Quality of conduct and reporting of propensity score methods in studies investigating the effectiveness of antimicrobial therapy

Eikenboom, A.M.; Cessie, S. le; Waernbaum, I.; Groenwold, R.H.H.; Boer, M.G.J. de

Open Forum Infectious Diseases

MAJOR ARTICLE

Quality of Conduct and Reporting of Propensity Score Methods in Studies Investigating the Effectiveness of Antimicrobial Therapy

Anna M. Eikenboom,[1] Saskia Le Cessie,[2,3] Ingeborg Waernbaum,[4] Rolf H. H. Groenwold,[2,3] and Mark G. J. de Boer[1]

[1]Department of Infectious Diseases, Leiden University Medical Centre (LUMC), Leiden, the Netherlands, [2]Department of Clinical Epidemiology, Leiden University Medical Centre (LUMC), Leiden, the Netherlands, [3]Department of Biomedical Data Sciences, Leiden University Medical Centre (LUMC), Leiden, the Netherlands, and [4]Department of Statistics, Umea University, Umea, Sweden

Background. Propensity score methods are becoming increasingly popular in infectious disease medicine to correct for confounding in observational studies. However, applying and reporting propensity score techniques correctly requires substantial knowledge of these methods. The quality of conduct and reporting of propensity score methods in studies investigating the effectiveness of antimicrobial therapy is yet undetermined.

Methods. A systematic review was performed to provide an overview of studies (2005–2020) on the effectiveness of antimicrobial therapy that used propensity score methods. A quality assessment tool and a standardized quality score were developed to evaluate a subset of studies in which antibacterial therapy was investigated in detail. The scale of this standardized score ranges between 0 (lowest quality) and 100 (excellent).

Results. A total of 437 studies were included. The absolute number of studies that investigated the effectiveness of antimicrobial therapy and that used propensity score methods increased 15-fold between the periods 2005–2009 and 2015–2019. Propensity score matching was the most frequently applied technique (65%), followed by propensity score–adjusted multivariable regression (25%). A subset of 108 studies was evaluated in detail. The median standardized quality score per year ranged between 53 and 61 (overall range: 33–88) and remained constant over the years.

Conclusions. The quality of conduct and reporting of propensity score methods in research on the effectiveness of antimicrobial therapy needs substantial improvement. The quality assessment instrument that was developed in this study may serve to help investigators improve the conduct and reporting of propensity score methods.

Keywords. propensity score methods; infectious diseases; antimicrobial therapy.

In infectious disease medicine, it is of great importance to investigate the effectiveness of antimicrobial therapies in an efficient and valid manner [1]. Increasing antimicrobial, and especially antibacterial, resistance and newly emerging infectious diseases such as severe acute respiratory syndrome coronavirus 2 (SAR-CoV-2) create the need for rapid development of antimicrobial therapy. Randomized controlled trials (RCTs) are considered the gold standard to investigate treatment effects [2, 3]. However, RCTs for many antimicrobial treatment decisions may not be feasible, may be unethical, or may be too costly or not timely enough [3]. Therefore, observational studies are commonly performed to investigate treatment effects of antimicrobial therapy [1]. However, because in observational studies treatment assignment is not a random process, direct comparison of treatment groups without taking baseline differences into account may lead to incorrect conclusions due to confounding. One of the approaches to correct for measured confounding is the use of propensity score methods [4–6]. These methods attempt to balance the observed baseline covariates between the treatment and control groups. Within participants with the same propensity score, the distribution of the observed covariates in the treated and untreated groups would be approximately the same, similar to an RCT [4].

Due to increasing antimicrobial resistance and rapidly evolving insights in antimicrobial therapy, it is not surprising that also in the field of infectious diseases the popularity of propensity score methods has increased greatly in the past 15 years [5, 7, 8]. However, applying propensity score techniques correctly requires substantial knowledge on propensity score analysis and its underlying assumptions. What's more, studies in which propensity methods are applied should report sufficient details of the analyses that were performed [4]. Therefore, a quality assessment instrument consisting of a set of quality

Received 30 December 2021; editorial decision 24 February 2022; accepted 6 March 2022; published online 7 March 2022.

Correspondence: Mark G. J. de Boer, MD, PhD, FIDSA, Leiden University Medical Center, Albinusdreef 2, 2333 ZA Leiden, the Netherlands (m.g.j.de_boer@lumc.nl).

Open Forum Infectious Diseases®2022

https://doi.org/10.1093/ofid/ofac110

criteria is needed to ensure quality of conduct and reporting of studies using propensity score methods [9–11]. In addition, it is still unknown how in the field of infectious disease medicine the quality of conduct and reporting of studies using propensity score methods has evolved over the past 15 years.

The aim of this study was to provide an overview of the use of propensity score methods in studies investigating the effectiveness of antimicrobial therapy between 2005 and 2020, to develop a quality assessment instrument for studies using propensity score methods, and to use this tool to assess the quality of conduct and reporting of the studies over time.

METHODS

The study consisted of a systematic review and development of a quality assessment tool as well as a quality assessment. The study structure and data flow are presented in Figure 1.

Systematic Review of the Use of Propensity Score Methods in Studies Investigating the Effectiveness of Antimicrobial Therapy

A systematic review was conducted following the PRISMA criteria (Supplementary Data 1) and was registered in the Prospero database (registration number: CRD42020210473) [12]. This systematic review describes (a) the number of studies investigating the effectiveness of antimicrobial therapy in which methods were applied that have been published in the past 15 years, (b) the infectious diseases that were investigated, and (c) how often different propensity score techniques have been used. PubMed was searched using a search strategy that was carefully designed in collaboration with a librarian (Supplementary Data 2). Titles and abstracts were screened to find studies that fulfilled the following eligibility criteria: original research of observational data; a main study aim of estimating the effectiveness (note: not safety) of antimicrobial therapy (ie, antibacterial, -fungal, -viral, and -parasitic therapy); use of propensity score methods in the analysis; published between September 1, 2005, and September 1, 2020; and written in English. Case reports, meta-analyses, reviews, abstracts, and protocols were excluded.

Quality Assessment of Conduct and Reporting of Studies Using Propensity Score Methods

A quality assessment instrument for studies using propensity score methods was developed inspired by the general principles of the Delphi method to reach consensus [13, 14]. The list of recommendations developed by Yao et al. was used as a starting point. On the basis of additional literature on propensity score methods, modifications were made to the quality criteria suggested by Yao et al., and quality criteria were added or removed from the list [4–6, 11, 15–24]. The tool was developed to assess if sufficient details were reported on the propensity score method used, whether assumptions of propensity score methods were discussed, and whether the balance of baseline variables before and after propensity score analysis was checked. The list was discussed among experts, and a quality assessment tool was drafted. Subsequently, 3 independent experts reviewed the

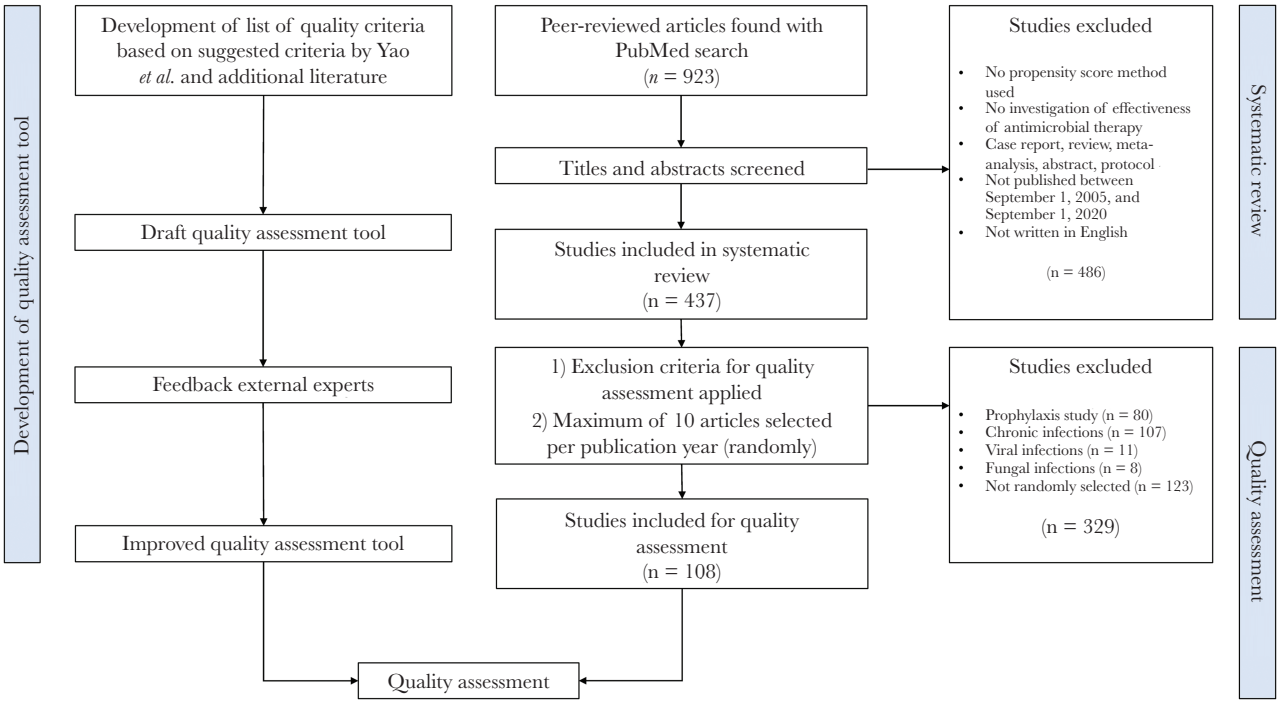

Figure 1. Study structure and data flow.

quality assessment tool using a feedback form (Supplementary Data 3). Following the principles of the Delphi method, improvements were made if similar feedback was provided by at least 2 out of 3 experts.

A subset of the papers that were included in the systematic review was evaluated in detail using the quality assessment instrument. The subset consisted of studies in which the effect of an antimicrobial therapy for (sub)acute bacterial infections was investigated. Studies that concerned fungal infections, parasitic infections, viral infections, and chronic infections and studies in which antibiotic prophylaxis was investigated were excluded from this part of the analyses. Due to time constraints, it was not feasible to include all eligible studies for quality assessment. If >10 studies met the inclusion criteria in 1 publication year, 10 studies were randomly selected. Every article received a random number, generated by using a random number generator. Then, articles were sorted by publication year and ascending random numbers. Per publication year, the 10 studies that received the lowest random numbers were included. Selected studies were reviewed and assessed by application of the quality assessment tool (by A.E.). The results were subsequently discussed in a larger team (by A.E., M.B., and S.C.). Because there were more quality criteria for studies that used propensity score matching and IPTW than for stratification and covariate adjustment using propensity scores, the maximum score depended on the propensity score method used. Therefore, total scores were standardized to a value between 0 and 100 by dividing the total score by the maximum score that could be achieved and then multiplying by 100. For the purpose of this study, the standardized score is further referred to as the Standardized Quality Score for Propensity score Methods (SQSPM).

Statistical Analyses

Categorical variables were reported using percentages; continuous variables were reported using means with standard deviations or, in case of skewed variables, medians with interquartile ranges. Articles included for systematic review were categorized per propensity score method used. The systematic review identified 4 propensity score techniques: matching, inverse probability of treatment weighting (IPTW or IPW), stratification, and propensity score–adjusted regression [4]. When the propensity score method used was not explicitly mentioned, the methods used were deducted from the "Methods" and "Results" sections of the paper. The number of studies using propensity score methods and the distribution of the propensity score methods used were both calculated per year.

The median SQSPM per year was determined, and the SQSPM was compared between different propensity score methods and types of infectious diseases. Summaries of the scores per criterion were provided to investigate which criteria were frequently met and which criteria need attention in future research. The degree of concordance between different items of the quality score was assessed using Pearson correlation coefficients and visualized using a heatmap. In this way, patterns of items (ie, clusters) that are often reported poorly could be observed. All analyses were conducted using SPSS, version 25.0 (IBM, Armonk, NY, USA).

RESULTS

Systematic Review

The systematic search strategy yielded 923 unique peer-reviewed studies. For the systematic review, 437 studies fulfilled the eligibility criteria, and a subset of 108 studies was included for quality assessment. A bibliography of the studies that were included in the systematic review and quality assessment can be found in Supplementary Data 4. The data flow is described in Figure 1.

Table 1 shows an overview of the number of studies in which propensity score methods were applied by infectious disease category and propensity score method. Overall, propensity score matching was the most frequently used propensity score method, with 65% of the studies overall using matching, followed by propensity score–adjusted regression, which was used in 25% of the included articles. In Figure 2A, the number of studies per year is reported.

The absolute number of studies investigating effectiveness of antimicrobial therapy and using propensity score increased 15-fold between the periods 2005–2009 and 2015–2019. In Figure 2B, the relative frequency of the use of the 4 propensity score methods per year is depicted. Initially, propensity score–adjusted regression was the most frequently used propensity score method, but its popularity decreased over time. Stratification was frequently applied in the early years as well but almost disappeared after 2017. IPTW and propensity score matching increased in popularity over the years.

Quality Assessment

The draft quality assessment tool consisted of twenty criteria of which criterion fourteen consisted of three sub-criteria (Supplementary Data 6). Based on the feedback from the independent experts, several adjustments were made. The definitive quality assessment tool (Table 2) consists of 18 criteria. For quality assessment, 108 studies were included. Table 3 shows that 66% of these studies received an SQSPM between 50 and 70. Overall, scores ranged between 33 and 88. The score category 80–90 was reached by 2 studies, in which propensity score matching was used. Both articles lacked a complete description of the influence of missing data on propensity score estimation (criterion 18) and the details of propensity score analysis (criterium 7). The discussion of the positivity assumption (ie, that all participants are able to receive both treatments, meaning that estimated propensity scores should not be too close to 0 or 1) (17b), the sensitivity analysis (10), and balance after propensity

Table 1. Number of Studies in Which Propensity Score Methods Were Applied per Infectious Disease Category and Propensity Score Method

Type of Infection	UTI (n[a] = 15), No. (%)	GI (n = 16), No. (%)	Pneumonia[b] (n = 58), No. (%)	Sepsis and BSI (n = 97), No. (%)	Prophylaxis[c] (n = 80), No. (%)	HIV (n = 23), No. (%)	Hepatitis B (n = 42), No. (%)	Hepatitis C (n = 37), No. (%)	Other (n = 69), No. (%)	Total (n = 437), No. (%)
Propensity score method										
Matching	8 (53.3)	11 (68.8)	32 (55.2)	57 (58.8)	50 (62.5)	11 (47.8)	38 (90.5)	33 (89.2)	42 (60.9)	282 (64.5)
IPTW	5 (33.3)	3 (18.8)	12 (20.7)	16 (16.5)	15 (18.8)	6 (26.1)	8 (19.0)	3 (8.1)	10 (14.5)	78 (17.8)
Stratification	1 (6.7)	1 (6.3)	7 (12.1)	3 (3.1)	4 (5.0)	2 (8.7)	0 (0.0)	0 (0.0)	9 (13.0)	27 (6.2)
Covariate adjustment using propensity scores	4 (26.7)	1 (6.3)	15 (25.9)	38 (39.2)	17 (21.3)	7 (30.4)	1 (2.4)	1 (2.7)	24 (34.8)	108 (24.7)

Abbreviations: BSI, bloodstream infection; GI, gastrointestinal infection; IPTW, inverse probability of treatment weighting; UTI, urinary tract infection.

[a]n = the number of studies within the infectious disease category. In multiple studies, >1 propensity score method was applied. Therefore, the numbers of the different propensity score methods within the infectious disease category add up to more than the total number of studies included in the infectious disease category. Percentages of different propensity score methods were calculated on the total number of studies within in the infectious disease category. Therefore, the percentages add up to >100%.

[b]Pneumonia and respiratory tract infections, including influenza infection.

[c]Prophylaxis: studies in which the effectiveness of prophylactic antimicrobial therapy was investigated.

score analysis (14b) were reported incompletely in both of the articles. In Figure 2B, the median SQSPM per year is depicted; it fluctuated between 53 and 61 from 2005 until 2020. There was no improvement of quality of conduct and reporting over the years. For every propensity score technique separately, a similar trend was observed (Supplementary Data 7). No differences in scores were observed between different types of infectious diseases (unpublished data).

Scores on Different Criteria and Concordance

In Figure 3, the percentages of articles that completely fulfilled the criteria are reported by criterion. The 5 criteria that were met most frequently were criteria that require description of the use of propensity score methods in the abstract or title (criterion 1), the propensity score method used (criterion 3), the statistical methods used to analyze the data after applying the propensity score method (criterion 9), the software used for analysis (criterion 11), and the sample size before and after matching (in matching studies; criterion 12). These criteria were all met in at least 90% of evaluated studies. The 5 criteria that were met least frequently were criteria that require description of the sensitivity analysis that was used (criterion 10), the distribution of propensity scores in both treatment groups (criterion 14c), the distribution of the weights (in IPTW studies; criterion 15), discussion of the positivity assumption (criterion 17b), and the influence of missing data on propensity score estimation (criterion 18). These criteria were met in ≤15% of evaluated studies.

The concordance in scores between criteria is reported in a heat map (Supplementary Data 8). A high level of concordance was observed between criteria that concern checking of balance and criteria that require discussion of underlying assumptions of propensity score methods. A high level of concordance was also observed between criteria requiring a detailed description of the propensity score model in the "Methods" section.

DISCUSSION

In the period of study, we observed a large absolute increase in the number of published peer-reviewed studies investigating the effectiveness of antimicrobial therapy in which propensity score methods were applied. The results of the quality assessment showed that the quality of conduct and reporting of propensity score methods in these studies was far from optimal, with the majority of the studies having a standardized quality score between 50 and 70 out of 100. We also found that the quality of conduct and reporting did not improve over time. Quality criteria that were more specific in studies using propensity score methods were less often met than more generic quality criteria. Furthermore, the concordance analysis showed that often many details of propensity score analysis were reported, or none were reported, and that often all assumptions of propensity score methods were discussed, or none were discussed. This indicates

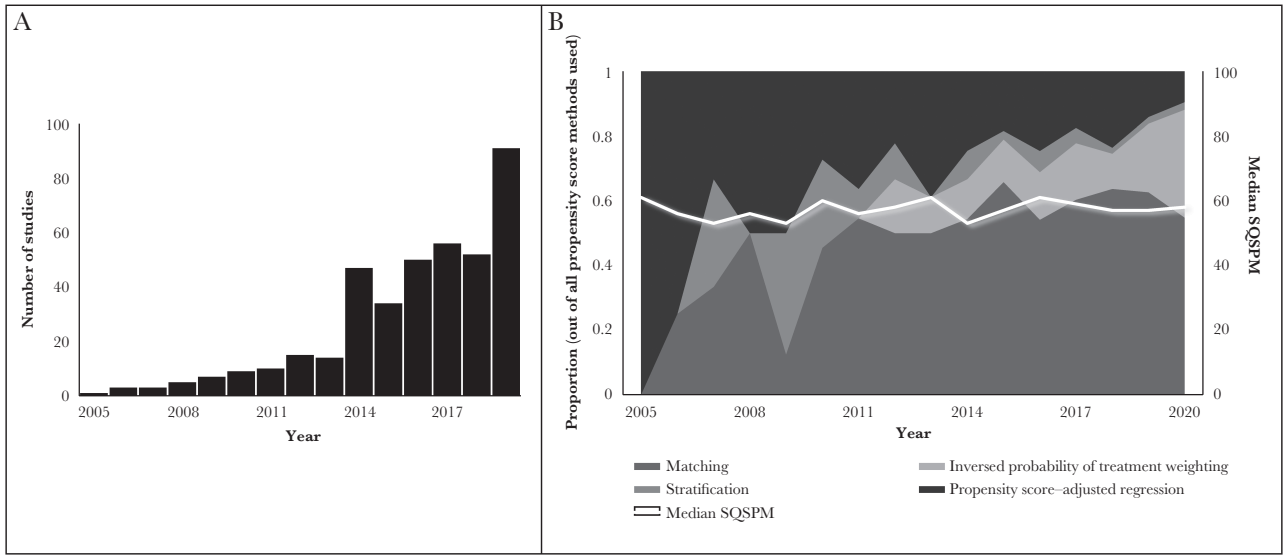

Figure 2. The frequency and quality of use of propensity score methods in studies investigating effectiveness of antimicrobial therapy. A, 2020 was not included in this figure, as we only included articles up to September 1, 2020. B, Distribution of propensity score methods used per year, median standardized quality scores per year in studies in which the effectiveness of antimicrobial therapy was investigated, and propensity score methods are applied. Abbreviation: SQSPM, Standardized Quality Score for Propensity score Methods.

that some researchers may not have been sufficiently aware of the concept of discussing underlying assumptions or providing details of the analysis methods that were used.

The fact that the quality of conduct and reporting of propensity score methods is far from optimal has been observed in other fields of research, for example, oncology [11], cardiovascular surgery [15], in high-ranked journals in different disciplines [10], and in a comprehensive quality assessment that included all fields of the medical literature [25]. In these papers, suggestions for improvement were provided, including describing the process of the propensity score model development and the propensity score analysis in detail, checking balance after applying propensity score methods, and discussion of assumptions of propensity score methods [10, 11, 15, 25]. Furthermore, previous research has shown that the quality of reporting of confounding in general was still far from ideal after the publication of the STROBE guideline [26].

Although in infectious disease research a systematic review on quality of conduct and reporting of propensity score methods has not previously been performed, attention to the importance of careful use of propensity score methods has been raised. For example, in a letter to the editor of this journal, Roth et al. argued that studies using propensity score methods in infectious disease medicine should be conducted and reported in a more standardized manner [9]. In an educational paper, Amoah et al. demonstrated in a case study how different propensity score methods and standard regression methods could be used correctly [27].

The quality assessment instrument that was developed in this study describes the spectrum of methodological and reporting standards. The instrument can be used by reviewers and researchers to improve quality of conduct and reporting of studies using propensity score methods in the field of antimicrobial therapy and in other fields of clinical research. It is important to emphasize that the instrument should not be seen as a scoring tool that provides an absolute indication of quality of conduct and reporting of such a study. In our quality assessment, all criteria had equal weight, and we used the SQSPM to count the number of criteria that were fulfilled and to compare these numbers between type of methods and over the years. However, to calculate a score that provides absolute indication of quality, several criteria should receive more weight than others. In particular, checking the balance after applying a propensity score method and discussion of the assumptions of propensity score methods should probably be weighted more heavily. Therefore, the quality assessment tool should rather be seen as a set of quality criteria that should all be fulfilled and can be used as a checklist to evaluate which items are still unaddressed. Still, even when all quality criteria would be fulfilled, propensity score methods only address measured confounders. Therefore, to ensure correct conduct of propensity score methods, the risk of unmeasured confounding should always be estimated before applying propensity score methods.

Strengths and Limitations

A strength of this study is that we first presented a detailed overview of the use of propensity score methods in studies investigating effectiveness on antimicrobial therapy, which is particularly helpful to put the results of the quality assessment into perspective. This outlined the relevance of improvement of

Table 2. Quality Assessment Tool for Studies Using Propensity Score Methods

	Criterion No.	Criterion[a]	Not Applicable to	Score
Title and abstract				
	1	The use of propensity score analysis is indicated with a commonly used term in the title or the abstract.		
		Yes		1
		No		0
Methods				
	2	Motivation[b] for using propensity score methods is indicated.		
		Yes		1
		No		0
	3	It is described which propensity method is used (if >1, consider the primary analysis).		
		Yes, matching		1
		Yes, weighting		1
		Yes, stratification		1
		Yes, covariate adjustment using propensity score		1
		No		0
	4	It is indicated which method is used to estimate the propensity score.		
		Yes, a logistic model		1
		Yes, boosting (meta-classifiers)		1
		Yes, decision trees		1
		Yes, other, namely:		1
		No		0
	5	The process of variable selection for the propensity score model is described.		
		Yes, variables are specified beforehand		1
		Yes, variables are selected with a statistical selection method		1
		Yes, some variables are specified beforehand, and others are selected with a statistical selection method		1
		Yes, other, namely:		1
		No		0
	6	The variables included in the propensity score model are described.		
		Yes		1
		No		0
	7	Details of propensity score analysis are described: a. Details that should be described for propensity score matching: matching algorithm, caliper, matching ratio, with/without replacement. b. Details that should be described for propensity score weighting: It is clear how weights are obtained. c. Details that should be described for propensity score stratification: The number of strata is provided; strata are defined clearly.		
		Yes		1
		Incomplete		0.5
		No		0
	8	Methods to assess comparability of baseline characteristics after applying a propensity score method are described.		
		Yes		1
		Incomplete		0.5
		No		0
	9	Statistical methods to analyze data after applying a propensity score method are described.		
		Yes		1
		Incomplete		0.5
		No		0
	10	Are sensitivity analyses performed to test the robustness of the propensity score method used?		
		Yes, complete description		1
		Yes, incomplete description		0.5
		No, no sensitivity analyses performed		0
	11	The software used for analysis is indicated. If propensity score matching was used: The package used to create matched sample is described as well.		

Table 2. Continued

Criterion No.	Criterion[a]	Not Applicable to	Score
	Yes, namely:		1
	No		0
Results			
12	Sample size for each treatment group before and after matching is reported.	IPTW, stratification, covariate adjustment using propensity score	
	Yes		1
	No		0
	Not applicable		0
13	The distribution of baseline characteristics for each group before propensity score analysis is described.		
	Yes		1
	Incomplete		0.5
	No		0
14a	After propensity score matching, weighting, or stratification: The distribution of baseline characteristics in the matched/weighted groups or in each stratum is reported.	Covariate adjustment using propensity score	
	Yes		1
	Incomplete		0.5
	No		0
	Not applicable		0
14b	After propensity score matching, weighting, or stratification: It has been checked whether sufficient balance has been achieved (love plot, standardized mean difference, etc.).		
	Yes		1
	Incomplete		0.5
	No		0
14c	After propensity score matching, weighting, or stratification: The distribution of the propensity scores in both treatment groups is described (in plot or text).		
	Yes		1
	Incomplete		0.5
	No		0
15	The distribution of the size of the weights is described.	PSM, stratification, covariate adjustment using propensity score	
	Yes		1
	No		0
	Not applicable		0
16	The number of patients with missing data for each variable of interest for the propensity score analysis is reported.		
	Yes		1
	No		0
Discussion			
17a	Modeling assumptions are met: The no unmeasured confounders assumption is discussed.		
	Yes		1
	No		0
17b	Modeling assumptions are met: It is discussed whether there is sufficient overlap to perform propensity score analysis (positivity assumption).		
	Yes		1
	No		0
18	The influence of missing data in propensity score estimation and missing data due to incomplete matching is discussed.		
	Yes		1
	Incomplete		0.5
	No		0

Abbreviations: IPTW, inversed probability of treatment weighting; PSM, propensity score matching.

[a]The quality assessment tool is based on the suggested quality criteria by Yao et al. [11], additional literature on propensity score methods, and discussion between experts.

[b]For this quality assessment, the use of propensity score methods was considered to be motivated when somewhere in the article it was at least mentioned that propensity score methods were used to address confounding.

Table 3. Standardized Quality Scores for Propensity Score Methods in Studies Investigating the Effectiveness of Antimicrobial Therapy per Propensity Score Method Category

SQSPM	All Studies (n = 108)	Propensity Score Method Used (No. of Studies)				
		Matching (n = 50)	IPTW (n = 11)	Stratification (n = 9)	Propensity Score Adjusted Regression (n = 36)	Propensity Score Method Not Specified (n = 2)
>30–40, No. (%)	5 (4.6)	0	3 (27.3)	0	1 (2.8)	1 (50.0)
>40–50, No. (%)	21 (19.4)	10 (20.0)	2 (18.2)	4 (44.4)	5 (13.9)	1 (50.0)
>50–60, No. (%)	37 (34.3)	16 (32.0)	2 (18.2)	3 (33.3)	15 (41.7)	0
>60–70, No. (%)	34 (31.5)	17 (34.0)	4 (36.4)	0	14 (38.9)	0
>70–80, No. (%)	9 (8.3)	5 (10)	0	2 (22.2)	1 (2.8)	0
>80–90, No. (%)	2 (1.9)	2 (4.0)	0	0	0	0
>90–100, No. (%)	0	0	0	0	0	0
Mean SQSPM (±SD)	58 (± 10)	61 (±10)	53 (±11)	56 (±12)	57 (±8)	40 (± 10)

Abbreviations: IPTW, inverse probability of treatment weighting; SQSPM, Standardized Quality Score for Studies using Propensity score Methods.

quality of conduct and reporting of propensity score methods and provided context for the quality assessment tool that was developed in this study. Furthermore, the quality assessment tool was developed in a structured and stepwise manner. This study was limited in a few respects. To keep the quality assessment tool practical and concise, several quality criteria had to be prioritized over others. For example, there are >2 assumptions of propensity score methods that could be discussed, but we included the 2 assumptions that were considered most important. Another limitation is that the initial quality assessment was performed by one of the investigators and that, as they become more experienced, investigators may become more (or less) strict in their assessments. However, after the quality assessment, 10 articles were re-assessed, and no major discrepancies were found. The above-named factors could have influenced the mean scores, but did not likely influence the overall outcome patterns. Of note, studies were assessed based on conduct and reporting, but it was not a study aim to evaluate if the obtained results were valid (eg, by assessing the likelihood that there would be unmeasured confounding). Previous research showed that methodological methods such as propensity scores may not always be mentioned in the title or abstract [28]. Therefore, it is possible that the quality of reporting of propensity score methods is even less optimal than reported here.

CONCLUSIONS

From the results of this study, the conclusion can be drawn that the quality of conduct and reporting of propensity score methods in studies investigating antimicrobial therapy needs substantial improvement. The quality assessment instrument constructed in this study can be used as a starting point for designing, conducting, and reporting a study in which propensity score methods are applied. Even so, the instrument can assist in reviewing an article of a study in which propensity score methods are used to evaluate if all requirements are met. Optimally, guidelines should be developed and incorporated in tools such as STROBE or ROBINS-I [29, 30]. By doing this, the quality of conduct and reporting of these increasingly popular statistical methods can be improved. This would structurally contribute to the validity of research on the effectiveness of antimicrobial therapy.

Supplementary Data

Supplementary materials are available at *Open Forum Infectious Diseases* online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

Acknowledgments

We would like to thank Dr. S. A. Swanson for providing feedback on the quality assessment tool and for commenting on the manuscript. We

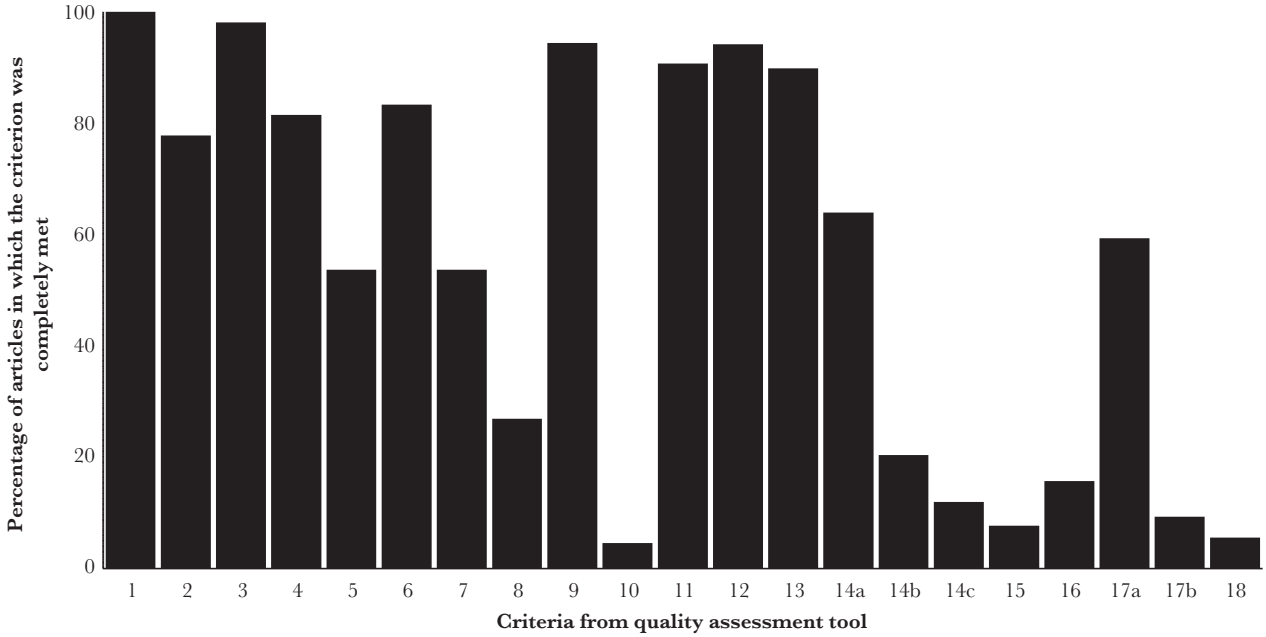

Figure 3. Percentage of studies that fulfilled the criteria of the quality assessment tool. See Table 2 for the definition of the criteria. For every criterion, only the studies to which the criterion applied were included.

would like to thank Mrs. E. P. Jansma for assisting with the literature search strategy.

Financial support. No external funding.

Potential conflicts of interest. All authors declare that there is no conflict of interest. All authors: no reported conflicts of interest. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

Author contributions. M.B. and S.C. conceived the idea of this study. A.E. performed the systematic review and the quality assessment. All authors contributed to the construction of the final quality assessment tool. The first version of the manuscript was written by A.E., and consecutive revisions were provided by M.B., S.C., R.G., and I.W.

Patient consent. Not applicable (this study does not include factors necessitating patient consent).

References

1. Frieden TR. Evidence for health decision making - beyond randomized, controlled trials. N Engl J Med **2017**; 377:465–75.
2. Evans I, Thornton H, Chalmers I, Glasziou P. Testing Treatments: Better Research for Better Healthcare. Pinter & Martin; **2011**.
3. Sibbald B, Roland M. Understanding controlled trials. Why are randomised controlled trials important? BMJ **1998**; 316:201.
4. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res **2011**; 46:399–424.
5. Beal S, Kupzyk K. An introduction to propensity scores: what, when, and how. J Early Adolesc **2014**; 34:66–92.
6. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika **1983**; 70:41–55.
7. Ventola CL. The antibiotic resistance crisis: part 1: causes and threats. P T **2015**; 40:277–83.
8. Michael CA, Dominey-Howes D, Labbate M. The antimicrobial resistance crisis: causes, consequences, and management. Front Public Health **2014**; 2:145.
9. Roth JA, Juchler F, Widmer AF, Battegay M. Plea for standardized reporting and justification of propensity score methods. Clin Infect Dis **2019**; 68:710–1.
10. Granger E, Watkins T, Sergeant JC, Lunt MA. Review of the use of propensity score diagnostics in papers published in high-ranking medical journals. BMC Med Res Methodol **2020**; 20:132.
11. Yao XI, Wang X, Speicher PJ, et al. Reporting and guidelines in propensity score analysis: a systematic review of cancer and cancer surgical studies. J Natl Cancer Inst **2017**; 109:djw323.
12. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ **2021**; 372:n71.
13. Helmer O. Analysis of the Future: The Delphi Method (P-3558). Santa Monica, CA: The RAND Corporation, **1967**.
14. Linstone HA, Turoff M. The Delphi Method: Techniques and Applications. Addison-Wesley Pub. Co., Advanced Book Program; **1975**.
15. Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. J Thorac Cardiovasc Surg **2007**; 134:1128–35.
16. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med **1998**; 17:2265–81.
17. Kuss O, Blettner M, Börgermann J. Propensity score: an alternative method of analyzing treatment effects. Dtsch Arztebl Int **2016**; 113:597–603.
18. Goetghebeur E, le Cessie S, De Stavola B, Moodie EE, Waernbaum I. Formulating causal questions and principled statistical answers. Stat Med **2020**; 39:4922–48.
19. Glynn RJ, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. Basic Clin Pharmacol Toxicol **2006**; 98:253–9.
20. Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. Med Decis Making **2009**; 29:661–77.
21. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Stat Med **2007**; 26:734–53.
22. Austin PC. A tutorial and case study in propensity score analysis: an application to estimating the effect of in-hospital smoking cessation counseling on mortality. Multivar Behav Res **2011**; 46:119–51.
23. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med **2009**; 28:3083–107.
24. Heinze G, Jüni P. An overview of the objectives of and the approaches to propensity score analyses. Eur Heart J **2011**; 32:1704–8.
25. Ali MS, Groenwold RH, Belitser SV, et al. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. J Clin Epidemiol **2015**; 68:112–21.
26. Pouwels KB, Widyakusuma NN, Groenwold RH, Hak E. Quality of reporting of confounding remained suboptimal after the STROBE guideline. J Clin Epidemiol **2016**; 69:217–24.
27. Amoah J, Stuart EA, Cosgrove SE, et al. Comparing propensity score methods versus traditional regression analysis for the evaluation of observational data: a

case study evaluating the treatment of gram-negative bloodstream infections. Clin Infect Dis **2020**; 71:e497–505.
28. Penning de Vries BBL, van Smeden M, Rosendaal FR, Groenwold RHH. Title, abstract, and keyword searching resulted in poor recovery of articles in systematic reviews of epidemiologic practice. J Clin Epidemiol **2020**; 121:55–61.
29. Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. PLoS Med **2007**; 4:e297.
30. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ **2016**; 355:i4919.