial support from the Research Foundation - Flanders (FWO) is gratefully acknowledged under grant number 11B5918N.

2016). However, R&D spillovers, in particular those from applied research, tend to be clustered and internalized strongly within collaborative networks (Jaffe, Trajtenberg and Henderson, 1993; Cassiman and Veugelers, 2002; Singh, 2005; Belenzon and Schankerman, 2013; Akcigit, Hanley and Serrano-Velarde, 2020). Accordingly, access costs to external knowledge increase with the dispersion and disconnect between scientific communities.

Most of the empirical literature studying the elasticity of access costs to existing knowledge on innovation outcomes has, either explicitly or implicitly, focused on variation in physical accessibility in (historical) contexts in which external information was scarce or available only at undue costs.[2] [3] Questions regarding the efficiency of search and information retrieval in the numerousness of accessible data points have received significantly less attention from the innovation literature. Furman and Stern (2011) show how specialized biological resource centres, providing access to certified bio-materials and lowering evaluation costs, can help amplifying cumulative impact of scientific discoveries. Similarly, Thompson and Hanley (2018) demonstrate in a randomized experiment that incorporating new scientific topics into Wikipedia articles enhances their diffusion in scientific literature. Zheng and Wang (2020) observe a decline in distant technological search for inventors located in China following the ban of Google's search engine in 2006.

In this paper, I ask whether universally accessible, topic-specific repositories of prior art can effectively decrease informational inefficiencies by reducing search

[2] Iaria, Schwarz and Waldinger (2018) study the disruption of international scientific cooperation induced by WWI and show how the exclusion from access to frontier research from opposite camps lastingly affected scientific productivity. Biasi and Moser (2020) and Bryan and Ozcan (2020) show that the authorized violation of copyrights to German scientific books in the U.S. during WWII and the 2008 open access mandate for NIH funded research publications enhanced the dissemination of scientific ideas, evidenced by citations from subsequent articles and patents. In recent working papers, Furman, Nagler and Watzinger (2018) and Berkes and Nencka (2020) document substantial increases in patenting activity following the opening of U.S. Patent Libraries 1975-1997 and the extensive roll-out of public 'Carnegie' libraries in the U.S. 1883-1991, respectively. Arts et al. (2020) provide concurring evidence due to the arrival of broadband internet to the U.K. in the early-2000s.

[3] A closely related stream of literature has investigated to what extent patent documents allow to disclose relevant scientific knowledge and affect follow-up inventions, providing ample confirming evidence: Graham and Hegde (2015); Hegde and Luo (2018); Baruffaldi and Simeth (2020); Hegde, Herkenhoff and Zhu (2020); Lück et al. (2019); de Rassenfosse, Pellegrino and Raiteri (2020)

costs for relevant prior art. Such costs derive from challenges of absorbing and filtering most relevant information out of the sheer mass of scientific knowledge produced, *conditional* on accessibility. Specifically, I contribute to the literature by disentangling the effect of access from the one of increased visibility of pieces of knowledge arising from the connection to a particular topic, established by the inclusion into a topic-targeted repository.

To investigate this, I study the launch of the International AIDS Patents Database (*AIDS DB*, hereafter) in 1994, the historically first publicly accessible online repository of patent full-texts and images. At the peak of the HIV/AIDS pandemic [4], the United States Patent and Trademark Office (USPTO) undertook an unprecedented effort to leverage the capacities of the new world wide web in the fight against the disease, by providing free of charge full-online access to all patents related to acquired immune deficiency syndrome, ranging from diagnostic testing to therapeutic treatments. This repository, hosting initially 1,500 U.S. patents, meant great improvement in the conditions of access and retrieval of information for researchers worldwide racing to develop effective technologies against HIV/AIDS. Before AIDS DB, interested inventors had comparably limited ways to screen, filter and efficiently rank weekly published new patents: Assessing the accurate technical content of a patent required to either inspect paper documents directly at the patent office, query computer terminals in patent depository libraries, or remote-order individual full-text copies via mail or fax. Moreover, adverse incentives in the drafting of patents, in particular of private firms, to inhibit effective disclosure of valuable information through the patent system, determine historically high and substantial efforts necessary for the retrieval and absorption of prior art from codified knowledge. The feature of being a centrally-maintained and expert-validated disease-targeted repository could decrease these search costs significantly, given that deposited patents span a broad range of technology classes and fields and many patent documents were, at first, not clearly recognizable as HIV-related from bibliographic searches. The AIDS DB was discontinued in March 1999, when the

[4] Incidences of HIV/AIDS-related infections had exponentially spread since the early-1980s, reaching a peak of > 20,000 cases in 1993 and a yearly mortality rate excess of almost 15,000 by 1995 in the U.S. alone (Source: U.S. Centres for Disease Control and Prevention (CDC)). As of 2020, about 38 million people worldwide are living with HIV, causing about 1.7% of deaths globally (Source: WHO).

USPTO launched its comprehensive online database including full-text and images of all patents, as broader bandwidths became available.

I empirically assess the marginal impact of the AIDS DB on cumulative inventive search costs relying on publicly available citation data, tracing references to patents in the repository from follow-on inventions in the worldwide patent universe. I, further, exploit data from USPTO examination procedures, patent front-page information as well as patent text to determine the specific technical content of inventions. In order to characterize inventor-level links, I rely on geolocalized addresses and assign inventors to scientific communities based on their prior collaborative activities in both basic and applied science in the universe of all USPTO patents and all biomedical scientific articles indexed in PubMed.

To mitigate concerns of positive selection and endogenous treatment, I design an empirical strategy relying on a within AIDS DB counterfactual: Exploiting idiosyncrasies of technology classes assigned to patents not being disease-specific, I estimate the elasticity of search costs on cumulative citations on AIDS DB patents for which the link to HIV/AIDS was, arguably, non-obvious to be detected through standard bibliographic search[5] prior to the repository inclusion. I compare the differential effects of database deposit for these patents to a control group of patents, also indexed in the AIDS DB and and comparable timing, technical content, institutional and scientific prior art background, for which the disease-link was explicit already pre-AIDS DB from the textual content of their front-pages, providing a baseline of online accessibility and the confounding factors relating to HIV-patents in the estimation

I find that, after online deposit, cumulative citations to AIDS DB patents without explicit front-page reference to HIV/AIDS subjects increased by around 26% relative to citations to the control group. Effects are particularly pronounced for external spillovers, i.e. on the share of cumulative citations originating from outside applicants' organizations, and within citations from inventors working on HIV-related treatments, providing support for the effectiveness of the disease-specific repository in line with the policy objective. In support of these results being causally related to lower search costs, I further find that the marginal impact of database deposit was contingent on how visible the HIV/AIDS link was on a patent front page:

[5] i.e. would have required accessing the full-text patent document

For patents mentioning applicability to HIV/AIDS only in the patent abstract the effect was equally positive, compared to those mentioning this in the title, but significantly smaller in size compared to patents without explicit reference. Moreover, citations unlikely to reflect knowledge spillovers – those from patents already under examination – were unaffected by database inclusion of cited patents. I further exclude that results are driven by individual patent examiners or within-changes in citation behaviour of examiners over time.

Differences in the rate of follow-up citations gradually increased and persisted for several years after online deposit, even as comprehensive online patent databases became available at the end of the 1990s, strengthening my belief that the results are, indeed, attributable to reductions in search costs, provided by the disease-specific link, rather than online accessibility. In line with predictions, I find strongly positive differential effects on cumulative citations to patents with intrinsically higher search costs: Private firm patents, recombinant patent, and patents introducing new medical subjects into technology classes. Furthermore, my results imply significant second order effects on the visibility and increased patent citation rates to scientific references in patents without front-page links to HIV/AIDS. I show robustness of findings using different estimations, different time windows, stricter control group definitions, and impact weighted citation counts.

In a set of additional results, I investigate repercussions on the intensive margin of knowledge spillovers generated among HIV researcher following the establishment of the AIDS DB, comparing changes in the relation between HIV/AIDS patents and their follow-up inventions over time. Relative to non-indexed cited references, I find indication for enhanced knowledge flows, evidenced by significantly higher rates of re-occurrence in citing patents of new words and novel scientific prior art references appearing in AIDS DB patents without front-page link to HIV, following the launch of the database. These effects were strongest among private firm citing inventors. I further estimate effects on the reach of spillovers generated across geographic boundaries and scientific collaboration networks. After AIDS DB deposit, international citations to indexed patents without previously obvious HIV/AIDS link increased substantially, in particular from academic inventors and public research institutes, while patents with HIV/AIDS references experienced a relative increase in domestic citations. Finally, based on changes in shortest path

length between cited and citing inventors in the universe of (author-)inventors and their scientific collaborations, I find evidence for a strong marginal impact of the AIDS DB on the diffusion of relevant knowledge across scientific community boundaries, in particular across previously disconnected communities, which was entirely driven by increased citations from private sector inventors to patents without previously explicit link to HIV/AIDS originating from distant network communities.

This paper intends to make several contributions to the existing literature. Primarily, my findings inform about how topic-specific repositories can enhance the cumulative impact of new scientific knowledge by reducing search and retrieval costs for researchers, adding to findings of prior studies by Furman and Stern (2011) and Thompson and Hanley (2018). In particular, I contribute by providing new evidence for the effectiveness of topic-specific online repositories to decrease search costs for follow-up invention and show that these conditions for knowledge accumulations are analogous between open science and applied technology.

My findings, further, speak to the growing body of prior work on the importance of access to existing knowledge for scientific production (Moser and Voena, 2012; Murray et al., 2016; Iaria, Schwarz and Waldinger, 2018), in particular on how access costs to information affect cumulative research impact (Bryan and Ozcan, 2020; Furman, Nagler and Watzinger, 2018; Biasi and Moser, 2020). Here, I confirm prior evidence that increasing accessibility to relevant prior art impacts subsequent invention and the diffusion of industrially applicable knowledge. Finally, my paper has close antecedents in prior work regarding the role of modern information technologies on knowledge diffusion, spillovers and collaboration in research and development (Agrawal and Goldfarb, 2008; Ding et al., 2010; Forman and van Zeebroeck, 2012; Bertschek, Cerquera and Klein, 2013; Forman and van Zeebroeck, 2019; Zheng and Wang, 2020), as well as in the broader literature on the impact of information technologies on economic progress (e.g. Czernich et al., 2011; Dittmar, 2011).

2.2 Background

2.2.1 External search and prior art search costs

The cumulativeness of R&D efforts is well documented in the innovation literature (e.g., Scotchmer, 1991; Galasso and Schankerman, 2015). Intertemporal spillovers from existing knowledge provide critical inputs for the direction of follow-up search, and spur the capacity of future advancement. Being non-rival in nature, these externalities generate social increasing returns to R&D investment (Griliches, 1991; Bloom, Schankerman and Van Reenen, 2013; Jones and Summers, 2020). In applied research, the primary channels, through which spillovers are internalized, rely on direct interaction, as knowledge flows tend to be intrinsically localized and strongly clustered among institutional networks (Jaffe, Trajtenberg and Henderson, 1993; Cassiman and Veugelers, 2002; Singh, 2005). When inventors conduct external search, i.e. attempt to source prior art information from outside their direct networks, important inputs are provided through patent documents and scientific publications. This is particularly given in fields in which these embody specific and valuable codified knowledge, such as chemical (including biomedical) technologies (e.g., Cohen, Nelson and Walsh, 2000; Jaffe, Trajtenberg and Fogarty, 2000; Giuri et al., 2007; Gambardella, Harhoff and S, 2011). Recent studies provide ample empirical support for the effective disclosure function of the patent system (e.g., Hegde, Herkenhoff and Zhu, 2020; Baruffaldi and Simeth, 2020; Lück et al., 2019; de Rassenfosse, Pellegrino and Raiteri, 2020).

A precondition for the efficient absorption of external codified knowledge is posed by accessibility. The elasticity of access costs to physical copies of scientific and technical literature on cumulative innovation has been found to be large and significant in prior studies (e.g., Iaria, Schwarz and Waldinger, 2018; Bryan and Ozcan, 2020; Biasi and Moser, 2020; Furman, Nagler and Watzinger, 2018). Such access costs, which constitute broader search costs from a point of view of microeconomic theory, have been historically substantial, but decreased drastically with the advent of modern information technologies, in particular broadband internet (Arts et al., 2020). Faster bandwidth, electronic file formats and online repositories have made scientific and technical information readily available and omni-accessible.

However, even conditional on full accessibility to prior art, inventors incur an additional and significant cost in capturing external knowledge spillovers: the search costs arising from the opportunity and mental effort necessary to screen the increasing bulk of information on new advances in a given technical domain, filter and rank these based on the relevance and usefulness for the inventor's specific inquiry, and find ways to integrate them in order to increase the value of a follow-up invention. For these characteristics, prior art search resembles a (non-stationary) sequential search problem with multiple periods (e.g., Pandora's problem in the model of Weitzman, 1979), where each new patent document issued represents a closed box at the beginning of each period, containing a potential reward in form of a knowledge spillover. The costs to open a box and internalize its content are paid during the search, while the reward is revealed and collected only afterwards, and with some delay. After each stage, the inventor decides whether to incur the cost of another search round or to use the fallback option, i.e. rely on local search (e.g., March, 1991; Cohen and Levinthal, 1990) and the knowledge acquired in previous rounds. The accuracy with which the inventor can delineate the sample of prior art documents to inspect with regards to their utility determines the overall efficiency of the search, by optimizing the reward/cost ratio. Ceteris paribus, we would expect the amount of spillovers generated to increase with the accuracy of up-front information provided.

Patent systems provide several remedies for searching inventors to facilitate processing the information overload that comes with disclosure on the front page of patent documents. The most important of these are the technology classes an invention is assigned to.[6] However, patent classes have been frequently questioned in the literature with regards to accurately delineating narrow technological fields (e.g., Thompson and Fox-Kean, 2005; Benner and Waldfogel, 2008; Arts, Cassiman and Gomez, 2018). Next to the device of technology classes, most patent offices have, for some time, offered Boolean search facilities to their databases (first, through patent library terminals, afterwards for their online repositories, see 2.2.2). The usefulness of these for effectively detecting prior art is, however, constrained by the fact that bibliographic patent text, in particular in U.S. patents, tends to be written

[6]Most importantly the International Patent Classification (IPC), or the USPC and CPC for U.S. patents.

in a highly abstract, legal jargon (e.g., Fromer, 2008; Ouelette, 2012; Lemley, 2012), making key word based searches prone to inaccuracy. These phenomena originate from private firms incentives in disclosing as little concrete information possible in patent documents, in order to conceal the nature of their inventions and protect from imitation (Risch, 2007; Devlin, 2009). Recent concurring evidence from computational linguistics by Kong et al. (2020) shows that private sector patents are significantly less readable than those of universities and public research institutions.

Despite the advent of modern information technology in bibliographic search[7], including tools and algorithms based on latent semantic analysis, page rank or applications of artificial intelligence, efficiency in information retrieval remains a concern, as evidenced by several platform initiatives launched within the course of the COVID-19 pandemic: The U.S. Centres of Diseases Control and Prevention (CDC), in 2020, launched a topic-specific online access repository collecting comprehensive scientific evidence, clinical trials and gene sequence databases related to the novel coronavirus, its diagnostics and treatments.[8] Simultaneously, the USPTO launched its platform *Patents4Partnerships*, an initiative targeted towards markets for technologies regarding the prevention, diagnosis and treatment of COVID-19, assembling a complete listing of inventions currently available related to the on-going public health crisis.[9]

In sum, while there is consensus that patent documents can transfer valuable and specific technical information, inventors face substantial search costs to identify such information from bibliographic archives.

2.2.2 The International AIDS Patent Database project

On October 26th 1994, the United States Department of Commerce announced the release of a new database allowing for immediate access to the full text and images of all U.S. patents related to the diagnostic testing and therapeutic treatment of

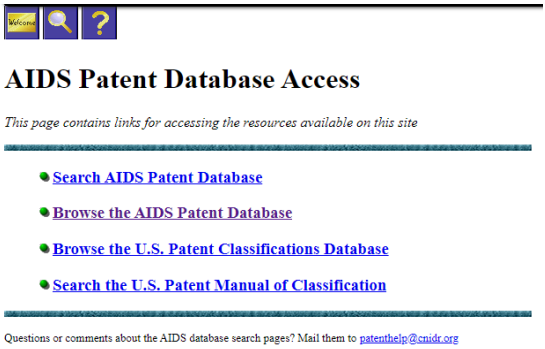
[7]Modern web search engines based on LSA and page rank algorithms, such as Google Patents, might prove useful in lowering these costs through automated retrieval of related patents to inventors' searches. For instance, Zheng and Wang (2020) observe a relative decline distant technology search of inventors located in China following the Chinese Google-ban.

[8]Online link to CDC COVID-19 database here (accessed 28/07/2021).

[9]Online link to USPTO Patents4Partnerships platform here (accessed 28/07/2021).

acquired immune deficieny syndrome (AIDS), the disease complex caused by infections with the *human immunodeficiency virus (HIV)*.

Figure 2.1. AIDS DB access page, fall 1996



Notes: The figure shows a screenshot to the access page to the AIDS Patent Database hosted on the CNIDR server in December 1996. Web-links to the page were prominently included on the home page of the USPTO and the National Science Foundation (NSF). The database included a search form (allowing for keyword, class and boolean search) as well as a browse page, including the full list and links to all hosted patents. Worldwide access to AIDS DB pages was possible with a dial-in modem and a telephone line. The data base included full-text and high-resolution images and drawings of all patents related to HIV/AIDS. Download pages were optimized for small (56k) bandwidths.

The AIDS DB was created as a joint effort by the United States Patent and Trademark Office (USPTO), the National Science Foundation (NSF), and the Clearinghouse for Network Information Discovery and Retrieval (CNIDR), by initiative of USPTO Commissioner Bruce Lehman.[10] Diagnosed HIV-1 infections had dramatically spread since the early-1980s, reaching a peak of > 20,000 cases in 1993 and causing a yearly mortality rate excess due to AIDS of almost 15,000 by 1995 in the U.S. alone.[11] Following the identification of the new human retrovirus found to be the etiological agent of AIDS in 1983, by late-1994 about 1,500 patents had been issued by the USPTO on technologies relating to HIV/AIDS. These were included in the initial launch version of the database, which was periodically updated with new HIV/AIDS-related patents issued until February 4th of 1997, to

[10] Sources: States News Service, October 26, 1994; Federal Technology Report, McGraw-Hill, November 10, 1994; USPTO Press Release #98-12.

[11] Source: U.S. Centres for Disease Control and Prevention (CDC).

host a final total of 2,916 patents. Figure 2.1 shows the access page to the AIDS DB which was provided through a link on the USPTO main website (see also Figure 2.B1 in Appendix 2.B).

After 1995, the project page also included links to the full-text of HIV/AIDS-related patents issued by the European and Japanese Patent Offices. The database system was designed and operated by CNIDR parent MCNC, a private, non-profit corporation located in Research Triangle Park, North Carolina. This included the development of a network-accessible search form and the optimization of electronic file formats allowing the inclusion of high-resolution images (complete with drawings, equations and diagrams) at enhanced compression.[12] Declared objective of the new online database was to connect and increase the informational efficiency between dispersed teams of researchers worldwide. Commerce Secretary Ronald Brown emphasized during the launch event on October 26th 1994:

> "This new online health care database is an excellent example of how our nation can best utilize the information superhighway to help improve people's lives, to help expand our knowledge and, most importantly, to connect ourselves with the resources and information previously beyond our reach." (Federal Technology Report, Nov. 10, 1994, p.4)

In fact, while all patents are by definition disclosed to the public, until then, researchers interested in technical information involving HIV/AIDS (or any other field) had to search paper files or local computer terminals at the patent office or the 78 depository libraries around the country [13], or rely on commercial services to conduct patent search surveys. The access to full patent documents from outside of library networks was even more difficult; The default remote delivery mode was ordering individual patent copies via mail or fax. The latter option was relatively faster, but also significantly more expensive, with delivery fees of several dozens of current USD per copy.[14] With the new online repository, the external search

[12] Source: PR Newswire, Oct. 26 1994, Financial News

[13] see Furman, Nagler and Watzinger, 2018 for an extensive discussion of the patent library system in the U.S.. In Europe, similar systems were in place in several countries, including the transnational PatLib library program from the European Patent Office.

[14] Source: Historical website of the USPTO, accessed Feb 12 1997.

costs to relevant prior art decreased suddenly for HIV/AIDS related knowledge, as Commissioner Lehman explained at the release ceremony:

> "Anybody with a [personal computer] and a modem will now be able to hook into this database anywhere in the country, and almost all scientific researchers in this area have that capability.[..] Researchers will be able to immediately get into the files and access that nugget of information they have been trying to get for years to complete their work." (States News Service, Oct. 26, 1994; Federal Technology Report, Nov. 10, 1994, p.4)

Projections regarding intensive usage seemed to be rapidly fulfilled: One and a-half years after launch, the AIDS DB recorded about 2.2 thousand requests per day on average (a total of 484 thousand request over a seven-month period). Moreover, these requests originated from a large number of connecting points worldwide, with a total of about 24 thousand distinct hosts served over a seven-month period.[15]

Next to providing electronic access to full patents, by being a disease-targeted repository, the AIDS DB also bore the potential to significantly lower search costs for HIV/AIDS researchers: As discussed in section 2.2.1, it is not-straightforward, even not for skilled inventors, to identify the applicability of a specific patent to a particular disease by inspecting the bibliographic information alone (e.g. provided on the patent front page or through bulletins/ newsletters). This is particularly challenging when searching patents of inventors out of the searching inventor's intellectual network and community. In fact, neither the USPC nor IPC patent classifications contain specific classes denoting HIV/AIDS (or other disease)-related inventions, as these span a broad range of different technology fields and domains. Moreover, the majority of AIDS DB patents did not include any textual reference to HIV/AIDS in title or abstract, which makes their retrieval through key word search comparably difficult. Therefore, the labeling as 'possibly relevant patent', by inclusion in the database, brought a major information retrieval advantage to AIDS DB patents over similar technologies.

For more than four years, the AIDS DB remained the only online repository for full patent information available on the world wide web. The USPTO started

[15] Source: CNIDR Web Server Statistics Dec 5 1996, accessed online here on June 11th 2020.

to expand its online holdings in 1996, offering a comprehensive web-searchable catalogue of bibliographic front cover information for over 2 million patents, across all fields. But it was not until late 1998 that the bulk volume of patents was made available online with their full text and images. Shortly after, in 1999, also the European Patent Office (EPO) launched its online platform, including full text information for all patents worldwide.[16] The AIDS DB project was discontinued in March 1999, and all hosted patents were included into the main USPTO database.

2.3 Data

2.3.1 Information about AIDS DB patents

I collect data from various sources. To retrieve the exact patents included in the AIDS DB, I web-scrape the archived historical pages of the CNIDR server. I recreate the full database content based on several snapshots of the AIDS DB browse pages containing the full list of US, EPO and JPO patents at different points in time between 1996 and 1998 (Figure 2.B3 provides an example of the layout of the scraped pages).[17] The archived snapshots also include links to the individual patent view pages, allowing to verify that the listed patents were indeed deposited with full text and images in the database (see Figure 2.B4 in Appendix 2.B).

While I know for each patent the database status at a certain point in time, the exact inclusion date is not recorded. Based on the most recent patents included in each recorded snapshot, I infer the average grant-to-database lag to be of 1-3 months, which is in line with historical information from the USPTO about the currentness of patents included in the database.[18] I am able to retrieve the patent numbers of all 2,916 U.S. patents that were deposited until February 1997 (1,668 of which were published at the time of the original launch in 1994, and 1,248 updated subsequently), as well as 695 European and 755 Japanese patents.

[16] Sources: Press releases/ historical archives of the USPTO and EPO websites.

[17] The chronologically first available snapshot containing the comprehensive database dates back to June 26th 1997 and was accessed here on June 11th 2020.

[18] See, e.g., here.

2.3.2 Patent universe data

I link the retrieved AIDS DB patent numbers to comprehensive information on the universe of patents worldwide in the EPO Patstat database (version: spring 2018), which constitutes the main data source for my analysis. Specifically, for each patent worldwide, this source provides information on filing, priority and publication date, documents part of the same international patent family, titles and abstracts, IPC technology classes and fields (based on Schmoch, 2008), raw inventor and assignee addresses, assignee sectors, as well as prior art references and citation links to all other patents. I supplement these data with specific information for U.S. patents concerning details on the patent prosecution process, namely examiners and examining art units (provided by Graham, Marco and Miller, 2015), and assigned USPC patent classes (Marco et al., 2015). To disambiguate inventor identities and geolocations for U.S. and European patents, I extensively rely on the data sets from Li et al. (2014) and Morrison, Riccaboni and Pammolli (2017). I assign patent locations to states, regions and metropolitan areas worldwide using geo-spatial boundary files provided by the United States Census Bureau, Eurostat and the OECD.[19] Further, I rely on the disambiguation of Marx and Fuegi (2020) to identify links to scientific publications referenced in U.S. patents. Ex-ante indicators of technological novelty are obtained from Verhoeven, Bakker and Veugelers (2016). To retrieve knowledge flows associated with the re-use of new keywords, I use the list of stemmed keywords in U.S. patents from Arts, Cassiman and Gomez (2018). Finally, I determine firm self-citations based on Bureau Van Dijk *Orbis* Intellectual Property Data, linking patents worldwide to consolidated ultimate owners.

2.3.3 Scientific publication universe data

In order to detect the broader scientific communities in which AIDS DB inventors are embedded in, I trace their publishing activities in the universe of scientific articles in the MEDLINE database indexed in PubMed.[20] For this purpose,

[19] Regions are aggregated to federal states in the U.S., Mexico and Australia, NUTS-1 regions in the E.U., prefectures for Japan, provinces for Korea and Canada, and districts in Israel. Metropolitan areas are based on Combined Statistical Areas (CSAs) for the U.S. and OECD Functional Urban Areas (FUAs) for the rest of the world.

[20] PubMed data is publicly available and can be accessed here: https://pubmed.ncbi.nlm.nih.gov/

I start from the author name disambiguation of all authors in PubMed, provided by Smalheiser and Torvik (2009) and Torvik and Smalheiser (2009), and links to their U.S. patents from Torvik (2018). Subsequently, I establish the link between author and inventor identities using a within-patent probabilistic matching procedure based on author-inventor name strings.[21] I further identify all publications of author-inventors relating to HIV/AIDS based on corresponding Medical Subject Headings (MeSH) terms assigned to the *human immune deficiency virus* and *acquired immunodeficiency syndrome* by the National Library of Medicine.[22] MeSH terms linked to publications are based on the MeSH tree version 2016.

2.3.4 Final database

The final assembled database contains detailed information about all patents deposited in the AIDS DB, and a large pool of comparison patents from the universe of similar technologies. It records all citation links to prior art and follow-up inventions, allowing to quantify the technological origins and cumulative impact of all patents included. Further, in order to evaluate the relation between cited and citing inventions in geographic and intellectual space, the data contains rich information about all inventors and assignees with their precise locations, prior patenting and publishing histories and embeddedness in the HIV/AIDS researching scientific communities.

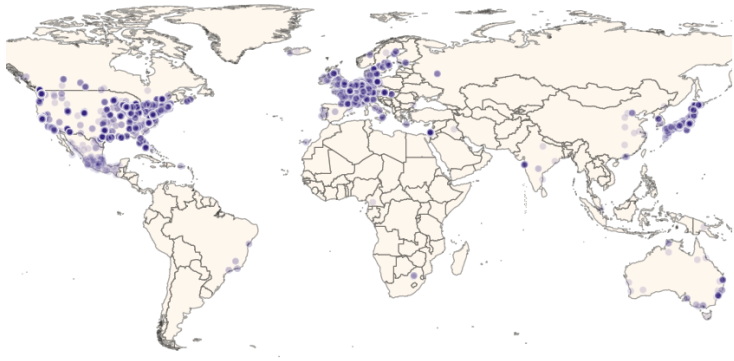
Figure 2.2 shows that HIV/AIDS-related technologies were developed almost exclusively in the Western hemisphere. Not surprisingly, the most active geographic hubs clustered in the North-East and West Coast of the U.S., Central and Western Europe and Japan. Further significant patenting activity originated

[21] I use a Jaro-Winkler similarity algorithm with varying acceptance thresholds. Random sample validation (N=200) of the matching approach yields a precision of 95,4% and recall of 97,8%.

[22] see: https://meshb.nlm.nih.gov/. Identified MeSH terms are: *AIDS; Acquired Immune Deficiency Syndrome; Acquired Immuno-Deficiency Syndrome; Immunodeficiency Syndrome, Acquired;AIDS Arteritis, Central Nervous System; AIDS Dementia Complex;AIDS Serodiagnosis; HIV Seropositivity; HIV Seroprevalence; Lymphoma, AIDS-Related; HTLV-III; Human Immunodeficiency Virus; Human T Cell Lymphotropic Virus Type III; Human T Lymphotropic Virus Type III; Human T-Cell Leukemia Virus Type III; Human T-Cell Lymphotropic Virus Type III; Human T-Lymphotropic Virus Type III; Immunodeficiency Virus, Human; Immunodeficiency Viruses, Human; LAV-HTLV-III; Lymphadenopathy-Associated Virus; Virus, Human Immunodeficiency; Human T-Cell Leukemia Virus*

from Israel and the East Coast of Australia.

Figure 2.2. Geographic Dispersion of AIDS DB Inventors

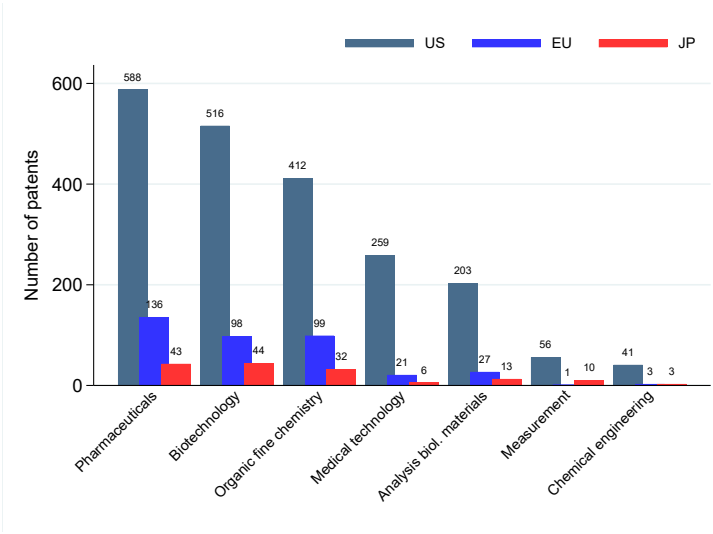


Notes: Shown locations are individual inventor addresses from all AIDS DB patents deposited from Oct 1994 until Feb 1997. Opacity grades indicate intensity of patenting activity. Geo-coded inventor addresses are provided by Morrison, Riccaboni and Pammolli (2017). Coordinates are geo-mapped using QGIS.

When looking at split counts of patents across main geographic areas and technology fields (see Figure 2.3), however, it becomes evident that the U.S. were by far the leading geographic area in HIV/AIDS-related treatments, accounting for the largest number of patents in each field. Pharmaceutical technologies constitute the largest field share in the database, closely followed by patents in biotechnology and organic fine chemistry. Further well represented fields were medical technology, analysis of biological materials, measurement and chemical engineering. The AIDS DB hosted HIV/AIDS-related patents from a total of 14 distinct technology fields, covering a broad range of inventive domains.

While geographically clustered, the HIV/AIDS inventor community was highly proliferated into small networks. A community detection based on Louvain modularity maximization (Blondel et al., 2008) reveals a large central community cluster, spanning the research groups of later Nobel laureates Françoise Barré-Sinoussi and Luc Montagnier at Institute Pasteur in Paris and the laboratory of Robert Gallo at the National Institutes of Health in Bethesda, MD, in the U.S., and a large number (> 300) of small and unconnected clusters a few researchers each. Moreover, my assembled data show that HIV/AIDS-technology research was strongly intertwined

Figure 2.3. Frequencies of AIDS DB Patents over Technology Fields and Geographic Origin



Notes: Bars show patent counts across technology fields by geographic areas. Counted are all patents deposited in the AIDS DB from Oct 1994 until Feb 1997. Fixed geographic areas are determined by the most represented geographical area among inventor locations of a patent (a random draw is taken in case of multiple). Technology fields are based on Schmoch (2008). Each patent is assigned to its most represented field (random draw in case of multiple). Seven further, less represented, fields are omitted in the graphic.

with advances in basic science; Many inventors listed on patents in the AIDS DB also ranked among the leading and most impactful fundamental science researchers in the area of the disease, as depicted in Table 2.A2 in Appendix 2.A; Several scientists (including Luc Montagnier, Samuel Broder, and William Haseltine) ranked both among the top-20 inventors and PubMed authors on HIV/AIDS-research by the end of 1996, and had large networks of collaborators (degree) in each sphere. This provides indication for a close connection between the technology and science frontier in the field.

2.4 Effects on Cumulative Patent Citations

This section outlines the empirical design and econometric results for the main analysis of the paper. Section 2.4.1 presents the sample construction and empirical model, and 2.4.2 the econometric results of the cumulative citation impact effects associated with the launch of the AIDS DB. Heterogeneous impact effects, in line with the mechanism of reduced search costs, are reported in 2.4.4.

2.4.1 Empirical strategy

As addressed above, the fundamental difficulty associated with identifying the marginal impact of an online repository, like AIDS DB, on cumulative invention arises from the need to isolate the intrinsic impact components of the embedded knowledge itself from the one of the access-enhancing institution. As discussed in Section 2.3.1, HIV/AIDS research originated a vibrant community of scientists, operating at the frontier of knowledge, and with a strong representation of public research institutions. Moreover, there was a concurring strong public interest in promoting and investment in HIV-related research. This suggests a positive selection of patents in the AIDS DB based on intrinsic quality and social value, which directly reflect on the potential to cumulative impact.

I solve the endogenous link problem by adopting a *within* AIDS DB comparison, which accounts for all unobserved factors related to selection into the database. In order to disentangle the effects of increased visibility (attention effect) from the one of better access, I estimate changes in cumulative impact across deposited patents as a function of additional up-front information on the technical content revealed by the explicit association with HIV/AIDS.

To determine the degree to which inclusion in the AIDS DB might have led to a shock in search costs, I leverage my knowledge about the conditions of external prior art search before the establishment of the AIDS DB. As discussed in Section 2.2.1, until then, inventors interested in retrieving patents related to HIV/AIDS could most exclusively rely on the bibliographic front-page information, provided in library terminals, bulletins or newsgroups, in order to identify those out of the bulk volume of existing patents and order full-text copies to verify their actual

content. While patent classes are not disease-specific, and associations from inventor names to HIV-research would typically require direct or indirect network ties, I focus particularly on the information conveyed in titles and abstracts; To detect whether a patent makes a clear reference to HIV/AIDS, I query all titles and abstract of AIDS DB patents for keywords of medical subject terms relating to HIV/AIDS, as defined by the National Library of Medicine.[23] The assumption behind this approach is that - while inventors are, without doubt, highly-educated specialists in their respective domains and perfectly capable to judge the precise content of pertinent patents upon closer inspection - the likelihood that they will detect a relevant HIV/AIDS-related patent out of a list of bibliographic information of numerous patents will be higher if a given patent makes a clear front-page reference to the disease. Accordingly, by marking all deposited patents as disease-related, the inclusion of a patent into the AIDS DB likely entailed a stronger reduction in search costs for patents not making front-page references to HIV/AIDS, compared to those making them, increasing the visibility of the former for related prior art search, conditional on same (online) accessibility. I subsequently divide AIDS DB patents into two categories: With vs. without front-page reference.[24]

While this within-comparison solves the positive selection of inclusion of patents into the repository, the criterion for unbiased inference requires, henceforth, these two groups to be comparable on all characteristics relating to cumulative diffusion except for the treatment status ("no reference"). To avoid comparing patents on different types of technologies within broader technological fields, which might have different dynamics of diffusion, I exploit the richness of information regarding the examination process of U.S. patents, and condition "no reference" and

[23] see: https://meshb.nlm.nih.gov/. Queried keywords are: *AIDS; Acquired Immune Deficiency Syndrome; Acquired Immuno-Deficiency Syndrome; Immunodeficiency Syndrome, Acquired; AIDS Arteritis, Central Nervous System; AIDS Dementia Complex;AIDS Serodiagnosis; HIV Seropositivity; HIV Seroprevalence; Lymphoma, AIDS-Related; HTLV-III; Human Immunodeficiency Virus; Human T Cell Lymphotropic Virus Type III; Human T Lymphotropic Virus Type III; Human T-Cell Leukemia Virus Type III; Human T-Cell Lymphotropic Virus Type III; Human T-Lymphotropic Virus Type III; Immunodeficiency Virus, Human; Immunodeficiency Viruses, Human; LAV-HTLV-III; Lymphadenopathy-Associated Virus; Virus, Human Immunodeficiency; Human T-Cell Leukemia Virus* I use the same list of terms to retrieve scientific articles relating to HIV in PubMed.

[24] Patents with front-page reference to HIV/AIDS accounted for about one third of AIDS DB patents (N=745).

control group ("with reference") patents to be examined in the same art unit.[25] Art units are the most granular inter-organizational units in the examination process of patent applications at the USPTO. Each art unit consists of a team of several patent examiners who specialize in a particular technology.[26] Using art units to detect technologically related patents has several advantages over the use of USPC patent classes in my study; First, given that patent applications are carefully screened and purposefully assigned to the competent art unit for examination, this avoids issues arising from randomness or misclassification that have been studied in the use of primary USPC (sub-)classes (Benner and Waldfogel, 2008; Arts, Cassiman and Gomez, 2018). Second, in particular for drugs and medical domains, art units provide a much more granular assessment device for specific technical content than the broader 3-digit classes, of which there are only eight in the USPC classification.

Next, within each AIDS DB patents - art units stratum, I further subdivide patent pairings based on whether they are assigned to a private firm or public institution, and whether they make prior art references to basic science, which have been widely shown to have significant influence on cumulative use and breadth of impact of technologies (Mansfield, 1995; Narin, Hamilton and Olivastro, 1997; Ahmadpoor and Jones, 2017). Finally, I pair patents, within these bins, based on coarsened invention filing and patent publication dates, and apply the weights of Iacus, King and Porro (2012) to ensure balance in the estimation.[27] By this, I keep all factors relating to the timing of invention, disclosure and online deposit constant across the sample groups. Several examples of patent pairs with vs. without front-page references are discussed extensively in Appendix 2.C.

To measure the realization of spillovers I predominantly rely on patent citations to AIDS DB patents as proxies. While these measures are widely established, they are known to be imperfect and noisy. In section 2.5.2 I, therefore, employ alternative metrics to identify knowledge flows. I use patent-level panel data to quantify the marginal effect of the AIDS DB on the cumulative rate of citations. Specifically,

[25] Given this constraint in data availability, I only consider patent family members filed at the USPTO.

[26] see Righi and Simcoe (2019) for an excellent discussion of the organization of art units at the USPTO.

[27] For ease of sample construction, I again assign a unique database deposit date to all patents in the same matched group. I based the unique date on the most frequent occurring, and earliest in case of multiple. Note, that the exact database deposit dates are only approximated.

I create a data set with yearly observations of citation counts for each patent in all years following its initial filing date. In line with prior literature [28], I remove inventor and applicant self-citations from the counts, as those do not reflect spillovers from external search. I am interested in determining the date closest in time to the inventive effort leading to the future patent. Therefore, I count a citation as a cumulative spillover with timing of the initial filing date of the citing patent. As initial filing date, I consider the priority date, for those patents with international priority, first or provisional filings, and the application filing date, for patents that are continuations or divisions of prior applications.

Table 2.1 reports summary statistics for the within-AIDS DB matched sample. My strict selection criteria allow to pair 1,367 AIDS DB patents.[29] A total of 11 technology fields and 57 art units are represented in the sample, suggesting that the latter are significantly more granular in technological scope then 3-digit classes. Most noticeable, the small difference in the level of citations received in the year pre-AIDS DB deposit across "no reference" and "reference" groups (see row (1), Table 2.1) is not statistically significant. Sample patents are highly comparable also on a broad range of relevant ex-ante patent-level characteristics, with no significant differences across the two groups, except for the small deviations regarding the mean number of prior art references and the scope of geographic patent protection (family size). In particular, patents without front-page references to HIV/AIDS do not appear to be more general in applicability, are not more often 'hit' patents (top-5% most highly cited in a given technology class-year cohort), nor more often recombinant (novel) patents.[30] They do not introduce more new keywords and do not reference scientific prior art in previously unconnected medical subjects to the technology class.[31] Patents across sample groups do also not significantly differ in terms of inventor team or assignee characteristics. Table 2.1 further shows the high

[28] e.g., Jaffe, Trajtenberg and Henderson (1993); Thompson (2006); Singh and Marx (2013)

[29] By this, my estimation sample covers about half of all originally deposited patents in the AIDS DB. Inference is limited to this subset.

[30] "Generality" is based on the index proposed by Trajtenberg, Henderson and Jaffe (1997). "Novel patents" are those making unprecedented combinations in technological prior art cited at the IPC-6-level, based on Verhoeven, Bakker and Veugelers (2016))

[31] New keywords are based on Arts, Cassiman and Gomez (2018). Medical subjects are retrieved based on MeSH terms assigned to scientific prior art references in PubMed (MeSH tree version 2018). SNPR disambiguation comes from Marx and Fuegi (2020)

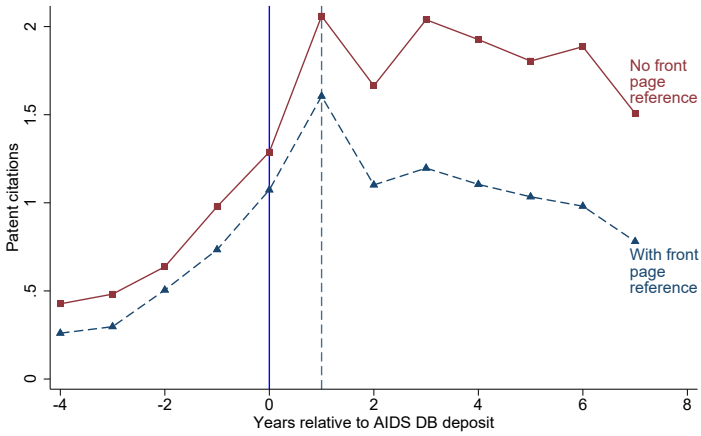
Table 2.1. Summary Statistics, Patents with vs. without Front Page Reference to HIV/AIDS

Within AIDS DB sample	No reference		With reference		Diff
	Mean	SD	Mean	SD	p-val.
Yearly patent family citations at AIDS DB deposit	1.29	2.77	1.07	2.17	.13
Generality index	.46	.25	.44	.25	.16
Share breakthroughs (top-5%)	.09		.09		.81
Share novel technologies	.25		.22		.21
Share introducing new words	.35		.31		.14
Share new-to-class medical subjects	.37		.34		.26
Number of patent references	9.25	9.57	7.56	8.12	.00
Share with scientific reference	.90		.90		.95
Number of scientific references	13.90	22.91	11.48	13.66	.03
Number of inventors	3.06	2.04	2.91	1.88	.18
Share of new inventors	.17		.18		.64
Share with author-inventors	.94		.95		.53
Number of author-inventors	2.62	1.90	2.49	1.65	.20
Patent family size	6.98	7.21	5.47	6.44	.00
Share private firm patents	.64		.64		.97
Assignee prior patent families	2.56k	11.15k	2.55k	5.29k	.98
DB-to-publication lag (m)	19.07	21.46	18.74	21.23	.79
DB-to-application lag (m)	55.81	27.94	56.15	28.47	.83
Number of patents	870		497		
Number of technology fields	11		8		
Number of examining art units	33		33		

Notes: Row (1) reports the group mean and standard deviation for yearly patent family citations to AIDS DB patents without vs. with front-page reference to HIV/AIDS for year t_0 relative to AIDS DB deposit. Inventor and applicant self-citations are removed from the counts. The following rows report ex-ante time-invariant characteristics. Control group patents consist of AIDS DB patents "with reference" to HIV/AIDS, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired "no reference" patents. Technology fields are based on Schmoch (2008). Sample observations are weighted according to Iacus, King and Porro (2012). Column (6) reports p-values from two-sample t-tests with unequal variances for differences in sample means. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *PubMed*, *BvD Orbis* and several disambiguations (links) between them (see Section 3.3 for details).

science-intensity of inventions in the sample: About 90% of included patents make prior art references to scientific publications, while the average within-patent share of inventors with prior scientific publications is even at 95%. Summary statistics also reveal strong involvement of public research institutions, accounting for one third of patent assignees. Average lags between deposit and patent publication (18 months) and application (55 months) indicate that I observe a large share of the sample patents for a significant time span before database inclusion.[32]

Figure 2.4. Group Means *within AIDS DB* Comparison Yearly Patent Citations



Notes: The figure plots trends in group means across AIDS DB patents without vs. with front-page reference to HIV/AIDS. Control group patents consist of AIDS DB patents "with reference" to HIV/AIDS, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired "no reference" patents. The *y*-axis scale reports levels of yearly patent family citations to patents in sample. Inventor self-citations are removed from the counts. The x-axis depicts years relative to online deposit (0). The dashed vertical line (1) indicates a 1-year lag of the database treatment, relative to deposit.The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *PubMed*, *BvD Orbis* and several disambiguations (links) between them (see Section 3.3 for details).

In order to evaluate whether small differences between the two sample groups might cause them to diverge dynamically, i.e. in pre-period trends of citations, I

[32] Online deposit of AIDS DB patents is assumed to take place on average one month after patent publication, compare Section 2.3.1

inspect group means over time prior to inclusion in the online database. Figure 2.4 plots almost perfectly parallel trends between "no reference" and "reference" patents in the relative periods until the AIDS DB deposit date, which is a necessary condition for inference of an average treatment effect on the treated, and suggests that the control group is well selected. Figure 2.4 also reveals that starting from $t+1$ differences in group means can be seen to substantially increase, while trends still follow largely parallel patterns. Note, that the spike in patent citations observed in the first period after deposit date is due to the drastic increase in overall patent filings at the USPTO immediately prior to the enactment of reforms to the U.S. patent system to implement provisions of the TRIPS agreement, that became effective on June 8th 1995 (see Figure 2.B5 in Appendix 2.B).[33]

I then compare within-patent changes in differences in citation rates across groups after AIDS DB deposit in a generalized difference-in-differences framework by estimating the following regression equation:

$$Y_{it} = \beta_1 * no\ reference_i \times post_{t-1} + patentFE_i + yearFE_t + \phi_y + \theta_{fy} + \epsilon_{it}, \qquad (2.1)$$

where i indexes patents, t indexes relative years to AIDS DB deposit, y indexes calendar years, and f indexes technology fields. The dependent variable measures the number of citations per relative year to deposit for each AIDS DB patent and, analogously, per relative year to deposit of the matched AIDS DB patent for each control patent. Control patents have the function to provide a reference level of citations that would have been received by a matched AIDS DB patent from database deposit in the absence of a shock to search costs, e.g. if the database would have been a non-disease specific online repository. The coefficient β_1 measures changes in citation rates to AIDS DB patents without front-page reference to HIV/AIDS, after deposit, relative to the group of patents making such references, which are the excluded reference category. The interacted $post_{t-1}$ indicator denotes the one-year lagged post-deposit status.[34] The regression model includes a full set of patent fixed effects. These control for all permanent differences across patents affecting

[33] These included, among others, a change in maximum patent term (from 17 years post-grant to 20 years post-filing), as well as the establishment of provisional applications, in line with the TRIPS agreement.

[34] Allocating the treatment to set in one year after database deposit seems a conservative lower bound for intertemporal spillovers from newly accessed prior art information to plausibly affect follow-up

the incoming citation patterns, for instance, the quality and complexity of an invention, its geographic origin, institutional context or technological field. This allows to keep constant a broad range of patent-level characteristics in a flexible and non-parametric manner. Note, that a group fixed effect (e.g., an indicator for all "no reference" patents) is omitted from the specification as it would be perfectly collinear with the sum of patent fixed effects of all patents in that group.

The model also includes fixed effects for relative years to the AIDS DB deposit date. These account for dynamic changes in the rate of citations over the life cycle common to all patents. Given that some patents enter the sample (i.e., are applied for and granted) several years before the AIDS DB launch, while others are deposited almost immediately after grant, this prevents results to be disproportionately driven by, e.g., more recently granted patents. Note, that the sum of pre-treatment and post-treatment relative year fixed effects is collinear to a $post_{t-1}$ period indicator, which is therefore also omitted from the specification. To control for the confounding influence of shocks possibly affecting citation rates over time in the overall economy or the patent system (e.g., the enactment of the TRIPS Agreement in 1995), the regression further includes a full set of calendar year fixed effects (captured by the parameter ϕ_y). Finally, in the preferred specification, I include linear field-year trends (θ_{fy}) to control for idiosyncratic variation in productivity of specific technology fields, for example, in up-rising biotechnology in the mid-1990s.[35]

I estimate regression (2.1) on a symmetric sample window of five years preceding and five years following the switching of the $post_{t-1}$ indicator, i.e., for example,

inventions. A one-year lag to assess the manifestation of technology spillovers is established also in prior work, e.g. Bloom, Schankerman and Van Reenen (2013).

[35] Recent contributions in the treatment effects literature have expressed concerns regarding the naive use of two-way fixed effects estimators with staggered treatment adoption arising from bias induced by unequal weighting of individual two-by-two estimators due to variance in treatment effects over time (Goodman-Bacon, 2021; De Chaisemartin and d'Haultfoeuille, 2020). In my specification, these concerns are mitigated in several ways: First, treated and control group units are fixed over time, without within-unit variation ("switching") in group status at any point of the sample. Second, to account for variance in treatment effects over time, next to event years, my model includes year dummies to capture absolute time-varying effects (as suggested by Goodman-Bacon (2021)) and units across groups are matched pairwise on timing. Third, the test suggested by De Chaisemartin and d'Haultfoeuille (2020) formally excludes the presence of any negative weights in the sum of all average treatment effects in my sample.

ranging from October 1991 to October 2000 for patents deposited in the initial launch of the AIDS DB database on October 26th 1994. The fact that some patents are represented only for later years in the sample window can be considered innocuous, as it is accounted for by the pairing (and weighting) of patents across groups, based on same filing/grant timing and individual year fixed effects. I report regression results of cumulative citation models, primarily, as OLS estimates.[36] For this, I standardize the number of yearly citations to mean zero and standard deviation one within technological fields.[37] This makes effect sizes on citation rates comparable across fields, despite the fixed functional form of the model, and avoids well-studied problems arising from the use of log-linearizations on distributions inflated with many zeros or the commonly used ln(n+1) transformation of the data (e.g., Silva and Tenreyro, 2006). For comparison and robustness, however, I complement all results with estimates from Poisson pseudo-maximum likelihood models.[38] Given the underlying count distribution, these are likely to be the most efficient estimator and model the conditional mean of citations most accurately. Moreover, they provide asymptotically correct standard errors even with over-dispersion, as citation counts are likely to exhibit (Silva and Tenreyro, 2011). Given the panel structure of my data, and the common concern of possibly serially correlated regression residuals leading to deflated OLS standard errors and resulting over-rejections of the null hypothesis (Bertrand, Duflo and Mullainathan, 2004), I cluster all standard errors at the patent-level.

2.4.2 Within AIDS DB Comparison: Effects on Patents without vs. with Front Page References to HIV/AIDS

Table 2.2 reports the econometric results from the estimation of regression (2.1) on the within AIDS DB sample. In column (1) of Table 2.2, I first estimate the model without field-specific linear trends. Starting from one year after deposit, patents without front-page reference to HIV/AIDS reveived on average .14 standard deviations more in cumulative citations relative to control group patents with front-page references (significant at the 1% level). Compared to the pre-deposit mean

[36] as in Furman and Stern (2011); Galasso and Schankerman (2015); Biasi and Moser (2020)
[37] compare Iaria, Schwarz and Waldinger (2018) for a similar approach
[38] similar to Bryan and Ozcan, 2020; Biasi and Moser, 2020

of citations, this implies a relative increase of .35 citations per year (about +29%) for the average "no reference" patent in the sample. The effect is slightly smaller (+.12 standard deviations, +26%) when including field-specific time trends, in my preferred specification in column(2) (but equally significant at 1%), suggesting these to explain about one tenth of the dynamic differential.[39]

I check the robustness of this finding across several alternative models: One caveat regarding the validity of these results might arise due to patents without specific references to HIV/AIDS covering more 'general' technologies, which intrinsically experience a broader diffusion outside of the HIV-research community, possibly explaining the positive delta. On the other hand, effects could also be driven by new entry of inventors with more diverse backgrounds. To investigate this, in columns (3) of Table 2.2, I re-estimate the model considering only follow-up citations originating from the community of established HIV/AIDS inventors, identified as those appearing on patents indexed in the original AIDS DB. The point estimate of the treatment parameter in column (3) indicates that effects were larger for this sub-group (+28%, significant at the 1% level), however accounting for the vast majority of incoming citations. Another concern as to which extent my results capture knowledge spillovers from external search might arise from the fact that HIV-research was predominantly conducted by large institutions. Hence, a significant part of the observed effect could be due to within-firm spillovers resulting from, e.g., an intensification of efforts in HIV/AIDS research and not be due to online deposit. In column (4) of Table 2.2, next to inventor and applicant self-citations, I therefore also remove ultimate-owner level firm self-references from the dependent variable citation counts.[40] Estimates show that the effect is magnified (increase of .15 standard deviations, significant at 1%) when excluding these citations.[41]

[39] In unreported results, I do not find any significant heterogeneity of effects conditional on pre-DB citation levels. In particular, I also don't find that inclusion into the AIDS DB would have impacted the likelihood for "no reference" patents of entering the top-impact ranks of a given field-year distribution. This is consistent with the view that patent that already had a strong visibility in a field, evidenced by prior received citations, would have not disproportionately benefited from online indexing.

[40] Consolidated firm-level self-citations are based on the *BvD Orbis* firm-patent link.

[41] Unfortunately, the Orbis firm-patent link information is available only for about 60% of firm patents in my sample. I, therefore, do not rely on these in the preferred specification.

Table 2.2. Effect for Patents without HIV/AIDS Front Page Reference, within AIDS DB

Dependent variable:	OLS					
Number of patent citations	(1)	(2)	(3)	(4)	(5)	(6)
No reference × post$_{t-1}$	0.135***	0.123***	0.134***	0.150***	0.181***	−0.046
	(0.044)	(0.044)	(0.042)	(0.044)	(0.046)	(0.044)
Abstr reference × post$_{t-1}$					0.101	
					(0.066)	
Patent fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Time/ year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Field time trends		Yes	Yes	Yes	Yes	Yes
DB inventor cites only			Yes			
Excl. firm self-cites				Yes		
In prosecution cites only						Yes
Observations	12,192	12,192	12,183	12,190	12,192	12,020
Number of patents	1,366	1,366	1,366	1,366	1,366	1,360
R^2	.388	.389	.433	.495	.389	.060
Mean at t_0	1.210	1.210	1.178	1.194	1.210	.016
SD at t_0	2.570	2.570	2.498	2.554	2.570	0.152

Notes: Each column reports parameter estimates of regression (2.1) for the matched panel of U.S. AIDS DB patents without vs. with front-page reference to HIV/AIDS. The dependent variable measures the yearly number of family citations for years $t-4$ to $t+5$ relative to the one-year lagged online date. Inventor and applicant self-citations are removed from the counts. The reference category consists of AIDS DB patents "with reference" to HIV/AIDS, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired "no reference" patents. The number of citations is standardized to mean zero and standard deviation one within technology fields (fields based on Schmoch (2008)). "DB inventor cites" are citations originating exclusively from HIV/AIDS inventors indexed in the AIDS DB. "Firm self cites" are self-citations at the ultimate owner-level. "In prosecution cites" are citations exclusively from patents already under examination at the time of DB inclusion of the cited patent. Sample observations are weighted based on Iacus, King and Porro (2012). Standard errors are clustered at the patent level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *BvD Orbis* and several disambiguations (links) between them (see Section 3.3 for details).

In column (5) of Table 2.2, I seek further support for lower search costs driving these results, by splitting up patents in the reference category based on how visible the HIV/AIDS reference was on their front page. Precisely, I distinguish between patents making a textual HIV/AIDS reference in the abstract section only, and those making a reference already in the document title. The idea behind this

is that retrieving information from patent abstracts required inventors to engage with a patent document already substantially more than just scrolling through lists of newly granted patents (including only titles, inventors, and classes) when searching for HIV-related prior art, implying somewhat higher search costs, and increasing the likelihood of overlooking a relevant patent making a reference only in the abstract. Hence, I add the category interaction of these patents to the model, comparing effects for lower search costs in cascading manner, relative to the background rate of patents with HIV/AIDS reference in the title. Results show the largest increase in citations for the "no reference" category (.18 standard deviations more, significant at 1%), while effects are also positive, and about 45% smaller in size (although not significant below the 10% level in the OLS estimation). These patterns are widely in line with a reduction of search costs as mechanisms driving my results.

Another alternative explanation for the observed pattern might simply be that patent examiners became more likely to add references to to certain AIDS DB patents, as their active involvement in the assembly process likely increased their attention to them as well.[42] I evaluate the severeness of this concern, in column (6) of Table 2.2, by re-estimating regression (2.1) *only* counting citations given from patents that were already under prosecution at the time of online deposit of the cited AIDS DB patent, i.e. filed before and granted after the AIDS DB deposit date.[43] These citations are very likely to be given by examiners rather than by the applicants.[44] They also cannot reflect knowledge spillovers from external search through online access to the AIDS DB, as patent applications were already pending and, accordingly, the inventive search process must have been terminated at the time

[42] Examiner citations typically account for about 40% of citations included in U.S. patent documents. As examiner added citation cannot reflect knowledge flows among inventors, they have the potential to introduce significant noise in these type of analyses (Alcacer and Gittelman, 2006; Thompson, 2006). Unfortunately, the precise information about examiner-added citations is given only for U.S. patents granted after January 2001 and, therefore, not available for the vast majority of citing patents in my sample. Note, however, that as long as their share does not unilaterally change over time for either HIV/AIDS or control group patents, the presence of examiner added citations is innocuous to my estimation given the specification I employ.

[43] In this case, deviant from my standard approach, I consider as citation date the grant date of a citing patent, which is arguably closest to the examination moment.

[44] see Arora, Belenzon and Lee (2018) for a similar approach

of online deposit. The point estimate of β_1 in column (6) for changes in citations added during prosecution after database deposit across "no reference" and "with reference" patents tends slightly negative, but is highly insignificant. This suggests that the launch of the AIDS DB had no influence on citation practices of patent examiners, at least not within database indexed patents.

Table 2.3. Effect for Patents without HIV/AIDS Front Page Reference, excluding Top-Citing Examiners

Dependent variable:	*Excl. top 1% citing examiners*		*Excl. top 5% citing examiners*	
Number of patent citations	(1)	(2)	(3)	(4)
No reference × post$_{t-1}$	0.132***	0.120**	0.146***	0.131**
	(0.047)	(0.047)	(0.053)	(0.051)
Patent fixed effects	Yes	Yes	Yes	Yes
Time/ year fixed effects	Yes	Yes	Yes	Yes
Field time trends		Yes		Yes
Observations	12,192	12,192	12,192	12,192
Number of patents	1,366	1,366	1,366	1,366
R^2	.412	.414	.359	.361
Mean at t_0	.962	.962	.709	.709
SD at t_0	2.345	2.345	2.190	2.190

Notes: Each column reports parameter estimates of regression (2.1) for the matched panel of U.S. AIDS DB patents without vs. with front-page reference to HIV/AIDS. The dependent variable measures the yearly number of family citations for years $t-4$ to $t+5$ relative to the one-year lagged online date, excluding citations from patents inspected by the top 1% (n=10) and top 5% (n=50) of examiners in number of patents citing the AIDS DB. Inventor and applicant self-citations are removed from the counts. The reference category consists of AIDS DB patents "with reference" to HIV/AIDS, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired "no reference" patents. The number of citations is standardized to mean zero and standard deviation one within technology fields (fields based on Schmoch (2008)). Sample observations are weighted based on Iacus, King and Porro (2012). Standard errors are clustered at the patent level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *BvD Orbis* and several disambiguations (links) between them (see Section 3.3 for details).

I further explore the possibility of confounding influences originating from

patent examiner behaviour in two ways: First, I re-estimate the main specification in Table 2.2 excluding from yearly citation counts those references made by patents inspected by USPTO examiners who accounted for a large number of citations to AIDS DB patents after launch of the database. By this, I attempt to rule out the competing explanation that higher citation rates to AIDS DB patents without front-page references to HIV/AIDS could have been driven by a few very actively citing examiners whose attention was drawn towards these previously less visible inventions. Table 2.3 provides the corresponding estimates: When excluding citations from patents under review by the ex-post 10 most citing examiners (out of 1,016 total citing examiners), representing the top 1% and accounting for about 20% of total yearly citations to AIDS DB patents, the estimated citation premium is qualitatively robust and with relatively +32% even slightly larger in size (column (1) of Table 2.3, significant at the 1% level). Accounting for variation in citations across fields over time in column (2) of Table 2.3 yields an adjustment of the effect to +29% (point estimate 0.12, significant at the 1% level), which is perfectly consistent with the main result in Table 2.2. The estimated yearly citation delta is even considerably higher (+ 40%) when excluding the 50 most frequently citing examiners, as reported in columns (3)-(4) in Table 2.3.

Second, I investigate changes in the individual citing behaviour in response to the AIDS DB launch across all patent examiners at the USPTO. For this, I construct a data set where each observation provides the count of citations from patents inspected by a given examiner in a given year (based on publication year) to a patent in the AIDS DB sample.[45] Including a full set of citing examiner fixed effects, results reported in Table 2.A1 in Appedix 2.A show no significant deviation for examiners to cite "no reference" compared to "with reference" AIDS DB patents post online-deposit, relative to their individual sample mean.[46] This is robust for

[45] To ensure a minimal relatedness to contents, I restrict the examiner sample to only examiners who inspected at least one patent with a prior art citation – either before or after database deposit – to an AIDS DB patent during the sample period.

[46] The idea behind this approach is that, as shown by prior literature, the assignment of patents to individual examiners can be considered as quasi-random within art units. Therefore, a large share of the variation stemming from increased applicant citations to "no reference" AIDS DB patents should be absorbed be the examiner fixed effects, unless there would be a significant change in citation behaviour of the examiners themselves.

citations from only examiners who were previously to online-deposit citing HIV patents, as well as from only those citing in the same technology field and art unit (columns (2)-(4) in Table 2.A1). Taken together, these sensitivity checks mitigate concerns that (changes in) examiner behaviour would have significantly affected the observed increase in citation rates to "no reference" patents in the AIDS DB, and support my interpretation of the estimated effect as elasticity of (a reduction in) search costs on the inventors' side.

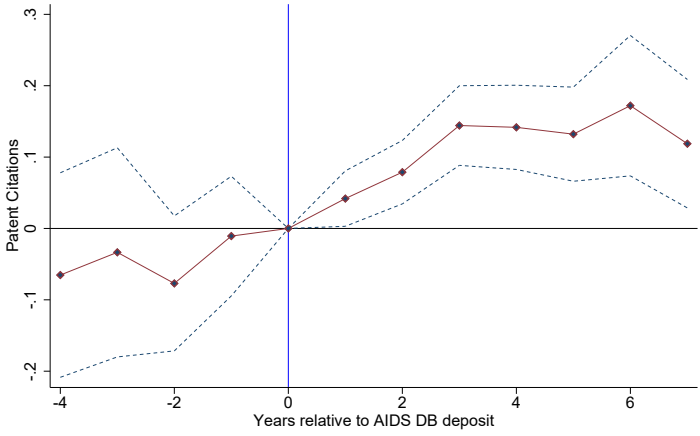
In Appendix 2.A is show further robustness of the entirety of the findings in this Section with quantitatively largely unchanged results: In Table 2.A3 in Appendix 2.A, I re-estimate all models with Poisson pseudo-maximum likelihood, which is the more efficient estimator given the count nature of the citation data, yielding slightly larger effect sizes. In Table 2.A4 I match and include only AIDS DB patents which are observed throughout all sample years (in order to form a balanced panel from $t-4$ to $t+5$), and show that results are robust and larger in magnitude for this sub-set. To address concerns that patent citations might exhibit exponential rather than linear cumulative growth rates, and accordingly small initial differences could result in large differences over time, fully or partially explaining the effect in the post-period, in Table 2.A5 of Appendix 2.A I show robustness and substantially larger estimates for a sub-set of "no reference" and "with reference" patents additionally matched on yearly pre-period citation levels (and, accordingly, trends). Finally, in Table 2.A6 of Appendix 2.A, I provide alternative results for effects on impact weighted forward citations, suggesting real economic effects behind the observed increased knowledge flows.

2.4.3 Timing of effects

To investigate the timing of the aggregate attention effects reported in Section 2.4.2, I re-estimate regression (2.1) with yearly coefficients, by interacting the "no reference" indicator with a set of individual year dummies for $t-4$ to $t+7$ relative to the database date (excluding the year of deposit as reference year). Figure 2.5 plots the corresponding point estimates within 95% confidence intervals. There are no significant differences estimated between citation trends of "no reference" and "with reference" group patents in the years prior to database inclusion, suggesting

that differences in pre-trends cannot explain the results. On the other hand, the figure shows a steep relative increase in the rate of citations to "no reference" patents in the years following their online availability, setting in highly significantly after one year, and reaching a plateau around the third year post-deposit and another subsequent peak after 6 years.

Figure 2.5. Yearly Effect for Patents without HIV/AIDS Front Page Reference



Notes: The figure plots parameter estimates from regression (2.1) with yearly coefficients for $t-4$ to $t+7$ relative to online deposit for the matched panel of U.S. AIDS DB patents without vs. with front-page reference to HIV/AIDS. The dependent variable measures the yearly number of family citations (inventor and applicant self-citations excluded). The year of deposit is omitted from the regression. The reference category consists of AIDS DB patents "with reference" to HIV/AIDS, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired "no reference" patents. Sample observations are weighted based on Iacus, King and Porro (2012). 95% confidence intervals are based on clustered standard errors. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *BvD Orbis* and several disambiguations (links) between them (see Section 3.3 for details).

Figure 2.4.2 also indicates a stable and persistent effect over time, as estimated differences, first, gradually increase and seem decline only towards the very end of the sample (indicating the years 2000-2003), following the discontinuation of the AIDS DB. Noticeably, estimated yearly differentials seem to remain mostly unaffected by the launch, first, of the comprehensive bibliographic online database

of the USPTO (in 1997, which corresponds to year $t + 3$ for the initial cohort of patents uploaded in 1994) and, second, the full-text and images online catalogue including all U.S. patents (in 1998) and EPO Espacenet (1999). In fact, these newly launched databases mostly levelled out differences in external access costs, but only to a lesser extent search costs, as, unlike the AIDS DB, these were not disease-specific repositories. This observation provides further support for the believe that the observed effects are, indeed, caused by a unilateral shock to search costs for "no reference" AIDS DB patents rather intrinsic quality differences or online electronic copy accessibility provided by means of the online repository.

2.4.4 Mechanism: Differential effects for patents with intrinsically higher search costs

I further investigate whether AIDS DB indexing particularly benefited the cumulative diffusion of patents that are associated with intrinsically higher search costs: Private firm patents, recombinant patents, and relying on medical subjects new to a technology class.

Table 2.4 shows estimation results for regression (2.1) as triple-differences for heterogeneity of marginal impact of the online repository for split-samples of these patents.[47] Columns (1)-(2) of Table 2.4 report differential effects for the group of private firm patents. As discussed in Section 2.2.1, these patents are subjected to adverse incentives of private firms against information disclosure towards rivals and particular prone to attempt to conceal the nature of the underlying inventions. Accordingly, I expect the shock to search costs from the disease-specific link to have been disproportionately higher for these patents. Results in Table 2.4 indicate that corporate assignee patents without front-pages references to HIV/AIDS received .17 standard deviations in citations more after AIDS DB deposit than "no reference" patents from public research institutions (e.g. universities, government research institutes, hospitals, etc.). Relative to the average patent in the sample this implies an additional increase of 35% (significant at the 10% level). Column (2) of Table 2.4

[47] This econometric specification compares changes in cumulative citations of, e.g., *private firm* patents with front-page references to HIV/AIDS to *private firm* patents without such references, and analogously for the sub-groups of novel patents and patents linking to new medical subjects.

Table 2.4. Differential Effects for Patents with Intrinsically Higher Search Costs, within AIDS DB

Dependent variable:	*Private firm patents*		*Novel patents*		*New-to-class MeSH*	
Number of patent citations	(1)	(2)	(3)	(4)	(5)	(6)
No reference × $post_{t-1}$ × *cat*	0.166*	0.119	0.252**	0.214*	0.291*	0.291*
	(0.086)	(0.086)	(0.109)	(0.109)	(0.176)	(0.156)
$Post_{t-1}$ × *cat*	-0.036	-0.008	-0.101	-0.092	-0.304**	-0.278**
	(0.072)	(0.068)	(0.070)	(0.076)	(0.148)	(0.129)
Main category interactions	Incl	Incl	Incl	Incl	Incl	Incl
Non-new MeSH interactions					Incl	Incl
Patent fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Time/ year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Field time trends	Yes	Yes	Yes	Yes	Yes	Yes
DB inventor cites only		Yes		Yes		Yes
Observations	12,192	12,183	12,192	12,183	12,192	12,183
Number of patents	1,366	1,366	1,366	1,366	1,366	1,366
R^2	.441	.433	.441	.434	.442	.434
Mean at t_0	1.210	1.178	1.210	1.178	1.210	1.178
SD at t_0	2.570	2.498	2.570	2.498	2.570	2.498

Notes: Each column reports parameter estimates of regression (2.1) split up as triple-differences for heterogeneity of effects on patents associated with intrinsically higher search costs in the matched panel of U.S. AIDS DB patents without vs. with front-page reference to HIV/AIDS. The $post_{t-1}$ parameter captures relative changes in citations to patents with front-page reference to HIV/AIDS in each split-sample category. Main category × $post_{t-1}$ interactions are included in all models. The dependent variable measures the yearly number of family citations for years $t-4$ to $t+5$ relative to the one-year lagged online date. The number of citations is standardized to mean zero and standard deviation one within technology fields (fields based on Schmoch (2008)). Inventor and applicant self-citations are removed from the counts. The reference category consists of AIDS DB patents "with reference" to HIV/AIDS, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired "no reference" patents. In columns (1)-(2) report differential effect estimates for the sub-sample of private firm patents. Columns (3)-(4) report corresponding estimates for novel patents. As "novel" are considered patents making novel combinations of technological prior art classes (IPC-6 level), following Verhoeven, Bakker and Veugelers 2016. Columns (5)-(6) show heterogeneous effects for the split-sample of patents referencing scientific prior art in medical subject terms (MeSH) that have not been previously linked to their respective technology class. A full set of interactions for patents making non-new-to-class medical subjects references are included. "DB inventor cites" are citations originating from HIV/AIDS inventors indexed in the AIDS DB. Sample observations are weighted based on Iacus, King and Porro (2012). Standard errors are clustered at the patent level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *BvD Orbis* and several disambiguations (links) between them (see Section 3.3 for details).

investigates robustness of this finding counting only incoming citations from the group of HIV/AIDS inventors, indexed in the AIDS DB. The estimated coefficient is equally positive, yet smaller in size and not significant below the 10% level. For both cases of citation counts, the relative changes after database deposit for private firm patents with front-page references to HIV/AIDS are close to zero and not statistically significant.[48]

Next, I compare differential effects for patents departing from existing trajectories of search, making them ceteris paribus more difficult to retrieve, e.g. if inventors aim to assess the relevance of new advances by inspecting the prior art cited.[49] First, I evaluate marginal impacts on recombinant (novel) patents (e.g., Fleming, 2001). I identify these as making new combinations of previously uncombined technology classes in the prior art they cite, using the measure suggested by Verhoeven, Bakker and Veugelers (2016) at the IPC-group level (IPC-6). Columns (3)-(4) of Table 2.4 show that effects on novel patents without front-page references to HIV/AIDS were .25 standard deviations larger compared to non-novel "no reference" patents (significant at the 5% level, + 54% relative to baseline) and that this pattern was robust for citations incoming from HIV/AIDS inventors (significant at the 10% level, given slightly reduced effect size). Novel patents *with* front-page reference to HIV/AIDS, at the same time, did not exhibit a significantly different change in citations relative to the background rate of non-novel "with reference" patents. Finally, I assess heterogeneity in impact for patents making scientific prior art references to articles in PubMed which were indexed in MeSH terms previously not linked to the technology class of the citing patent. Coefficient estimates in columns (5)-(6) of Table 2.4 show similar patterns for "no reference" patents making such new connections to scientific underpinnings, significant at the 10% level in both

[48] In unreported additional estimation results, I do not find significant heterogeneity of effects for other institutional categories of patent applicants, namely universities, hospitals and government research organizations. In particular government institutions, such as e.g. the National Institutes of Health in the U.S. or Institut Pasteur in France, held a leading and exposed role in the development of HIV-related treatments. Therefore, their patents enjoyed high visibility irrespective of the AIDS DB. These observations are consistent with the view that database inclusion would have disproportionately benefited the diffusion of private sector invention. Note however, that for some of the non-corporate assignee categories my sample size is very small, e.g. hospitals (n = 18), which makes it statistically difficult to estimate differential effects.

[49] Prior art references are also included on the front page of patents documents.

models, while the differential effect of database deposit for this group was opposite in patents with front-page HIV/AIDS references. These patterns are in line with my predictions of these groups of patents experiencing larger marginal impact from the disease-specific link established by AIDS DB indexing, given previously higher retrieval costs.

Also the recency since availability of the online deposited patents may affect search costs associated with their retrieval. As age of knowledge is intrinsically linked to higher diffusion levels, patents that had already been granted several years before their AIDS DB inclusion might have experienced relatively lower rates of excess citations from online indexing, ceteris paribus. On the other hand, due to the short-term higher visibility of new inventions associated with their recent publication and announcement in the *USPTO Patent Gazzette*, it is also thinkable that more dated patents would have benefited relatively more from the additional attention drawn to them. In line with these contradicting predictions, I do not find any differential effects conditional on recency since first publication of the patented knowledge. As reported in the results of Table 2.A7 in Appendix 2.A, the citation premium to "no reference" patents relative to patents with front-page reference to HIV/AIDS is estimated to be homogeneous across the patent age distribution.

My investigation of heterogeneity of effects also does not yield any significant differences based individual countries or aggregated geographic areas of origin of patents. Obviously, variation in our sample on this dimension is relatively contained, as HIV research was strongly clustered and concentrated mostly in the U.S., and to a much lesser extent in Europe and Japan (see 2.3). This makes it difficult to meaningfully estimate differential impact effects for sub-samples of locations outside of these hubs. I further do not find heterogeneity based on individual level features of pre-database centrality and degree of connectedness of inventors within the HIV-research community (see Table 2.A2 in Appendix 2.A), in particular no disproportional gains from higher visibility of patents from more peripheral inventors to the community.[50] This motivates the further inquiry of differences in impact of the AIDS DB at the receiving end of the knowledge spillover, conducted in Section 2.5.3.

[50] Estimation results unreported.

2.4.5 Second order effects on citations to referenced scientific publications

Finally, I investigate whether the higher visibility associated with inclusion in the database of patents with previously higher search costs generated second order effects on visibility of the scientific prior art applied in these patents. The showcasing of "hidden" technologies linked to the treatment and diagnostics of HIV/AIDS may have further led to a socially desirable display of useful scientific knowledge and revealed potential for new modes of application of fundamental insights. This seems particularly standing to reason, given the closeness and strong reliance on science of inventions in the fields relevant to HIV/AIDS research.

To estimate second order effects on the subsequent use in technology of papers included among scientific references in patents in the AIDS DB, I compute yearly cumulative patent citation rates to each PubMed article cited by a patent in the sample.[51] In a next step, I link these reference citations back to the characteristics and timing of AIDS DB deposit of the focal patent(s) and construct an SNPR panel, where each observation is a PubMed article referenced in an AIDS DB patent cited in a given year relative to the online deposit (of the AIDS DB patent).[52]

Table 2.5 reports results for the estimation of equation 2.1 on this SNPR-level panel. The coefficient estimate in column (1) of Table 2.5 implies that scientific articles referenced in a AIDS DB patent without front-page reference to HIV/AIDS experience an additional .071 standard deviations in yearly citations after online deposit of the referencing patent compared to patents with obvious front-page links to HIV/AIDS (signifcant at the 1% level). This corresponds to an increase of + 17.9% in patent references to these articles relative to the pre-treatment level. When accounting for technology field trends, in column (2) of Table 2.5, this effect is estimated to be even larger, at + 18.6% (significant at the 1% level). In column (4) of Table 2.5, I include scientific publication fixed effects in the model, in order to account for idiosyncatic, time-invariant differences in potential of technological

[51] Patent-to-article citations are sourced from Marx and Fuegi (2020)

[52] Note that in this case, unlike in the main estimation, it is possible that scientific articles appear both in the treated and control group, as they may be referenced by multiple patents. In the case of articles cited simultaneously in both a "no reference" and "with reference" patents, inference will be limited to variation stemming from different timing and/ or references unique to one of the two groups.

Table 2.5. Second Order Effect on SNPRs in Patents without HIV/AIDS Front Page Reference

Dependent variable:	OLS			
Number of patent citations to scientific references	(1)	(2)	(3)	(4)
No reference × post$_{t-1}$	0.071***	0.074***	0.074***	0.069**
	(0.026)	(0.026)	(0.028)	(0.027)
Patent fixed effects	Yes	Yes	Yes	Yes
Time/ year fixed effects	Yes	Yes	Yes	Yes
Field time trends		Yes	Yes	Yes
Referenced paper fixed effects			Yes	Yes
Paper year time trends				Yes
Observations	96,532	96,532	96,532	96,472
Number of patents	1,086	1,086	1,086	1,086
Number of scientific papers	7,848	7,848	7,848	7,843
R^2	.318	.318	.690	.692
Mean at t_0	2.464	2.464	2.464	2.464
SD at t_0	6.197	6.197	6.197	6.197

Notes: Each column reports parameter estimates of regression (2.1) for the matched panel of scientific references to articles in PubMed in U.S. AIDS DB patents without vs. with front-page reference to HIV/AIDS. The dependent variable measures the yearly number of family citations to scientific articles referenced in AIDS DB patents for years $t-4$ to $t+5$ relative to the one-year lagged online date. Inventor and applicant self-citations are removed from the counts. The reference category consists of AIDS DB patents "with reference" to HIV/AIDS, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired "no reference" patents. The number of citations is standardized to mean zero and standard deviation one within technology fields (fields based on Schmoch (2008)). Sample observations are weighted based on Iacus, King and Porro (2012) and 1/ number of SNPRs in each patent. Standard errors are clustered at the patent level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *BvD Orbis* and several disambiguations (links) between them (see Section 3.3 for details).

applicability across scientific discoveries. The coefficient estimate is robust and significant at the 1% level. We additionally controlling for variation of citation rates over the life-cycle of publications, with publication year × citing year fixed effects, in column (4) of Table 2.5, the relative effect size is minimally reduced to + 17.4% (significant at the 1% level). This results indicate a strong and robust second order

effect of the AIDS DB establishment on the visibility and subsequent use of the scientific knowledge components linked to patents with previously non-obvious link to HIV/AIDS, which provides further evidence in support of the effectiveness of the online repository in line with the policy objective.

2.5 Changes in the Quality and Reach of Knowledge Spillovers

In this section, I investigate whether the increase in informational efficiency due to the establishment of the AIDS DB had repercussions on the intensive margin of knowledge spillovers generated among researchers, which was a declared policy objective behind the database project. Precisely, I evaluate changes in likelihood of transfer of new knowledge elements from indexed patents to citing follow-up applications (Section 2.5.2), as well as in the reach of citation links across geographic distances and HIV-researcher community boundaries (Section 2.5.3). In each analysis, I further assess heterogeneity of effects between private sector and public research institute inventors, in order to investigate to which degree they benefited (differently) from access to the online repository. The distinction between corporate vs. academic inventors is in this case particularly interesting given that private sector researchers faced substantially higher access barriers to external patent documents, while inventors from public research institutes were embedded in more sophisticated and far reaching information systems and communication channels other than AIDS DB (e.g. BITNET, patent libraries, etc.), as discussed in Section 2.2.1.

2.5.1 Citation-level estimation model

The estimation of treatment effects on diffusion patterns is challenging insofar as it, primarily, requires to isolate the natural (intrinsic) diffusion of a specific piece of knowledge in time from the influence of the institution, or policy, assessed. In the following, I move my inquiry from a patent-level to a citation link-level analysis *within* citing applications (patents), thereby holding constant all factors impacting the context and dynamics at the receiving end of knowledge spillovers.

Specifically, I determine differences between AIDS DB patents and other references of same timing cited within the same citing patent, and compare changes in these differences over time across AIDS DB patents without vs. with front-page reference to HIV/AIDS, to evaluate the marginal impact of the online repository.[53] For this, I create a citation-level data set containing one observation for each cited patent-citing patent link made from any follow-up patent to each AIDS DB patent during a five-year period before and after the one-year lagged launch of the database, i.e. between 1991 and 2000.[54] For each citing patent and cited AIDS DB patent, I further include one observation for each non-AIDS DB patent, referenced as prior art in the citing patent, that was granted and first published in the same year as the paired AIDS DB patent. I again remove inventor and applicant self-citations between cited and citing patents. For each within citing patent-cited year group, I assign equal weight of .5 to both the sum of all AIDS DB and all non-AIDS DB cited patent observations, in order to give each cited year the same weight within citing patents.[55] Finally, I balance the data set by giving each citing patent a weight of one, in order not to overweight the importance of citing applications with many references to prior art.

I then investigate how the relation to prior AIDS DB vs. non-AIDS DB patents, cited within same applications, changed after the establishment of the online repos-

[53] Another possible approach to estimating this would be to compare changes in citation-level quality and reach *within* cited patents over time, similar to the approach used in Section 2.4), thus controlling for all time invariant heterogeneity across AIDS DB patents. Empirically however, given that by construction it conditions on receiving citations in both periods of analysis, which leads to very few patents, and accordingly observations ($N < 100$), in several of the split-samples studied in this section, this would make statistical inference difficult due to very large confidence intervals estimated.

[54] To ensure results are based on the same sample of patents, I only include citations links to AIDS DB patents included in the external matched control sample utilized in Section 2.4.

[55] So, e.g., if a citing patent - cited year group contains 1 AIDS DB and 2 non-AIDS DB cited patents, the assigned individual weights are .5, for the AIDS DB patent, and, respectively, .25 for each non-AIDS DB patent.

itory by estimating the following type of regressions:

$$\begin{aligned} Y_{ijy} =& \beta_1 * no\ reference_j \times post1994_{t-1} \\ &+ \beta_2 * with\ reference_j \times post1994_{t-1} \\ &+ \beta_3 * no\ reference_j + \beta_4 * with\ reference_j \\ &+ citPatent \times citedYearFE_{i \times y} + \theta_{tr} + \epsilon_{ij}, \end{aligned} \tag{2.2}$$

where i indexes citing patents, t indexes years, y indexes cited years and r indexes citing geographic regions. Y_{ijy} is the generic dependent variable quantifying the quality and reach of knowledge flow associated with a citation. The coefficients β_1 and β_2 measure the change in outcomes after one-year lagged online deposit (i.e. after 1995) between AIDS DB and non-AIDS DB control patents for patents without vs. with front-page references to HIV/AIDS, while β_3 and β_4 capture the respective pre-AIDS DB differences.

The regression model includes a fixed effect for each citing patent × cited year pair. These fixed effects control for all differences regarding the context of invention of the citing patent that might affect knowledge flows, for instance, the identity of the citing inventor, the citing institution, their scientific networks and quality. Specifically, they also account for all unobserved shocks affecting knowledge flows and information access channels of citing inventors, for example, increased resources for specific research lines, as these are held constant within a citing patent. Similarly, the fixed effects control for all permanent differences in access to knowledge across geographic regions, or permanent differences in citation patterns across technological fields. Moreover, they account for age differences across cited prior art, by holding constant all changes due to the natural diffusion of knowledge over time within a citing patent. Note, that the sum of all citing patents fixed effects would be collinear to a $post_{t-1}$ period indicator, which is therefore omitted from the specification. The regressions further include region-specific time trends, absorbing changing intrinsic components of knowledge agglomeration in specific geographic areas over time, for example, Maryland in the U.S. becoming more central to the global HIV-researcher community over the years.[56] To account for

[56] Regions are aggregated to federal states in the U.S., Mexico and Australia, NUTS-1 regions in the E.U., prefectures for Japan, provinces for Korea and Canada, and districts in Israel.

potential correlations of regression residuals regarding the presence of unobserved random shocks to knowledge production (e.g., national R&D policies or related specific developments), I cluster standard errors at the citing patent country level.[57] Summary statistics for the cited-citing level sample are reported in Table 2.A8 in Appendix 2.A.

2.5.2 Quality of knowledge flows

To evaluate changes in the intensity of knowledge spillovers associated with citations, I proxy the quality of knowledge flow by the re-occurrence in subsequently citing patents of new knowledge elements originally appearing on cited patent documents: New words in patent text, and novel scientific references.

Table 2.6 reports the results from the estimation of regression (2.2) for the re-occurrence of new words and novel SNPRs in citing patents (columns (1) and (3)), split up as triple-differences for citations from private firm patents in columns (2) and (4).[58] I identify "new words" as unique keywords on a specific patent that appear for the first time in the universe of all U.S. patents since 1976, using the patent text data provided by Arts, Cassiman and Gomez (2018). New keywords are a suitable measure to proxy the transfer of unique knowledge elements, as they occur comparatively often in patents, in about 30% of patents in my sample (compare Table 2.1). Furthermore, new scientific and technical words have been used as an alternative measure to patent citations for tracing knowledge flows in prior literature (e.g., Iaria, Schwarz and Waldinger, 2018; de Rassenfosse, Pellegrino and Raiteri, 2020; Baruffaldi and Pöge, 2020). The likelihood of re-use of new words introduced increased for both AIDS DB patents without and with front-page reference after database deposit, as shown in column (1) of Table 2.6. This appears to have been much more pronounced for the "with reference" group of patents, suggesting gains from online accessibility to likely be driving these results. Indeed, the split-sample differences estimated in column (2) reveal that effects on "with reference" patents

[57] I assign each patent to a unique country, based on the most frequent occurring inventor country-location listed on the patent. In the rare case of multiple equally frequent countries, I chose a random location.

[58] All regression models control for the number of new words introduced and scientific references made by each cited patents.

Table 2.6. Effects on Quality of Generated Spillovers

Dependent variable:	*New word*		*Novel SNPR*	
Probability of re-occurrence	(1)	(2)	(3)	(4)
No reference × post1994$_{t-1}$	0.012*	0.027**	0.031***	-0.021**
	(0.007)	(0.012)	(0.006)	(0.008)
With reference × post1994$_{t-1}$	0.068***	0.016	-0.011	-0.139***
	(0.006)	(0.019)	(0.030)	(0.028)
No reference × post1994$_{t-1}$		-0.008		0.084***
× private firm citing		(0.068)		(0.014)
With reference × post1994$_{t-1}$		0.075***		0.188***
× private firm citing		(0.025)		(0.028)
Main category interactions	Incl	Incl	Incl	Incl
Non-new/novel interactions	Incl	Incl	Incl	Incl
Citing patent × cited year FE	Yes	Yes	Yes	Yes
Citing region time trends	Yes	Yes	Yes	Yes
Observations	36,690	36,618	36,684	36,612
Number of citing clusters	8,573	8,554	8,570	8,551
R^2	.261	.263	.359	.368
Conditional mean at t_0	.028	.015	.095	.060

Notes: Each column reports parameter estimates of regression (2.2) on citing patents-cited year pairs for the subset of cited patents introducing new words (columns (1)-(2)) and referencing novel scientific publications (columns (3)-(4)). "New words" are unique keywords appearing for the first time on a patent in the universe of U.S. patents since 1976. "Novel SNPR" are non-patent references to scientific publications in PubMed which are in the top-5% of the distribution of new medical subject term combinations in a their respective year and field of publication, following Boudreau et al. (2016). The dependent variable measures the probability of re-occurrence of a new word and novel scientific reference in citing U.S. patents filed between 1991 and 2000. Inventor and applicant self-citations are excluded. Displayed are parameter estimates for patents introducing new words and referencing novel science only. Full triple interactions for patents with no new words and non-novel science are included. Controls for the number of new words and the number of scientific references are included. The reference category consists of cited non-AIDS DB patents, published in the same year, within the same application. Sample observations are weighted in order to give equal weight to each citing patent. Standard errors are clustered at citing country-level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *BvD Orbis*, *PubMed* and several disambiguations (links) between them. Patent key words are obtained from Arts, Cassiman and Gomez (2018) (see Section 3.3 for details).

were large and significant (at the 1% level) only for the sub-set of citing corporate patents, while not different from the control group for public research institutions' citations (captured by the main category interactions in column (2)). On the other hand, the latter seemed to have benefited more from enhanced access to external patents more difficult to identify as HIV-related. I re-investigate this pattern for the likelihood of re-occurrence of novel scientific references. I measure "novel SNPRs" as references to scientific publications that score among the top-5% of publications in a scientific field-year cohort regarding the share of unprecedented combinations of MeSH terms assigned in the universe of all biomedical research articles.[59] Technological usefulness and application potential of such recombining scientific advances is, arguably, harder to realize and less visible outside of the sphere of basic science and fundamental research. Findings reported in columns (3)-(4) of Table 2.6 support this prediction, as estimates point to a disproportionate increase in likelihood of re-citation of novel scientific advances in patents without front-page reference to HIV/AIDS, and large and positive effects concentrated in the sub-set of private sector citations which, in the case of "no reference" patents, are about 4 times larger than for academic citations.

2.5.3 Geographic and social distance of spillovers

I further investigate changes in international citations and knowledge flows across network boundaries of scientific communities following the launch of the AIDS DB, which were particularly emphasized as leading objectives behind the repository (compare Section 2.2.2). To measure the geographic spillover distance, for each pair of cited and citing patents, I determine the share of overlap between geographic locations of all citing inventors and all cited inventors.[60] The share of international citation links is then simply given by the inverse of the overlap, e.g. 0% if all citing and cited inventors are located in the same country, and vice versa.[61] To determine

[59] Compare, e.g., Boudreau et al. (2016) for a similar measure of scientific novelty.

[60] for each patent-inventor instance, I use all disambiguated locations for the same inventor listed on patent documents for the patent family (see Section 2.3.1 for details)

[61] Following the same reasoning, if a patent with two inventors, one located in the U.S. and the other in France, cites a prior patent with equally two inventors, one located in France and the other in Japan, the share of international citations will be: $(1 \times .5 + 1 \times .5) \times .5 + (1 \times .5 + 0 \times .5) \times .5 = .75$

the social distance between inventor communities, I consider each patenting inventor as a node in a dynamic, undirected social network of researchers, whose edges (connections) are based on observed prior collaborations between these inventors at a given point in time.[62] Subsequently, I determine, for each pair of citing and cited inventors in the data, the shortest path in the network graph of all > 5 million inventors of USPTO patents and their existing collaborative ties (as evidenced by co-appearance on prior patents) at the moment of filing of the citing patent. I consider as *minimal social distance* between a citing and cited patent, the shortest of all paths between any inventor pair involved. To account for the fact that, in any finite network, the existence of a network tie between a citing and cited patent increases stochastically in the number of inventors, I control for the count of inventors on the cited patent in all specifications. Given that AIDS DB inventors were strongly intertwined with the community of basic science authors (as shown in Section 2.3.4), I further consider their existing collaborative ties in the universe of fundamental science, based on prior scientific co-authorships on biomedical publications, and determine the minimal social distance between any pair of cited and citing patents' inventors based on the comprehensive author-inventor network graph consisting of the union of all > 5 million inventors on U.S. patents and > 16 million authors indexed in PubMed and their realm-transcending collaborative ties.[63]

Table 2.7 compares results for the estimation of regression (2.2) with the share of international citations as well as the likelihood of citation to an entirely unconnected community (social distance = ∞, no finite shortest path) as outcome variables. In the main effect specifications (columns (1) and (3)) of Table 2.A9, AIDS DB patents with front-page references to HIV/AIDS received relatively less citations from outside of geographic and social network boundaries after database inclusion (significant at the 1% level). This suggests a thickening of citation clusters within these boundaries following online accessibility. For AIDS DB patents without front-page references, on the other hand, I observe opposite patterns, suggesting a positive influence of

[62] Knowledge flows are found to be naturally clustered alongside these collaborative network ties (e.g., Singh, 2005).

[63] Inventor and author identities are disambiguated based on Li et al. (2014); Morrison, Riccaboni and Pammolli (2017); Smalheiser and Torvik (2009); Torvik and Smalheiser (2009). Inventor information covers years 1976–2011, author information years 1858–2009. For details, see Section 3.3.

Table 2.7. Effects on the Reach of Generated Spillovers

Dependent variable:	*International*		*Detached community*	
Probability of distant citation	(1)	(2)	(3)	(4)
No reference × post1994$_{t-1}$	0.020**	0.058***	0.007*	-0.012**
	(0.009)	(0.006)	(0.004)	(0.006)
With reference × post1994$_{t-1}$	-0.026***	-0.110***	-0.021**	-0.036***
	(0.009)	(0.022)	(0.008)	(0.010)
No reference × post1994$_{t-1}$ × private firm citing		-0.049***		0.027***
		(0.016)		(0.008)
With reference × post1994$_{t-1}$ × private firm citing		0.119***		0.023
		(0.043)		(0.015)
Main category interactions		Incl		Incl
Citing patent × cited year FE	Yes	Yes	Yes	Yes
Citing region time trends	Yes	Yes	Yes	Yes
Observations	36,690	36,618	36,684	36,612
Number of citing clusters	8,573	8,554	8,570	8,551
R^2	.556	.556	.731	.730
Mean at t_0	.352	.352	.201	.201

Notes: Each column reports parameter estimates of regression (2.2) on citing patents-cited year pairs. The main category parameter is included. The dependent variable in columns (1)-(2) measures shares of international citations between all pairwise links of citing and cited inventor locations for AIDS DB and control group patents in citing patents between 1991 and 2000. The dependent variable in columns (3)-(4) measures the probability that a citation originates from a research team which is entirely unconnected to the networks of direct and indirect collaborators (social distance =) of any cited inventor at the time of filing of the citing patent. Displayed are parameter estimates for the post-period only. Main category parameters and full sets of interactions are included. Inventor and applicant self-citations are excluded. The reference category consists of non-AIDS DB patents, published in the same year, cited within the same application. Sample observations are weighted in order to give equal weight to each citing patent. Standard errors are clustered at the citing country-level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, and *BvD Orbis*. Geo-coordinates and inventor/ author identities are disambiguated based on input data from Li et al. (2014); Morrison, Riccaboni and Pammolli (2017); Smalheiser and Torvik (2009); Torvik and Smalheiser (2009) (see Section 3.3 for details).

DB indexing for patents previously more difficult to detect as HIV-related to be referenced across geographic and scientific network boundaries. Effect magnitudes indicate a relative average increase of + 20% (for international citations) and + 58%

(for across-community citations) compared to pre-AIDS DB levels.[64]

When looking at heterogeneity at the receiving end, estimates in column (2) of Table 2.7 show that the impact on international spillovers from "no reference" patents was much smaller for private firm inventors (-.05 percentage points, - 85%), suggesting that gains in enhanced retrieval of HIV-relevant prior art with higher search costs from abroad were particularly driven and internalized by academic inventors. Private sector researchers exhibited a strong and positive heterogeneous increase in foreign citations to patents with front-page references to HIV/AIDS, corroborating the prediction of stronger benefits of online accessibility for this category.

Split-sample results in column (4) of Table 2.7 show, on the other hand, that positive effects on citations to detached scientific communities were strongly driven by corporate inventors; Their likelihood of citation to external patents without HIV/AIDS front-page references outside of the network of direct or indirect collaborators increased by .03 percentage points compared to the background rate of control group references, cited by the same patent, which was about three times larger than the corresponding effect for non-firm inventors (significant at the 1% level). For citations to prior AIDS DB patents with HIV/AIDS front-page reference, instead, there were no significant differential effects for private firm inventors.

For the findings on the reach of spillovers generated, I provide more results on different sub-level of geographic and social distance in Tables 2.A9 and 2.A10 in Appendix 2.A. I further show robustness of findings for social distance metrics based exclusively on inventor network graphs in Table 2.A11 in Appendix 2.A. While results are qualitatively robust, these show that shortest paths based exclusively on inventor networks drastically overstate the true distances between researchers in strongly science-intensive environments, like in this case, where patents are only a partial indicator of research output. This can be seen also from summary statistics in Table 2.A8 in Appendix 2.A: Shortest paths between cited and citing patents in the inventor graph are, on average, 6.8 degrees long for AIDS DB patents prior to AIDS DB launch (6.9 for control patents). When considering prior collaborations on basic research articles, however, this distance decreases to only 3.5 degrees on average (3.6 for control group patents). Depending on the exact conceptualization

[64] Pre-AIDS DB level estimates not reported in the table.

of social distance, this might have important implications for estimating the true extent of distance of knowledge flows.

Taken together, these results suggest a strong positive impact of the institution of the AIDS DB on facilitating the flow and diffusion of, HIV/AIDS related, technical knowledge across dispersed communities of inventors, in particular for those technologies that were more difficult to detect as HIV/AIDS-related in pre-AIDS DB external search efforts. Combined with the results in Section 2.4, these findings provide credible support that part of the cumulative impact of the AIDS DB online repository can be attributed to broader diffusion across distant teams of researchers, both in geographic and social terms.

2.6 Conclusion

Access to existing knowledge is a crucial input for technical progress and economic growth. However, due to constraints of bounded rationality, the costs to filter out relevant knowledge inputs in light of a growing abundance and general availability of information increase for inventors and other scientists alike. Many examples of public and private sector institutions and devices have emerged over the past three decades, in the form of search engines, structured databases and platforms, but we still know very little about their repercussions on scientific production, and the underlying mechanisms governing these. Yet, the problem of 'too-many-giants', on whose shoulders to stand, on has relevance and important implications far beyond prior art search, but becomes salient also with regards to questions like media literacy and public political opinion-forming (e.g., Bimber 2001; Gavazza, Nardotto and Valletti 2019).

The case of the 1994 AIDS Patent Database, as an early modern-era information-enhancing institution, enables me to study these two concurring mechanisms separately: On the one hand, the online repository provided broad accessibility at minimal cost to the full body of technical prior art related to the deadly infectious disease behind the HIV-pandemic. Patent documents, despite their abstract jargon and strategic motives of patent holders to 'conceal' the nature of the underlying invention, have been shown by prior literature to be important carriers of codified knowledge and relevant channels of knowledge transfer between distant inventors

(e.g., Furman, Nagler and Watzinger, 2018; Hegde, Herkenhoff and Zhu, 2020).

The main stand-alone contribution of my paper arises with regards the design of such institutions. The disease-specific connection, established by inclusion in the AIDS Patent Database, appears to have disproportionately benefited the visibility and subsequent diffusion of technical advances that were more difficult to identify as related to HIV/AIDS with the previous capabilities of external prior art search, based most exclusively on bibliographic information. The stronger reduction in search costs explains 30% of the variation in cumulative diffusion between these patents and those making clear front-page references to the disease. This speaks to prior findings by Thompson and Hanley (2018), who show a causal increase in follow-up citations to scientific articles appearing in topic-specific pages in Wikipedia. The catalytic effect of the topic-connection in the online repository, in my analysis is strongest for the cumulative impact of technologies embodying new ideas and novel concepts. These are particularly vulnerable to barriers affecting knowledge flows and, at the same time, need often parallel experimentation in order to prevail (Murray et al., 2016). In my analyses it shows that, not only did patents with higher up-front retrieval costs experience the relatively strongest increase in cumulative impact, but the effects on scientific and geographic community-crossing citations were also strongly concentrated in these patents, and disproportionately benefited private firm inventors. Considering this remarkable effectiveness of a comparatively low-cost policy measure, that is an online database, this has important and corroborating implications for public and private sector decision makers regarding the imperative of free access to prior art for the productivity of researchers and makes a powerful argument for the establishment of access-providing institutions.

My findings, therefore, speak in particular to the effective organization and design of patent search devices, which are historically structured based on technology classes. Complementary categorizations, such as use-indexed headings (based on, for instance, medical subjects or specific diseases), could provide useful tools for prior art searching inventors to process and condense the thousands of patents granted every year even in the most narrow technology classes. Nevertheless, for the interpretation and evaluation of transferability of these findings it should, obviously, be taken into account that HIV/AIDS research constituted a very particular and dynamic domain, spanning the frontier of many sub-disciplines both of basic science

and technological knowledge, especially at the time it is observed in my empirical setting. Similar to other studies focusing on nascent and highly-innovative domains, it should, therefore, be subject to further discussion to which extent these findings can be transferred to other contexts and different circumstances of inventive search.

References

Agrawal, Ajay, and Avi Goldfarb. 2008. "Restructuring research: Communication costs and the democratization of university innovation." *American Economic Review*, 98(4): 1578–90.

Ahmadpoor, Mohammad, and Benjamin F Jones. 2017. "The dual frontier: Patented inventions and prior scientific advance." *Science*, 357(6351): 583–587.

Akcigit, Ufuk, Douglas Hanley, and Nicolas Serrano-Velarde. 2020. "Back to Basics: Basic Research Spillovers, Innovation Policy and Growth." *The Review of Economic Studies*.

Alcacer, Juan, and Michelle Gittelman. 2006. "Patent citations as a measure of knowledge flows: The influence of examiner citations." *The Review of Economics and Statistics*, 88(4): 774–779.

Arora, Ashish, Sharon Belenzon, and Honggi Lee. 2018. "Reversed citations and the localization of knowledge spillovers." *Journal of Economic Geography*, 18(3): 495–521.

Arts, Sam, Bruno Cassiman, and Juan Carlos Gomez. 2018. "Text matching to measure patent similarity." *Strategic Management Journal*, 39(1): 62–84.

Arts, Sam, Mattia Nardotto, Thomas Schaper, and Jesse Wursten. 2020. "The Impact of Broadband Internet on Invention: Evidence from the U.K." *Working Paper*.

Baruffaldi, Stefano, and Felix Pöge. 2020. "A Firm Scientific Community: Industry Participation and Knowledge Diffusion." *IZA Discussion Paper No. 13419, Available at SSRN: https://ssrn.com/abstract=3643183*.

Baruffaldi, Stefano, and Markus Simeth. 2020. "Patents and Knowledge Diffusion: The Effect of Early Disclosure." *Research Policy*, forthcoming.

Belenzon, Sharon, and Mark Schankerman. 2013. "Spreading the word: Geography, policy, and knowledge spillovers." *Review of Economics and Statistics*, 95(3): 884–903.

Benner, Mary, and Joel Waldfogel. 2008. "Close to you? Bias and precision in patent-based measures of technological proximity." *Research Policy*, 37(9): 1556–1567.

Berkes, Enrico, and Peter Nencka. 2020. "Knowledge Access: The Effects of Carnegie Libraries on Innovation." *Working Paper*.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How much should we trust differences-in-differences estimates?" *The Quarterly journal of economics*, 119(1): 249–275.

Bertschek, Irene, Daniel Cerquera, and Gordon J Klein. 2013. "More bits–more bucks? Measuring the impact of broadband internet on firm performance." *Information Economics and Policy*, 25(3): 190–203.

Biasi, Barbara, and Petra Moser. 2020. "Effects of Copyrights on Science - Evidence from the US Book Republication Program." *American Economic Journal: Microeconomics*, forthcoming.

Bimber, Bruce. 2001. "Information and political engagement in America: The search for effects of information technology at the individual level." *Political Research Quarterly*, 54(1): 53–67.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast unfolding of communities in large networks." *Journal of statistical mechanics: theory and experiment*, 2008(10): P10008.

Bloom, Nicholas, Mark Schankerman, and John Van Reenen. 2013. "Identifying technology spillovers and product market rivalry." *Econometrica*, 81(4): 1347–1393.

Boudreau, Kevin J, Eva C Guinan, Karim R Lakhani, and Christoph Riedl. 2016. "Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science." *Management Science*, 62(10): 2765–2783.

Bryan, Kevin A, and Yasin Ozcan. 2020. "The impact of open access mandates on invention." *The Review of Economics and Statistics*, forthcoming.

Cassiman, Bruno, and Reinhilde Veugelers. 2002. "R&D Cooperation and Spillovers: Some Empirical Evidence from Belgium." *American Economic Review*, 92(4): 1169–1184.

Cohen, Wesley M, and Daniel A Levinthal. 1990. "Absorptive capacity: A new perspective on learning and innovation." *Administrative science quarterly*, 128–152.

Cohen, Wesley M, Richard R Nelson, and John P Walsh. 2000. "Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not)." National Bureau of Economic Research Working Paper 7552.

Czernich, Nina, Oliver Falck, Tobias Kretschmer, and Ludger Woessmann. 2011. "Broadband infrastructure and economic growth." *The Economic Journal*, 121(552): 505–532.

De Chaisemartin, Clément, and Xavier d'Haultfoeuille. 2020. "Two-way fixed effects estimators with heterogeneous treatment effects." *American Economic Review*, 110(9): 2964–96.

de Rassenfosse, G., G. Pellegrino, and E. Raiteri. 2020. "Do patents enable disclosure? Evidence from the 1951 Invention Secrecy Act." Ecole polytechnique fédérale de Lausanne.

Devlin, Alan. 2009. "The misunderstood function of disclosure in patent law." *Harvard Journal of Law & Technology*, 23: 401.

Ding, Waverly W, Sharon G Levin, Paula E Stephan, and Anne E Winkler. 2010. "The impact of information technology on academic scientists' productivity and collaboration patterns." *Management Science*, 56(9): 1439–1461.

Dittmar, Jeremiah E. 2011. "Information technology and economic change: the impact of the printing press." *The Quarterly Journal of Economics*, 126(3): 1133–1172.

Fleming, Lee. 2001. "Recombinant uncertainty in technological search." *Management science*, 47(1): 117–132.

Forman, Chris, and Nicolas van Zeebroeck. 2012. "From wires to partners: How the Internet has fostered R&D collaborations within firms." *Management science*, 58(8): 1549–1568.

Forman, Chris, and Nicolas van Zeebroeck. 2019. "Digital technology adoption and knowledge flows within firms: Can the Internet overcome geographic and technological distance?" *Research policy*, 48(8): 103697.

Fromer, Jeanne C. 2008. "Patent disclosure." *Iowa L. Rev.*, 94: 539.

Furman, Jeffrey L., and Scott Stern. 2011. "Climbing atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research." *American Economic Review*, 101(5): 1933–63.

Furman, Jeffrey L, Markus Nagler, and Martin Watzinger. 2018. "Disclosure and subsequent innovation: Evidence from the patent depository library program." National Bureau of Economic Research.

Galasso, Alberto, and Mark Schankerman. 2015. "Patents and cumulative innovation: Causal evidence from the courts." *The Quarterly Journal of Economics*, 130(1): 317–369.

Gambardella, A, D Harhoff, and Nagaoka S. 2011. "The Social Value of Patent Disclosure." LMU Munich.

Gavazza, Alessandro, Mattia Nardotto, and Tommaso Valletti. 2019. "Internet and politics: Evidence from UK local elections and local government policies." *The Review of Economic Studies*, 86(5): 2092–2135.

Giuri, Paola, Myriam Mariani, Stefano Brusoni, Gustavo Crespi, Dominique Francoz, Alfonso Gambardella, Walter Garcia-Fontes, Aldo Geuna, Raul Gonzales, Dietmar Harhoff, and Karin Hoisl. 2007. "Inventors and invention processes in Europe: Results from the PatVal-EU survey." *Research Policy*, 36(8): 1107–1127.

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing." *Journal of Econometrics*.

Graham, S, and D Hegde. 2015. "Disclosing patents' secrets." *Science*, 347(6219): 236–237.

Graham, Stuart JH, Alan C Marco, and Richard Miller. 2015. "The USPTO patent examination research dataset: A window on the process of patent examination." *Georgia Tech Scheller College of Business Research Paper No. WP*, 43.

Griliches, Zvi. 1991. "The search for R&D spillovers." National Bureau of Economic Research.

Hegde, Deepak, and Hong Luo. 2018. "Patent publication and the market for ideas." *Management Science*, 64(2): 652–672.

Hegde, Deepak, Kyle Herkenhoff, and Chenqi Zhu. 2020. "Patent disclosure and innovation." *Available at SSRN 3158031*.

Iacus, Stefano M, Gary King, and Giuseppe Porro. 2012. "Causal inference without balance checking: Coarsened exact matching." *Political analysis*, 1–24.

Iaria, Alessandro, Carlo Schwarz, and Fabian Waldinger. 2018. "Frontier Knowledge and Scientific Production: Evidence from the Collapse of International Science*." *The Quarterly Journal of Economics*, 133(2): 927–991.

Jaffe, Adam B, Manuel Trajtenberg, and Michael S Fogarty. 2000. "Knowledge spillovers and patent citations: Evidence from a survey of inventors." *American Economic Review*, 90(2): 215–218.

Jaffe, Adam B, Manuel Trajtenberg, and Rebecca Henderson. 1993. "Geographic localization of knowledge spillovers as evidenced by patent citations." *The Quarterly Journal of Economics*, 108(3): 577–598.

Jones, Benjamin F, and Lawrence H Summers. 2020. "A Calculation of the Social Returns to Innovation." National Bureau of Economic Research Working Paper 27863.

Jones, Eric. 2003. *The European miracle: environments, economies and geopolitics in the history of Europe and Asia.* Cambridge University Press.

Kong, Nancy, Uwe Dulleck, Adam B Jaffe, Sowmya Vajjala, et al. 2020. "Linguistic Metrics for Patent Disclosure: Evidence from University Versus Corporate Patents." *NBER Working Paper*, , (w27803).

Lemley, Mark A. 2012. "The myth of the sole inventor." *Mich. L. Rev.*, 110(5): 709.

Li, Guan-Cheng, Ronald Lai, Alexander D'Amour, David M Doolin, Ye Sun, Vetle I Torvik, Z Yu Amy, and Lee Fleming. 2014. "Disambiguation and co-authorship networks of the US patent inventor database (1975–2010)." *Research Policy*, 43(6): 941–955.

Lück, Sonja, Benjamin Balsmeier, Florian Seliger, and Lee Fleming. 2019. "Early disclosure law reduces duplication in the US and European patent systems." Northwestern Pritzker School of Law.

Mansfield, Edwin. 1995. "Academic research underlying industrial innovations: sources, characteristics, and financing." *The review of Economics and Statistics*, 55–65.

March, James G. 1991. "Exploration and exploitation in organizational learning." *Organization science*, 2(1): 71–87.

Marco, Alan C, Michael Carley, Steven Jackson, and Amanda Myers. 2015. "The uspto historical patent data files: Two centuries of innovation." *Available at SSRN 2616724*.

Marx, Matt, and Aaron Fuegi. 2020. "Reliance on science: Worldwide front-page patent citations to scientific articles." *Strategic Management Journal*, 41(9): 1572–1594.

Mokyr, Joel. 2005. "Long-term economic growth and the history of technology." In *Handbook of economic growth*. Vol. 1, 1113–1180. Elsevier.

Morrison, Greg, Massimo Riccaboni, and Fabio Pammolli. 2017. "Disambiguation of patent inventors and assignees using high-resolution geolocation data." *Scientific data*, 4: 170064.

Moser, Petra, and Alessandra Voena. 2012. "Compulsory Licensing: Evidence from the Trading with the Enemy Act." *American Economic Review*, 102(1): 396–427.

Murray, Fiona, Philippe Aghion, Mathias Dewatripont, Julian Kolev, and Scott Stern. 2016. "Of Mice and Academics: Examining the Effect of Openness on Innovation." *American Economic Journal: Economic Policy*, 8(1): 212–52.

Narin, Francis, Kimberly S Hamilton, and Dominic Olivastro. 1997. "The increasing linkage between US technology and public science." *Research policy*, 26(3): 317–330.

Ouelette, Lisa L. 2012. "Do patents disclose useful information?" *Harvard Journal of Law & Technology*, 25(2): 545–608.

Righi, Cesare, and Timothy Simcoe. 2019. "Patent examiner specialization." *Research Policy*, 48(1): 137–148.

Risch, Michael. 2007. "The Failure of Public Notice in Patent Prosecution." *Harv. JL & Tech.*, 21: 179.

Rosenberg, Nathan. 1976. *Perspectives on technology.* CUP Archive.

Schmoch, Ulrich. 2008. "Concept of a technology classification for country comparisons." *Final report to the world intellectual property organisation (wipo), WIPO*.

Scotchmer, Suzanne. 1991. "Standing on the shoulders of giants: cumulative research and the patent law." *Journal of economic perspectives*, 5(1): 29–41.

Silva, JMC Santos, and Silvana Tenreyro. 2006. "The log of gravity." *The Review of Economics and statistics*, 88(4): 641–658.

Silva, JMC Santos, and Silvana Tenreyro. 2011. "Further simulation evidence on the performance of the Poisson pseudo-maximum likelihood estimator." *Economics Letters*, 112(2): 220–222.

Singh, Jasjit. 2005. "Collaborative networks as determinants of knowledge diffusion patterns." *Management science*, 51(5): 756–770.

Singh, Jasjit, and Matt Marx. 2013. "Geographic constraints on knowledge spillovers: Political borders vs. spatial proximity." *Management Science*, 59(9): 2056–2078.

Smalheiser, Neil R, and Vetle I Torvik. 2009. "Author name disambiguation." *Annual review of information science and technology*, 43(1): 1.

Thompson, Neil C., and Douglas Hanley. 2018. "Science is Shaped by Wikipedia: Evidence From a Randomized Control Trial." *Available at SSRN 3039505*.

Thompson, Peter. 2006. "Patent citations and the geography of knowledge spillovers: evidence from inventor-and examiner-added citations." *The Review of Economics and Statistics*, 88(2): 383–388.

Thompson, Peter, and Melanie Fox-Kean. 2005. "Patent citations and the geography of knowledge spillovers: A reassessment." *American Economic Review*, 95(1): 450–460.

Torvik, Vetle I. 2018. "Author-Linked data for Author-ity 2009."

Torvik, Vetle I, and Neil R Smalheiser. 2009. "Author name disambiguation in MEDLINE." *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(3): 1–29.

Trajtenberg, Manuel, Rebecca Henderson, and Adam Jaffe. 1997. "University versus corporate patents: A window on the basicness of invention." *Economics of Innovation and new technology*, 5(1): 19–50.

Verhoeven, Dennis, Jurriën Bakker, and Reinhilde Veugelers. 2016. "Measuring technological novelty with patent-based indicators." *Research Policy*, 45(3): 707–723.

Weitzman, Martin L. 1979. "Optimal search for the best alternative." *Econometrica: Journal of the Econometric Society*, 641–654.

Zheng, Yanfeng, and Qinyu Wang. 2020. "Shadow of The Great Firewall: The Impact of Google Blockade on Innovation in China." *Strategic Management Journal, Forthcoming*.

Appendices

2.A Appendix tables

Table 2.A1. Responsiveness to AIDS DB Deposit of Citation Rates from Patents within Individual Examiners

Dependent variable:	OLS			
Number of patent citations	(1)	(2)	(3)	(4)
No reference × post$_{t-1}$	0.002	-0.002	-0.015	-0.049
	(0.003)	(0.010)	(0.021)	(0.053)
Patent/ year/ field fixed effects	Yes	Yes	Yes	Yes
Citing examiner fixed effects	Yes	Yes	Yes	Yes
Pre-deposit DB cites		Yes	Yes	Yes
Pre-deposit same IPC cites			Yes	
Pre-deposit same art unit cites				Yes
Observations	10,501,376	2,377,280	928,913	296,069
Number of citing examiners	1,016	230	230	230
R^2	.005	.009	.018	.041
Mean at t_0	.0001	.0007	.0017	.0053
SD at t_0	.014	.029	.046	.081

Notes: Each column reports parameter estimates of regression (2.1) for the matched panel of U.S. AIDS DB patents without vs. with front-page reference to HIV/AIDS. The dependent variable measures the yearly number of family citations by individual examiners at the USPTO for years $t-4$ to $t+5$ relative to the one-year lagged online date. I have no information whether citations are given by the examiners or applicants. Inventor and applicant self-citations are removed from the counts. The reference category consists of AIDS DB patents "with reference" to HIV/AIDS, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired "no reference" patents. The number of citations is standardized to mean zero and standard deviation one within technology fields (fields based on Schmoch (2008)). "Pre-deposit DB cites" indicates a sample restriction to exclusively citations of examiners with prior citations to patents in the AIDS DB before DB deposit. "Same IPC" indicates a sample restriction to citations of examiners pre-DB citing AIDS DB patents in the same technological field as a cited patent. "Same art unit" indicates a sample restriction to citations of examiners pre-DB citing AIDS DB patents in the art unit. Sample observations are weighted based on Iacus, King and Porro (2012). Standard errors are clustered at the patent level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *BvD Orbis* and several disambiguations (links) between them (see Section 3.3 for details).

Table 2.A2. Top-20 HIV/AIDS Inventors vs. Authors, by the End of 1996

Top-20 AIDS DB inventors					
Name	Location	↓Patents	Citations	Degree	Centrality
Montagnier, Luc	Paris, France	36	557	52	.091
Mueller, Richard A.	Chicago, IL, U.S.	28	378	22	.005
Schinazi, Raymond F.	Atlanta, GA, U.S.	20	379	19	.008
Gallo, Robert C.	Bethesda, MD, U.S.	20	626	42	.254
Sessler, Jonathan L.	San Francisco, CA, U.S.	18	853	18	.001
Paessens, Arnold	Duesseldorf, Germany	17	129	33	.004
Hargrave, Karl D.	Ridgefield, CT, U.S.	16	93	12	.000
Chu, Chung K.	Atlanta, GA, U.S.	16	234	11	.004
Norbeck, Daniel W.	Chicago, IL, U.S.	15	498	19	.000
Carter, William A.	Philadelphia, PA, U.S.	15	157	0	-
Fleet, George W.	Oxford, U.K.	14	81	10	.005
Clavel, Francois	Paris, France	13	174	13	.000
Geutard, Denise	Paris, France	13	174	13	.000
Hoshino, Hiroo	Takasaki, Japan	13	68	28	.004
Krenitsky, Thomas A.	Chapel Hill, NC, U.S.	13	136	9	.000
Hemmi, Gregory W.	San Francisco, CA, U.S.	12	653	7	.000
Broder, Samuel	Bethesda, MD, U.S.	12	214	18	.248
Alizon, Marc	Paris, France	12	202	14	.001
Rideout, Janet L.	Chapel Hill, NC, U.S.	12	175	13	.001
Haseltine, William A.	Boston, MA, U.S.	12	203	19	.002
Top-20 AIDS DB PubMed authors					
Name	Location	↓HIV articles	Citations	Degree	Centrality
Levy, Jay A.	San Francisco, CA, U.S.	154	9,417	271	.087
Ho, David D. A.	New York, NY, U.S.	107	16,748	289	.125
Montagnier, Luc	Paris, France	106	3,752	307	.121
Sodroski, Joseph G.	Boston, MA, U.S.	105	9,146	216	.047
Mitsuya, Hiroaki	Bethesda, MD, U.S.	103	5,089	220	.051
Baba, Masanori	Fukushima, Japan	86	3,748	165	.023
Broder, Samuel	Bethesda, MD, U.S.	84	5,003	183	.040
Matthews, Thomas J.	Durham, NC, U.S.	82	4,796	266	.072
Haseltine, William A.	Boston, MA, U.S.	81	4,862	128	.027
Nakashima, Hideki	Tokyo, Japan	77	1,700	211	.022
Nara, Peter L.	Frederick, MD, U.S.	75	3,590	244	.069
Bolognesi, Dani P.	Durham, NC, U.S.	73	3,827	164	.022
Chermann, Jean-Claude	Paris, France	71	1,335	177	.061
Guertler, Lutz G.	Munich, Germany	70	813	175	.039
Weiss, Robert A.	London, U.K.	67	5,452	176	.047
Yarchoan, Robert	Bethesda, MD, U.S.	67	3,605	182	.035
Berzofsky, Jay A.	Bethesda, MD, U.S.	64	4,089	148	.015
Busch, Michael P.	San Francisco, CA, U.S.	63	2,708	188	.047
Lane, H. Clifford	Bethesda, MD, U.S.	62	3,018	194	.026
De Rossi, Anita	Padua, Italy	62	1,494	170	.057

Notes: Inventor betweenness centrality values multiplied by ×100, author centrality by by ×10

Table 2.A3. Effect for Patents without HIV/AIDS Front Page Reference, within AIDS DB, PPML Estimates

Dependent variable:	Poisson					
Number of patent citations	(1)	(2)	(3)	(4)	(5)	(6)
No reference × post$_{t-1}$	0.330***	0.301***	0.297***	0.304***	0.521***	-0.925**
	(0.109)	(0.112)	(0.110)	(0.111)	(0.142)	(0.413)
Abstr reference × post$_{t-1}$					0.352*	
					(0.189)	
Patent fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Time/ year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Field time trends		Yes	Yes	Yes	Yes	Yes
DB inventor cites only			Yes			
Excl. firm self-cites				Yes		
In prosecution cites only						Yes
Observations	10,727	10,717	10,624	10,684	10,717	2,761
Number of patents	1,196	1,195	1,185	1,191	1,195	309
Mean at t_0	1.210	1.210	1.178	1.194	1.210	.016

Notes: Each column reports parameter estimates of regression (2.1) for the matched panel of U.S. AIDS DB patents without vs. with front-page reference to HIV/AIDS from a Poisson pseudo-maximum likelihood estimation (correct for over-dispersion by construction) to address the count nature of citations. The dependent variable measures the yearly number of family citations for years $t-4$ to $t+5$ relative to the one-year lagged online date. Inventor and applicant self-citations are removed from the counts. The reference category consists of AIDS DB patents "with reference" to HIV/AIDS, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired "no reference" patents. "DB inventor cites" are citations originating exclusively from HIV/AIDS inventors indexed in the AIDS DB. "Firm self cites" are self-citations at the ultimate owner-level. "In prosecution cites" are citations exclusively from patents already under examination at the time of DB inclusion of the cited patent. Sample observations are weighted based on Iacus, King and Porro (2012). Standard errors are clustered at the patent level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *BvD Orbis* and several disambiguations (links) between them (see Section 3.3 for details).

TABLE 2.A4. EFFECT FOR PATENTS WITHOUT HIV/AIDS FRONT PAGE REFERENCE, WITHIN AIDS DB, PATENTS OBSERVED IN ALL SAMPLE YEARS

Dependent variable:	OLS					
Number of patent citations	(1)	(2)	(3)	(4)	(5)	(6)
No reference × post$_{t-1}$	0.154**	0.140**	0.158***	0.169***	0.212***	-0.142**
	(0.062)	(0.062)	(0.060)	(0.063)	(0.066)	(0.056)
Abstr reference × post$_{t-1}$					0.122	
					(0.091)	
Patent fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Time/ year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Field time trends		Yes	Yes	Yes	Yes	Yes
DB inventor cites only			Yes			
Excl. firm self-cites				Yes		
In prosecution cites only						Yes
Observations	6,535	6,535	6,514	6,525	6,535	6,240
Number of patents	655	655	653	654	655	624
R^2	.428	.431	.477	.469	.431	.052
Mean at t_0	1.707	1.707	1.640	1.680	1.707	.038
SD at t_0	3.271	3.271	3.168	3.251	3.271	.232

Notes: Each column reports parameter estimates of regression (2.1) for the matched panel of U.S. AIDS DB patents without vs. with front-page reference to HIV/AIDS only for patents observed in each year of the sample period. The dependent variable measures the yearly number of family citations for years $t-4$ to $t+5$ relative to the one-year lagged online date. Inventor and applicant self-citations are removed from the counts. The reference category consists of AIDS DB patents "with reference" to HIV/AIDS, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired "no reference" patents. The number of citations is standardized to mean zero and standard deviation one within technology fields (fields based on Schmoch (2008)). "DB inventor cites" are citations originating exclusively from HIV/AIDS inventors indexed in the AIDS DB. "Firm self cites" are self-citations at the ultimate owner-level. "In prosecution cites" are citations exclusively from patents already under examination at the time of DB inclusion of the cited patent. Sample observations are weighted based on Iacus, King and Porro (2012). Standard errors are clustered at the patent level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *BvD Orbis* and several disambiguations (links) between them (see Section 3.3 for details).

Table 2.A5. Effect for Patents without HIV/AIDS Front Page Reference, within AIDS DB, Matched on Pre-Trends

Dependent variable:	OLS					
Number of patent citations	(1)	(2)	(3)	(4)	(5)	(6)
No reference × post$_{t-1}$	0.155***	0.147**	0.124**	0.147***	0.239***	0.030
	(0.059)	(0.058)	(0.057)	(0.056)	(0.053)	(0.047)
Abstr reference × post$_{t-1}$					0.166**	
					(0.081)	
Patent fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Time/ year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Field time trends		Yes	Yes	Yes	Yes	Yes
DB inventor cites only			Yes			
Excl. firm self-cites				Yes		
In prosecution cites only						Yes
Observations	6,774	6,774	6,818	6,829	6,774	6,661
Number of patents	775	775	781	781	775	768
R^2	.225	.233	.261	.278	.232	.022
Mean at t_0	.492	.492	.493	.499	.492	.006
SD at t_0	1.035	1.035	1.040	1.038	1.035	.080

Notes: Each column reports parameter estimates of regression (2.1) for the matched panel of U.S. AIDS DB patents without vs. with front-page reference to HIV/AIDS, additionally matched on pre-period yearly citation levels. The dependent variable measures the yearly number of family citations for years $t-4$ to $t+5$ relative to the one-year lagged online date. Inventor and applicant self-citations are removed from the counts. The reference category consists of AIDS DB patents "with reference" to HIV/AIDS, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired "no reference" patents. The number of citations is standardized to mean zero and standard deviation one within technology fields (fields based on Schmoch (2008)). "DB inventor cites" are citations originating exclusively from HIV/AIDS inventors indexed in the AIDS DB. "Firm self cites" are self-citations at the ultimate owner-level. "In prosecution cites" are citations exclusively from patents already under examination at the time of DB inclusion of the cited patent. Sample observations are weighted based on Iacus, King and Porro (2012). Standard errors are clustered at the patent level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *BvD Orbis* and several disambiguations (links) between them (see Section 3.3 for details).

TABLE 2.A6. EFFECT FOR PATENTS WITHOUT HIV/AIDS FRONT PAGE REFERENCE, WITHIN AIDS DB, WEIGHTED CITATIONS

Dependent variable:	OLS					
Log(impact weighted citations)	(1)	(2)	(3)	(4)	(5)	(6)
No reference × post$_{t-1}$	0.159**	0.138*	0.137*	0.142*	0.226**	−0.014
	(0.078)	(0.078)	(0.076)	(0.078)	(0.099)	(0.018)
Abstr reference × post$_{t-1}$					0.155	
					(0.132)	
Patent fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Time/ year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Field time trends		Yes	Yes	Yes	Yes	Yes
DB inventor cites only			Yes			
Excl. firm self-cites				Yes		
In prosecution cites only						Yes
Observations	12,192	12,192	12,183	12,190	12,192	12,070
Number of patents	1,367	1,366	1,366	1,366	1,366	1,366
R^2	.435	.443	.407	.404	.442	.082
Mean at t_0	20.207	20.207	19.954	19.974	20.207	.159

Notes: Each column reports parameter estimates of regression (2.1) for the matched panel of U.S. AIDS DB patents without vs. with front-page reference to HIV/AIDS. The dependent variable measures the yearly number of 10-year impact weighted family citations for years $t-4$ to $t+5$ relative to the one-year lagged online date. Weighted citation counts are transformed to their natural logarithm, adding a small quantity (+1) to each count. Inventor and applicant self-citations are removed from the counts. The reference category consists of AIDS DB patents "with reference" to HIV/AIDS, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired "no reference" patents. The number of citations is standardized to mean zero and standard deviation one within technology fields (fields based on Schmoch (2008)). "DB inventor cites" are citations originating exclusively from HIV/AIDS inventors indexed in the AIDS DB. "Firm self cites" are self-citations at the ultimate owner-level. "In prosecution cites" are citations exclusively from patents already under examination at the time of DB inclusion of the cited patent. Sample observations are weighted based on Iacus, King and Porro (2012). Standard errors are clustered at the patent level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *BvD Orbis* and several disambiguations (links) between them (see Section 3.3 for details).

TABLE 2.A7. DIFFERENTIAL EFFECTS FOR PATENTS BASED ON RECENCY OF GRANT, WITHIN AIDS DB

Dependent variable:	*Top 25%*	*Top 50%*	*Oldest 25%*	*Oldest 10%*
Number of patent citations	(1)	(2)	(3)	(4)
No reference × $post_{t-1}$	0.194***	0.167**	0.136***	0.151***
	(0.056)	(0.071)	(0.047)	(0.044)
$Post_{t-1}$ × *recent*	0.300***	0.284***	-0.236***	-0.167*
	(0.095)	(0.081)	(0.078)	(0.091)
No reference × $post_{t-1}$ × *recent*	-0.093	-0.038	0.027	-0.101
	(0.083)	(0.089)	(0.104)	(0.153)
Patent fixed effects	Yes	Yes	Yes	Yes
Time/ year fixed effects	Yes	Yes	Yes	Yes
Field time trends	Yes	Yes	Yes	Yes
Observations	12,192	12,183	12,192	12,183
Number of patents	1,366	1,366	1,366	1,366
R^2	.507	.508	.507	.507
Mean at t_0	1.210	1.210	1.210	1.210
SD at t_0	2.570	2.570	2.570	2.570

Notes: Each column reports parameter estimates of regression (2.1) split up as triple-differences for heterogeneity of effects on patents based on grant recency in the matched panel of U.S. AIDS DB patents without vs. with front-page reference to HIV/AIDS. "Top 25%" most recent patents are granted less than one year prior to online deposit. "Oldest 10%" patents are granted more than 4 years prior to online deposit. The $post_{t-1}$ parameter captures relative changes in citations to patents with front-page reference to HIV/AIDS in each split-sample category. The dependent variable measures the yearly number of family citations for years $t-4$ to $t+5$ relative to the one-year lagged online date. The number of citations is standardized to mean zero and standard deviation one within technology fields (fields based on Schmoch (2008)). Inventor and applicant self-citations are removed from the counts. The reference category consists of AIDS DB patents "with reference" to HIV/AIDS, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired "no reference" patents. Sample observations are weighted based on Iacus, King and Porro (2012). Standard errors are clustered at the patent level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *BvD Orbis* and several disambiguations (links) between them (see Section 3.3 for details).

Table 2.A8. Summary Statistics, Quality and Reach of Spillovers Pre-AIDS DB

Within citing patents comparison	Control	No reference		With reference		Diff
	Mean	Mean	SD	Mean	SD	p-val.
Re-occurrence of new words	.01	.00	.05	.01	.12	.00
Re-occurrence of novel SNPR	.02	.01	.10	.02	.15	.00
Share of international citations	.40	.29	.44	.38	.47	.00
Share of interregional citations	.90	.89	.28	.90	.28	.04
Share of intermetropolitan citations	.90	.90	.29	.90	.28	.48
Shortest path, author-inventor graph	3.58	3.47	1.48	3.49	1.49	.45
Shortest path, inventor graph	6.94	6.82	3.13	6.80	2.85	.71
Share shortest path > 1 (auth-invt)	.95	.94	.24	.95	.22	.01
Share shortest path > 2 (auth-invt)	.84	.84	.36	.83	.38	.13
Share shortest path > 3 (auth-invt)	.63	.60	.49	.60	.49	.78
Share shortest path = ∞ (auth-invt)	.24	.25	.43	.22	.42	.00
Number of citing patents	6,716	6,064		6,716		

Notes: The table reports group mean summary statistics for variable relating to the quality and reach of spillovers between between all pairwise links of citing and cited inventor locations for AIDS DB and control group patents in citing patents between 1991 and 1995. Inventor and applicant self-citations are excluded. Control group patents in column (2) are non-AIDS DB cited patents of similar timing within the same citing application. Columns (3) - (6) report means for AIDS DB patents without (3-4) and with (5-6) front-page reference to HIV/AIDS. Column (7) reports p-values from two-sample t-tests with unequal variances for differences in sample means. "New words" are unique keywords appearing for the first time on a patent in the universe of U.S. patents since 1976. "Novel SNPR" are non-patent references to scientific publications in PubMed which are in the top-5% of the distribution of new medical subject term combinations in a their respective year and field of publication, following Boudreau et al. (2016). Shares of international, -regional, and -metropolitan are the share of links in all pairwise links between inventor locations on a citing and cited patent that cross the respective geographic area for citing inventors. The bottom panel of the table reports group means for minimal social distance between all shortest paths of citing and cited inventors based on the network graph of all > 5mio inventors on U.S. patents and all > 16mio authors indexed in PubMed and their prior collaboration ties at the filing time of the citing patent. Citations to patents with all first-time inventors/authors are excluded. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT* and *PubMed*. Patent key words are obtained from Arts, Cassiman and Gomez (2018). Scientific non-patent references come from ?. Geo-coordinates and inventor/ author identities are disambiguated based on input data from Li et al. (2014); Morrison, Riccaboni and Pammolli (2017); Smalheiser and Torvik (2009); Torvik and Smalheiser (2009) (see Section 3.3 for details).

TABLE 2.A9. EFFECTS ON THE GEOGRAPHIC REACH OF GENERATED SPILLOVERS

Dependent variable:	*International*		*Interregional*		*Intermetropolitan*	
Probability of distant citation	(1)	(2)	(3)	(4)	(5)	(6)
No reference × post1994$_{t-1}$	0.020**	0.058***	−0.011***	0.021***	−0.020***	0.004
	(0.009)	(0.006)	(0.002)	(0.005)	(0.003)	(0.004)
With reference × post1994$_{t-1}$	−0.026***	−0.110***	0.040***	−0.015**	0.028***	−0.029***
	(0.009)	(0.022)	(0.012)	(0.006)	(0.006)	(0.009)
No reference × post1994$_{t-1}$ × private firm citing		−0.049***		−0.047***		−0.036***
		(0.016)		(0.006)		(0.009)
With reference × post1994$_{t-1}$ × private firm citing		0.119***		0.083***		0.083***
		(0.043)		(0.020)		(0.015)
Main category interactions		Incl		Incl		Incl
Citing patent × cited year FE	Yes	Yes	Yes	Yes	Yes	Yes
Citing region time trends	Yes	Yes	Yes	Yes	Yes	Yes
Observations	36,310	36,238	36,158	36,086	27,894	27,837
Number of citing clusters	8,524	8,505	8,509	8,490	6,986	6,972
R^2	.556	.556	.369	.369	.371	.372
Mean at t_0	.352	.352	.897	.897	.908	.908

Notes: Each column reports parameter estimates of regression (2.2) on citing patents-cited year pairs. The main category parameter is included. The dependent variables measure shares of international, interregional, and intermetropolitan citations between all pairwise links of citing and cited inventor locations for AIDS DB and control group patents in citing patents between 1991 and 2000. Displayed are parameter estimates for the post-period only. Main category parameters and full sets of interactions are included. Inventor and applicant self-citations are excluded. The reference category consists of non-AIDS DB patents, published in the same year, cited within the same application. Sample observations are weighted in order to give equal weight to each citing patent. Standard errors are clustered at the citing country-level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, and *BvD Orbis*. Geo-coordinates of inventor locations are based on input data from Li et al. (2014); Morrison, Riccaboni and Pammolli (2017) (see Section 3.3 for details).

TABLE 2.A10. EFFECTS ON THE SOCIAL DISTANCE (SD) REACH OF GENERATED SPILLOVERS, AUTHOR-INVENTOR GRAPH

Dependent variable:	*SD* > 1	*SD* > 2	*SD* > 3	*SD* ∞	*Log(SD)*
Shortest path of citation	(1)	(2)	(3)	(4)	(5)
No reference × post1994$_{t-1}$	0.004	0.030***	0.030***	0.007*	0.030***
	(0.003)	(0.004)	(0.005)	(0.004)	(0.005)
With reference × post1994$_{t-1}$	0.077***	0.060*	0.031*	-0.021**	0.079**
	(0.025)	(0.031)	(0.018)	(0.008)	(0.029)
Citing patent × cited year FE	Yes	Yes	Yes	Yes	Yes
Citing region time trends	Yes	Yes	Yes	Yes	Yes
Observations	36,933	36,933	36,933	36,933	31,055
Number of citing clusters	8,630	8,630	8,630	8,630	7,548
R^2	.441	.450	.595	.731	.584
Mean at t_0	.954	.870	.631	.201	1.250

Notes: Each column reports parameter estimates of regression (2.2) on citing patents-cited year pairs. The main category parameter is included. The dependent variables measures the probability that a citation originates from a research team at at different degrees of minimal social distance (shortest path) in the networks of direct and indirect prior collaborators of any cited inventor at the time of filing of the citing patent, based on the universe of all > 5mio inventors on U.S. patents and all > 16mio authors indexed in PubMed. The sample consists for AIDS DB and control group patents in citing patents between 1991 and 2000. Displayed are parameter estimates for the post-period only. Main category parameters and full sets of interactions are included. Inventor and applicant self-citations are excluded. The reference category consists of non-AIDS DB patents, published in the same year, cited within the same application. Sample observations are weighted in order to give equal weight to each citing patent. Standard errors are clustered at the citing country-level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, and *BvD Orbis*. Inventor and author identities are disambiguated based on input data from Li et al. (2014); Morrison, Riccaboni and Pammolli (2017); Smalheiser and Torvik (2009); Torvik and Smalheiser (2009) (see Section 3.3 for details).

Table 2.A11. Effects on the Social Distance (SD) Reach of Generated Spillovers, Inventor Graph

Dependent variable:	*SD* > 1	*SD* > 2	*SD* > 3	*SD* ∞	*Log(SD)*
Shortest path of citation	(1)	(2)	(3)	(4)	(5)
No reference × post1994$_{t-1}$	0.010***	0.026***	0.024***	0.020***	0.041***
	(0.002)	(0.004)	(0.008)	(0.006)	(0.013)
With reference × post1994$_{t-1}$	0.042*	0.053*	−0.000	−0.023**	0.111*
	(0.020)	(0.027)	(0.018)	(0.011)	(0.057)
Citing patent × cited year FE	Yes	Yes	Yes	Yes	Yes
Citing region time trends	Yes	Yes	Yes	Yes	Yes
Observations	36,925	36,925	36,925	36,925	24,621
Number of citing clusters	8,628	8,628	8,628	8,628	6,111
R^2	.410	.397	.389	.714	.507
Mean at t_0	.971	.951	.928	.439	1.794

Notes: Each column reports parameter estimates of regression (2.2) on citing patents–cited year pairs. The main category parameter is included. The dependent variables measures the probability that a citation originates from a research team at at different degrees of minimal social distance (shortest path) in the networks of direct and indirect prior collaborators of any cited inventor at the time of filing of the citing patent, based on the universe of all > 5mio inventors on U.S. patents. The sample consists for AIDS DB and control group patents in citing patents between 1991 and 2000. Displayed are parameter estimates for the post-period only. Main category parameters and full sets of interactions are included. Inventor and applicant self-citations are excluded. The reference category consists of non-AIDS DB patents, published in the same year, cited within the same application. Sample observations are weighted in order to give equal weight to each citing patent. Standard errors are clustered at the citing country-level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, and *BvD Orbis*. Inventor identities are disambiguated based on input data from Li et al. (2014); Morrison, Riccaboni and Pammolli (2017) (see Section 3.3 for details).

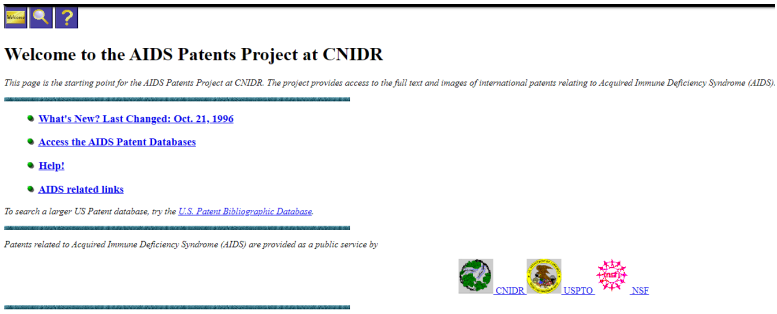
2.B Appendix figures

FIGURE 2.B1. AIDSDB LINK ON USPTO WEBSITE, FALL 1996



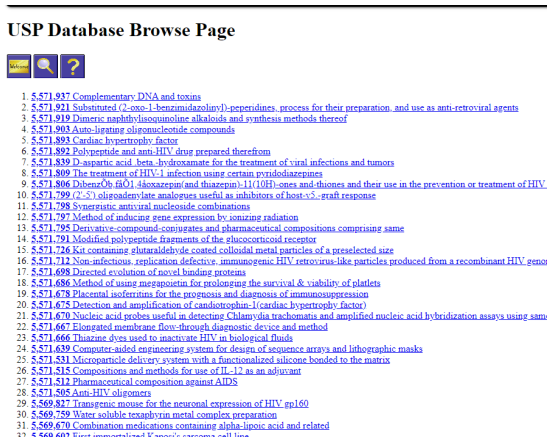
Notes: The figure shows a screenshot of the website of the USPTO in late 1996. The home page included a prominently positioned button linking to the access page of AIDS Patent Database hosted on the CNIDR server (see center-right). Wordwide access to both websites was possible with a dial-in modem and a telephone line.

FIGURE 2.B2. AIDSDB ACCESS PAGE II, FALL 1996



Notes: The figure shows a screenshot to the access page to the AIDS Patent Database hosted on the CNIDR server in December 1996. The database included a search form (allowing for keyword, class and boolean search) as well as a browse page, including the full list and links to all hosted patents. The data base included full-text and high-resolution images and drawings of all patents related to HIV/AIDS. Download pages were optimized for small (56k) bandwidths.

FIGURE 2.B3. AIDSDB ACCESS PAGE III, FALL 1996

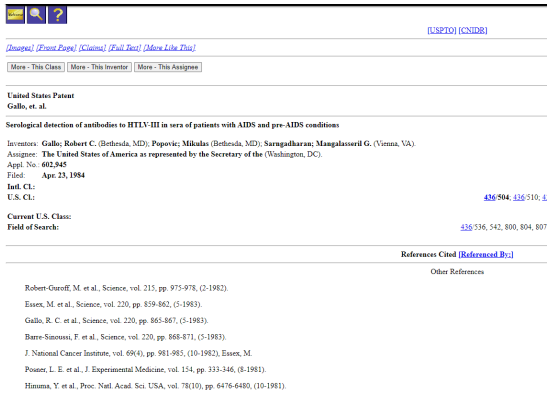


USP Database Browse Page

1. 5,571,937 Complementary DNA and toxins
2. 5,571,921 Substituted (2-oxo-1-benzimidazolinyl)-peperidines, process for their preparation, and use as anti-retroviral agents
3. 5,571,919 Dimeric naphthylisoquinoline alkaloids and synthesis methods thereof
4. 5,571,903 Auto-ligating oligonucleotide compounds
5. 5,571,893 Cardiac hypertrophy factor
6. 5,571,892 Polypeptide and anti-HIV drug prepared therefrom
7. 5,571,839 D-aspartic acid ,beta.-hydroxamate for the treatment of viral infections and tumors
8. 5,571,809 The treatment of HIV-1 infection using certain pyridodiazepines
9. 5,571,806 DibenzÕb,fâÕ[4âoxazepin(and thiazepin)-11(10H)-ones and-thiones and their use in the prevention or treatment of HIV
10. 5,571,799 (2'-5') oligoadenylate analogues useful as inhibitors of host-vS.-graft response
11. 5,571,798 Synergistic antiviral nucleoside combinations
12. 5,571,797 Method of inducing gene expression by ionizing radiation
13. 5,571,795 Derivative-compound-conjugates and pharmaceutical compositions comprising same
14. 5,571,791 Modified polypeptide fragments of the glucocorticoid receptor
15. 5,571,726 Kit containing glutaraldehyde coated colloidal metal particles of a preselected size
16. 5,571,712 Non-infectious, replication defective, immunogenic HIV retrovirus-like particles produced from a recombinant HIV genor
17. 5,571,698 Directed evolution of novel binding proteins
18. 5,571,686 Method of using megapoietin for prolongeing the survival & viability of platlets
19. 5,571,678 Placental isoferritins for the prognosis and diagnosis of immunosuppression
20. 5,571,675 Detection and amplification of cardiotrophin-1(cardiac hypertrophy factor)
21. 5,571,670 Nucleic acid probes useful in detecting Chlamydia trachomatis and amplified nucleic acid hybridization assays using same
22. 5,571,667 Elongated membrane flow-through diagnostic device and method
23. 5,571,666 Thiazine dyes used to inactivate HIV in biological fluids
24. 5,571,639 Computer-aided engineering system for design of sequence arrays and lithographic masks
25. 5,571,531 Microparticle delivery system with a functionalized silicone bonded to the matrix
26. 5,571,515 Compositions and methods for use of IL-12 as an adjuvant
27. 5,571,512 Pharmaceutical composition against AIDS
28. 5,571,505 Anti-HIV oligomers
29. 5,569,827 Transgenic mouse for the neuronal expression of HIV gp160
30. 5,569,759 Water soluble texaphyrin metal complex preparation
31. 5,569,670 Combination medications containing alpha-lipoic acid and related
32. 5,569,607 First immortalized Kaposi's sarcoma cell line

Notes: The figure shows a screenshot to the browse view page to the AIDS Patent Database hosted on the CNIDR server in December 1996. Patents listed were sorted in descending order based on latest issue date. Links were clickable and allowed direct access to the patent view pages. The data base included full-text and high-resolution images and drawings of all patents related to HIV/AIDS. Download pages were optimized for small (56k) bandwidths.

FIGURE 2.B4. AIDSDB ACCESS PAGE IV, PATENT VIEW, FALL 1996



[USPTO] [CNIDR]

[Images] [Front Page] [Claims] [Full Text] [More Like This]

More - This Class | More - This Inventor | More - This Assignee

United States Patent
Gallo, et. al.

Serological detection of antibodies to HTLV-III in sera of patients with AIDS and pre-AIDS conditions

Inventors: **Gallo; Robert C.** (Bethesda, MD); **Popovic; Mikulas** (Bethesda, MD); **Sarngadharan; Mangalasseril G.** (Vienna, VA).
Assignee: **The United States of America as represented by the Secretary of the** (Washington, DC).
Appl. No.: **602,945**
Filed: **Apr. 23, 1984**
Intl. CL:
U.S. CL: 436/504; 436/510; 4
Current U.S. Class:
Field of Search: 436/536, 542, 800, 804, 807

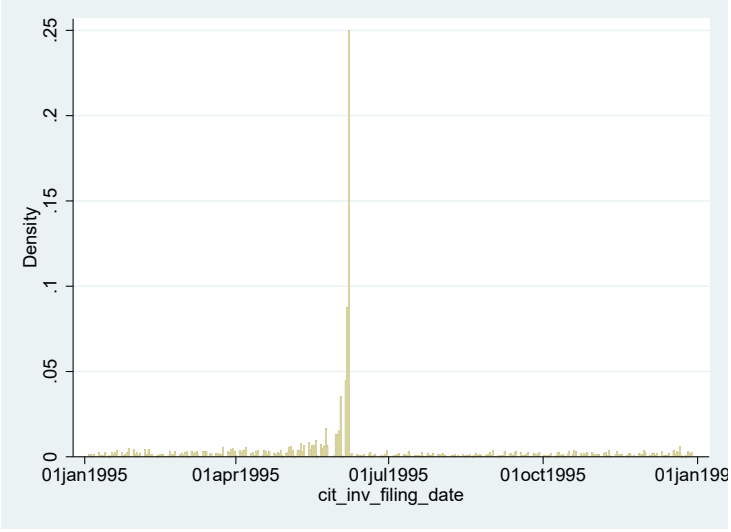
References Cited [Referenced By:]

Other References

Robert-Guroff, M. et al., Science, vol. 215, pp. 975-978, (2-1982).
Essex, M. et al., Science, vol. 220, pp. 859-862, (5-1983).
Gallo, R. C. et al., Science, vol. 220, pp. 865-867, (5-1983).
Barre-Sinoussi, F. et al., Science, vol. 220, pp. 868-871, (5-1983).
J. National Cancer Institute, vol. 69(4), pp. 981-985, (10-1982), Essex, M.
Poiter, L. E. et al., J. Experimental Medicine, vol. 154, pp. 333-346, (8-1981).
Hinuma, Y. et al., Proc. Natl. Acad. Sci. USA, vol. 78(10), pp. 6476-6480, (10-1981).

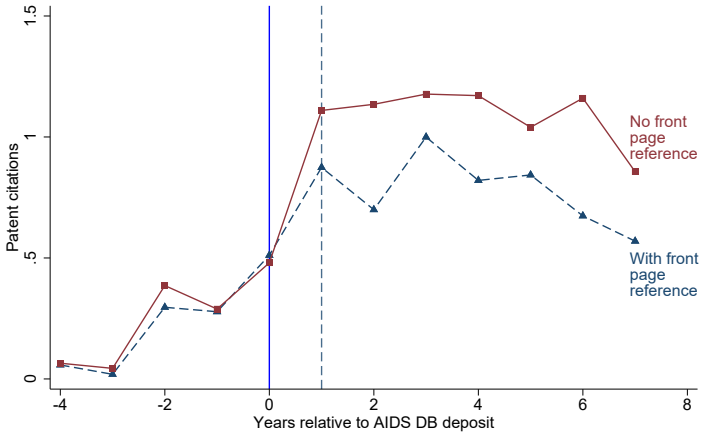
Notes: The figure shows a screenshot to the AIDS DB patent view page of patent # US 4,520,113, by Robert Gallo and co-inventors, granted in 1985, assigned to the U.S. Department of Health and Human Services in Washington, DC. The patent view page included further links to access full-text and high-resolution images and drawings, to this and all patents related to HIV/AIDS. Wordwide access to both websites was possible with a dial-in modem and a telephone line. Download pages were optimized for small (56k) bandwidths.

Figure 2.B5. Spike in Daily U.S. Patent Filings prior to TRIPS Reforms Enactment on June 8th 1995



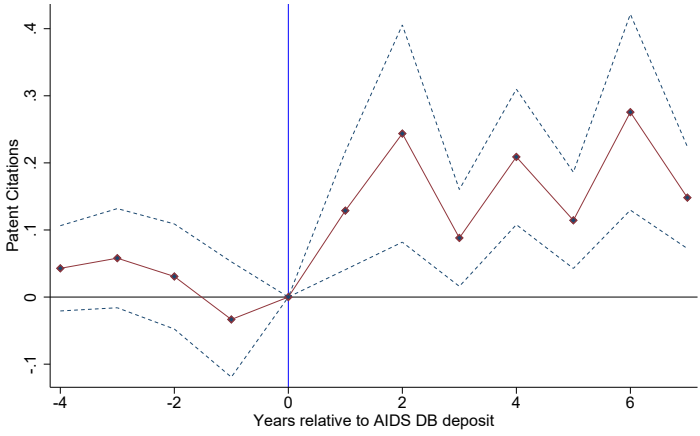
Notes: The figure shows spikes in the number of patent filings to the USPTO in the days and weeks leading up to June 8th of 1995, on which the U.S. patent system amended a significant part of its provision to be in line with the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS), which was negotiated at the end of the Uruguay Round of the General Agreement on Tariffs and Trade (GATT) between 1989 and 1990. The most important amendments included a change in patent term from 17 years after grant to 20 years after filing date of the application, the implementation of a domestic priority system in form of 'provisional applications', and the requirement to recognize foreign priority filings (Source: USPTO).

Figure 2.B6. Group Means *within AIDS DB* Comparison Yearly Patent Citations, Matched on Pre-Trends



Notes: The figure plots trends in group means across AIDS DB patents without vs. with front-page reference to HIV/AIDS, additionally matched on pre-period yearly citation levels. Control group patents consist of AIDS DB patents "with reference" to HIV/AIDS, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired "no reference" patents. The *y*-axis scale reports levels of yearly patent family citations to patents in sample. Inventor self-citations are removed from the counts. The x-axis depicts years relative to online deposit (0). The dashed vertical line (1) indicates a 1-year lag of the database treatment, relative to deposit.The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *PubMed*, *BvD Orbis* and several disambiguations (links) between them (see Section 3.3 for details).

Figure 2.B7. Yearly Effect for Patents without HIV/AIDS Front Page Reference, Matched on Pre-Trends



Notes: The figure plots parameter estimates from regression (2.1) with yearly coefficients for $t-4$ to $t+7$ relative to online deposit for the matched panel of U.S. AIDS DB patents without vs. with front-page reference to HIV/AIDS, additionally matched on pre-period yearly citation levels. The dependent variable measures the yearly number of family citations (inventor and applicant self-citations excluded). The year of deposit is omitted from the regression. The reference category consists of AIDS DB patents "with reference" to HIV/AIDS, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired "no reference" patents. Sample observations are weighted based on Iacus, King and Porro (2012). 95% confidence intervals are based on clustered standard errors. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *BvD Orbis* and several disambiguations (links) between them (see Section 3.3 for details).

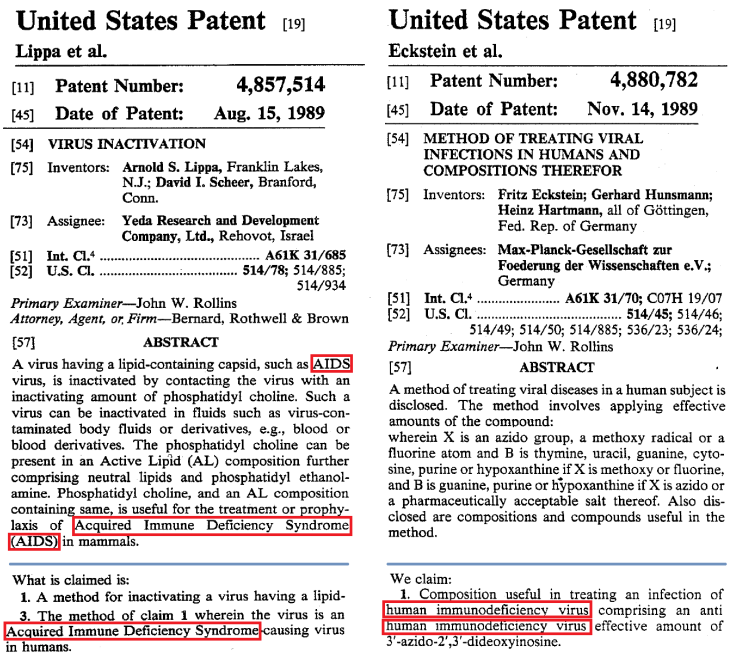
2.C Examples of AIDS DB patents with vs. without front-page reference to HIV/AIDS

In this appendix sections, I present several examples of cases of AIDS DB patents, that are paired in the sample by the matching approach outlined in Section 2.4.2, that differ in the extent to which the disease-link to HIV/AIDS applications becomes obvious from the front-page information. [65] This evidence is, of course, anecdotal and non-systematic. It is important to highlight that my design neither claims nor requires these patent pairs to be (almost) identical in technical content nor that a specialized inventor would not recognize the link to HIV/AIDS of those patents without obvious front-page references upon more careful inspection. Merely, my assumption is that the front-page reference of specific terms relating to the disease increases the likelihood that an inventor, searching for prior art art inputs, would retrieve a patent as relevant out of a bibliographic index containing numerous, diverse patents.

The first example, shown in Figure 2.C1, is the case of two patents granted in the second half of 1989 on methods, and related compounds, inhibiting viral replication after HIV-infection, targeted specifically at the prodromal (pre-AIDS) stage of disease. The patent on the left-hand side (patent #: US 4,857,514) has been invented by Arnold Lippa and David Sheer (both based in the U.S.) and has been first filed in 1985. It is assigned to the "Yeda Research and Development Company Ltd.", which is a technology transfer organization of the Weizmann Institute of Science, in Rehovot, Israel, one of the leading public research institutions in natural and exact sciences. The patent on the right in Figure 2.C1 (patent #: US 4,880,782) is an invention of Fritz Eckstein, Gerhard Hunsmann and Heinz Hartmann (all

[65] To identify this in the paper, I query the front-pages of AIDS DB patents for explicit textual references to terms relating to HIV/AIDS, as defined by the National Library of Medicine (see: https://meshb.nlm.nih.gov/).Queried keywords for HIV/AIDS-related terms are: *AIDS; Acquired Immune Deficiency Syndrome; Acquired Immuno-Deficiency Syndrome; Immunodeficiency Syndrome, Acquired; AIDS Arteritis, Central Nervous System; AIDS Dementia Complex;AIDS Serodiagnosis; HIV Seropositivity; HIV Seroprevalence; Lymphoma, AIDS-Related; HTLV-III; Human Immunodeficiency Virus; Human T Cell Lymphotropic Virus Type III; Human T Lymphotropic Virus Type III; Human T-Cell Leukemia Virus Type III; Human T-Cell Lymphotropic Virus Type III; Human T-Lymphotropic Virus Type III; Immunodeficiency Virus, Human; Immunodeficiency Viruses, Human; LAV-HTLV-III; Lymphadenopathy-Associated Virus; Virus, Human Immunodeficiency; Human T-Cell Leukemia Virus*

FIGURE 2.C1. EXAMPLE 1- LIPPA ET AL. VS. ECKSTEIN ET AL. (1989)



United States Patent [19]

Lippa et al.

[11] **Patent Number: 4,857,514**

[45] **Date of Patent: Aug. 15, 1989**

[54] **VIRUS INACTIVATION**

[75] Inventors: **Arnold S. Lippa,** Franklin Lakes, N.J.; **David I. Scheer,** Branford, Conn.

[73] Assignee: **Yeda Research and Development Company, Ltd.,** Rehovot, Israel

[51] **Int. Cl.**[4] **A61K 31/685**

[52] **U.S. Cl.** **514/78;** 514/885; 514/934

Primary Examiner—John W. Rollins
Attorney, Agent, or Firm—Bernard, Rothwell & Brown

[57] **ABSTRACT**

A virus having a lipid-containing capsid, such as AIDS virus, is inactivated by contacting the virus with an inactivating amount of phosphatidyl choline. Such a virus can be inactivated in fluids such as virus-contaminated body fluids or derivatives, e.g., blood or blood derivatives. The phosphatidyl choline can be present in an Active Lipid (AL) composition further comprising neutral lipids and phosphatidyl ethanolamine. Phosphatidyl choline, and an AL composition containing same, is useful for the treatment or prophylaxis of Acquired Immune Deficiency Syndrome (AIDS) in mammals.

What is claimed is:

1. A method for inactivating a virus having a lipid-

3. The method of claim **1** wherein the virus is an Acquired Immune Deficiency Syndrome-causing virus in humans.

United States Patent [19]

Eckstein et al.

[11] **Patent Number: 4,880,782**

[45] **Date of Patent: Nov. 14, 1989**

[54] **METHOD OF TREATING VIRAL INFECTIONS IN HUMANS AND COMPOSITIONS THEREFOR**

[75] Inventors: **Fritz Eckstein; Gerhard Hunsmann; Heinz Hartmann,** all of Göttingen, Fed. Rep. of Germany

[73] Assignees: **Max-Planck-Gesellschaft zur Foederung der Wissenschaften e.V.;** Germany

[51] **Int. Cl.**[4] **A61K 31/70;** C07H 19/07

[52] **U.S. Cl.** **514/45;** 514/46; 514/49; 514/50; 514/885; 536/23; 536/24;

Primary Examiner—John W. Rollins

[57] **ABSTRACT**

A method of treating viral diseases in a human subject is disclosed. The method involves applying effective amounts of the compound:
wherein X is an azido group, a methoxy radical or a fluorine atom and B is thymine, uracil, guanine, cytosine, purine or hypoxanthine if X is methoxy or fluorine, and B is guanine, purine or hypoxanthine if X is azido or a pharmaceutically acceptable salt thereof. Also disclosed are compositions and compounds useful in the method.

We claim:

1. Composition useful in treating an infection of human immunodeficiency virus comprising an anti human immunodeficiency virus effective amount of 3'-azido-2',3'-dideoxyinosine.

Notes: The upper part of the figure shows (shortened) front-page information of the original patent documents. The left-hand side patent contains explicit textual references to terms relating to HIV/AIDS, as defined by the National Library of Medicine (see: https://meshb.nlm.nih.gov/). The patent on the right does not contain such front-page references. The lower part of the figure (below the blue separating line) contains relevant excerpts from the full-patent text, which was not retrievable by inventors from standard bibliographic search.

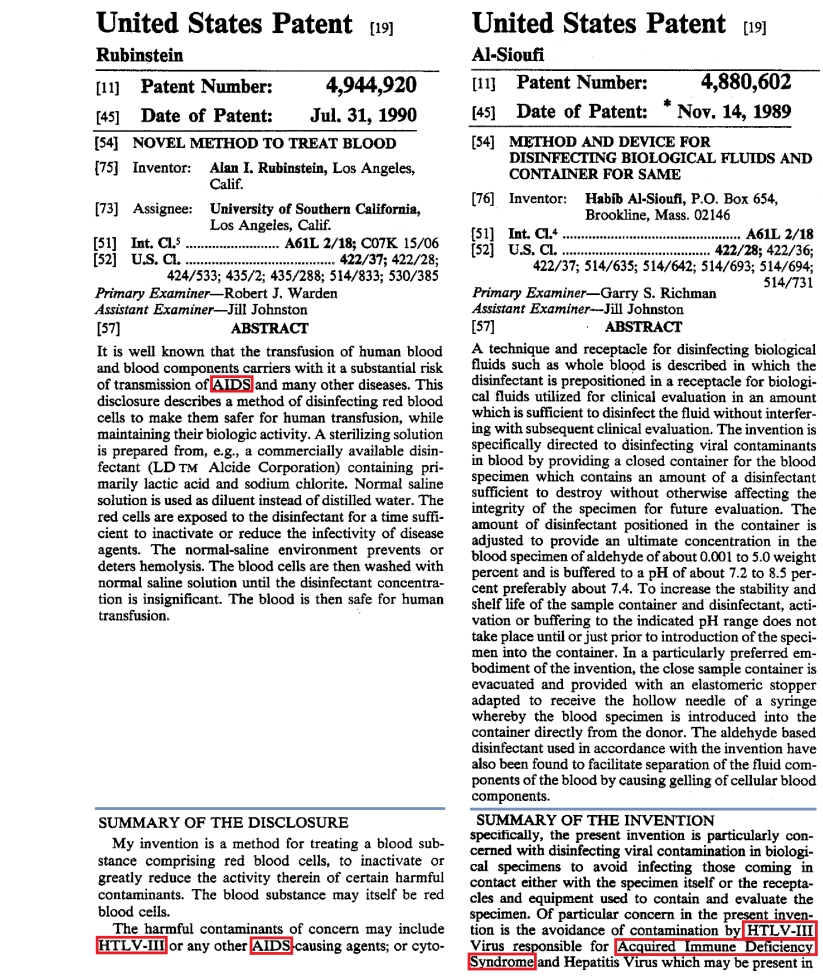
based in Germany), filed first in 1986, and is assigned to the Max-Planck Society and the German Primate Research Centre, both renown non-profit public research organizations located in Göttingen, Germany. The technological similarity of the inventions can been seen from the assigned patent classes in Figure 2.C1: Both share the same 6-digit IPC class ("A61K 31"), as well as the same USPC 3-digit class ("514"). This similarity is further evidenced in Figure 2.C1 by the fact that both patents have been reviewed by the same primary USPTO patent examiner (John W. Rollins). Each of the two patents reports several scientific publications as prior art references. Figure 2.C1 further shows that, while the Lippa et al. patent makes two explicit references to "*AIDS*" and the usefulness of the invention for "*the*

treatment and prophylaxis of Acquired Immune Deficiency Syndrome (AIDS)", the front page of the patent by the German inventors does not include any such reference regarding the applicability to HIV/AIDS, but remains very technical and focused primarily on the chemical compounds related to the invention.

The bottom part of Figure 2.C1, below the inserted blue line, shows information extracted from the full text body of the two patents, which has not been visible to inventors in standard bibliographic prior art search. It can be clearly seen that both inventions apply to the treatment of the AIDS-causing, human immunodeficiency virus. Most noticeably, this becomes visible from the inspection of the patent claims – which represent the core of the invention for which the patent is claimed; In the case of the Eckstein et al. patent, which included no front page reference to HIV/AIDS, already the first (main) claim made reflects the usefulness of the invention for "*treating an infection of human immunodeficiency virus*". Analogously, though only in the 3rd claim, this can be seen for the Lippa et al. patent.

The second example, shown in Figure, is a patent pair from 1990 of two U.S. based single-inventors; Patent # US 4,944,920 of Alan Rubinstein, assigned to the University of Southern California in Los Angeles, CA, and patent # US 4,880,602 of Habib Al-Sioufi, a pathologist at the University of Massachusetts Hospital in Boston, MA. Both inventions relate to methods for disinfecting viral contaminants from HIV in human blood and other body fluids, without affecting the integrity of the blood specimen for transfusion or further clinical evaluation. In each of the two cases, the usefulness of the invention is targeted at the avoidance of contamination from HIV, but not limited to this specific viral agent. Similar to the previous case, the strong technological relatedness between the two patents is shown in Figure 2.C2, namely by sharing the same IPC-class ("A16L 2/18") and USPC (sub-)classes ("422", "514"), as well as the same assistant USPTO patent examiner (Jill Johnston). While the Rubinstein patent makes a clear reference to the "*risk of transmission of AIDS*" in the abstract on the front page, no such reference is found for the Al-Sioufi patent cover page, which remains relatively unspecific with regard to usefulness and areas of application (see Figure 2.C2). However, when inspecting the remainder of the patent document, when outlying the content of the invention, also the latter patent points out that "*of particular concern in the present invention is the avoidance of contamination by HTLV-III virus responsible for Acquired Immune*

FIGURE 2.C2. EXAMPLE 2- RUBINSTEIN VS. AL-SIOUFI (1990)



United States Patent [19]

Rubinstein

[11] **Patent Number: 4,944,920**

[45] **Date of Patent: Jul. 31, 1990**

[54] **NOVEL METHOD TO TREAT BLOOD**

[75] Inventor: **Alan I. Rubinstein**, Los Angeles, Calif.

[73] Assignee: **University of Southern California**, Los Angeles, Calif.

[51] **Int. Cl.**[5] **A61L 2/18**; C07K 15/06

[52] **U.S. Cl.** **422/37**; 422/28; 424/533; 435/2; 435/288; 514/833; 530/385

Primary Examiner—Robert J. Warden

Assistant Examiner—Jill Johnston

[57] **ABSTRACT**

It is well known that the transfusion of human blood and blood components carriers with it a substantial risk of transmission of AIDS and many other diseases. This disclosure describes a method of disinfecting red blood cells to make them safer for human transfusion, while maintaining their biologic activity. A sterilizing solution is prepared from, e.g., a commercially available disinfectant (LD TM Alcide Corporation) containing primarily lactic acid and sodium chlorite. Normal saline solution is used as diluent instead of distilled water. The red cells are exposed to the disinfectant for a time sufficient to inactivate or reduce the infectivity of disease agents. The normal-saline environment prevents or deters hemolysis. The blood cells are then washed with normal saline solution until the disinfectant concentration is insignificant. The blood is then safe for human transfusion.

SUMMARY OF THE DISCLOSURE

My invention is a method for treating a blood substance comprising red blood cells, to inactivate or greatly reduce the activity therein of certain harmful contaminants. The blood substance may itself be red blood cells.

The harmful contaminants of concern may include HTLV-III or any other AIDS-causing agents; or cyto-

United States Patent [19]

Al-Sioufi

[11] **Patent Number: 4,880,602**

[45] **Date of Patent: * Nov. 14, 1989**

[54] **METHOD AND DEVICE FOR DISINFECTING BIOLOGICAL FLUIDS AND CONTAINER FOR SAME**

[76] Inventor: **Habib Al-Sioufi**, P.O. Box 654, Brookline, Mass. 02146

[51] **Int. Cl.**[4] **A61L 2/18**

[52] **U.S. Cl.** **422/28**; 422/36; 422/37; 514/635; 514/642; 514/693; 514/694; 514/731

Primary Examiner—Garry S. Richman

Assistant Examiner—Jill Johnston

[57] **ABSTRACT**

A technique and receptacle for disinfecting biological fluids such as whole blood is described in which the disinfectant is prepositioned in a receptacle for biological fluids utilized for clinical evaluation in an amount which is sufficient to disinfect the fluid without interfering with subsequent clinical evaluation. The invention is specifically directed to disinfecting viral contaminants in blood by providing a closed container for the blood specimen which contains an amount of a disinfectant sufficient to destroy without otherwise affecting the integrity of the specimen for future evaluation. The amount of disinfectant positioned in the container is adjusted to provide an ultimate concentration in the blood specimen of aldehyde of about 0.001 to 5.0 weight percent and is buffered to a pH of about 7.2 to 8.5 percent preferably about 7.4. To increase the stability and shelf life of the sample container and disinfectant, activation or buffering to the indicated pH range does not take place until or just prior to introduction of the specimen into the container. In a particularly preferred embodiment of the invention, the close sample container is evacuated and provided with an elastomeric stopper adapted to receive the hollow needle of a syringe whereby the blood specimen is introduced into the container directly from the donor. The aldehyde based disinfectant used in accordance with the invention have also been found to facilitate separation of the fluid components of the blood by causing gelling of cellular blood components.

SUMMARY OF THE INVENTION

specifically, the present invention is particularly concerned with disinfecting viral contamination in biological specimens to avoid infecting those coming in contact either with the specimen itself or the receptacles and equipment used to contain and evaluate the specimen. Of particular concern in the present invention is the avoidance of contamination by HTLV-III Virus responsible for Acquired Immune Deficiency Syndrome and Hepatitis Virus which may be present in

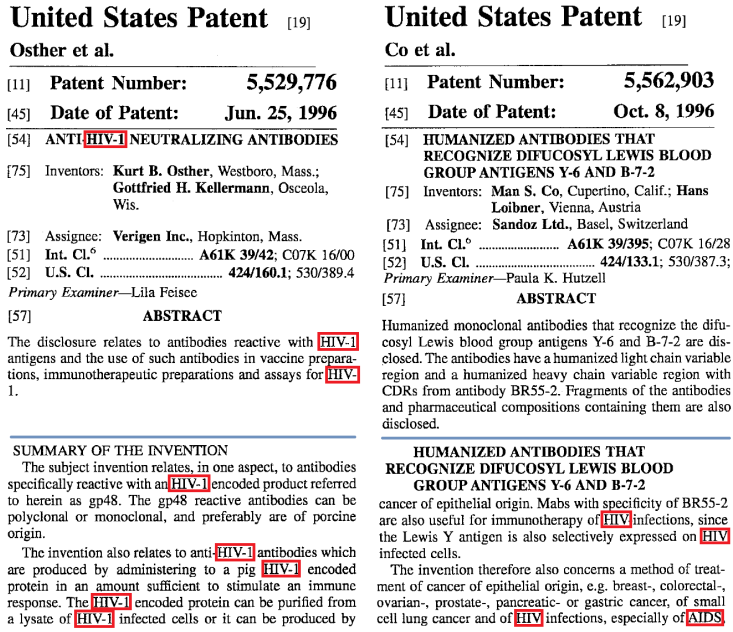
Notes: The upper part of the figure shows (shortened) front-page information of the original patent documents. The left-hand side patent contains explicit textual references to terms relating to HIV/AIDS, as defined by the National Library of Medicine (see: https://meshb.nlm.nih.gov/). The patent on the right does not contain such front-page references. The lower part of the figure (below the blue separating line) contains relevant excerpts from the full-patent text, which was not retrievable by inventors from standard bibliographic search.

Deficiency Syndrome(..)"[66], as reported in Figure 2.C2 below the blue separation line.

[66]HTLV-III is the abbreviation for "Human T-cell Lymphotropic Virus Type-III", which is the term used to denote the causative agent of AIDS in the original paper by Gallo et al. *Science* 220:865-867 (1983),

On the left hand side at the bottom of Figure 2.C2 it can be seen that the usefulness for decontamination from HIV-agents is outlined in the Rubinstein patent full text with almost identical wording.

FIGURE 2.C3. EXAMPLE 3- OSTHER ET AL. VS. CO ET AL. (1996)



United States Patent [19]

Osther et al.

[11] **Patent Number: 5,529,776**

[45] **Date of Patent: Jun. 25, 1996**

[54] **ANTI HIV-1 NEUTRALIZING ANTIBODIES**

[75] Inventors: **Kurt B. Osther**, Westboro, Mass.; **Gottfried H. Kellermann**, Osceola, Wis.

[73] Assignee: **Verigen Inc.**, Hopkinton, Mass.

[51] **Int. Cl.**[6] **A61K 39/42**; C07K 16/00

[52] **U.S. Cl.** **424/160.1**; 530/389.4

Primary Examiner—Lila Feisee

[57] **ABSTRACT**

The disclosure relates to antibodies reactive with HIV-1 antigens and the use of such antibodies in vaccine preparations, immunotherapeutic preparations and assays for HIV-1.

SUMMARY OF THE INVENTION

The subject invention relates, in one aspect, to antibodies specifically reactive with an HIV-1 encoded product referred to herein as gp48. The gp48 reactive antibodies can be polyclonal or monoclonal, and preferably are of porcine origin.

The invention also relates to anti-HIV-1 antibodies which are produced by administering to a pig HIV-1 encoded protein in an amount sufficient to stimulate an immune response. The HIV-1 encoded protein can be purified from a lysate of HIV-1 infected cells or it can be produced by

United States Patent [19]

Co et al.

[11] **Patent Number: 5,562,903**

[45] **Date of Patent: Oct. 8, 1996**

[54] **HUMANIZED ANTIBODIES THAT RECOGNIZE DIFUCOSYL LEWIS BLOOD GROUP ANTIGENS Y-6 AND B-7-2**

[75] Inventors: **Man S. Co**, Cupertino, Calif.; **Hans Loibner**, Vienna, Austria

[73] Assignee: **Sandoz Ltd.**, Basel, Switzerland

[51] **Int. Cl.**[6] **A61K 39/395**; C07K 16/28

[52] **U.S. Cl.** **424/133.1**; 530/387.3;

Primary Examiner—Paula K. Hutzell

[57] **ABSTRACT**

Humanized monoclonal antibodies that recognize the difucosyl Lewis blood group antigens Y-6 and B-7-2 are disclosed. The antibodies have a humanized light chain variable region and a humanized heavy chain variable region with CDRs from antibody BR55-2. Fragments of the antibodies and pharmaceutical compositions containing them are also disclosed.

HUMANIZED ANTIBODIES THAT RECOGNIZE DIFUCOSYL LEWIS BLOOD GROUP ANTIGENS Y-6 AND B-7-2

cancer of epithelial origin. Mabs with specificity of BR55-2 are also useful for immunotherapy of HIV infections, since the Lewis Y antigen is also selectively expressed on HIV infected cells.

The invention therefore also concerns a method of treatment of cancer of epithelial origin, e.g. breast-, colorectal-, ovarian-, prostate-, pancreatic- or gastric cancer, of small cell lung cancer and of HIV infections, especially of AIDS.

Notes: The upper part of the figure shows (shortened) front-page information of the original patent documents. The left-hand side patent contains explicit textual references to terms relating to HIV/AIDS, as defined by the National Library of Medicine (see: https://meshb.nlm.nih.gov/). The patent on the right does not contain such front-page references. The lower part of the figure (below the blue separating line) contains relevant excerpts from the full-patent text, which was not retrievable by inventors from standard bibliographic search.

A final example of AIDS DB patent pairs in the sample with vs. without front-page reference to HIV/AIDS is provided in Figure 2.C3 comparing the 1996 patents of Osther et al. (patent # US 5,529,776) and Co et al. (patent # US 5,562,903). The are two private firm patents, the former assigned to Verigen Inc., of Hopkinton, MA, in the U.S., and the latter to Sandoz Ltd., of Basel in Switzerland.

Both patented inventions relate to the production of specific antibodies, of porcine or murine chimeric origin, useful as immunotherapeutics to humans in-

later adopted to be referred to as Human Immunodeficiency Virus (HIV) by the scientific community.

fected by HIV. These can be further used to isolate HIV antigens, particularly useful for vaccine development. Each of the two documents constitutes a continuation of previous applications, with filing year 1993. As Figure 2.C3 shows, they are assigned to the same IPC-6 classes ("A61K 39") and USPC 3-digit classes ("424", "530"). They were examined by the same art unit, but with different examiners. While the Osther et al. patent makes several explicit front-page references to HIV, in both title and abstract, the Co et al. patent is entirely missing such an indication (see Figure 2.C3). When inspecting the full text body of the latter patent, however, the link to HIV/AIDS becomes clearly evident starting from the first paragraph of patent, containing the summary of the invention (excerpts of which can be seen in the bottom-right part of Figure 2.C3, below the blue separating line). The bottom-left part of Figure 2.C2 shows the corresponding text-excerpt from the invention summary of the Osther et al. patent, with several references regarding the applicability to HIV infections of the produced antibodies, in line with the explicit indication provided on the patent front page.

2.D External matched control group results

Complementing the main findings of the paper on internal search costs, in this appendix section, I attempt to establish a baseline of the marginal impact of *online access* on cumulative citations by comparing AIDS DB patents to a control group of external patents with same timing, which were not included in the repository. To account for the selection at the researcher level and the particular institutional environments of HIV-research, I sample the control group exclusively from the population of non-AIDS DB patents of the same inventors whose patents were deposited in the database. To avoid confounding influences stemming from productivity changes or switching research focus over time, I only consider those patents by HIV-inventors that were filed at the same time or since the filing of their first included AIDS DB patent.[67] Further, I only consider patent family members filed at the USPTO. To avoid comparing patents on different types of technologies within broader technological fields, of different institutional context, scientific background and timing, I apply the same selection criteria to this control group as in the main analysis and apply the weights of Iacus, King and Porro (2012) to ensure balance in the estimation (see also Section 2.4.1).

Table 2.D1 summarizes the main invention-level characteristics of patents in the matched sample. My strict selection criteria yield to at least one control patent for 1,979 AIDS DB patents.[68] A total of 17 technology fields and 57 art units are represented in the sample. Previous to deposit, AIDS DB patents in the sample receive on average 1.3 citations per year, which is a difference in citation levels of about .2 compared to the control group (significant at 1%). The sample groups are highly comparable on a range of ex-ante characteristics, however, some small, significant, differences persist, namely in the number of scientific references and the inventor team composition. Average lags between deposit and patent publication (18 months) and application (55 months) indicate that I observe a large share of the

[67] Unfortunately, it is not possible to compare patents within exactly the same inventor, given that my design requires matching timing of invention, and only a very small group of inventors files more than one patent within the same narrow time window.

[68] By this, my sample covers 65.5% of all originally deposited patents in the AIDS DB. Inference is limited to this subset.

Table 2.D1. Summary Statistics, until Year of AIDS DB Deposit

Matched sample	AIDSDB patents		Control group		Diff
	Mean	SD	Mean	SD	p-val.
Yearly patent family citations until AIDSDB deposit	1.348	2.909	1.147	2.599	.009
Number of patent references	9.145	10.729	8.768	10.028	.196
Share with scientific reference	.798		.799		.930
Number of scientific references	11.636	21.403	9.271	15.110	.000
Share of novel technologies	.247		.227		.096
Share introducing new words	.311		.304		.592
Number of inventors	2.996	2.110	3.373	2.004	.000
Share with author-inventors	.916		.951		.000
Number of author-inventors	2.469	1.826	2.872	1.798	.000
Patent family size	6.300	6.898	6.271	6.885	.882
Share private firm patents	.667		.666		.940
Assignee prior patent families	2.989k	10.436k	3.554k	9.123k	.056
DB-to-publication lag (m)	18.533	20.997	18.010	21.921	.393
DB-to-application lag (m)	54.746	27.297	54.639	27.714	.891
Number of patents	1,979		3,361		
Number of technology fields	14		17		
Number of examining art units	57		57		

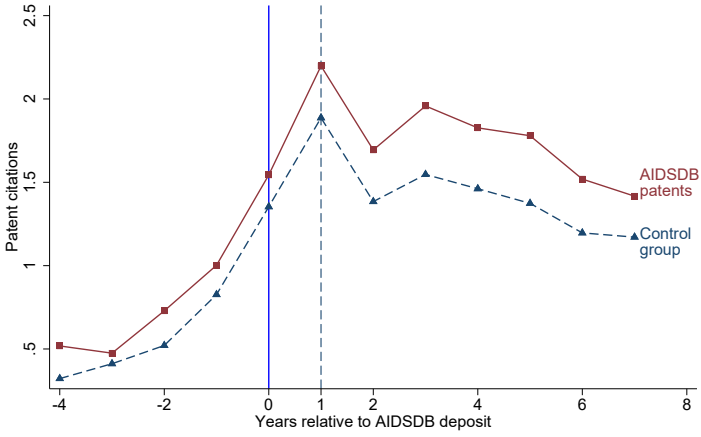
Notes: Row (1) reports the group mean and standard deviation for yearly patent family citations to deposited and control patents between year $t-4$ and t relative to AIDS DB deposit. Inventor and applicant self-citations are removed from the counts. The following rows report ex-ante time-invariant characteristics. Control group patents consist of not-online deposited patents of AIDS DB inventors, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired AIDS DB patents. Technology fields are based on Schmoch (2008). Sample observations are weighted according to Iacus, King and Porro (2012). Column (6) reports p-values from two-sample t-tests with unequal variances for differences in sample means. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *PubMed*, *BvD Orbis* and several disambiguations (links) between them (see Section 3.3 for details).

sample patents for a significant time span before database inclusion.[69]

In order to evaluate whether the significant initial difference in levels of yearly patent citations between the two groups (as reported in Table 2.D1) is stable in the pre-period, I inspect trends in groups means over time for the years preceding

[69] Online deposit of AIDS DB patents is assumed to take place on average one month after patent publication, compare Section 2.3.1

Figure 2.D1. Group Means Comparison Yearly Patent Citations



Notes: The figure plots trends in group means across AIDS DB deposited and control patents. Control group patents consist of *not* online-deposited patents of AIDS DB inventors, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired AIDS DB patents. The *y*-axis scale reports levels of yearly patent family citations to patents in sample. Inventor and applicant self-citations are removed from the counts. The x-axis depicts years relative to online deposit (0). The dashed vertical line (1) indicates a 1-year lag of the database treatment, relative to deposit.The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *PubMed*, *BvD Orbis* and several disambiguations (links) between them (see Section 3.3 for details).

the database deposit in Figure 2.D1. The graph shows that, despite the constant small difference in levels, trends of group means between AIDS DB and control patents remain basically parallel over the entire pre-period from $t-4$ to t (with the minor exception from $t-4$ to $t-3$), which is a necessary condition for inference of an average treatment effect on the treated, and suggests that the control group is well selected. Figure 2.D1 also reveals that starting from $t+1$ differences in group means can be seen to substantially increase, while trends still follow largely parallel patterns.

I then compare within-patent changes in differences in citation rates across groups after AIDS DB deposit in a generalized difference-in-differences framework by estimating the following regression equation, similar to the one in the main

analysis (cf. 2.1):

$$Y_{it} = \beta_1 * AIDSDB_i \times post_{t-1} + patentFE_i + yearFE_t + \phi_y + \theta_{fy} + \epsilon_{it}, \qquad (2.3)$$

where i indexes patents, t indexes relative years to AIDS DB deposit, y indexes calendar years, and f indexes technology fields. The dependent variable measures the number of citations per relative year to deposit for each AIDS DB patent and, analogously, per relative year to deposit of the matched AIDS DB patent for each control patent. Control patents have the function to provide a reference level of citations that would have been received by a matched AIDS DB patent in the absence of the database treatment. Therefore, it is necessary to assign each control patent to a 'fictitious' deposit date, relative to which the treatment effect is observed. I assign to each group of two or more matched patents a unique database deposit date, based on the most frequent occurring actual deposit date in the group.[70] The coefficient β_1 measures changes in citation rates to AIDS DB patents, after deposit, relative to the control group patents, which are the excluded reference category. The interacted $post_{t-1}$ indicator denotes the one-year lagged post-deposit status. The regression model includes a full set of patent fixed effects and fixed effects for relative years to the AIDS DB deposit date. To control for the confounding influence of shocks possibly affecting citation rates over time in the overall economy or the patent system (e.g., the enactment of the TRIPS Agreement in 1995), the regression further includes a full set of calendar year fixed effects (captured by the parameter ϕ_y). Finally, I include linear field-year trends (θ_{fy}) to control for idiosyncratic variation in productivity of specific technology fields. I estimate regression (2.3) on a symmetric sample window of five years preceding and five years following the switching of the $post_{t-1}$ indicator, i.e., for example, ranging from October 1991 to October 2000 for patents deposited in the initial launch of the AIDS DB database on October 26th 1994. Finally, I cluster all standard errors at the patent-level. For a more extensive discussion of the econometric framework, see Section 2.4.1.

Table 2.D2 includes the main econometric results from the estimation of regression (2.3) on the matched sample. Column (1) of Table 2.D2 shows that, starting

[70] In case of multiple, I assign the earliest date. By construction determined deposit dates within matched groups are very close to each other.

from one year after deposit, AIDS DB patents received on average .04 standard deviations in cumulative citations more relative to control group patents (significant at the 10% level). Compared to the pre-deposit mean of citations (see Table 2.D1), this implies a relative increase of .13 citations per year (roughly +10%) for the average AIDS DB patent. When including technology field specific time trends (Table 2.D2, column (2)), the results become slightly larger and more significant (+.05 std. deviation, p-value < .05). This constitutes my preferred specification and provides the first main finding of the paper.

Table 2.D2. Main Effect on Patent Citations, Matched Sample

Dependent variable:	OLS				Poisson
Number of patent citations	(1)	(2)	(3)	(4)	(5)
AIDSDB × post$_{t-1}$	0.044*	0.049**	0.050**	−0.004	0.052**
	(0.024)	(0.024)	(0.024)	(0.025)	(0.022)
Patent fixed effects	Yes	Yes	Yes	Yes	Yes
Time/ year fixed effects	Yes	Yes	Yes	Yes	Yes
Field time trends		Yes	Yes	Yes	Yes
Excl. firm self-cites			Yes		
Prosecution cites only				Yes	
Observations	45,818	45,818	45,731	44,997	39,862
Number of patents	5,339	5,339	5,339	5,329	4,618
Mean dependent					1.395

Notes: Each column reports parameter estimates of regression (2.3) for the matched panel of U.S. AIDS DB and control patents. The dependent variable measures the yearly number of family citations for years $t-4$ to $t+5$ relative to the one-year lagged online date. Inventor self-citations are removed from the counts. Control group patents are the reference category and consist of *not* online-deposited patents of AIDS DB inventors, filed and granted at the same time, with similar institutional and scientific background, and examined by the same art unit as paired AIDS DB patents. In columns (1)-(4), the number of citations is standardized to mean zero and standard deviation one within technology fields (fields based on Schmoch (2008)). Column (5) estimates Poisson pseudo-maximum likelihood regressions (correct for over-dispersion by construction) to address the count nature of citations. Sample observations are weighted based on Iacus, King and Porro (2012). Standard errors are clustered at the patent level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *BvD Orbis* and several disambiguations (links) between them (see Section 3.3 for details).

I check the robustness of this finding across several alternative models: In col-

umn (3) of Table 2.D2, I re-estimate regression (2.3) excluding all firm-level self-citations from the citation counts. The point estimate in column (3) shows that this further increases the result slightly (significant at the 5% level), suggesting the effect to be largely due to an increase in external spillovers. In column (4) of Table 2.D2, I re-estimate regression (2.3) *only* counting citations given from patents that were already under prosecution at the time of online deposit of the cited AIDS DB patent (or its linked control patent), i.e. filed before and granted after the AIDS DB deposit date.[71] These citations are very likely to be given by examiners rather than by the applicants.[72] They also cannot reflect knowledge spillovers from external search through online access to the AIDS DB, as patent applications were already pending and, accordingly, the inventive search process must have been terminated at the time of online deposit. Results in column (4) show no significant difference in citations added during prosecution after database deposit across AIDS DB and control patents. The point estimate for β_1 is even slightly negative, but close to zero and highly insignificant. This suggests that the launch of the AIDS DB had no influence on citation practices of patent examiners. Finally, in Table 2.D2 column (5), I estimate robustness of the preferred specification with Poisson pseudo-maximum likelihood, in order to address the count nature of citations. The Poisson estimates fully confirm the main result, indicating a relative increase in citations to AIDS DB patents of 5.3% (+.074 citations for the average patent) compared to control group patents (significant at the 5% level), starting from one year after online deposit, which is slightly smaller in magnitude compared to the marginal impact estimated with OLS.[73]

In order to investigate the timing of these effects, I re-estimate regression (2.3) with yearly coefficients, by interacting the *AIDSDB* indicator with a set of individual year dummies for $t-4$ to $t+7$ relative to the database date (excluding the year of deposit as reference year). Figure 2.D2 plots the corresponding point estimates within 95% confidence intervals. There are no significant differences estimated between citation trends of AIDS DB and control group patents in the years prior

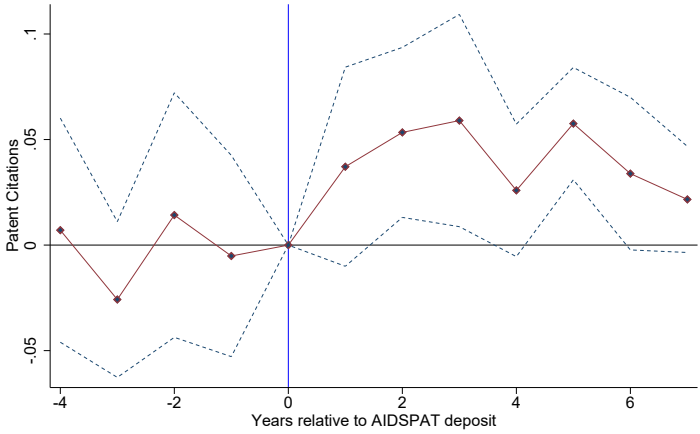
[71] In this case, deviant from my standard approach, I consider as citation date the grant date of a citing patent, which is arguably closest to the examination moment.

[72] see Arora, Belenzon and Lee (2018) for a similar approach

[73] The percentage increase associated with the point estimate of β_1 in the Poisson pseudo-maximum likelihood model is given by e^{β_1}

to database inclusion, suggesting that differences in pre-trends cannot explain the results. On the other hand, the figure shows a steep relative increase in the rate of citations to AIDS DB patents in the years following their online availability, setting in highly significantly after one year, and reaching a peak around the third year post-deposit.

Figure 2.D2. Yearly Effect on Patent Citations



Notes: The figure plots parameter estimates from regression (2.3) with yearly coefficients for $t-4$ to $t+7$ relative to online deposit for the matched panel of U.S. AIDS DB and control patents. The dependent variable measures the yearly number of family citations (inventor self-citations excluded). The year of deposit is omitted from the regression. Control group patents consist of *not* online-deposited patents of AIDS DB inventors, filed and granted at the same time, with similar institutional and scientific background, and examined by the same art unit as paired AIDS DB patents. Sample observations are weighted based on Iacus, King and Porro (2012). 95% confidence intervals are based on clustered standard errors. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *BvD Orbis* and several disambiguations (links) between them (see Section 3.3 for details).

Figure 2.D2 also indicates a non-monotonic effect over time, as estimated differences, first, decline sharply for one period after four years, and, subsequently, seem to gradually level out starting from the fifth year post-AIDS DB. For most patents in the sample (those of the initial cohort of patents uploaded in 1994), these periods coincide with the launch, first, of the comprehensive bibliographic online database of the USPTO (in 1997) and, second, the full-text and images online catalogue including all U.S. patents (in 1998) and EPO Espacenet (1999). This provides further

support for the believe that the observed effects are, indeed, caused by enhanced-access to AIDS DB patents by means of the online repository.

Chapter 3

The Impact of Broadband Internet on Invention: Evidence from the U.K.[1]

3.1 Introduction

In the endogenous growth theory literature, cumulative innovation constitutes a fundamental underpinning of economic progress (Romer, 1990; Aghion and Howitt, 1992; Jones, 1995; Weitzman, 1998). Prior knowledge and ideas are the main input in the production of new technological inventions. Edison's light bulb, Mullis' polymerase chain reaction to multiply DNA, or Google's PageRank are just a few famous examples. Larry Page, for instance, heavily relied on prior art ranking scientific publications in order to create Google's search algorithm to rank web pages. In this regard, scholars have argued that technological progress and innovation heavily depend on the ability and cost to access relevant information

[1] This chapter is based on joint work with Sam Arts, Mattia Nardotto and Jesse Wursten. I thank the members of my doctoral committee, Geert Dhaene and seminar participants at KU Leuven for many useful comments and suggestions. Financial support from the Research Foundation - Flanders (FWO) is gratefully acknowledged under grant numbers 11B5918N and 1125818N.

(Jones, 2003; Mokyr, 2005).

Different large scale surveys across Europe and the U.S. suggest that technical knowledge from patent documents represents the most important source of codified information for inventors, next to scientific publications and tacit knowledge from interpersonal networks (Cohen, Nelson and Walsh, 2000; Jaffe, Trajtenberg and Fogarty, 2000; Giuri et al., 2007; Devlin, 2009; Sampat, 2010; Ouelette, 2012). While the patent system gives private incentives to incur the cost of innovation, the social value of granting patents is also contingent on the extent to which the publication of patent documents triggers knowledge spillovers and cumulative innovation (Machlup and Penrose, 1950; Scotchmer and Green, 1990; Scotchmer, 1991). A large body of prior work confirms that patent disclosure stimulates knowledge diffusion and technological progress (Graham and Hegde, 2015; Hegde and Luo, 2018; Furman, Nagler and Watzinger, 2018; Baruffaldi and Simeth, 2020; Hegde, Herkenhoff and Zhu, 2020; Lück et al., 2019; de Rassenfosse, Pellegrino and Raiteri, 2020). However, there is very little research and evidence on how inventors can access and search technical information in patent documents, and whether reducing the access cost triggers knowledge diffusion and innovation. One noticeable exception is Furman, Nagler and Watzinger (2018), who document a significant increase in patenting in the 48 regions who opened a new U.S. patent library, i.e. a central repository of physical patent documents, between 1975 and 1997.

In this paper, we contribute to this line of research by studying how the arrival of ADSL broadband internet to the United Kingdom in the early 2000s facilitated access to distant technical knowledge and led to a corresponding increase in innovation. Information technologies, like the printing press, the telephone, or computer networks, provide efficient means of storing and disseminating information (e.g. Dittmar, 2011). ADSL broadband, the first generation of high-speed and 'always-on' internet, entailed an exponential improvement in download and upload rates compared to basic cable internet. Broadband allowed inventors to efficiently and timely access codified technical information from the world wide web for the first time, in particular information stored in full-text online patent databases of both major patent offices (USPTO and EPO) and commercial providers. This made it much easier and less costly to access and search technical knowledge, particularly knowledge stemming from geographically distant regions and outside inventors'

own social network and firm.

Linking proprietary, high-resolution data on the UK telecommunications market to public patent records, we evaluate how the switch from dial-up to broadband internet impacted knowledge sourcing and inventive activity in Great Britain. To do so, we exploit geographical discontinuities in activation status among 2,029 local exchange (LE) areas due to gradual delays in market entry of high-speed internet service providers (ISP) between 2000 and 2007. The roll-out of broadband happened through the upgrading of existing telephony nodes, called local exchanges, which had to be upgraded individually. Admittedly, the choice of market entry by ISPs in a LE is likely driven by firm establishments and local inventive workforce, even within narrowly defined areas. Therefore, we supplement our regional fixed effects estimations with an instrumental variables approach (IV) based on within region variation of weather conditions and areal demand mismatch between firm and local exchange positions that exogenously shift the likelihood of ISP entry, but are unrelated to firm/inventor establishment.

The launch of broadband in the UK provides a particularly well-suited research setting. This is because the roll-out of the infrastructure was not instantaneous, due to demand-side and supply-side factors, thus generating variation in internet availability between areas, even within small and homogeneous regions. The main causes of these initial delays in broadband expansion were due on the one hand to the uncertainty faced by the ISPs about the demand of internet services from private households, and on the other hand to the technical difficulties that the nation-wide roll-out of a new series of technologies naturally contemplated. In our empirical analysis we exploit these factors to develop a set of exogenous shifters that we use in an instrumental variable setting, to study the effect of internet access on innovation. The first instrument hinges on the negative impact of adverse weather conditions on broadband take up and on service quality (Gavazza, Nardotto and Valletti, 2019). Weather conditions have been shown to have an important weight in ISPs entry decision, as broadband equipment were (and still are) particularly sensitive to adverse weather, which strongly impacted maintenance costs and service quality. Thus, this class of instruments uses variation in weather conditions, namely rainfall, between local exchange areas to predict broadband activation in the following year. The identifying assumption behind our approach is that differences in weather

conditions between LE areas *within NUTS3 regions* are orthogonal to the inventive firms' and inventors' characteristics established in these areas.[2] The second class of instruments exploits potential mismatches between demographic - hence demand relevant - characteristics of local exchange catchment areas and (partially) overlapping Census 2001 Middle Layer Super Output Areas (MSOA).[3] The intuition behind this approach is that firms and inventors base their establishment decisions *within NUTS3 regions* on current population characteristics of MSOAs, while they are unaware of the extension of the (historical) catchment areas of the nodes of the telecommunications network, which are relevant to the internet service providers.

As main result, we find that the activation of broadband increased yearly patenting in a LE area between 10-55%, and forward citation-weighted patenting by %, depending on the specification employed. The latter finding suggests broadband particularly resulted in higher quality and more impactful inventions. We relate these extensive margin effects to the reduced costs of accessing and searching technical knowledge from abroad. In particular, we show that broadband access increased the likelihood to cite foreign technical prior art by 4-14%. Moreover, broadband enabled inventors to search and build on more recent foreign technical prior art, suggesting it allows inventors to keep up to date on the knowledge frontier and state of the art in given field, particularly if new developments originate in remote areas. Apart from sourcing (recent) foreign knowledge, we find no significant effect of broadband on sourcing distant knowledge from outside inventors' own social network or firm, or from different technology fields. Furthermore, besides affecting knowledge sourcing, high-seed internet might also influence technological progress and innovation by changing the organization of R&D departments and teams, e.g. through a reduction of communication costs and resource-sharing across multiple branches. However, we find no significant increase in team size, geographically dispersed teams, cross-disciplinary teams, or new collaborations among inventors. These findings strengthen our belief that a better and timely access to foreign technical knowledge triggered the increase in innovation. This finding is

[2] NUTS3 regions correspond to territorial subdivisions of between 150,000 and 800,000 inhabitants. In total there are 174 NUTS-3 regions in the U.K, 167 of which generate patents in our data.

[3] An MSOA is a statistical area covering a population of between 5,000 7,200 inhabitants. There are 2,029 LEs and 3,374 patent generating MSOAs in our data.

closely in line with (Iaria, Schwarz and Waldinger, 2018), who show that World War I caused a significant decline in international scientific references in papers, and consequently reduced scientific progress and output.

To further strengthen the causal claim that early broadband increased patenting through a reduction of access costs to (recent) foreign knowledge, we conduct a number of additional tests. First, broadband could also affect innovation in LE areas through the opportunity it creates, for instance by attracting new companies and inventors.[4] Reassuringly, we find that our results hold when we restrict our sample to firms established in a region before the arrival of broadband, ruling out that our effects are driven by the location choices of firms. Moreover, the introduction of broadband did not attract a significant number of new inventors to a LE area.

Second, we show that broadband access in LEs did not significantly affect public sector patenting (i.e. by universities, government research labs or hospitals). Indeed, prior literature highlights that these institutions were pre-embedded into different types of information networks (e.g. BitNet, ARPAnet, physical libraries), making them arguably exposed to a much lesser shock when broadband arrived.[5] Third, we find no effect of broadband on the characteristics of examiner-given citations. Indeed USPTO patent examiners work from the patent office in Virginia in the US and should not be affected by the arrival of broadband in the UK. Fourth, we observe no change in the overall number of prior art references per patent, allowing us to exclude that the effect of broadband was merely one of patent attorneys or inventors becoming more aware of which patents they would need to cite for legal reasons, without the actual manifestation of knowledge spillovers. Finally, we do not see an increase in references to scientific publication in patents. While patent documents became publicly available online in 1999, accessing scientific publications still required expensive subscriptions to publishers or library access (e.g. through academic affiliations).

However, a series of historical placebo tests reveal significant pre-trends and higher historical patenting rates of LE areas within our chosen exclusion restrictions. This warrants caution for the causal interpretability of results from our (current)

[4] see, e.g. Cheyre, Klepper and Veloso (2015); Akcigit, Baslandze and Stantcheva (2016) regarding determinants of inventor mobility.

[5] for a discussion on this, see e.g. MacKie-Mason and Varian (1996); Spear (1996)

IV-estimation. We discuss motives and consequences of this limitation and discuss a series of potential remedies and alternative approaches devised in the attempt to re-establish causal identification.

Our paper has antecedents in the empirical literature regarding the importance of access to existing knowledge for subsequent scientific and technological progress (Moser and Voena, 2012; Williams, 2013; Iaria, Schwarz and Waldinger, 2018). In particular, we contribute to the debate about how access costs to prior information influence research productivity and shape trajectories of scientific and technological inquiry (Furman and Stern (2011); Bryan and Ozcan (2016); Biasi and Moser (2018)). Our findings suggest that the elasticity of the cost of access on technological progress and innovation is significant and large, even in the 21st century, and that it has real economic effects. Moreover, we provide evidence in support of the argument that access costs constrain the diffusion of technological knowledge embodied in patent documents. In this regard, our results also closely relate to recent empirical work showing how the disclosure of patent documents fosters knowledge spillovers and cumulative innovation (Graham and Hegde, 2015; Hegde and Luo, 2018; Furman, Nagler and Watzinger, 2018; Baruffaldi and Simeth, 2020; Hegde, Herkenhoff and Zhu, 2020; Lück et al., 2019; Zheng and Wang, 2020; de Rassenfosse, Pellegrino and Raiteri, 2020). While our data does not allow us to make inferences about whether inventors acquired technical information from patent documents through online patent databases or other channels, we causally show that broadband access enabled inventors to become aware of and built on technical prior art they would have otherwise not.

Our findings further contribute to the literature on the effects of ICT adoption and internet on firm performance (Forman and van Zeebroeck, 2012; Bertschek, Cerquera and Klein, 2013; Dhyne et al., 2018; Forman and van Zeebroeck, 2019). These studies show internet adoption impacts intra-firm knowledge diffusion, R&D collaboration and productivity. Other related work has documented the positive effects of BitNet, a predecessor of cable internet, on academic research collaborations and productivity (Agrawal and Goldfarb, 2008; Ding et al., 2010). Finally, given the comprehensive, country-wide evidence of our study, our results also speak to the wider literature on the impacts of information technologies on aggregate growth, employment and productivity (e.g., Dewan and Kraemer, 2000; Black and Lynch,

2001; Stiroh, 2002; Dittmar, 2011), and of broadband infrastructure in particular (e.g., Czernich et al., 2011; Hjort and Poulsen, 2019).

3.2 The UK Broadband Roll-Out and the Search for Technical Prior Art

In this section, we outline the institutional context of our investigation by discussing the evolution of the UK broadband market, followed by a depiction of how inventors can search technical information from patent documents, before and after the arrival of broadband internet.

3.2.1 The UK broadband market

At the beginning of 2020s, thirty years after the advent of the internet, the market for internet services relies on a mix of technologies and networks – xDSL, Cable, Fiber to the Home, or Mobile internet – and it hosts a good number of Telcom companies competing for internet users. At its infancy in the 1990s however, internet services were provided only through one technology and one network: the copper lines of the telephone network. In the UK, this network was owned and managed by British Telcom (BT), the formerly state-owned telecommunication company which in the 1930s set up the network of nodes (called Local Exchanges, LEs) and links that connect each household to the rest of the country and, through undersea connections, to the rest of the world.

The telephone line proved to be a viable option for the transmission of a digital signal, thanks to an upgrade to the old-fashioned telephone connection. This network, although it was not meant for that purpose, had the advantage of being already deployed and covering virtually the whole UK (due to the universal coverage obligation). Until the end of the 1990s tough, i.e. until the moment when the first family of Digital Subscriber Line (DSL) technologies were introduced, internet users could only connect to the internet trough dial-up (modem) connections, which at most could reach a (download) speed of 64 kbit/s. The introduction of DSL marked the start of a new era due to its large improvements in data transmission speed. With the first generation of DSL technologies introduced in the early 2000s

in the UK, the speed of internet connection increased up to the (theoretical) speed of 8 Mbit/s, i.e. by a factor of 125.[6]

This fundamental technological step happened right after the release of online patent databases by the major patent offices and commercial providers, which enabled inventors to efficiently and timely search and download the full text of virtually the entire patent record of the major patent offices (see further, section 3.2.2). However, internet access at sufficient speed however strongly depended on two conditions that had to be met at the level of the LE to which the inventor was connected: i) The LE being upgraded by BT to host ADSL connections; ii) The LE receiving the investments of Internet Services Providers (ISPs) to offer such high-speed ADSL connections. In order to understand the relevance of these two conditions, it is worth reconstructing how the broadband market evolved in the UK during the 1990s and the first decade of the 2000s. When the internet era started, the privatization process of BT was almost completed.[7] Before its privatization, BT was vertically integrated, meaning that it was both managing the telephone network, and selling telephone services to the final users. This was a source of concern for policy makers once the social and economic potential of the internet became clear. The reason was simply that BT's market power in the telephone services market could be transferred to the new market of internet services. This led to various actions to vertically separate the two levels of the market and to facilitate entry of ISPs in the downstream market of internet services.[8] BT was separated into two companies, Openreach (in charge of managing the network) and BT Retail (investing in the LEs and selling internet access to final users), and a set of *open access* policies were introduced to boost entry of other ISPs in the downstream market. The market for broadband services could then profit from an investment-friendly

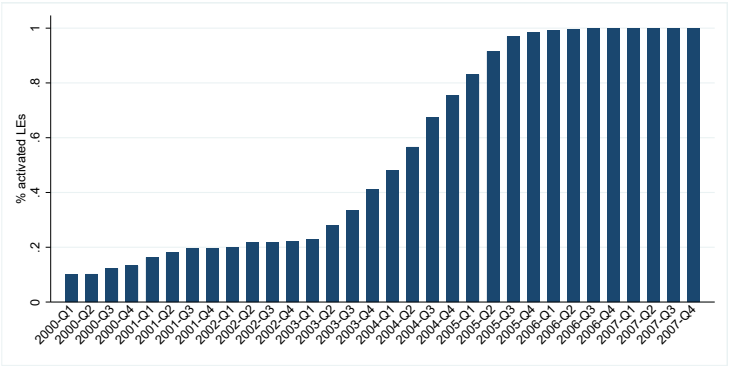
[6] As documented in Nardotto, Valletti and Verboven (2015), and Ahlfeldt, Koutroumpis and Valletti (2017), and several other studies, actual speed was much lower than the theoretical speed that could be reached by the technology due to factors, such as the physical distance to the node, the weather, and other conditions, which determine a deterioration of the quality of the signal and thus a drastic decrease in the actual speed of the connection.

[7] The company started its process of privatization in the mid-1980s, and concluded it in the mid-1990s.

[8] The actions taken in the UK were in line with those taken in the rest of EU countries, which had to confront a very similar market structure characterized by the presence of old incumbents which were formerly state-owned companies. A relevant role in shaping the institutional framework was played by the European Commission through its EC Directives.

regulatory framework where BT, and subsequently Openreach, was mandated to quickly enable the LEs to host ADSL investments. BT charged fees in exchange for access to the network and ISPs could freely enter the downstream market by investing in the necessary technology at the level of a LE in order to serve internet users, both private and corporate.

Figure 3.1. Share of LEs with Broadband Access by Quarter



The combination of investment-friendly regulation and growing demand for internet access determined the rapid diffusion of the internet. Broadband penetration, measured as the share of households with a broadband connection, rose from less than 10% in 2000, to more than 80% in 2010. However, the market did not grow at a constant rate, with modest growth rates until 2005, to effectively take-off in the second part of the 2000s when private internet users massively wanted to be online.[9] In the end of the 1990s and the first part of the 2000s the process of upstream LE activation by BT was much ahead of the actual downstream investment rate of ISPs because of the high installation costs[10] and a relatively weak demand for internet services.

This slow take up of investments by ISPs is reflected in Figure 3.1 which reports

[9] Private users in the second part of the 2000s could finally enjoy a combination of newer/better ADSL technologies, and the arrival of new content such as Youtube (founded in 2005), Facebook (introduced in UK in 2007) and other successful platforms.

[10] According to Nardotto, Valletti and Verboven (2015), who examine the period between 2005 and 2009 the cost to enable a LEs halved, thanks to technological improvements and learning. This figure is probably much larger if one considers the costs in the early 2000s.

the share of LEs which received ADSL investments of at least one ISP between 2000 and 2007. The roll-out of the technology started at a slow pace and between the end of the 1990s and mid 2003 only 20% of LEs provided ADSL broadband connections to private users and firms. In 2004 the diffusion finally took-off, and in the following 3 years the market was fully covered. This slow diffusion of internet access due to ISPs' investment decisions is at the core of our identification strategy, which exploits the timing of the introduction of broadband services between different LE areas (see section 3.4.1).

3.2.2 The search for technical prior art online

Existing knowledge and prior art are a fundamental input for the creation of new technologies. Industrial R&D in particular relies on external knowledge spillovers, i.e. knowledge stemming from outside the firm and team of inventors (e.g., Jaffe, 1986; Bloom, Schankerman and Van Reenen, 2013). During the inventive search process, inventors can rely on codified knowledge from predominantly scientific publications and patent documents, or tacit knowledge stemming from personal interactions and social networks. Prior empirical work shows that interpersonal networks of inventors indeed foster the diffusion of knowledge, and that these networks explain the frequently observed localization of knowledge flows within firms and geographic regions (Jaffe, Trajtenberg and Henderson, 1993; Verspagen and Schoenmakers, 2004; Singh, 2005; Singh and Marx, 2013). Despite the importance of external knowledge for industrial R&D, there is less evidence on the channels through which inventors access and search codified knowledge outside their own social network, firm, and geographic region. Large scale inventor surveys in Europe and the U.S. point out that patent documents are the most important source of codified knowledge next to scientific publications and conference proceedings (Giuri et al., 2007; Cohen, Nelson and Walsh, 2000; Jaffe, Trajtenberg and Fogarty, 2000).[11] More recent studies confirm that the public disclosure of patent documents triggers external knowledge spillovers, technological progress, and cumulative in-

[11] In particular legal scholars have often questioned the usefulness of patent documents as sources for information for inventors (e.g. Devlin, 2009; Ouelette, 2012; Lemley, 2012). Main concerns raised in this regard relate to the little incentives inventors have to disclosure actual valuable information in their own patents and to consult other patents, due to the risk of committing "willful infringement".

novation (Graham and Hegde, 2015; Hegde and Luo, 2018; Furman, Nagler and Watzinger, 2018; Baruffaldi and Simeth, 2020; Hegde, Herkenhoff and Zhu, 2020; Lück et al., 2019; de Rassenfosse, Pellegrino and Raiteri, 2020)

Despite the importance of open and early access to patent documents, corporate inventors initially faced considerable hurdles and costs to access and search technical information from patents. Before the internet, the default was to keep track of the state of the art and new developments in a field by screening periodical bulletins containing titles and abstracts of newly granted patents, and subsequently ordering individual copies of selected patent documents from the central patent office via mail or fax. This search process was inefficient, slow, and costly. For instance, for UK-based inventors, the fee for regular mail delivery for a single copy of one patent document from the USPTO ranged between $5-19 current USD in 1997, with a delivery time of 4-6 weeks.[12] Fax delivery was much faster but also more expensive with fees ranging between $41-60 current USD for a single patent copy international delivery. [13] Besides ordering individual copies by mail and fax, people could also consult physical patent document repositories in patent libraries or information centres of the UK patent office located in large metropolitan areas across the country.[14] Notice that there were fewer than 10 of such patent libraries across the UK in 1995.[15] Finally, a number of university libraries in the UK were also connected to the CASSIS system through the ARPAnet, enabling access to the electronic repository of patent documents from the USPTO MacKie-Mason and Varian (1996); Furman, Nagler and Watzinger (2018).

Dial-up internet made many things easier for inventors in the mid-1990s.. Commercial service providers, such as Questel-Orbit and Derwent, were the first to realize the potential of the world wide web to offer online patent information and sell electronic full text copies of patents at comparatively lower prices.[16] The ma-

[12] We determine inflation adjusted prices in current $USD based on Consumer Price Index (CPI) data from the U.S. Labor Department's Bureau of Labor Statistics (cumulative inflation rate through February 2020).

[13] Source: Historical website of the USPTO, accessed Feb 12 1997.

[14] The PatLib program was an initiative by the EPO and national patent offices targeted at providing access to patents and trademarks to individual inventors and businesses.

[15] Source: WIPO 2015.

[16] For an extensive discussion of patent information available online to UK inventors in 1996, see Spear (1996).

jor patent offices started their free online databases with some delay. In 1996, the USPTO was the first patent office to offer a comprehensive online catalogue with bibliographic front cover information and abstracts for over 2 million patents.[17] This was followed by an update in late 1998, making the full text and images of all US patents available online. In 1999, the EPO launched its online platform *Espacenet*, which included full text information for all patents worldwide.[18]

Despite the online availability of patent information from commercial providers and patent offices, the critical problem with searching and retrieving patent documents was the limited bandwidth and slow download rates of 56k-modems (see section 3.2.1). Searching for, selecting, and opening a full text patent document from a dial-up connection meant, in the best case, several minutes of waiting time until all pages and drawings were downloaded. In many instances, though, web pages were simply not reachable, as the early infrastructure of the net was not capable of handling the torrent of user requests during peak times (Spear, 1996). Taken together, accessing and searching patent information online remained a tedious, costly, and at best complementary tool in the late-1990s.

Broadband internet dramatically decreased the costs of accessing information online and improved the speed and easy of use. Patent documents, technical reports, and scientific publications could now be browsed and downloaded within seconds. Moreover, DSL-technologies enabled parallel use of telephone, fax and internet for the first time. This 'always-on' feature of broadband enabled internet search to be fully integrated as a simultaneous functionality into many research and development activities.

3.3 The Data

In this section we present the data sources that we employ in the empirical analyses: i) data on the broadband market; ii) patent data; and iii) socio-demographic data.

[17] This was preceded by a pilot project offering full text and image access to a limited number of HIV/AIDS related patents in 1994 on the USPTO website (Source: USPTO).

[18] Sources: Press releases/ historical archives of the USPTO and EPO websites.

3.3.1 Broadband market data.

Data on the evolution of the broadband market, which we discussed in Section 3.2.1, have been provided by Ofcom, the UK regulation and competition authority for the broadcasting, telecommunications and postal industries. The data includes the year and month of arrival of ADSL for each of the approximately 5,500 LEs in the UK, covering the countries of England, Wales and Scotland. The data also contains the exact geographical catchment area covered by each LE. We use this information to determine for each inventor the LE she/he is connected to and the start date at which she/he could access the internet at broadband speed. Notice that despite the high granularity of our data on micro-market entry of broadband service providers in a LE, we do not observe actual broadband adoption at the inventor level.

3.3.2 Data on patenting activity

We rely on public patent records to quantify the yearly inventive output in a LE and to characterize the knowledge and prior art on which the inventors relied during the inventive search process.

We start by collecting information on all patents granted by the USPTO since 1976 from *PATSTAT 2018* (spring edition). We rely on data from the USPTO because US patents provide the most comprehensive coverage of all inventive activity worldwide. The majority of all patented inventions (approximately 60%) has a USPTO patent family member. Moreover, UK firms predominantly file patents at the USPTO rather than at the UK or European Patent Office Griffith, Harrison and Van Reenen (2006). Therefore, US patents provide a better measure of the inventive output of UK based firms. Finally, the USPTO provides the best and most reliable coverage of information on citing and cited patents' locations, that is indispensable for our analyses.

First, we collect information on the geographic location of assignees and inventors for all patents granted by the USPTO.[19] From the bulk of US patents, we restrict the sample to corporate patents having the assignee and at least one

[19] Note that, while a large body of work exists in the literature using geo-locations of US inventors, only less than 5% of all USPTO patents actually contain fine-grained address information of inventors (and applicants).

inventor located in the United Kingdom. [20] [21] Because the USPTO does not provide fine-grained address information for all inventors and assignees, we replenish coarse geo-locations with high-resolution address strings sourced from international patent family members, i.e. patents issued by other national and regional patent offices covering the same invention as the US patent Morrison, Riccaboni and Pammolli (2017); de Rassenfosse, Kozak and Seliger (2019). First, we rely on European Patent Office (EPO) family members, which have the highest-precision and most comprehensive address string coverage. % of all US patents in our sample have an EP family member. Second, for the remaining US patents without an EP family member (%), we obtained additional information on applicant and inventor addresses for UK patent family members directly from the United Kingdom Intellectual Property Office (IPO).[22] Next, for the few patents in our sample without an EP or IPO family member (% of our sample), we rely on address information from patent family members at different patent offices, obtained from de Rassenfosse, Kozak and Seliger (2019)[23] We drop % of the patents in our sample for which we have no fine-grained location information. Finally, we geo-disambiguate all raw address strings in our data using the *HERE Developer* API, a commercial mapping and location data service.[24] Figure 3.2 provides an overview of the dispersion of inventive activity across the UK between 1975 and 2010. Historically, patenting appears to be clustered geographically, and concentrated in the south and south-east of Great Britain around the metropolitan and industrial centers of London, Liverpool, Manchester and Birmingham. Clusters of inventive activity are also visible in the north of England, between Newcastle and Middlesborough, as well as around Glasgow and Edinburgh in Scotland.

In a next step, we are interested in determining whether or not the inventors of a patent had access to ADSL broadband internet during the inventive search process.

[20] We exclude public sector patents, i.e. patents filed by universities, public hospitals, government or military institutions, as these are typically embedded in different types of knowledge networks (e.g.BitNet) and inventive search is conducted under different boundary conditions (public funding, non-profit orientation,...)

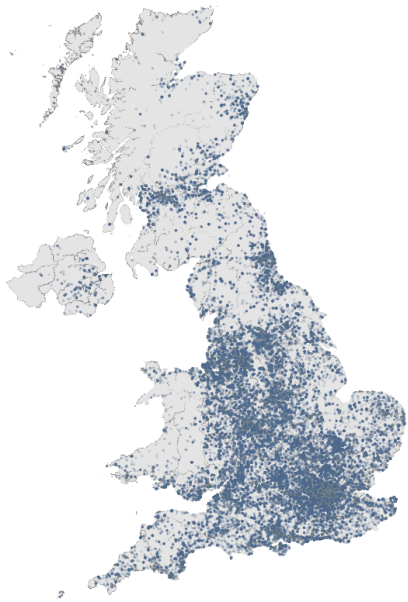
[21] We determine assignee types based on the PSN sector allocation in PATSTAT.

[22] We thank the IPO for providing us timely with the requested data sets and very useful further explanations.

[23] We thank Florian Seliger for providing us with the input data.

[24] API interface available at: https://developer.here.com/

Figure 3.2. Geolocation of Inventors across Great Britain 1975-2010



To determine broadband activation status, we use assignee address information because 1) inventor addresses usually correspond to the private home rather than the workplace location (Pavitt and Patel, 1998; Griliches, 1990; Verspagen and Schoenmakers, 2004), because the workplace address most likely reflects the location of inventive activity, and because companies were typically faster to adopt broadband compared to private households. However, prior literature has warranted caution in using addresses of firms listed as assignee on a patent, as these might reflect the address of firm headquarters or the address of the legal entity that owns a specific patent, e.g. the parent firm holding (Jaffe, Trajtenberg and Henderson, 1993; Singh, 2008; Singh and Marx, 2013). To increase the likelihood that the assignee address corresponds to the actual workplace of the inventors, we restrict our sample to patents for which at least 75% of the inventor addresses are within a 75km radius around the assignee address, which we consider as an upper bound for plausible commuting distances. More than 90% of all patents in our sample fall within this definition, suggesting that the address of UK firms listed on US patents are likely to

correspond to the workplace of the inventors. Since we only dispose of broadband activation information for UK addresses, to further avoid measurement error, we restrict our sample (in our main specification) to only those patents having all inventors located in the UK, i.e. excluding international collaborations. Notice that our results are robust for the subset of patents with 100% of inventors within the same 75km radius around the assignee address, as well as for 50km and 25km radii.[25] In case a patent is assigned to two assignee firms (less than 10% of patents), we consider only those assignee addresses for which the 75/75 definition is valid.

3.3.3 Census data

We collect additional information on the socio-demographic characteristics of different subsets of the population from the UK Census. The UK national census takes place during the first year of each decade. Given that our analysis spans the period between 2000 and 2007, we consider the Census of 2001. We collect both standard socio-demographic information – such as age structure, education, and unemployment rate – which will be used as controls, and variables – such as population density – which will be used to construct instrumental variables, as we discuss in Section 3.4.1. To reconstruct demographic characteristics at the level of the LE, we downloaded information at the Output Area (OA) level, which is the smallest geographical unit at which the UK Census is available.[26]

3.3.4 Descriptive statistics

Table 3.1 provides descriptive statics for our final sample of 6,961 patents, which represents about half of all corporate USPTO patents with all assignee and inventors located in the UK during the same time window. To evaluate how biting the con-

[25] From our sample of patents for which 75% of the inventor addresses are located within 75km from the assignee address, 95% also fulfill the stricter 50km, and 83% the 25km criterion.

[26] There are approximately 216,000 OAs in England, Wales and Scotland in the 2001 Census. The minimum population size for an OA was 40 resident households, but the vast majority, accounting for 79.6% of the total, lies between 110 and 139 households. LEs instead are 5,500 thus a LE contains on average almost 40 OAs. As the boarders of the two geographies do not perfectly match, we input the demographics of the areas that overlap with a boarder pro-quota based on the proportion of the overlapping area.

Table 3.1. Patent level summary statistics

	in-sample				not in-sample			
	p50	Mean	SD	Max	p50	Mean	SD	Max
Citations [10y]	11	18.24	24.03	281	12	19.86	30.53	502
Patent References	15	22.01	21.43	330	13	18.69	22.62	589
Scientific References	0	3.12	13.58	363	0	2.68	12.19	363
Inventors	2	2.06	1.43	16	2	2.14	1.52	17
Co-Patent [01]	0	0.02	0.15	1	0	0.04	0.20	1
Assignee 5y Patents	5	76.46	362.74	5.1k	8	141.31	453.83	5.6k
Assignee 5y Citations	36	721.68	2.9k	65k	80	1.6k	4.7k	102k
N of Patents	6,961				6,703			

Notes: In-sample patents are patents with high-resolution address information allowing to map them accurately to Local Exchange catchment areas. Patent data are sourced from PATSTAT, USPTO, and BvD Orbis (assignees). Additional applicants and inventors address information was obtained from the U.K. IPO and de Rassenfosse, Kozak and Seliger (2019).

straint of having high-resolution address information is for how generalizable our results are for the underlying population, we compare characteristics of in-sample to not in-sample patents in Table 3.1. The distributions across most patenting indicators are heavily skewed. The average patent in our sample lists 2 inventors, includes 22.01 prior art references to patents and 3.12 references to scientific publications. Over a period of 10 years, patents receive on average 18.24 citations by follow-up inventions (patent families). About 2% of patents are assigned to multiple corporate or institutional assignees.

The comparison with patents not in our sample reveals significant differences with regards to the distribution of technological impact (10-year forward citations), the number of patent references and, in particular, the number of 5-year patenting activity by the same corporate assignees. The median firm in our sample files 5 patents in 5 years, compared to 8 patents filed by not in-sample firms, for mean values this differences is about twice as large. We interpret this as a sign that large multinational firms are possibly underrepresented in our data, as they are more likely to not file patents directly through the national UK IPO, and less likely to assign patents to inventor workplace addresses. The remaining indicators show roughly comparable scores.

Table 3.2 provides yearly summary statistics of inventive activity at the LE

area level. We register patents originating from 2,029 LE areas during our sample window. As seen in Figure 3.2, patenting is not evenly distributed across nodes. While the most active LE generates 114 patents in one year, the average number of filings per LE is merely 0.55. Around 62% of patents are assigned to companies established in the region before the arrival of broadband. The amount of citation weighted patents per LE for pre- versus newly established applicants is proportional to this division. About half of the LE have no patents in the 5 years before the arrival of ADSL.

Table 3.2. Node (LE) level yearly summary statistics

	Mean	SD	Max
Patents	0.55	2.27	114
Citation weighted Patents [10y]	9.86	68.44	4.9k
Patents (pre-BB Firms)	0.34	1.91	114
Citation weighted Patents (pre-BB Firms)	5.99	47.55	2.5k
Public Patentee [01]	0.10	0.30	1
Pre-BB Patenting Node [01]	0.49	0.50	1
N of Nodes	2,029		

Notes: Nodes (LEs) are Local Exchange Points geographical catchment areas (N = 5,500) across the U.K.. Data on LEs and ADSL activation are sourced from Ofcom.

When we descriptively evaluate note-level characteristics against the timing of broadband arrival, it becomes well clear that the ADSL-rollout cannot be considered as a random process: Local exchanges activated in the very first cohort, before the end of the year 2000, showed well above-average levels of inventive activity, generating about 1.17 patents per year receiving on average 13 patent follow-up citations. LEs activated only a year after, in 2001, had significantly lower inventive output, with 0.37 yearly patent filings and less than 4 forward citations received. With an average total population of 38,435 and density of 35 inhabitants per sq. kilometer, these early activated nodes were particularly crowded, while 2001 follow-up nodes had about 20% less in population size and considerably less population density (16/sq. km). In terms of age structure of population, the first cohort of LEs had a particularly high share of young adults under 30 (+28% more compared to the 2001-cohort) and relatively fewer inhabitants above the age of 45. These differences were magnified when comparing early ADSL nodes to those only activated in later

years, i.e. post-2001.

3.4 Empirical Analysis & Results

In this section we present our empirical analysis. We start by the identification strategy, and we discuss the challenges posed by the potential sources of endogeneity that are related to the roll-out of the ADSL infrastructure. We then show our findings and we present evidence that earlier internet access led to more patents, which also receive more citations. The main mechanisms appear to be more international and faster knowledge sourcing.

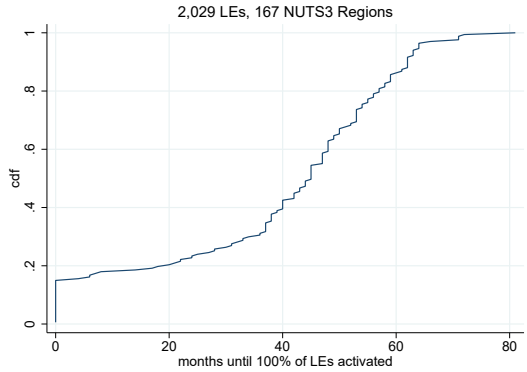
3.4.1 Identification strategy

The goal of this paper is to document the effect of the internet on patenting. To do so, our identification strategy exploits the variation in timing of introduction of broadband internet in local areas. As discussed in Section 3.2.1, the roll-out of ADSL nationwide was not instantaneous, and both demand and cost factors affected the decisions of ISPs to invest in the LEs. Importantly, this process of progressive roll-out was also happening at the local level. Figure 3.3 reports the average time (in months) needed to reach different levels of ADSL coverage within the NUTS 3 regions, i.e. a similar exercise to the one in Figure 3.1 in Section 3.2.1, but performed at the local level.

As Figure 3.3 shows, the roll-out was not only a slow process from a nationwide perspective, but also within regions, thus generating the variation in timing of introduction that we exploit to estimate the effect of internet on patenting activity.

However, we cannot consider this roll-out process fully exogenous. In particular, the core endogeneity concern is that certain areas are more attractive to both ISPs and innovative firms. For example, we would expect both more patenting activity and earlier internet activity in the urbanized web around London than in the scenic vistas of Cornwall. We address this concern with two sets of instruments, which are based on two different mechanisms driving ISPs' investments decision, in a way that they are not correlated with firms' R&D or location decisions conditional on our controls. Before discussing the instruments, it is important to notice that

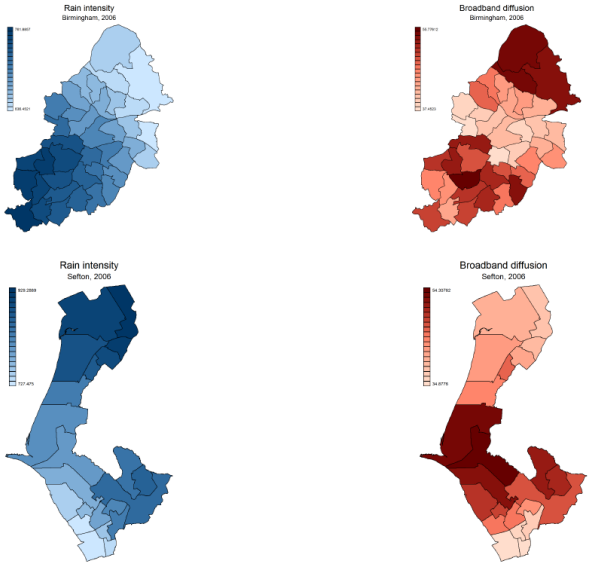
Figure 3.3. Cumulative of *within NUTS3* ADSL Activation Lag across Great Britain



the decision of an ISP to invest in a specific LE was not driven by the demand coming from the firms we are considering in this study. Instead, the main source of revenues, and thus the key determinant of an ISP's decision are the expected private internet connections, as they outnumbered by far the rest of the market. Thus, factors that affected the economic trade-off of serving the local internet markets can be exploited once we control for the characteristics that are attractive for firms and inventors in general.

The first set of instruments follows the identification strategy in Gavazza, Nardotto and Valletti (2019). Thus, we exploit weather conditions as a source of variation for the timing of investments in the LEs. In particular, we use the amount of rainfall in a LE, which affects both the cost of upgrading a LE for the ISPs and local demand for internet services, as more rain reduces service quality, as argued in great detail by Gavazza, Nardotto and Valletti (2019). They note that there is an 0.8 correlation between rainfall and network fault report volumes as "these conditions led to more underground structures flooding and more faults due to water ingress into failed joints and cables" (Ofcom, 2014). Moreover, adverse weather conditions lead to reduced reliability and speed of the internet service, in turn reducing local demand (Bouckaert, van Dijk and Verboven, 2010; Nardotto, Valletti and Verboven, 2015; Gavazza, Nardotto and Valletti, 2019).

Figure 3.4. Areal Variation in Rainfall and ADSL Activation



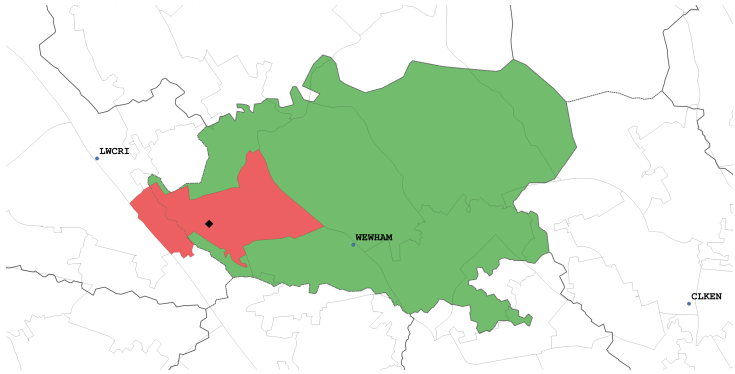
Notes: Figure from Gavazza, Nardotto and Valletti (2019).

We include NUTS3 fixed effects, therefore the exclusion restriction holds unless innovative firms are sensitive to rain conditions within NUTS3 regions. That would require two strong assumptions. First, that firms are aware of differences in rainfall within relative small regions and second that this is a meaningful determinant of their location choices. The first is implausible, as rainfall data have only recently become publicly available and are not trivial in their use. Moreover, the difference in rainfall within a NUTS3 region is sufficiently large to matter for the highly sensitive electrical equipment of ISPs, but not so large that you would notice it through anecdotal experience. Figure 3.4 shows variation in rainfall for urban Birmingham and rural Sefton. The difference between the driest and wettest point is respectively 20% and 27%, large enough to affect the vulnerability of the ISP equipment, but not enough to be noticeable without access to detailed rain data.

Second, firms being aware of differences in rainfall would only affect our estimates if they base their location decisions on this information. Yet, there do not

appear to be any reasons to assume mild differences in rainfall would matter to businesses. Instead, a survey by Keeble and Nachum (2001) suggests that firms mainly care about proximity to clients and talent and does not even mention weather conditions. As a placebo test, we examine whether rainfall predicts public sector patenting. Public institutions largely have their own internet networks and are thus not affected by ISP decisions, but if firms care about local rain conditions, it stands to reason that public institutions do too. However, as we will show in Table 3.6 we find a precisely estimated null effect, providing suggestive evidence that variations in rainfall within NUTS3 regions truly are exogenous.

Figure 3.5. Mismatch between MSOA and LE Catchment Areas



Notes: Example of LE catchment area and UK's MSOA. The dark solid lines are the LEs' boarders while light solid lines are the boarders of the MSOAs. LEs' locations are identified by the blue dots (in the figure we have the LEs named WEWHAM, LWCRI, and CLKEN). In this example, a firm located in the black diamond, is part of the red MSOA and it is served by the LE WEWHAM, whose catchment area is the one in green.

The second set of instruments exploits the as-good-as-randomly defined catchment areas of the local exchanges which are determined by the British Telecom's telephone network as it was constructed in 1930 and are not used for any other purpose, and have not been changed since (Gavazza, Nardotto and Valletti, 2019). Identification is then driven exclusively by the mismatch between the area of relevance for ISPs and for firms. We use Figure 3.5 to illustrate this identification approach.

In the figure, blue dots indicate the position LEs, while their catchment area is delimited by the dashed thick line. The solid gray lines are the boarders of the

MSOAs. An MSOA comprises on average 6,500 families, which means, on average 15,000 inhabitants. MSOAs are census areas and are wide enough to capture well the characteristics of the section of the city or of the town where the firm and the inventor is operating. In the figure we have a fictitious firm located at the diamond. In our data we have exact firm's location so, for each firm, we know exactly the relevant MSOA, in red in our example, and the relevant LE, which we depict in green in our example. The core of the identification strategy relies on the mismatches between the characteristics of the MSOA, relevant for the firm, and those of the LE's catchment area, which are relevant for the ISPs. So, while controlling for both the characteristics of the MSOA and of the wider NUTS3 region, we use the demographics of the remaining part of the LE as instrument for the arrival of broadband investments. This instrument resembles the one proposed by Hausman (1996) and later applied by Gentzkow and Shapiro (2010), or Fan (2013) which hinge on the fact that the demographic characteristics of the neighboring areas of a given area i provide incentives for the supply-side to take actions (pricing in their case, investing in ours) which can affect also the area i. Clearly, this supply-side connection is present in our case, as the investment decisions taken by ISPs are based on the characteristics of the (entire) catchment area of the LE, but then they affect every building which is located on top of it, so all MSOAs that are covered by the LE.

One limitation of this method is that while we know exactly which area matters for the ISP, we do not have that information for firms. This is less of an issue than one might expect as the NUTS3 fixed effects control for characteristics of the broader region. Therefore, any endogeneity would have to arise from innovative firms basing their location decisions on some other subset of the NUTS3 region that overlaps more strongly with the (random) local exchange catchment area than with the carefully constructed MSOAs. Moreover, there would need to be a link between the number of innovative firms and the chosen demographics which is not a trivial assumption.

The instruments we use are population density, a set of age share variables and the total number of people, each measured within the catchment area.[27] We include the same variables as controls, but this time measured at the MSOA level. We also

[27] Ages are split 0-14-19-29-44-59, with 60+ as the reference group.

include NUTS3 fixed effects (MSOAs are smaller than NUTS3 regions).

3.4.2 Inventive activity by LEs

Our first goal is to determine whether broadband access lead to increased patenting. The timing of this process is non-trivial. We know the filing date of each patent[28], but not necessarily when the main inventive search period took place. Moreover, we know when the first ISP entered the LE, but not when the firm actually buys broadband access, nor how quickly its use spread throughout the organisation. Therefore, we define our broadband variable as the number of years since activation to allow for timing flexibility.

We describe our headline regression in equations 3.1–3.3, where subscript n and y refer to nodes and years respectively. We limit our sample to nodes that patented at least once during our 2000–2007 time frame in order to minimise noise.[29]

$$pc_{ny} = \beta * BB_{n,y-1} + \phi_y + \epsilon_{ny} \qquad \text{Year FE} \quad (3.1)$$

$$pc_{ny} = \beta * BB_{n,y-1} + \phi_y + \theta_{nuts3} + \epsilon_{ny} \qquad \text{Region \& Year FE (YnR FE)} \quad (3.2)$$

$$BB_{n,y-1} = \gamma * RAIN_{n,y-2} + \phi_y + \theta_{nuts3} + \eta_{ny} \qquad \text{Rain IV} \quad (3.3)$$

pc_{ny} is the log patent count[30] and $BB_{n,y-1}$ is the number of years since broadband was activated, lagged one year and set to zero before activation. ϕ_y is a set of year dummies, present in all regressions, ϵ_{ny} is the error term. We will refer to the sparse specification in Equation 3.1 as the Year FE model.

In equation 3.2 we add a set of NUTS3 fixed effects, θ_{nuts3}. We will refer to this specification as YnR FE, for year and region fixed effects. Equations 3.2 and 3.3 together define our rainfall instrumental variables approach, where $RAIN_{n,y-2}$ is the annual rainfall in a particular node n, lagged one year more for than the broadband variable as we assume investment decisions were based on past rainfall data.

[28] We use the so-called *patent priority filing date*, which is the first time the applicant filed for a patent at *any* patent office.

[29] Note that the first LEs were activated in 2000 and by 2005 broadband diffusion exceeded 80% (see Figure 3.1).

[30] We add one to each count to retain observations with zero patenting that year.

Table 3.3. Yearly Patents per Node (Log), Rainfall Sample

	(1) Year FE	(2) YnR FE	(3) Rain IV
Years Broadband$_{t-1}$	0.07***	0.07***	0.17***
	(0.01)	(0.01)	(0.04)
Constant	0.11***	0.11***	
	(0.01)	(0.01)	
N	17,152	17,152	17,152
Year FE	✓	✓	✓
Region FE	×	✓	✓
First Stage F-Stat.			71.77

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

We present results in Table 3.3, where we find positive coefficients for both fixed effects models (+0.07, columns 1 and 2), and an IV estimate that is considerably larger, at +0.17 (se 0.04). The IV coefficient implies that nodes with broadband access become 18.5% more prolific every year. Extrapolated to five years of access would mean an increase of $5 * 18.5 = 92.5\%$. Ten years: 185%. Clearly it is not reasonable to extrapolate this linear estimate until infinity.[31] Instead, we limit our extrapolation to observed lead times in broadband access as our coefficient is de facto identified through differences in broadband activation.

For each node n, we calculate the share of other nodes in the NUTS3 region that were activated g years later.[32] Taking averages over all nodes, we find that in 68% the comparison node activated in the same year or earlier. Of the remaining third that actually identifies the coefficient estimate, we find that almost 80% represents three or fewer years lead time. As a result, we do not feel confident to extrapolate the linear estimate more than two to three years, which corresponds to a 37–55% increase in patenting activity due to broadband access. Although this number appears large, it is not out of line with the literature. For example, Hegde, Herkenhoff and Zhu (2020) find that simply revealing patent information twelve months earlier already

[31] Unfortunately we can only identify one variable with one instrument, making it impossible to include e.g. a squared broadband access variable.

[32] We add nodes that activated earlier than the focal node in the zero gap category to avoid double counting.

leads to an increase in patenting activity of 21%. Similarly, Bertschek, Cerquera and Klein (2013) estimate an increase of 56-63% in the probability of innovating due to broadband internet use in German firms. A potential explanation is that baseline patenting rates are very small - our predictions correspond to an average increase of just 0.6-0.9 patents per node per year.

As a robustness check, we turn to the demographic instruments Z_n in equation 3.7. While the instruments are aggregated at the node level, we add the same set of characteristics aggregated at the MSOA level as control variables X_m, where subscript m refers to an MSOA. As a result, our sample goes from being node-year level to node-MSOA-year level.

$$pc_{mny} = \beta * BB_{n,y-1} + \phi_y + \epsilon_{mny} \quad \text{Year FE} \quad (3.4)$$

$$pc_{mny} = \beta * BB_{n,y-1} + \phi_y + \theta_{nuts3} + \epsilon_{mny} \quad \text{YnR FE} \quad (3.5)$$

$$pc_{mny} = \beta * BB_{n,y-1} + X_m\Lambda_2 + \phi_{y,2} + \theta_{nuts3,2} + \epsilon_{mny} \quad \text{Demo IV } 2^{nd} \quad (3.6)$$

$$BB_{m,n,y-1} = Z_n\Gamma + X_m\Lambda_1 + \phi_{y,1} + \theta_{nuts3,1} + \eta_{mny} \quad \text{Demo IV } 1^{st} \quad (3.7)$$

The fixed effects regressions of equations 3.4 and 3.5 are identical to their rainfall sample equivalent, with the difference that the patent count pc_{mny} is now calculated at the intersection of nodes n and MSOA m. The instrumental variables setup now includes the node-level demographic instruments Z_n in the first stage and the MSOA-level demographic controls in both first and second stage (Equations 3.6 and 3.7).

Table 3.4. Yearly Patents per Node-MSOA (Log), Demographics Sample

	(1) Year FE	(2) YnR FE	(3) Demo IV
Years Broadband$_{t-1}$	0.02***	0.02***	0.05***
	(0.00)	(0.00)	(0.01)
N	31,168	31,168	31,168
Year FE	✓	✓	✓
Region FE	×	✓	✓
First Stage F-Stat.			338.89

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.4 shows that while the estimates remain qualitatively similar, they are

about a factor three smaller with a coefficient of +0.05 (se 0.01, column 3) for the demographic IV setup. Following the same logic as above, this implies an increase in yearly patenting of 10–15%. We are more confident in the exogeneity of the rainfall instrument and thus expect the true effect to be closer the rainfall estimate than the demographics ones, however it is reassuring that both lead to positive and significant results.

Table 3.5. Yearly Citation-Weighted Patents per Node (Log). Rainfall and Demographics Samples

	(1) Year FE	(2) YnR FE	(3) Rain IV	(4) Demo IV
Years Broadband$_{t-1}$	0.17***	0.16***	0.50***	0.14***
	(0.01)	(0.02)	(0.10)	(0.02)
N	17 152	17 152	17 152	31 168
Year FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
MSOA controls	×	×	×	✓
First Stage F-Stat.			71.77	338.89

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Next, we weight patents by the number of forward citations, which are commonly used as a proxy for patent value. Table 3.5 shows that this increases the coefficients roughly threefold to +0.5 (se 0.1, Rain IV) and +0.14 (se 0.02, Demo IV). These coefficients combine the effect of broadband access on the number of yearly patents and the number of citations those patents receive. This suggests that not only does broadband lead to more patenting activity, it also leads to patents that are cited more widely, implying these are more relevant than patents generated without broadband access.

As a second robustness check we examine whether broadband access also increased the number of public patents (Table 3.6). These are patents generated by universities, government research labs and hospitals, which were already pre-broadband integrated into alternative international information system networks of academic institutions (BitNet, ARPAnet) and should thus be not, or less, affected by reductions in information access costs stemming from improvements in private network technology. We find estimates very close to zero for all specifications, with

estimates that are significantly positive only in the pure fixed effects regression.

Table 3.6. Yearly *Public* Patents per Node (Log). Rainfall and Demographics Samples

	(1) Year FE	(2) YnR FE	(3) Rain IV	(4) Demo IV
Years Broadband$_{t-1}$	0.04***	0.02***	0.02	-0.00
	(0.00)	(0.00)	(0.03)	(0.01)
N	17 152	17 152	17 152	31 168
Year FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
MSOA controls	×	×	×	✓
First Stage F-Stat.			71.77	335.4

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

In the Appendix we show that our positive impact of broadband on patent counts and citations also holds when we restrict the sample to nodes with active patenting before the introduction of broadband (Tables 3.A1 and 3.A3) or analogously to patents by firms with patenting activity pre-broadband (Table 3.A2 and 3.A4).

3.4.3 Historical placebo and concerns about instrument exogeneity

In order to assess validity and investigate empirical support for the exogeneity of our instruments, we conduct a series of historical placebo tests on our data. If inventive activity in a local exchange area is perfectly orthogonal to broadband activation within the bounds of our exclusion restriction, our IV-approach should not predict significant differences for patenting across LE nodes in the absence of ADSL technology. For confirm this with a placebo test, we exploit the long trail of patent data available, going back to several decades before the arrival of broadband internet, and regress lagged, historical patenting in LEs on instrumented (future) broadband activation during the years 2000–2007.

Specifically, we use lagged patent applications filed ten and twenty years prior to the actual observation period. The results of this placebo test are reported in Table 3.7. Instrumenting earlier broadband activation within NUTS3-region based

Table 3.7. Historical Placebo Test for Yearly Patents per Node (Log). Rainfall and Demographics Samples

	10y-lagged # patents		*20y-lagged # patents*	
	(1) Rain IV	(2) Demo IV	(3) Rain IV	(4) Demo IV
Years Broadband$_{t-1}$	0.10***	0.06***	0.43***	0.04***
	(0.03)	(0.01)	(0.11)	(0.01)
N	17 152	31 168	17 152	31 168
Year FE	✓	✓	✓	✓
Region FE	✓	✓	✓	✓
MSOA controls	×	✓	×	✓
First Stage F-Stat.	71.77	113.69	71.77	113.69

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

on variation in weather conditions yields predictions of higher patenting activity in LEs already a decade before the actual arrival of ADSL at +0.1 (se 0.03, Rain IV) and even larger differences when looking at patenting 20-years prior to activitation (point estimate +0.4, se 0.1, Rain IV (3)). Also in models exploiting mismatch in demographic pull factors attracting ISPs between LEs and overlapping MSOAs as exclusion restriction, significantly higher rates of historical patenting are estimated for activated LEs (+0.06, se 0.01 for 10-year lagged patents and +0.04, se 0.01, for 20-year lags, Demo IV). The failure of this placebo test raises concerns regarding the exclusion restriction behind our identification strategy and warrants caution for the causal interpretation of our findings.

We consider possible explanations for this problem. One issue might be the strong concentration of patenting activity in relatively densely populated areas, which are on average less rainy. While this relationship is not necessarily causal, it might generate spurious correlations between our instruments and patenting in parts of our finite sample. This does not strictly need to imply a violation of exogeneity, however it is likely to decrease the power of instruments particularly in those areas were there is the highest variation in patenting. Our second concern relates to difficulties of our econometric IV specification to instrument the timing of treatment in our nodel-level panel: While our endogenous measure of years since broadband activation implies a time-varying intensity of treatment, the prediction

out the first stage in our 2SLS model is de-facto cross-sectional, i.e. has the same value within a node in pre- and post-treatment stages, given our cross-sectional instruments. Using time-varying covariates in the first stage, on the other hand, would make the source of identification in our model unclear. Related to this, it might be problematic with regards to our linear specification that the treatment of broadband arrival is binary but irreversible: Being an attractive state, no LE reverts back to ISDN status after ADSL is introduced. Hence, in our panel the treatment is always 0 until it becomes 1, and always 1 afterwards within each node. Simple linear regression, as we use it in our first stage, might not be well suited for such type of data.

We attempt to solve these concerns in several ways. First, we devise an alternative version of our simple rainfall instrument, based on the interaction of LE area average precipitation with the share of overall broadband activated LEs in the economy in a given year. The idea behind this approach is to take into account yearly investment decisions of ISPs in broadband expansion in the overall U.K. market, which should be uncorrelated to the characteristics of a specific LE area micro-market, but affect LE entry differently based on attractiveness. This individual LE specific likelihood further varies from period to period as a function of national allocated budget and the already activated nodes. Second, instead of predicting *years since* broadband with 2SLS, we conduct a stepwise regression approach in which we, first, predict the actual year of activation based on our instruments and use this prediction to compute time since (predicted) activation in each year of observation for each node, and manually correct standard errors to account for this two-stage procedure in the second stage regression. Finally, in order to accommodate the specific nature of the treatment our data in a more suitable econometric specification, we devise a control function approach which uses a non-linear first stage based on a duration model, where the failure state is represented by the activation of broadband technology.[33] Specifically, we estimate a Cox proportional hazard

[33] In the literature, such control function approaches are often found with reversed stages, e.g. where a linear first stage predicts a certain treatment and a non-linear second stage is used to estimate survival of a patient conditional on receiving the treatment. In our case, however, the duration model appears more suitable to estimate the first stage, given the time delays in activation across nodes and the resulting staggered treatments.

model using our exclusion restrictions to estimate ADSL failure duration for each LE, and use the computed inverse Mills ratios from this stage to derive a correction term for the endogeneity affecting years since broadband activation in the second stage.

However, none of the above attempts is able to solve the problematic pre-trends detected through the historical placebo tests on our instrumental variable approach.[34] While similar types of instruments are widely established in the literature, this persisting problem constitutes an unsolved limitation regardig the causal interpretability of IV estimates, to be considered also in the remainder of the current version of this paper.

3.4.4 Mechanisms

As main mechanism we put forward that broadband access facilitates the inventive search process, in particular the acquisition of foreign knowledge becomes much easier with high speed internet as it allows inventors to browse through the body of international patents and technical news reports, as described in section 3.2.2. In this section we show that patents originating from broadband activated nodes cite significantly more geographically distant prior art. Moreover, the international patent literature they cite is considerably more recent than that cited by non-broadband patents.

We start by constructing a dataset linking focal citing patents to applicant-added cited patents.[35] We limit our set of focal patents to those granted by the USPTO to avoid issues arising from differences in citation practices between offices. We consider the full universe of globally granted patents for the cited side to avoid selecting only highly internationalised patents. For each focal and cited patent, we obtain address information from the USPTO, EP and IPO, retaining the most

[34] All estimation results from these attempts unreported in the paper.

[35] In line with Alcacer and Gittelman (2006), we only consider applicant-added citations to reflect knowledge flows. We will use examiner-added citations as a placebo test. We are able to distinguish between applicant and examiner added citations for 99.31% of the patents in our sample period. We exclude the remaining 0.7%. We also exclude inventor and firm self citations based on PSN inventors names in PATSTAT and BvD Orbis firm identifiers, as those reflect other types of knowledge flows than the ones we're interested in (similar to, e.g., Thompson, 2006; Singh and Marx, 2013).

precise location.[36]

From this dataset, we generate our dependent variable $match_{ij}^{country}$, which equals one when the focal patent i and cited patent j are both from the same country. As discussed in chapter 3.3, the location of a patent is an ambiguous construct given that patent inventors can reside in multiple places. To solve this, we assign patents to countries by randomly selecting one listed inventor and using their location.[37] Our results are robust to using a more conservative spillover measure that is based on the closest distance between locations across all possible combinations of cited-citing inventor pairs.

$$match_{ij}^{country} = \beta * BB_{i,y-1} + \phi_y + \psi_c + \epsilon_{ij} \quad \text{Class and Year FE (CY FE)} \tag{3.8}$$

$$match_{ij}^{country} = \beta * BB_{i,y-1} + \phi_y + \psi_c + \theta_{nuts3} + \epsilon_{ij} \quad \text{CY and Region FE (CYR FE)} \tag{3.9}$$

$$BB_{i,y-1} = \gamma * RAIN_{i,y-2} + \phi_y + \psi_c + \theta_{nuts3} + \eta_{ij} \quad \text{Rain IV, } 1^{st} \tag{3.10}$$

$$match_{ij}^{country} = \beta * BB_{n,y-1} + X_{i_m}\Lambda_2 + \phi_{y,2} + \psi_{c,2} + \theta_{nuts3,2} + \epsilon_{ij} \quad \text{Demo IV } 2^{nd} \tag{3.11}$$

$$BB_{m,n,y-1} = Z_{i_n}\Gamma + X_{i_m}\Lambda_1 + \phi_{y,1} + \psi_{c,1} + \theta_{nuts3,1} + \eta_{ij} \quad \text{Demo IV } 1^{st} \tag{3.12}$$

Equations 3.8–3.12 describe our regression equations. We add primary 3-digit USPC technology class fixed effects ψ_c to absorb common differences in citing behaviour between technological domains. The rainfall and demographics IV now share the same sample, as both the node n and the MSOA m are fully determined by the location of the focal patent i.

Table 3.8 shows the results. We weight each observation such that all focal patents get a combined weight of one, regardless of the number of patent they cite.[38] With coefficients of -0.05 (se 0.03) and -0.02 (se 0.01) in the rainfall and demographics

[36] For example, many UK-based patents provide only city information in their USPTO document, whereas the IPO patent has street-level information. For non-UK patents this is less important, as there we require only country information.

[37] This approach is established practice in the literature starting with Jaffe, Trajtenberg and Henderson (1993). We follow Thompson (2006); Singh and Marx (2013) in randomly selecting one inventor location in case a patent lists multiple.

[38] This avoids giving excess weight to focal patents that cite more often than others.

IV respectively, we find that access to broadband reduces the probability of citing domestic rather than foreign patents by 4–14%.[39]

Table 3.8. Probability of Citing a Same Country Patent

	(1) CY FE	(2) CYR FE	(3) Rain IV	(4) Demo IV
Years Broadband$_{t-1}$	-0.01***	-0.01	-0.05*	-0.02*
	(0.00)	(0.00)	(0.03)	(0.01)
N	63,633	63,633	63,633	63,633
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
MSOA controls	×	×	×	✓
First Stage F-Stat.			18.31	31.16

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The international prior art that they cite also appears to be more recent as we show in Table 3.9. The dependent variable is now $citage_{ij}$, the number of years between the application date of the focal patent i and each of its cited patents j. The coefficient of the rainfall IV in column (3) suggests that inventors with broadband access cite foreign patents which are 2–3 years more recent. This would explain why we find such a large increase in the number of yearly patents, given that Hegde, Herkenhoff and Zhu (2020) find a 21% increase in patenting due to a one-year reduction in the citation lag and our reduction is 2–3 times larger.[40]

We turn to examiner added citations to provide further evidence that these effects are truly causal. As the examiners are located in the art units of the USPTO, their behaviour should not be affected in any way by the broadband access of the inventor. This is exactly what we find in Tables 3.10 and 3.11.[41] We find insignificant results across the board, both for rainfall and the demographics IVs. The null effect on geographical spillovers is very precisely estimated. Unfortunately the citation lag

[39] The lower bound is the Demo IV estimate times two years, the upper bound is the Rain IV estimate times three years. See Section 3.4.2 for a discussion on why we do not extrapolate the results beyond 2–3 years.

[40] Note also that roughly 88% of all cited patents of UK focal patents are foreign, suggesting that most knowledge is sourced abroad.

[41] We include art unit and USPTO filing year fixed effects in all specifications referring to examiner-added citations.

Table 3.9. Citation Lag in Years, Foreign Cited Patents Only

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demo IV
Years Broadband$_{t-1}$	-0.12**	-0.11*	-0.93**	-0.05
	(0.05)	(0.06)	(0.47)	(0.16)
N	52,788	52,788	52,788	52,788
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
MSOA controls	×	×	×	✓
First Stage F-Stat.			20.57	30.14

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

effects have larger standard errors. It is nevertheless reassuring that the coefficient from the rainfall IV is almost cut in half for the examiner-added patents.

Table 3.10. Probability of *Examiner* Referencing a Same Country Patent

	(1) CY FE	(2) CYR FE	(3) Rain IV	(4) Demo IV
Years Broadband$_{t-1}$	-0.00	-0.00	-0.02	0.01
	(0.00)	(0.00)	(0.02)	(0.01)
N	34,155	34,155	34,155	34,155
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
MSOA controls	×	×	×	✓
First Stage F-Stat.			21.80	38.65

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

In the Appendix, we show that there is no corresponding reduction in the citation lag to same country patents (Table 3.A5), suggesting that local knowledge already flowed relatively smoothly between inventors. We also do not find an increase in the total number of references (Table 3.A6), which rules out inventors simply cite more international patents for legal reasons. Instead, this finding is fully consistent with our hypothesis of increased patenting activity through quicker access to foreign knowledge.

Table 3.11. Citation Lag of *Examiner Added Citations* in Years, Foreign Cited Patents Only

	(1) CY FE	(2) CYR FE	(3) Rain IV	(4) Demo IV
Years Broadband$_{t-1}$	−0.05	−0.08	−0.50	−0.21
	(0.06)	(0.06)	(0.51)	(0.21)
N	29,534	29,534	29,534	29,534
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
MSOA controls	×	×	×	✓
First Stage F-Stat.			20.00	35.98

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Our results hold when we restrict the sample to firms already patenting before broadband (Table 3.A7, examiner robustness check in Table 3.A8), ruling out that our results are driven by firms moving to internet activated regions. The geographic spillover effects are strongest in small firms, although the estimates are imprecise and only significant at 5% when we use the demographic instruments (Table 3.A9).

We do not find any impact on the propensity to cite science (Table 3.A10, only pre broadband firms in Table 3.A11). This is reassuring, as scientific knowledge remained locked behind paywalls forming considerable barriers to entry.

We also do not find an increase in the number of inventors per patent, ruling out that the increase in patenting activity is driven by easier collaboration (Tables 3.A12 and 3.A13). This is in line with expectations, as the real increase in communication options arrived only with the second and third generation of broadband, while we only study the first generation. In the same vein, we do not find any effect on the number of international inventor teams (Tables 3.A17 and 3.A18).

We do find weak evidence that internet diffusion led to an increase in the number of new inventors, although the coefficient is only significant at 5% for the rainfall IV (Tables 3.A14 and 3.A15). This effect appears to be driven by small firms (Table 3.A16). This would suggest broadband reduced the entry barriers for innovation - where before inventors needed a large personal network to get access to the relevant information, now the internet allows newcomers to find that information online more rapidly.

In Tables 3.A21 - 3.A32 in Appendix, we document that increases in citations to foreign prior art were particularly strong for references to EP patents. This suggests that the additional knowledge that inventors acquired from foreign patents due to broadband access was extracted more than expected from European Patent Office documents, and less from USPTO patents. This is particularly interesting, given the historically higher reliance of U.K. inventors on the USPTO patent body, as stressed by prior literature.

In additional, unreported, estimations we find no evidence for a significant reduction of inventor or firm self-citations in prior art references. We further cannot find evidence that broadband activation would have benefited access from firm inventors to university or public research organization patents. While effective knowledge transfer from academic research to industry is a particular concern to public R&D policy, prior literature has pointed out knowledge in patents from academic institutions to be particularly difficult to absorb for firms, which would be in line with our observation. Unlike expected, we also do not find any evidence that inventors would explore new knowledge outside of their domains or would become more likely to build on more distant fields after broadband access. This is in contrast to findings of related studies by Furman and Stern (2011); Furman, Nagler and Watzinger (2018); Biasi and Moser (2018); Iaria, Schwarz and Waldinger (2018); Hegde, Herkenhoff and Zhu (2020). One attempt for explanation of this pattern could be that ADSL access might have become more inadvertently integral part of inventor search routines, while incurring the cost of visiting a library, or a biological resource centre, would usually precondition a specific problem requiring targeted search into a specific, perhaps more distant or unfamiliar domain.

3.5 Conclusion

Inventors are standing on the shoulders of giants. Understanding the importance of access to prior knowledge, and in particular the cost of accessing this information, is a major quest of the empirical literature on cumulative innovation.

In this paper, we investigate the role of information technologies in facilitating the assimilation of distant prior art, and their repercussions on inventive activity. We exploit idiosyncrasies of the roll-out of high-speed internet in the UK to showcase

the impact of early ADSL diffusion on inventive search and local R&D productivity in 2,029 patent generating catchment areas of local exchange nodes of the British telecommunications distribution network between 2000 and 2007. Using complementary identification strategies based on push-pull factors to ISP market entry, orthogonal to inventive firm or inventor establishment, we find that access to broadband services increased yearly patenting by up to 50% in the first 2-3 years after broadband arrival, alongside with a similar gain future technological impact. We can ascribe this increase in patenting largely to easier access to geographically distant knowledge through broadband, as evidenced by an increase in the share of applicant added foreign citations, and a steep decrease in average citation lags to prior art patents from abroad. We further show that these changes are not reflected in public sector patenting and examiner-added citations, and exclude a number of competing explanations underlying the extensive margin effect.

Our findings highlight a strong influence of broadband technology on the efficiency of inventive search and the accrual of cross-border knowledge spillovers. This is consistent with the view that information access plays and important role for subsequent invention, and shows that the elasticity of the cost of access to external knowledge is large and substantial, even in the middle of the information age, where most relevant codified knowledge was online available.

Limitations of our study are mainly imposed - at the current stage - by the historical pre-trends in patenting activities detected wihtin our instrumental variable apporach, which warrant caution for the causal interpretability of our results, and require further assessment. Other limitations relate to the comparatively short sample window, as the U.K. early broadband roll-out was essentially completed within a few years. Therefore, we are bound to extrapolating from a linear effect over this period, but our research setting does not allow to identify non-linearity in the impact of broadband internet over a longer time frame. Furthermore, our instrumental variable approach unfortunately lacks the power to investigate heterogeneity of results in greater detail. Alternative settings and different identification will be needed to explore the multifacetedness of main effects further. Finally, the generalizability of our findings is confined by the composition of our sample, which is tributary to the accuracy of patent address information and limited to inventive activity geographically concentrated in the UK.

References

Aghion, Philippe, and Peter Howitt. 1992. "A model of growth through creative destruction." *Econometrica*, 60(2): 323–351.

Agrawal, Ajay, and Avi Goldfarb. 2008. "Restructuring research: Communication costs and the democratization of university innovation." *American Economic Review*, 98(4): 1578–90.

Ahlfeldt, Gabriel, Pantelis Koutroumpis, and Tommaso Valletti. 2017. "Speed 2.0: Evaluating access to universal digital highways." *Journal of the European Economic Association*, 15(3): 586–625.

Akcigit, Ufuk, Salome Baslandze, and Stefanie Stantcheva. 2016. "Taxation and the international mobility of inventors." *American Economic Review*, 106(10): 2930–81.

Alcacer, Juan, and Michelle Gittelman. 2006. "Patent citations as a measure of knowledge flows: The influence of examiner citations." *The Review of Economics and Statistics*, 88(4): 774–779.

Baruffaldi, Stefano, and Markus Simeth. 2020. "Patents and Knowledge Diffusion: The Effect of Early Disclosure." *Research Policy*, forthcoming.

Bertschek, Irene, Daniel Cerquera, and Gordon J Klein. 2013. "More bits–more bucks? Measuring the impact of broadband internet on firm performance." *Information Economics and Policy*, 25(3): 190–203.

Biasi, Barbara, and Petra Moser. 2018. "Effects of Copyrights on Science - Evidence from the US Book Republication Program." National Bureau of Economic Research Working Paper 24255.

Black, S E, and L M Lynch. 2001. "How to compete: The impact of workplace practices and information technology on productivity." *Review of Economics and Statistics*, 83(3): 434–445.

Bloom, Nicholas, Mark Schankerman, and John Van Reenen. 2013. "Identifying technology spillovers and product market rivalry." *Econometrica*, 81(4): 1347–1393.

Bouckaert, Jan, Theon van Dijk, and Frank Verboven. 2010. "Access regulation, competition, and broadband penetration: An international study." *Telecommunications Policy*, 34(11): 661–671.

Bryan, Kevin A, and Yasin Ozcan. 2016. "The impact of open access mandates on invention." *Mimeo, Toronto*.

Cheyre, Cristobal, Steven Klepper, and Francisco Veloso. 2015. "Spinoffs and the mobility of US merchant semiconductor inventors." *Management Science*, 61(3): 487–506.

Cohen, Wesley M, Richard R Nelson, and John P Walsh. 2000. "Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not)." National Bureau of Economic Research Working Paper 7552.

Czernich, Nina, Oliver Falck, Tobias Kretschmer, and Ludger Woessmann. 2011. "Broadband infrastructure and economic growth." *The Economic Journal*, 121(552): 505–532.

de Rassenfosse, G., G. Pellegrino, and E. Raiteri. 2020. "Do patents enable disclosure? Evidence from the 1951 Invention Secrecy Act." Ecole polytechnique fédérale de Lausanne.

de Rassenfosse, G., J. Kozak, and F Seliger. 2019. "Geocoding of worldwide patent data." *Nature Scientific Data*, 6(260).

Devlin, Alan. 2009. "The misunderstood function of disclosure in patent law." *Harvard Journal of Law & Technology*, 23: 401.

Dewan, S, and KL Kraemer. 2000. "Information technology and productivity: Evidence from country-level data." *Management Science*, 46(4): 548–562.

Dhyne, Emmanuel, Jozef Konings, Jeroen Van den bosch, and Stijn Vanormelingen. 2018. "The Return on Information Technology: Who Benefits Most?" CEPR Discussion Paper No. DP13246.

Ding, Waverly W, Sharon G Levin, Paula E Stephan, and Anne E Winkler. 2010. "The impact of information technology on academic scientists' productivity and collaboration patterns." *Management Science*, 56(9): 1439–1461.

Dittmar, Jeremiah E. 2011. "Information technology and economic change: the impact of the printing press." *The Quarterly Journal of Economics*, 126(3): 1133–1172.

Forman, Chris, and Nicolas van Zeebroeck. 2012. "From wires to partners: How the Internet has fostered R&D collaborations within firms." *Management science*, 58(8): 1549–1568.

Forman, Chris, and Nicolas van Zeebroeck. 2019. "Digital technology adoption and knowledge flows within firms: Can the Internet overcome geographic and technological distance?" *Research policy*, 48(8): 103697.

Furman, Jeffrey L., and Scott Stern. 2011. "Climbing atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research." *American Economic Review*, 101(5): 1933–63.

Furman, Jeffrey L, Markus Nagler, and Martin Watzinger. 2018. "Disclosure and subsequent innovation: Evidence from the patent depository library program." National Bureau of Economic Research.

Gavazza, Alessandro, Mattia Nardotto, and Tommaso Valletti. 2019. "Internet and politics: Evidence from UK local elections and local government policies." *The Review of Economic Studies*, 86(5): 2092–2135.

Giuri, Paola, Myriam Mariani, Stefano Brusoni, Gustavo Crespi, Dominique Francoz, Alfonso Gambardella, Walter Garcia-Fontes, Aldo Geuna, Raul Gonzales, Dietmar Harhoff, and Karin Hoisl. 2007. "Inventors and invention processes in Europe: Results from the PatVal-EU survey." *Research Policy*, 36(8): 1107–1127.

Graham, S, and D Hegde. 2015. "Disclosing patents' secrets." *Science*, 347(6219): 236–237.

Griffith, Rachel, Rupert Harrison, and John Van Reenen. 2006. "How special is the special relationship? Using the impact of US R&D spillovers on UK firms as a test of technology sourcing." *American Economic Review*, 96(5): 1859–1875.

Griliches, Zvi. 1990. *Patent statistics as economic indicators: A survey, part 2.* NBER.

Hegde, Deepak, and Hong Luo. 2018. "Patent publication and the market for ideas." *Management Science*, 64(2): 652–672.

Hegde, Deepak, Kyle Herkenhoff, and Chenqi Zhu. 2020. "Patent disclosure and innovation." *Available at SSRN 3158031*.

Hjort, Jonas, and Jonas Poulsen. 2019. "The arrival of fast internet and employment in Africa." *American Economic Review*, 109(3): 1032–79.

Iaria, Alessandro, Carlo Schwarz, and Fabian Waldinger. 2018. "Frontier Knowledge and Scientific Production: Evidence from the Collapse of International Science*." *The Quarterly Journal of Economics*, 133(2): 927–991.

Jaffe, Adam B. 1986. "Technological opportunity and spillovers of R&D: evidence from firms' patents, profits and market value." national bureau of economic research.

Jaffe, Adam B, Manuel Trajtenberg, and Michael S Fogarty. 2000. "Knowledge spillovers and patent citations: Evidence from a survey of inventors." *American Economic Review*, 90(2): 215–218.

Jaffe, Adam B, Manuel Trajtenberg, and Rebecca Henderson. 1993. "Geographic localization of knowledge spillovers as evidenced by patent citations." *The Quarterly Journal of Economics*, 108(3): 577–598.

Jones, Charles I. 1995. "R&D-based models of economic growth." *Journal of Political Economy*, 103(4): 759–784.

Jones, Eric. 2003. *The European miracle: environments, economies and geopolitics in the history of Europe and Asia.* Cambridge University Press.

Keeble, David, and Lilach Nachum. 2001. "Why Do Business Service Firms Cluster? Small Consultancies, Clustering and Decentralisation in London and Southern England." Centre for Business Research, University of Cambridge Working Papers wp194.

Lemley, Mark A. 2012. "The myth of the sole inventor." *Mich. L. Rev.*, 110(5): 709.

Lück, Sonja, Benjamin Balsmeier, Florian Seliger, and Lee Fleming. 2019. "Early disclosure law reduces duplication in the US and European patent systems." Northwestern Pritzker School of Law.

Machlup, Fritz, and Edith Penrose. 1950. "The patent controversy in the nineteenth century." *The Journal of Economic History*, 10(1): 1–29.

MacKie-Mason, Jeffrey K, and Hal R Varian. 1996. "Some economics of the Internet."

Mokyr, Joel. 2005. "Long-term economic growth and the history of technology." In *Handbook of economic growth*. Vol. 1, 1113–1180. Elsevier.

Morrison, Greg, Massimo Riccaboni, and Fabio Pammolli. 2017. "Disambiguation of patent inventors and assignees using high-resolution geolocation data." *Scientific data*, 4: 170064.

Moser, Petra, and Alessandra Voena. 2012. "Compulsory Licensing: Evidence from the Trading with the Enemy Act." *American Economic Review*, 102(1): 396–427.

Nardotto, Mattia, Tommaso Valletti, and Frank Verboven. 2015. "Unbundling the incumbent: Evidence from UK broadband." *Journal of the European Economic Association*, 13(2): 330–362.

Ofcom. 2014. "Fixed access market reviews: wholesale local access, wholesale fixed analogue exchange lines, ISDN2 and ISDN30 (Annexes)."

Ouelette, Lisa L. 2012. "Do patents disclose useful information?" *Harvard Journal of Law & Technology*, 25(2): 545–608.

Pavitt, Keith, and Pari Patel. 1998. "The international distribution and determinants of technological activities." *Oxford Review of Economic Policy*, 4(4): 35–55.

Romer, Paul M. 1990. "Endogenous technological change." *Journal of political Economy*, 98(5, Part 2): S71–S102.

Sampat, Bhaven N. 2010. "When do applicants search for prior art?" *The Journal of Law and Economics*, 53(2): 399–416.

Scotchmer, Suzanne. 1991. "Standing on the shoulders of giants: cumulative research and the patent law." *Journal of economic perspectives*, 5(1): 29–41.

Scotchmer, Suzanne, and Jerry Green. 1990. "Novelty and disclosure in patent law." *The RAND Journal of Economics*, 131–146.

Singh, Jasjit. 2005. "Collaborative networks as determinants of knowledge diffusion patterns." *Management science*, 51(5): 756–770.

Singh, Jasjit. 2008. "Distributed R&D, cross-regional knowledge integration and quality of innovative output." *Research Policy*, 37(1): 77–96.

Singh, Jasjit, and Matt Marx. 2013. "Geographic constraints on knowledge spillovers: Political borders vs. spatial proximity." *Management Science*, 59(9): 2056–2078.

Spear, Brian. 1996. "Patent information on the internet." *World Patent Information*, 18(1): 33–37.

Stiroh, Kevin J. 2002. "Information technology and the U.S. productivtiy revival: What do the industry data say?" *American Economic Review*, 92(5): 1559–1576.

Thompson, Peter. 2006. "Patent citations and the geography of knowledge spillovers: evidence from inventor-and examiner-added citations." *The Review of Economics and Statistics*, 88(2): 383–388.

Verspagen, B, and W Schoenmakers. 2004. "The spatial dimension of patenting by multinational firms in Europe." *Journal of Economic Geography*, 4(1): 23–42.

Weitzman, Martin L. 1998. "Recombinant growth." *The Quarterly Journal of Economics*, 113(2): 331–360.

Williams, Heidi L. 2013. "Intellectual Property Rights and Innovation: Evidence from the Human Genome." *Journal of Political Economy*, 121(1): 1–27.

Zheng, Yanfeng, and Qinyu Wang. 2020. "Shadow of The Great Firewall: The Impact of Google Blockade on Innovation in China." *Strategic Management Journal, Forthcoming*.

Appendices

3.A Appendix tables

Table 3.A1. Yearly Patents per Node (Log), Rainfall Sample - Pre-Broadband Active LEs

	(1) Class+Year FE	(2) Region FE	(3) Rain IV
Years Broadband$_{t-1}$	0.08***	0.09***	0.23***
	(0.01)	(0.01)	(0.07)
Constant	0.21***	0.19***	
	(0.02)	(0.02)	
N	8712	8712	8712
Year FE	✓	✓	✓
Region FE	×	✓	✓
First Stage F-Stat.			40.51

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A2. Yearly Patents per Node (Log), Rainfall Sample - Pre-Broadband Established Firms

	(1) Class+Year FE	(2) Region FE	(3) Rain IV
Years Broadband$_{t-1}$	0.04***	0.04***	0.11***
	(0.01)	(0.01)	(0.03)
Constant	0.07***	0.06***	
	(0.01)	(0.01)	
N	17,152	17,152	17,152
Year FE	✓	✓	✓
Region FE	×	✓	✓
First Stage F-Stat.			71.77

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A3. Yearly Patents per Node (Log), Rainfall Sample - Pre-Broadband Active LEs

	(1) Class+Year FE	(2) Region FE	(3) Rain IV
Years Broadband$_{t-1}$	0.19***	0.21***	0.70***
	(0.02)	(0.03)	(0.19)
Constant	0.60***	0.57***	
	(0.05)	(0.05)	
N	8,712	8,712	8,712
Year FE	✓	✓	✓
Region FE	×	✓	✓
First Stage F-Stat.			40.51

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A4. Yearly Citation Weighted Patents per Node (Log), Rainfall Sample - Pre-Broadband established firms

	(1) Class+Year FE	(2) Region FE	(3) Rain IV
Years Broadband$_{t-1}$	0.09***	0.09***	0.33***
	(0.01)	(0.01)	(0.08)
Constant	0.20***	0.19***	
	(0.02)	(0.02)	
N	17,152	17,152	17,152
Year FE	✓	✓	✓
Region FE	×	✓	✓
First Stage F-Stat.			71.77

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A5. Citation Lag in Years, Domestic Cited Patents Only

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demogs IV
Years Broadband$_{t-1}$	0.01	0.01	0.10	0.02
	(0.01)	(0.01)	(0.08)	(0.03)
Constant	1.56***	1.56***		
	(0.02)	(0.02)		
N	6,402	6,399	6,397	6,397
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
First Stage F-Stat.			18.85	31.83

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A6. Applicant-Added Citations per Patent (Log)

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demogs IV
Years Broadband$_{t-1}$	0.01	0.01	-0.16	-0.09*
	(0.02)	(0.02)	(0.13)	(0.06)
Constant	1.96***	1.96***		
	(0.05)	(0.04)		
N	6,402	6,399	6,397	6,397
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
First Stage F-Stat.			18.85	31.83

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A7. Probability of Citing a Same Country Patent – Pre-Broadband Established Firms

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demogs IV
Years Broadband$_{t-1}$	-0.01***	-0.01***	-0.07**	-0.05***
	(0.00)	(0.00)	(0.03)	(0.02)
N	40,398	40,398	40,398	40,398
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
First Stage F-Stat.			15.39	14.85

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A8. Probability of Examiner Adding a Same Country Patent Reference – Pre-Broadband Established Firms

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demogs IV
Years Broadband$_{t-1}$	-0.00*	-0.00	-0.03	0.01
	(0.00)	(0.00)	(0.04)	(0.02)
N	20,760	20,760	20,760	20,760
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
First Stage F-Stat.			15.39	14.85

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A9. Firm Size Splits (above Median) for Same Country Citations

	(1) Rain IV Small	(2) Rain IV Large	(3) Demogs IV Small	(4) Demogs IV Large
Years Broadband$_{t-1}$	-0.07	0.01	-0.04**	0.01
	(0.05)	(0.03)	(0.02)	(0.01)
N	33,458	30,164	33,458	30,169
Year + Class FE	✓	✓	✓	✓
Region FE	✓	✓	✓	✓
First Stage F-Stat.	13.65	12.67	11.68	31.06

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A10. Probability of Reliance on Science [0,1]

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demogs IV
Years Broadband$_{t-1}$	0.00	0.01	-0.05	-0.03
	(0.01)	(0.01)	(0.04)	(0.02)
Constant	0.29***	0.28***		
	(0.01)	(0.02)		
N	6,402	6,399	6,397	6,397
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
First Stage F-Stat.			18.85	31.83

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A11. Probability of Reliance on Science [0,1] - Pre-Broadbrand Established Firms

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demogs IV
Years Broadband$_{t-1}$	0.01	0.01	-0.03	-0.04
	(0.01)	(0.01)	(0.06)	(0.03)
Constant	0.30***	0.29***		
	(0.02)	(0.02)		
N	4,027	4,017	4,015	4,015
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
First Stage F-Stat.			18.85	31.83

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A12. Number of Inventors (Log) per Patent

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demogs IV
Years Broadband$_{t-1}$	0.00	0.01	0.03	0.03
	(0.01)	(0.01)	(0.06)	(0.03)
Constant	0.62***	0.61***		
	(0.03)	(0.02)		
N	8,301	8,300	8,298	8,298
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
First Stage F-Stat.			23.83	31.89

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A13. Number of Inventors (Log) per Patent – Pre-Broadband Established Firms

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demogs IV
Years Broadband$_{t-1}$	-0.00	0.01	-0.03	0.03
	(0.01)	(0.02)	(0.08)	(0.04)
Constant	0.69***	0.67***		
	(0.03)	(0.03)		
N	5,116	5,108	5,106	5,106
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
First Stage F-Stat.			14.46	16.54

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A14. Probability of New Inventor Entry [01]

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demogs IV
Years Broadband$_{t-1}$	0.01*	0.00	0.16**	0.03
	(0.01)	(0.01)	(0.06)	(0.02)
Constant	0.64***	0.66***		
	(0.02)	(0.02)		
N	6402	6399	6397	6397
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
First Stage F-Stat.			18.85	31.83

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A15. Probability of New Inventor Entry [01] - Pre-Broadband Established Firms

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demogs IV
Years Broadband$_{t-1}$	0.02*	0.01	0.20**	0.01
	(0.01)	(0.01)	(0.08)	(0.04)
Constant	0.61***	0.62***		
	(0.02)	(0.03)		
N	4,027	4,017	4,015	4,015
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
First Stage F-Stat.			15.36	14.17

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A16. Firm Size Splits for New Inventor Entry [01]

	(1) Rain IV Small	(2) Rain IV Large	(3) Demogs IV Small	(4) Demogs IV Large
Years Broadband$_{t-1}$	0.28**	0.09	0.04	0.02
	(0.12)	(0.07)	(0.05)	(0.03)
N	3325	2976	3325	2976
Year + Class FE	✓	✓	✓	✓
Region FE	✓	✓	✓	✓
First Stage F-Stat.	11.84	13.92	11.78	30.54

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A17. Probability of Geographic Co-Location of Inventor Teams (Same Country)

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demogs IV
Years Broadband$_{t-1}$	-0.01	-0.01	-0.02	-0.01
	(0.01)	(0.01)	(0.06)	(0.03)
Constant	0.25***	0.24***		
	(0.02)	(0.02)		
N	8,301	8,300	8,298	8,298
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
First Stage F-Stat.			23.83	31.89

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A18. Probability of Geographic Co-Location of Inventor Teams (Same Country) - Pre-Broadband Established Firms

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demogs IV
Years Broadband$_{t-1}$	-0.02*	-0.01	-0.05	-0.03
	(0.01)	(0.01)	(0.08)	(0.04)
Constant	0.25***	0.24***		
	(0.03)	(0.02)		
N	5,116	5,108	5,106	5,106
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
First Stage F-Stat.			14.46	16.54

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A19. Probability of New Inventor Collaboration

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demogs IV
Years Broadband$_{t-1}$	0.00	0.01	0.06	0.06***
	(0.01)	(0.01)	(0.05)	(0.02)
Constant	0.50***	0.49***		
	(0.02)	(0.02)		
N	8,301	8,300	8,298	8,298
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
First Stage F-Stat.			23.83	31.89

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A20. Probability of New Inventor Collaboration – Pre-Broadband Established Firms

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demogs IV
Years Broadband$_{t-1}$	0.00	0.01	-0.00	0.07**
	(0.01)	(0.01)	(0.07)	(0.03)
Constant	0.54***	0.52***		
	(0.02)	(0.03)		
N	5,116	5,108	5,106	5,106
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
First Stage F-Stat.			14.46	16.54

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A21. Probability of Applicant Citing a US Patent Document

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demo IV
Years Broadband$_{t-1}$	0.00	-0.00	-0.07*	-0.00
	(0.00)	(0.00)	(0.04)	(0.01)
N	63,638	63,638	63,638	63,638
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
MSOA controls	×	×	×	✓
First Stage F-Stat.			18.31	31.16

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A22. Probability of Applicant Citing a *Foreign* US Patent Document

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demo IV
Years Broadband$_{t-1}$	-0.00	-0.00	-0.10**	-0.02
	(0.00)	(0.00)	(0.04)	(0.01)
N	56,028	56,028	56,028	56,028
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
MSOA controls	×	×	×	✓
First Stage F-Stat.			18.77	30.26

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A23. Probability of *Examiner Added* Citation to US Patent Document

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demo IV
Years Broadband$_{t-1}$	-0.00	-0.00	0.01	-0.00
	(0.00)	(0.00)	(0.01)	(0.01)
N	34,161	34,161	34,161	34,161
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
MSOA controls	×	×	×	✓
First Stage F-Stat.			21.80	38.65

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A24. Probability of *Examiner Added* Citation to *Foreign* US Patent Document

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demo IV
Years Broadband$_{t-1}$	-0.00	-0.00	-0.00	-0.00
	(0.00)	(0.00)	(0.01)	(0.01)
N	31,876	31,876	31,876	31,876
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
MSOA controls	×	×	×	✓
First Stage F-Stat.			19.31	38.67

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A25. Probability of Applicant Citing a US Patent Document - Pre-Broadband Established Firms Only

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demo IV
Years Broadband$_{t-1}$	0.00	-0.00	-0.03	-0.00
	(0.01)	(0.01)	(0.04)	(0.02)
N	40,402	40,402	40,402	40,402
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
MSOA controls	×	×	×	✓
First Stage F-Stat.			15.39	14.85

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A26. Probability of Applicant Citing a *Foreign* US Patent Document – Pre-Broadband Established Firms Only

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demo IV
Years Broadband$_{t-1}$	-0.00	-0.01	-0.07	-0.03
	(0.01)	(0.01)	(0.04)	(0.02)
N	35,659	35,659	35,659	35,659
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
MSOA controls	×	×	×	✓
First Stage F-Stat.			15.86	14.60

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A27. Probability of Applicant Citing an EP Patent Document

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demo IV
Years Broadband$_{t-1}$	0.00	0.00	0.04	0.01
	(0.00)	(0.00)	(0.02)	(0.01)
N	63,638	63,638	63,638	63,638
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
MSOA controls	×	×	×	✓
First Stage F-Stat.			18.31	31.16

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A28. Probability of Applicant Citing a *Foreign* EP Patent Document

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demo IV
Years Broadband$_{t-1}$	-0.00	0.00	0.05**	0.01
	(0.00)	(0.00)	(0.03)	(0.01)
N	56,028	56,028	56,028	56,028
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
MSOA controls	×	×	×	✓
First Stage F-Stat.			18.77	30.26

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A29. Probability of *Examiner Added* Citation to EP Patent Document

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demo IV
Years Broadband$_{t-1}$	-0.00	-0.00	0.01	-0.00
	(0.00)	(0.00)	(0.01)	(0.01)
N	34,161	34,161	34,161	34,161
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
MSOA controls	×	×	×	✓
First Stage F-Stat.			21.80	38.65

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A30. Probability of *Examiner Added* Citation to *Foreign* EP Patent Document

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demo IV
Years Broadband$_{t-1}$	-0.00	-0.00	-0.01	0.00**
	(0.00)	(0.00)	(0.01)	(0.00)
N	31,876	31,876	31,876	31,876
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
MSOA controls	×	×	×	✓
First Stage F-Stat.			19.31	38.67

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A31. Probability of Applicant Citing an EP Patent Document - Pre-Broadband Established Firms Only

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demo IV
Years Broadband$_{t-1}$	0.00	0.00	0.05*	0.02
	(0.00)	(0.00)	(0.03)	(0.01)
N	40,402	40,402	40,402	40,402
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
MSOA controls	×	×	×	✓
First Stage F-Stat.			15.39	14.85

Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.A32. Probability of Applicant Citing a *Foreign* EP Patent Document - Pre-Broadband Established Firms Only

	(1) Class+Year FE	(2) Region FE	(3) Rain IV	(4) Demo IV
Years Broadband$_{t-1}$	-0.00	0.00	0.07*	0.03*
	(0.00)	(0.01)	(0.04)	(0.02)
N	35,659	35,659	35,659	35,659
Year + Class FE	✓	✓	✓	✓
Region FE	×	✓	✓	✓
MSOA controls	×	×	×	✓
First Stage F-Stat.			15.86	14.60

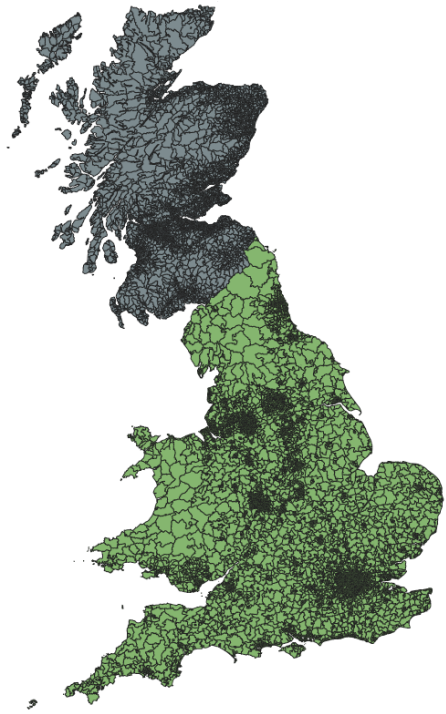
Notes: Standard errors clustered at the LE-level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.B Appendix figures

Figure 3.B1. NUTS3 Regions (Counties) of Great Britain (N=167)



Figure 3.B2. Census 2001 MSOAs Across Great Britain (N=3,374)



General Conclusion

Access to the existing stock of knowledge is essential for cumulative innovation, which grounds on the coupled interplay of scientific progress and technical change, fueling prosperity and welfare of societies as a key engine of economic growth. However, our quest for understanding the economic impacts of knowledge access on innovation reveals that it is a far more complex and multifaceted phenomenon than what the opening example of Gutenberg's book press and Newton's famous metaphor may suggest.

Starting from the presumption that the value of knowledge for innovation is not constant, scholars in the economics of science and innovation have started to explicate this relationship more carefully and to inquire questions about effective access and its repercussions on the unfolding of innovation. A growing consensus in the debate seems to arise with regards to the realization that the timing, access modes, absorptive capacity as well as the very nature of the accessed knowledge matter for this connect. The literature has addressed this complex from several different angles: Access to scientific knowledge and its impact on technology, access-enhancing institutions, both informal or formal types, and their impact on the amplification of intertemporal knowledge spillovers, as well as the role of modern information technologies in shaping this relationship.

However, our understanding of the working mechanisms of effective access and its elasticity on innovation is still limited and only starting to be explored. This dissertation revisits findings of prior research and contributes to the literature on knowledge access and cumulative innovation with regards to a number of so far open questions: What is the importance for scientific human capital of inventors

for access to frontier knowledge for new technologies? Which types of inventors contribute the most to frontier science access for technological progress, in light of significant heterogeneity of scientific footprints? How can we design effective institutions to lower search costs for the highly-skilled workforce and direct their attention in order to access specific types of relevant knowledge in the sheer bulk of daily information overflow surrounding not just consumers but also scientists and engineers alike? What are effective mechanisms of access in the light of colliding incentives of increasing social returns from non-rivalry and private returns in knowledge production based on excludability and internalization of spillovers? What are the benefits for knowledge access and innovative outcomes when introducing new generations of information technology with significantly faster?

This dissertation sheds lights into these questions and informs the scholarly debate centering around these aspects of knowledge access, science and invention in three empirical essays. The following sections recapitulate on the main contributions of my doctoral research, develops on implications of my findings for policy and corporate decision making, and advances further avenues for future research in the domains and questions behind these essays.

Main Findings

Chapter 1 addresses a central question in the literature on industry-science linkages which is: What is the contribution of science participation of inventors for their technological outcomes? Prior work (e.g. Gittelman and Kogut, 2003) has remained rather general on this relationship both regarding the type of involvement and the access mechanism to science. We contribute to this question by adding new conceptual discussion and, most relevantly, empirical insight on three important levels of heterogeneity in these links: First, regarding the specific type of scientific involvement. Second, differences in the type of scientific knowledge that is accessed by different author types. And third, a significant heterogeneity in technological outcomes, especially breakthrough technologies, depending on the author and knowledge types accessed in patents.

Based on our comprehensive analysis of author-inventor links and patent-paper trails in the universe of biomedical research, our main finding shows that inven-

tors at the frontier of knowledge, i.e. involved in recent scientific breakthrough, push technological progress by generating significantly higher cumulative impact and, in particular, top5% hit patents. This is an important and non-obvious insight, following a controversial debate on different skill-sets of scientists vs. engineers in the literature (e.g., Allen et al., 1977). We further find this relationship to be crucially dependent on both the quality of scientific output as well as the recency of involvement, highlighting the underlying frontier nature of the relationship. Analysing references to scientific articles made in frontier author patents, we contribute further evidence for a significantly higher reliance on frontier knowledge in the patents of such authors, and show that these links account for great part of the frontier author technological impact premium, in particular with regards to the generation of breakthrough technology. These results contribute especially to the relatively recent literature highlighting the importance of early access to scientific knowledge (e.g., Ahmadpoor and Jones, 2017; Baruffaldi and Pöge, 2020; Arora, Belenzon and Dionisi, 2021, in particular to recent top-science (e.g., Iaria, Schwarz and Waldinger, 2018; Poege et al., 2019, by advancing a mechanism of effective access based on individual scientific human capital. Moreover, by showing that access gains to frontier science are strongest for knowledge originating in the close proximity to the authors (self- and network-citations), these new insights further support the idea of increased absorptive capacity through participation, which is core to an even broader literature on industry-science links (e.g.,Cohen and Levinthal, 1989; Rosenberg, 1990; Cockburn and Henderson, 1998; Arora, Belenzon and Sheer, 2021).

Chapter 2 shifts the focus from human capital externalities to access-enhancing institutions and shows that topic-specific repositories may constitute effective instruments to increase the dissemination of dispersed relevant knowledge, in line with a policy objective of intensifying spillovers across disconnected communities of researchers. My analyses of the cumulative technological impact of the CNIDR AIDS Patents Database show that topic-specific collections (or databases) allow to decrease search costs for prior art, by indexing pieces of knowledge for which the link to a specific topic was previously non-obvious in external technological search. By this, the HIV/AIDS-specific index led to a +30% higher citation rate to patents associated with previously higher retrieval costs. These patterns are not explained

by differences in technology types nor do they reflect changes in patent examiner behavior, suggesting that the database proved highly effective in shifting the attention of researchers towards this type of more 'hidden' advances in the domain of HIV/AIDS related treatments and diagnostics. This is further evidenced by higher citation rates coming in from geographically more distant inventors and previously entirely disconnected collaborative communities.

Findings in this chapter speak mostly to the recent body of work on access-enhancing institutions (e.g., Furman and Stern, 2011; Thompson and Hanley, 2018; Furman, Nagler and Watzinger, 2018; Berkes and Nencka, 2020, and make a particular contribution to this literature by highlighting new insights on the working mechanism and effective design of such institutions.

Finally, chapter 3 contributes new causal evidence regarding the role of modern information technology in shaping cumulative innovation through their effect on access costs in knowledge sourcing. The study of ADSL broadband internet provides a particularly useful case to investigate this question given it provided faster – i.e. at lower costs – access to by then fully disclosed and available patent information on the internet. Thereby, broadband provides a more comprehensive and more inclusive experiment to test the elasticity of access costs, compared to e.g. libraries concentrated in specific hub areas. Based on our 2SLS estimation of the effect of moving from basic internet to broadband activation, which exploits geographic variation in activation status for local exchange points within NUTS3 regions provided by delays in market entry of internet service providers, we find a significant and large baseline increase of patent filings following the 2–3 years after broadband activation. This was paralleled by an even larger growth in incoming follow-up citation rates to new inventions originating from broadband activated areas, suggesting a true economic effect of the ICT. This main finding contributes evidence to the literature on the effect of broadband technology on productivity (through innovation) (e.g., Czernich et al., 2011; Bertschek, Cerquera and Klein, 2013). We add to these prior studies with insights on the working mechanism of variation in access costs to more distant information driving these effects, by showing changes in the composition of new patents filed after broadband in the geographic scope and, in particular, the recency of foreign prior art referenced. This insight allows to extrapolate prior findings by Forman and van Zeebroeck 2012, 2019 on within-

firm knowledge flows following ICT adoption to a national economy-wide level, highlighting the usefulness of ADSL on external knowledge acquisition.

Implications

The scientific findings of the individual essays of this dissertation allow to formulate a number of normative yet concrete implications for current and future policy as well as corporate decisions. The question of knowledge access is, after all, an important issue that pertains to the effective craft of economic systems, well beyond the niche of science and technology progress touched upon by this dissertation. It is more a critical issue at the intersection of many areas, ranging from innovation, to inequality, democratic participation, education, and general economic development, with obvious interconnections between these.

Chapter 1 of the dissertation makes a flaming statement for the importance of frontier research, i.e. recent top-science, on one hand, and technological breakthrough and industrial progress, on the other, as well as for the individuals connecting these two phenomena. Frontier knowledge and its pivotal contribution for progress is not only receiving more attention in the academic literature, but also a key concern for R&D and industrial policy, as already highlighted by the European Commission's pertaining High-Level Expert Group in 2005 (Commission, 2005). Frontier research has immense potential for prosperity and technology leadership, but it also involves significant risk and, therefore, special necessities when it comes to the design of public subsidies, funding for excellence research and, in general, the design of knowledge-transfer systems when it comes to maximizing the social returns to investments in frontier research. Here the empirical evidence presented in Chapter 1 makes a compelling case for the effectiveness of access to frontier science in technology through the channel of the individuals involved in the creation of scientific breakthrough. For policy, this becomes particularly important with regards to incentives for collaboration of scientists in universities and public research institutions with private sector and industry. While our paper does not explicit distinguish between institutional affiliations of frontier scientists, it aligns

with the vast body of research in innovation suggesting high-benefits for both private firms and socially-relevant knowledge-transfer through interpersonal ties and collaborations of industry with university researchers, which are found to generate mutual positive effects on their academic work as well.

This links directly to implications of these findings for corporate decision making and the organization of firm R&D. We show that the involvement of corporate inventors (or inventors patenting with private corporations) in recent top-scientific research is directly linked access to frontier knowledge and to high-impact technology, which translates not only into large social but also substantial private value for firms. While prior academic research shows that corporations may be reluctant to invest in basic research for which they cannot prevent spillouts (free-riding) (e.g., Rosenberg, 1990; Arora, Belenzon and Patacconi, 2018; Arora, Belenzon and Sheer, 2021), our finding make a strong case for the importance of corporate participation. This is particularly relevant for corporate management considering the crucial role of priority and early access for the recoupment of private value out of scientific knowledge, which has been recently documented by Arora, Belenzon and Dionisi (2021). While our findings do not indicate that frontier authors are necessarily more often the first to access frontier science, they make access to this knowledge much more, and when they are first, these links are associated with particularly higher breakthrough rates. Breaking up ties with scientific research might endanger the functioning of such fragile yet high-gain connections to frontier knowledge, that firm R&D strategy may need to be aware and cautious of.

Chapter 2 bears important implications, in particular, with regards to the effectiveness of access-enhancing systems and the design of such institutions. On one hand, for policy the findings of the chapter show that topic-specific collections (which are becoming seemingly more popular, also in response to the COVID-2019 pandemic) are an effective an comparatively 'low-cost' instrument for intervention into R&D search and drawing the attention of researchers to a particular topic. This may be an especially useful policy instrument in times of acute intensification of disease-targeted research activities, as during pandemics of public health crises. But perhaps also with regards to particular topics (e.g. health conditions) on which further or more research may be socially desirable. Obviously depending on the reasons for a possible underlying market failure, topic-specific item collections may

be effective as a more cost-efficient substitute or complement to disease-targeted funding opportunities, as evaluated in recent work by Myers (2020). Furthermore, my insights on the elasticity of search costs on cumulative knowledge diffusion may also entail implications for the design of patent classes and bibliographic search tools provided by patent offices (or commercial providers). While patent classes are traditionally not use-oriented, a disease-(or topic-)specific indexing of patent records may constitute a useful instrument to optimize knowledge diffusion through the patent system.

For corporate R&D the implications of Chapter 2 appear somewhat more ambivalent. On one hand, it is vital for private firms to protect their intellectual assets, which is the reason corporate patents are often found do be written in a particular abstract jargon aimed at making them more difficult to be retrieved. Strictly interpreted, my finding on the usefulness of topic-repositories for uncovering such hidden patents may imply for firm IP management to refrain from patenting at all for technologies for which they see a risk of spillout of relevant proprietary information due to publication. On the other hand, however, topic-specific collections may also provide a an interesting intelligence system for external knowledge retrieval and in markets for technology. This is particularly indicated by my finding that private firm inventors seemed benefit disproportionately from the reduction in prior art search costs induced by the repository. Furthermore, this may have influences on signaling of occupied market niches and deter industry participants to enter a certain technological field, which may entail adverse anti-commons effects that would be, again, of potential concern for policy. However, the concrete eventuality of disease-indexing having severe adverse effects on reducing essential parallel experimentation appears fairly low.

Chapter 3 provides new evidence regarding the importance of broadband technology, specially with regards to inventive activity. The main finding of the chapter, showing that higher bandwidth internet has real effects on the production of R&D, has important implications for policy, particularly with regards to ongoing current debates around digital backlog (e.g. in the U.S. and several European states) as well as to the digital divide between rural and urban areas and the need for broadband infrastructure expansion in areal countries. Showing the mechanism of increased patenting as a result for enhanced knowledge access, our study appears particularly

informative for policy in this regard, considering that in many regions in question low access to relevant information persists independent of missing broadband access. Based on our findings in Chapter 3, we would expect inventors nowadays located outside of knowledge hubs to likely experience the strongest gains from DSL activation. Chapter 3 provides the perhaps strongest links among essays of this dissertation to aspects of inequality and development, relevant in the context of knowledge access.

The insights from our results in Chapter 3 bare important implications obviously also for corporate strategy. As a matter of fact, our analysis shows effects on the population of private sector patents. Private firms were, indeed, among the strongest lobbyists for broadband access and towards accelerating the ADSL roll-out in the U.K., as they were fearing competitive disadvantages from being cut-off this new technology. Our results substantiate these concerns. They make a compelling case not only for the effectiveness of ICT for enhanced access to distant critical knowledge in firms, but they also suggest a true increase potential on corporate R&D productivity. By this, we add to prior findings by Forman and van Zeebroeck (2019), highlighting the usefulness of ICT adoption on facilitating knowledge exchange and intra-firm spillovers.

Limitations and Avenues for Future Research

A quality of good research consists in the acknowledgment of its limitations. The chapters of this dissertation are empirical essays which are intentionally constrained in the scope of their analyses under the scientific imperative of assuring validity of results. Some of the limitations arising from choices on empirical designs, sample delineation, or outcome measures pre-determine avenues for future research, other potentially interesting questions and opportunity for follow-up inquiry have arisen during the crafting of the chapters of the dissertation, by the nature of research typically only towards the end the process.

Chapter 1 is the only essay of the dissertation which is based on a (repeated) cross-sectional design and, thereby, cannot draw on sources of dynamic variation in parameters of the right-hand side of its empirical equation. While this allows us to evaluate the pattern of interest in its entirety and full scope of the universe of frontier

author, SNPR, and patent links in biomedical science for a long sample window, obviously what we can observe in this setting are only equilibrium-outcomes, baring residual concerns of a multitude of endogenous matches. These may become relevant when attempting to gauge implications for concrete policy instruments. Here, dynamic evaluation studies of specific policy interventions, or variation provided by natural experiments, may be important future complements to our findings and help to causally validate some of the mechanisms that our broad evidence suggest on concrete cases. There may be multiple options to identify sources of such intervention occurring in our sample window. Therefore, this seems like a viable and interesting avenue for future research. Furthermore, our paper in Chapter 1 takes the specific angle of links maximizing social technological value, which we assess through follow-up citations (i.e. cumulative impact). An interesting alternative perspective in this regard is, of course, the one of the private value of frontier science access. While high-impact patents will be, by definition, strongly linked to private value, a more accurate assessment of the private returns of frontier authors in corporate technology may help reconciling these traditionally colliding incentive dimensions in the context of knowledge access with new evidence. In this line of thought, also a more accurate analysis of eventual heterogeneity of effects for different backgrounds of authors appearing on corporate patents, based on their institutional affiliations, may be a highly-relevant open question to be investigated further. A promising avenue, building on a related but slightly different stream of literature, further consists in the assessment of a potential reciprocity of effects of technology involvement on the scientific output of frontier authors, perhaps helping to advance our understanding on the 'dual-frontier' of knowledge further.

As for Chapter 2, a certain drawback of my evaluation of the effectiveness of disease-specific repositories is perhaps the limitation of the analysis to the assessment of knowledge diffusion, while the paper is not able to say anything about effects of the online database on spurring knowledge production, e.g. in certain areas that were previously more 'hidden' in the HIV inventor community. While the focus on cumulative citations, as outcome measure of interest, is common to most related studies in the literature, it would be an interesting quest to pursue for future research to evaluate beneficial outcome effects of this or similar repositories on actual inventive activity. Furthermore, the AIDS database launch, situated in

the mid-1990s, has certainly the characteristics more of an historical case study than a of current setting. While it is an implicit assumption of the study, it would be particularly interesting to evaluate the persistence in relevance (and usefulness) of topic-specific collections (like the aforementioned COVID-19 databases, or the USPTO Patents4Partnership program) also 25 years later, in the era of more sophisticated search algorithms and artificial intelligence tools being increasingly employed in science and technology to predict useful combinations.

Finally, with regards to Chapter 3, an empirical limitation of our analysis relates to the few degrees of freedom that our 2SLS design leaves for the assessment of heterogeneity, which is partly related to the weakness of our instruments in smaller sub-samples. As discussed in the previous section, however, a particularly interesting research question may arise for heterogeneity of effects of broadband access on innovation in more remote areas. Solid evidence into this pattern would be warranted to infer more about the concrete potential of broadband technology to reduce the digital divide (beyond the availability of ICT) and, accordingly, the inequality between regions. An evaluation of such 'catching-up' mechanisms would be especially informative in large areal countries with a stronger divide between regions, such as the U.S. or Germany. A further promising avenue for research in the context of broadband and access to knowledge may arise with regards to entry and the policy-relevant question of 'who becomes an inventor?', particularly regarding issues of underrepresentation and efficiency in the allocation of talent in the ideas sector. This question could be relatively straightforward to assess, given the richness of information that fine-grained ICT data, patent records, and Census data provide. Lastly, a somewhat surprising insight emerging from our results that we do not find a significant impact of broadband access on communication and the formation new R&D collaborations. While first generation ADSL was not suitable for having video-calls or sending larger data files through the internet, based on observations of prior studies (e.g., on BitNet), we would have, to some extent expected, that - following the access to more distant codified knowledge - broadband would have exerted a sizeable effect also on the formation of new collaborations between more distant inventors. Revisiting and further explicating this result may provide an interesting question to be address for future work.

References

Ahmadpoor, Mohammad, and Benjamin F Jones. 2017. "The dual frontier: Patented inventions and prior scientific advance." *Science*, 357(6351): 583–587.

Allen, Thomas J, et al. 1977. "Managing the flow of technology: Technology transfer and the dissemination of technological information within the R&D organization." *MIT Press Books*, 1.

Arora, Ashish, Sharon Belenzon, and Andrea Patacconi. 2018. "The decline of science in corporate R&D." *Strategic Management Journal*, 39(1): 3–32.

Arora, Ashish, Sharon Belenzon, and Bernardo Dionisi. 2021. "First-mover Advantage and the Private Value of Public Science." National Bureau of Economic Research.

Arora, Ashish, Sharon Belenzon, and Lia Sheer. 2021. "Knowledge spillovers and corporate investment in scientific research." *American Economic Review*, 111(3): 871–98.

Baruffaldi, Stefano, and Felix Pöge. 2020. "A Firm Scientific Community: Industry Participation and Knowledge Diffusion." *IZA Discussion Paper No. 13419, Available at SSRN: https://ssrn.com/abstract=3643183*.

Berkes, Enrico, and Peter Nencka. 2020. "Knowledge access: The effects of Carnegie libraries on innovation."

Bertschek, Irene, Daniel Cerquera, and Gordon J Klein. 2013. "More bits–more bucks? Measuring the impact of broadband internet on firm performance." *Information Economics and Policy*, 25(3): 190–203.

Cockburn, Iain M, and Rebecca M Henderson. 1998. "Absorptive capacity, coauthoring behavior, and the organization of research in drug discovery." *The journal of industrial economics*, 46(2): 157–182.

Cohen, Wesley M, and Daniel A Levinthal. 1989. "Innovation and learning: the two faces of R & D." *The economic journal*, 99(397): 569–596.

Commission, European. 2005. "Frontier Research: The European Challenge. High-Level Expert Group Report." European Commission.

Czernich, Nina, Oliver Falck, Tobias Kretschmer, and Ludger Woessmann. 2011. "Broadband infrastructure and economic growth." *The Economic Journal*, 121(552): 505–532.

Forman, Chris, and Nicolas van Zeebroeck. 2012. "From wires to partners: How the Internet has fostered R&D collaborations within firms." *Management science*, 58(8): 1549–1568.

Forman, Chris, and Nicolas van Zeebroeck. 2019. "Digital technology adoption and knowledge flows within firms: Can the Internet overcome geographic and technological distance?" *Research policy*, 48(8): 103697.

Furman, Jeffrey L, and Scott Stern. 2011. "Climbing atop the shoulders of giants: The impact of institutions on cumulative research." *American Economic Review*, 101(5): 1933–63.

Furman, Jeffrey L, Markus Nagler, and Martin Watzinger. 2018. "Disclosure and subsequent innovation: Evidence from the patent depository library program." National Bureau of Economic Research.

Gittelman, Michelle, and Bruce Kogut. 2003. "Does good science lead to valuable knowledge? Biotechnology firms and the evolutionary logic of citation patterns." *Management Science*, 49(4): 366–382.

Iaria, Alessandro, Carlo Schwarz, and Fabian Waldinger. 2018. "Frontier Knowledge and Scientific Production: Evidence from the Collapse of International Science*." *The Quarterly Journal of Economics*, 133(2): 927–991.

Myers, Kyle. 2020. "The elasticity of science." *American Economic Journal: Applied Economics*, 12(4): 103–34.

Poege, Felix, Dietmar Harhoff, Fabian Gaessler, and Stefano Baruffaldi. 2019. "Science quality and the value of inventions." *Science advances*, 5(12): eaay7323.

Rosenberg, Nathan. 1990. "Why do firms do basic research (with their own money)?" *Research Policy*, 19(2): 165–174.

Thompson, Neil C., and Douglas Hanley. 2018. "Science is Shaped by Wikipedia: Evidence From a Randomized Control Trial." *Available at SSRN 3039505*.

Faculty of Economics and Business

Department of Management, Strategy and Innovation

Naamsestraat 69
3000 LEUVEN, België
tel. +32 16 326 700
thomas.schaper@kuleuven.be
www.kuleuven.be