assessment of low to very low. The authors did not report any factors warranting the rating up of quality to moderate. In addition, the sample sizes of the included randomised controlled trials for the outcomes of complete clinical improvement and bacterial eradication were relatively small; thus, we think that these two bodies of evidence should be rated as low quality and not moderate quality due to imprecision, in addition to the risk of bias and indirectness initially considered by the authors (appendix).

The GRADE system is used to rate the quality of bodies of evidence in systematic reviews and practice guidelines, and has been applied by more than 100 organisations and institutions worldwide.[4] However, GRADE is a specialised and complex approach, requiring training and experience in its proper application. Thus, GRADE is frequently misused by systematic reviewers and guideline developers.[5] Users should follow the *GRADE Handbook* and seek appropriate training when they use the GRADE approach to rate the quality of evidence in systematic reviews or practice guidelines.

We declare no competing interests.

*Meng Lv, Xufei Luo, *Yaolong Chen*
chenyaolong@lzu.edu.cn

Chevidence Lab of Child and Adolescent Health (ML, YC) and National Clinical Research Center for Child Health and Disorders, Ministry of Education Key Laboratory of Child Development and Disorders, China International Science and Technology Cooperation Base of Child Development and Critical Disorders (ML, YC), Children's Hospital of Chongqing Medical University, Chongqing, China; Chongqing Key Laboratory of Pediatrics, Chongqing 400014, China (ML, YC); School of Public Health (XL) and Research Unit of Evidence-Based Evaluation and Guidelines, Chinese Academy of Medical Sciences, School of Basic Medical Sciences (YC), Lanzhou University, Lanzhou, China

1 Uyttebroek S, Chen B, Onsea J, et al. Safety and efficacy of phage therapy in difficult-to-treat infections: a systematic review. *Lancet Infect Dis* 2022; published online March 3. https://doi.org/10.1016/S1473-3099(21)00612-5.

2 Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008; **336:** 924–26.

3 Schünemann H, Brożek J, Guyatt G, et al. GRADE handbook. October, 2013. https://gdt.gradepro.org/app/handbook/handbook.html (accessed March 11, 2022).

4 Alonso-Coello P, Schünemann HJ, Moberg J, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: introduction. *BMJ* 2016; **353:** i2016.

5 Gordon M, Guyatt G. Assessment of evidence quality in inflammatory bowel disease guidance: the use and misuse of GRADE. *Gastroenterology* 2020; **159:** 1209–15.

Authors' reply

We would like to thank Meng Lv and colleagues for their feedback on our systematic review.[1] Their Correspondence queried our assessment of the level of evidence of the included studies, which was done according to the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach.[2] GRADE is supported by the *Cochrane Handbook for Systematic Reviews of Interventions* and provides a framework for rating the quality of studies, presenting a summary of findings, and developing recommendations.[3] In the context of writing a systematic review, the *GRADE Handbook*[2] provides step-by-step guidance on presenting results in an evidence table.

In our systematic review, two researchers independently screened the included studies for risk of bias, inconsistency, imprecision, indirectness, and publication bias, as prescribed by GRADE. During this process, we were critical towards the included studies due to the paucity of high-quality randomised controlled trials, an issue that we also extensively disclosed in our discussion. We understand the comments made by Lv and colleagues with respect to our assessment of the observational studies. We agree that during this process the level of evidence might have been overestimated for these studies. Although GRADE has many advantages compared with previous grading systems (eg, improved transparency), the GRADE Working Group states that the GRADE approach does not preclude the use of authors' own judgment and highlights that categorisation involves some arbitrariness: "Therefore, GRADE is not a quantitative system for grading the quality of evidence. Each factor for downgrading or upgrading reflects not discrete categories but a continuum within each category and among the categories. When the body of evidence is intermediate with respect to a particular factor, the decision about whether a study falls above or below the threshold for up- or downgrading the quality (by one or more factors) depends on judgment".[2]

See **Online** for appenidx

We also acknowledge that we did not consider imprecision as a limiting factor to determine the quality of one randomised controlled trial on bacterial eradication, as appropriate power calculations were done by the authors.[4] Nevertheless, the conclusions and secondary clinical implications of our work are not affected by the proposed corrections as we concluded that the overall quality of the evidence was very low to moderate and strong conclusions on safety and efficacy could not be made. We thank Lv and colleagues for the valuable feedback and opening this very interesting discussion on the integration of the GRADE approach in systematic reviews.

We declare no competing interests.

*Saartje Uyttebroek, Jolien Onsea, Laura Van Gerven, *Willem-Jan Metsemakers*
willem-jan.metsemakers@uzleuven.be

Department of Otorhinolaryngology (SU, LVG) and Department of Trauma Surgery (JO, W-JM), UZ Leuven, Leuven 3000, Belgium; Department of Neurosciences, Experimental Otorhinolaryngology, Rhinology Research (SU, LVG), Department of Development and Regeneration, Locomotor and Neurological Disorders (JO, W-JM), and Department of Microbiology, Immunology and Transplantation, Allergy and Clinical Immunology Research Group (LVG), KU Leuven, Leuven, Belgium

1 Uyttebroek S, Chen B, Onsea J, et al. Safety and efficacy of phage therapy in difficult-to-treat infections: a systematic review. *Lancet Infect Dis* 2022; published online March 3. https://doi.org/10.1016/S1473-3099(21)00612-5.

2 Schünemann H, Brożek J, Guyatt G, et al. GRADE handbook. October, 2013. https://gdt.gradepro.org/app/handbook/handbook.html (accessed March 22, 2022).

3 Schünemann HJ, Higgins JPT, Vist GE, et al. Chapter 14: completing 'summary of findings' tables and grading the certainty of the evidence. Cochrane handbook for systematic reviews of interventions, version 6.3. February, 2022. https://training.cochrane.org/handbook/current/chapter-14 (accessed March 22, 2022).

4 Wright A, Hawkins CH, Anggård EE, Harper DR. A controlled clinical trial of a therapeutic bacteriophage preparation in chronic otitis due to antibiotic-resistant *Pseudomonas aeruginosa*; a preliminary report of efficacy. *Clin Otolaryngol* 2009; **34:** 349–57.

Evaluating BLOOMY and SOFA scores in hospitalised patients

We congratulate Evelina Tacconelli and colleagues[1] on developing the Bloodstream Infection due to Multidrug-resistant Organisms: Multicenter Study on Risk Factors and Clinical Outcomes (BLOOMY) prediction scores. Among patients admitted to hospital with bloodstream infection, the BLOOMY 14-day score had a C statistic of 0·873 for mortality, while the simplified quick BLOOMY score had a C statistic of 0·828. Strengths of this commendable study include prospective multicentre data collection.

However, the Article raises important questions. First, because of variable patient-level baseline risks, subgroup analyses are essential to determine the degree of heterogeneity in these variables' predictive performance across different populations. Although the Methods describe subgroup analyses, we could not find these results in the main Article or its appendix. Second, the BLOOMY and quick BLOOMY scores were compared only indirectly to the Sequential Organ Failure Assessment (SOFA) and quick SOFA (qSOFA) scores[1] due to unavailability of respiratory rate. Such indirect comparisons are often not valid due to differential case mix and differences in clinical practices between model development populations.[2]

We are particularly interested in BLOOMY's performance among patients with cancer, because of the prevalence of bloodstream infection in this group[3] and the potential for short-term risk estimates to influence decisions on cancer-directed therapies and supportive care. We applied BLOOMY, quick BLOOMY, SOFA, and qSOFA to electronic health record data from a single-centre cohort of oncology patients meeting BLOOMY inclusion criteria from June 1, 2018, to June 30, 2021.[4] We compared 14-day mortality C statistics (BLOOMY 14-day *vs* SOFA; quick BLOOMY *vs* qSOFA).

Of 844 patients, 33 (4%) died within 14 days of blood culture collection. C statistics for 14-day mortality did not differ between BLOOMY (0·734 [95% CI 0·659–0·810]) and SOFA (0·721 [0·637–0·804]; p=0·75) or between quick BLOOMY (0·739 [0·664–0·813]) and qSOFA scores (0·712 [0·629–0·794]; p=0·30).

Our findings have important implications. First, although cancer-related bloodstream infection has been identified as a risk factor for mortality,[3] mortality was lower in our cohort than Tacconelli and colleagues' study. Second, we found lower discrimination for BLOOMY than for SOFA and quick BLOOMY than for qSOFA in our cohort. These findings probably indicate a so-called dataset shift—ie, differential case mix, epidemiology, and practices between cohorts.[5] We hope Tacconelli and colleagues can report their malignancy-specific results to contextualise our findings. Finally, BLOOMY and quick BLOOMY did not outperform SOFA and qSOFA in our cohort. The newly developed scores, despite using many of the same predictors, are more complex than SOFA and qSOFA. Without improvement within the context of bloodstream infection, the value of using such models is unclear. Thus, we urge further external validation of BLOOMY and quick BLOOMY, particularly among patients with cancer, before widespread adoption.

We declare no competing interest. The code for this project is freely available upon request to the corresponding author.

*Nicole Benzoni, Alice F Bewley, M Cristina Vazquez-Guillamet, *Patrick G Lyons*
plyons@wustl.edu

Department of Medicine (NB, AFB, MCV-G, PGL) and Siteman Cancer Center (PGL), Washington University School of Medicine in St Louis, St Louis, MO 63110, USA; Healthcare Innovation Lab, BJC HealthCare, St Louis, MO, USA (PGL)

1 Tacconelli E, Göpel S, Gladstone BP, et al. Development and validation of BLOOMY prediction scores for 14-day and 6-month mortality in hospitalised adults with bloodstream infections: a multicentre, prospective, cohort study. *Lancet Infect Dis* 2022; published online Jan 19. https://doi.org/10.1016/S1473-3099(21)00587-9.

2 Collins GS, Moons KGM. Comparing risk prediction models. *BMJ* 2012; **344:** e3186.

3 Hensley MK, Donnelly JP, Carlton EF, Prescott HC. Epidemiology and outcomes of cancer-related versus non-cancer-related sepsis hospitalizations. *Crit Care Med* 2019; **47:** 1310–16.

4 Lyons PG, Klaus J, McEvoy CA, Westervelt P, Gage BF, Kollef MH. Factors associated with clinical deterioration among patients hospitalized on the wards at a tertiary care hospital. *J Oncol Pract* 2019; **15:** e652–65.

5 Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. *N Engl J Med* 2021; **385:** 283–86.

Authors' reply

We thank Nicole Benzoni and colleagues for their Correspondence and for sharing the results of an assessment of the 14-day mortality Bloodstream Infection due to Multidrug-resistant Organisms: Multicenter Study on Risk Factors and Clinical Outcomes (BLOOMY) score[1] in a retrospective cohort of US-based hospitalised patients with cancer and bloodstream infections. We are pleased to see that in the C statistics the 14-day BLOOMY score in the assessed population was slightly better than the Sequential Organ Failure Assessment (SOFA) score. Although retrospective assessment of the score might imply inadequate or missing data or ambiguous use of parameters from a different timepoint other than day 3 (the timepoint of application of the BLOOMY 14-day