Received: 13 November 2022 | Revised: 13 June 2023 | Accepted: 20 June 2023

DOI: 10.1002/jrsm.1653

REVIEW

Research Synthesis Methods WILEY

Population adjusted-indirect comparisons in health technology assessment: A methodological systematic review

Bang Truong[1,2] | **Lan-Anh T. Tran**[3] | **Tuan Anh Le**[4] | **Thi Thu Pham**[5] | **Tat-Thang Vo**[6]

[1]Faculty of Pharmacy, HUTECH University, Ho Chi Minh City, Vietnam

[2]Department of Health Outcomes Research and Policy, Auburn University Harrison College of Pharmacy, Auburn, Alabama, USA

[3]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

[4]Department of Biology, KU Leuven, Leuven, Belgium

[5]Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[6]Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Correspondence
Tat-Thang Vo, Department of Statistics and Data Science, The Wharton School, 265 South 37th Street, Philadelphia, PA 19104-1686, USA.
Email: tatthang@wharton.upenn.edu

Abstract

In health technology assessment (HTA), population-adjusted indirect comparisons (PAICs) are increasingly considered to adjust for the difference in the target population between studies. We aim to assess the conduct and reporting of PAICs in recent HTA practice, by performing, a methodological systematic review of studies implementing PAICs from PubMed, EMBASE Classic, Embase/Ovid Medline All, and Cochrane databases from January 1, 2010 to Feb 13, 2023. Four independent researchers screened the titles, abstracts, and full-texts of the identified records, then extracted data on methodological and reporting characteristics of 106 eligible articles. Most PAIC analyses (96.9%, $n = 157$) were conducted by (or received funding from) pharmaceutical companies. Prior to adjustment, 44.5% of analyses ($n = 72$) (partially) aligned the eligibility criteria of different studies to enhance the similarity of their target populations. In 37.0% of analyses ($n = 60$), the clinical and methodological heterogeneity across studies were extensively assessed. In 9.3% of analyses ($n = 15$), the quality (or bias) of individual studies was evaluated. Among 18 analyses using methods that required an outcome model specification, results of the model fitting procedure were adequately reported in three analyses (16.7%). These findings suggest that the conduct and reporting of PAICs are remarkably heterogeneous and suboptimal in current practice. More recommendations and guidelines on PAICs are thus warranted to enhance the quality of these analyses in the future.

KEYWORDS

health technology assessment, indirect treatment comparisons, matching-adjusted indirect comparison, population adjustment, simulated treatment comparison

Highlights

What is already known about the topic?

- Population-adjusted indirect comparisons (PAICs) are increasingly used to adjust for the difference in the target population between trials in health technology assessment.

What is new?

- The conduct and reporting of PAICs are remarkably heterogeneous and suboptimal in current practice.

Potential impact for *Research Synthesis Methods* readers?

- More recommendations and guidelines on methodological and reporting standards are warranted to enhance the quality of PAIC analyses in the future.

1 | INTRODUCTION

In the absence of head-to-head clinical trials, indirect comparisons are increasingly used to evaluate the relative treatment effect of medical interventions in health technology assessment (HTA). Traditional indirect comparison techniques and network meta-analysis compare two or more interventions by using aggregate data (AgD) from eligible trials, based on the assumption that effect modifiers are evenly distributed across different study populations.[1,2] On many occasions, this assumption might be violated and lead to biased treatment effect estimates.[3] To overcome this challenge, many statistical approaches have been proposed to adjust for the difference between studies in the effect modifiers' distributions, when the AgD are available for some studies and the individual participant data (IPD) are available for others.[4] Among these methods, Matching-Adjusted Indirect Comparisons (MAIC) and Simulated Treatment Comparisons (STC) are the most commonly used.[3,5,6] Both methods allow to assess how results of a trial with IPD would look like, were such trial conducted in the target population of another trial with only AgD accessible. While MAIC is based on propensity score weighting,[3,4] STC focuses on modeling the outcome generating mechanism that is supposed to be the same across trials.[3,7] For example, consider two studies assessing treatments *A* and *B* (i.e., trial *AB*), and *B* and *C* (i.e., trial *BC*) in two different patient populations. The interest lies in the relative effect of treatment *A* versus treatment *C* in the target population of trial *AB*. To account for the differences between populations, results of trial *BC* (with IPD) are standardized over the case-mix of the target population of trial *AB* (with AgD) by using MAIC or STC. Treatment *A* is then indirectly compared to treatment *C* via the common comparator, *B* (i.e., an anchored comparison). When no common comparator is available (*A* and *C* are directly compared after population adjustment), one has an unanchored comparison, which is more prone to bias due to confounding.[3] While MAIC and STC can only be used to assess two treatments (or three in anchored comparisons) and two studies, novel methods based on multilevel network meta-regression (ML-NMR) have also been proposed to evaluate simultaneously multiple trials and/or multiple treatments.[8]

In a recent paper, Phillippo et al. characterized 18 studies using population-adjusted indirect comparison (PAIC) methods from the United Kingdom National Institute for Health and Care Excellence (NICE).[4] The authors highlighted an increased use of these methods in HTA practice. A shortcoming of this paper, however, is that it only described the characteristics of a small subset of technology appraisals submitted to NICE, hence the findings could not be generalized to all PAICs conducted in the literature. Besides, some important concerns that could affect the validity of PAIC findings were not assessed in this first review. For instance, it remains unclear how bias and heterogeneity assessments are often performed in PAIC reports. These are critically important because studies included in a PAIC analysis are also required to be of high methodological quality, or to be relatively homogeneous in the eligibility criteria, in the common comparator (for anchored comparisons) and in the outcome measurement. Besides, adequate reporting of the modeling strategy and of the model fitting results (e.g., in STC and ML-NMR) is essential to assess the validity of the obtained findings,[7,9] but this was not assessed in Reference 4.

Considering continued methodological advances in population adjustment methods and the rapid increases in publications in HTA, a more comprehensive systematic review of PAIC practice is strongly warranted. In this study, we aim to extensively assess how PAICs were conducted and reported in the general literature. By shedding light on the current practice of PAIC, this review could pave the way for future recommendations and guidelines on methodological standards of PAIC, thus enhancing the utility of these methods in HTA.

2 | METHODS

2.1 | Study design and search strategies

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.[10]

Since this review does not focus on health outcomes, the protocol was not registered in the International Prospective Register of Systematic Reviews (PROSPERO).[11] We systematically searched for eligible articles using the following keywords: "transportability," "direct standardization," "population adjustment," "external control," "simulated treatment comparison(s)," "population-adjusted indirect comparison(s)," and "matching-adjusted indirect comparison(s)" on PubMed, EMBASE Classic, Embase/Ovid Medline All, and Cochrane databases from January 1, 2010 (the year when the first population-adjusted method was proposed in the literature) to February 13, 2023. The complete search strategies can be found in Supplementary Table **S1** (available online). Additionally, articles were also identified through manual searches from reference lists of included eligible articles.

2.2 | Eligibility criteria

We included all papers that reported the use of a population-adjusted indirect comparison method to account for the difference between study populations prior to assessing the treatment effectiveness. We excluded non-English publications, articles without full-text or full-text not accessible even after having contacted the first author, and appraisals not conducted on human subjects. In addition, we removed articles in which the conducted PAIC was to illustrate the development of new statistical approaches, or articles that had access to the IPD of all individual studies in the analysis. This is because we wanted to focus on the most commonly seen setting in HTA, where the submitting company had the IPD available from their own trial or trials, but very often only published AgD from those of their competitors.[4]

2.3 | Study selection

All search records were imported into Endnote X9 (Clarivate Analytics), and duplicated records were removed before all references were imported into MS Excel documents. The titles and abstracts of all records were then screened by three independent researchers (LATT, TAL, TTP). The full-text copies of potentially eligible reports were also obtained and independently examined for further assessment if needed. In the case of disagreement, a fourth reviewer was consulted (TTV) (Supplementary Table S2). The result of this process was reported through a PRISMA flowchart.

2.4 | Data extraction

We extracted the following information from eligible articles: (i) general characteristics, (ii) methodological characteristics of the conducted PAIC analysis, and (iii) reporting and discussions of the PAIC results. The data extraction form is available online (Supplementary Table S3).

For general characteristics, we extracted information on the clinical area of the study (based on the ICD-10 code), treatment/exposure type (pharmacological or non-pharmacological), outcome type (continuous, binary, or time-to-event), number and type of studies included in the PAIC analysis (randomized controlled trials, single-arm or observational studies), number of treatments being compared, and sponsorship (industry or academia).

For methodological characteristics of the PAIC analysis, we extracted information on the PAIC method being used (MAIC, STC, or others), comparison type (anchored or unanchored), number of adjusted covariates and covariate selection methods (e.g., based on the availability of covariates in each study, experts' opinions, or results of any statistical methods). We further evaluated how an analysis with multiple (>2) studies and/or with multiple treatments (>2 for unanchored and >3 for anchored comparisons) was handled, whether the eligibility criteria of different studies were aligned before adjustment, whether the clinical heterogeneity between studies was assessed, whether the methodological quality of individual studies was assessed, whether any sensitivity analysis was conducted, and whether/how the case-mix overlap between populations was evaluated (e.g., by the effective sample size or by examining the presence of extreme weights in MAIC). Regarding heterogeneity assessment, we further determined whether this was done in a formal and systematic way (e.g., by describing the characteristics of eligible studies in a table, and then comparing studies to determine whether they were similar enough for evidence synthesis—which is recommended by the Cochrane handbook).[12] In some PAIC analyses, the authors did not perform such a rigorous heterogeneity assessment. However, they mentioned the potential difference or the level of similarity between individual studies when discussing the PAIC findings. We considered this as an informal heterogeneity assessment.

For the reporting and discussion of PAIC results, we determined whether the covariate distribution in each individual study before adjustment (and also after adjustment if MAIC was used) was fully reported, partially reported or not reported. In addition, we identified whether estimates and corresponding uncertainty measures of the coefficients in the outcome regression model

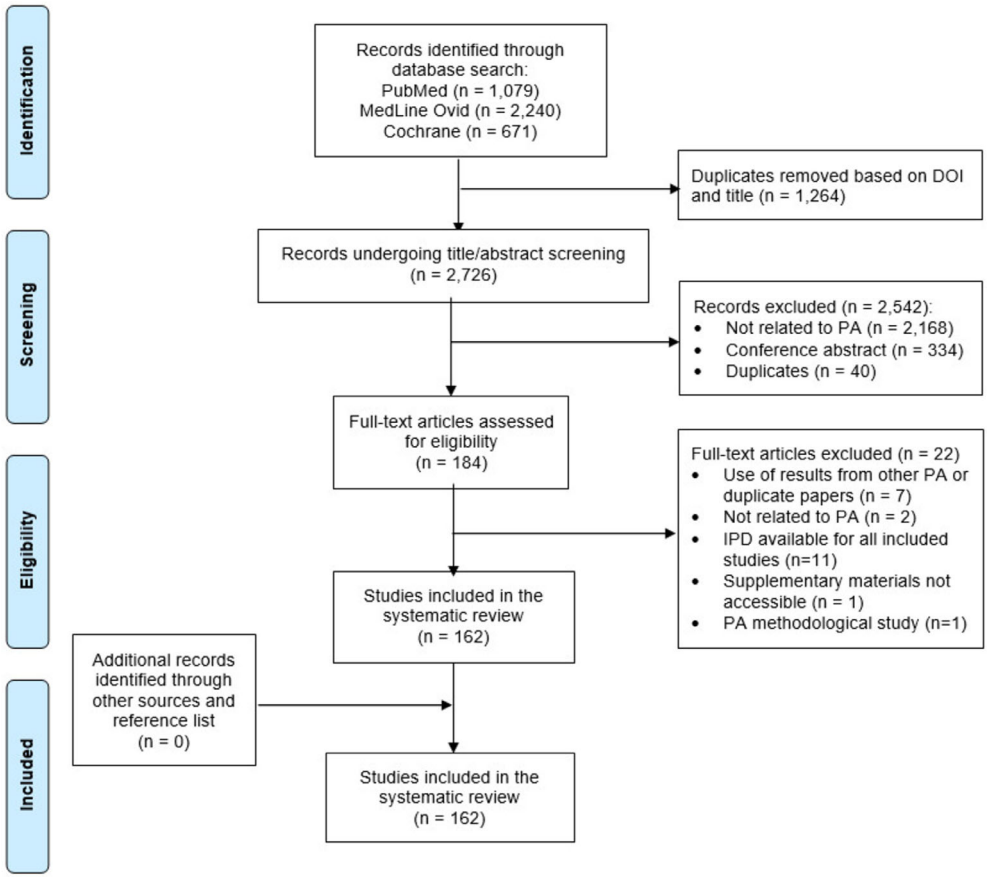

FIGURE 1 PRISMA diagram of included studies of population adjustment methods in health technology assessment. 11 PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses. DOI, digital object identifier. [Colour figure can be viewed at wileyonlinelibrary.com]

were adequately reported, if such a model was specified in the PAIC analysis (i.e., when STC or ML-NMR was used). Next, we determined whether there was a comparison of the clinical results before versus after population adjustment, and if such a comparison was considered, whether the authors discussed the clinical relevance of the change in the results. Finally, we assessed if there was any limitation acknowledged by the authors.

When multiple PAICs were implemented to evaluate multiple treatment arms (>2 for unanchored and >3 for anchored comparisons) or several outcomes, we only focused on the first pairwise comparison and the first outcome reported. Data were extracted by one researcher (BT) and then independently double-checked by another researcher (LATT, TAL, and TTP). If there was any disagreement during pair discussion, a consensus was reached by consulting a third reviewer (TTV) (Supplementary Table S4).

2.5 | Data synthesis

Categorical data were summarized using frequencies and percentages. Continuous data were summarized using median and interquartile range. Data analysis was conducted using MS Excel 2016 (Supplementary Table S5).

3 | RESULTS

3.1 | Characteristics of eligible studies

Figure 1 shows the PRISMA flow diagram summarizing the study selection process. Of 3990 identified records, 2726 underwent title and abstract screening after removing duplicates. We further excluded 2564 records due to the following reasons: conference abstracts only ($n = 334$) or supplementary materials ($n = 1$) not achievable even after having contacted the authors, not a PAIC analysis ($n = 2170$), duplicate articles or articles using results from other PAIC studies already included in our review ($n = 47$), IPD accessible for all individual studies in the PAIC analysis ($n = 11$), or studies focusing on PAIC methods ($n = 1$). No additional record was identified by retrieving the bibliography of the included studies.

In summary, we identified 162 eligible records. As seen in Table 1, more than half of these records are in oncology (58.0%, $n = 94$). Most PAICs were conducted to

TABLE 1 Characteristics of eligible studies.

Characteristics ($N = 162$)	Statistic
Year of publication, *N* (%)	
2011–2015	10 (6.2)
2016–2020	65 (40.1)
2021	38 (23.5)
2022	40 (24.7)
2023	9 (5.5)
Clinical area, *N* (%)	
Infectious and parasitic diseases	8 (4.9)
Neoplasm	94 (58.0)
Diseases of the blood and blood-forming organs	8 (4.9)
Endocrine, nutritional and metabolic diseases	5 (3.1)
Mental, behavioral and neurodevelopmental disorders	4 (2.5)
Diseases of the nervous system	16 (9.9)
Diseases of the eye and adnexa	2 (1.2)
Diseases of the circulatory system	4 (2.5)
Diseases of the respiratory system	2 (1.9)
Diseases of the skin and subcutaneous tissue	11 (6.8)
Diseases of the musculoskeletal system and connective tissue	6 (3.7)
Diseases of the digestive system	1 (0.6)
Type of treatment/exposures, *N* (%)	
Medication	156 (96.3)
Non-pharmacological treatments	6 (3.7)
Type of outcome, *N* (%)	
Continuous	20 (12.4)
Binary	76 (46.9)
Time-to-event	66 (40.7)
Number of studies included in the analysis, *N* (%)	
Two studies	74 (45.7)
Three studies	26 (16.0)
More	62 (38.3)
Number of treatment arms, *N* (%)	
Two	108 (66.7)
Three	23 (14.2)
More	31 (19.1)
Number of studies with IPD, *N* (%)	
One	120 (74.1)
Two	30 (18.5)
Three or more	12 (7.4)
Number of studies with AgD, *N* (%)	
One	125 (77.2)
Two	23 (14.2)
Three or more	14 (8.6)
Type of studies, *N* (%)	
RCT (IPD) to RCT (AgD)	95 (58.6)
RCT (IPD) to single arm (AgD)	8 (4.9)
Single arm (IPD) to RCT (AgD)	21 (13.0)
Single arm (IPD) to single arm (AgD)	28 (17.3)
Single arm (IPD) to OS (AgD)	4 (2.5)
OS (IPD) to RCT (AgD)	6 (3.7)
Source of funding, *N* (%)	
Industry	157 (96.9)
Academia	5 (3.1)

Abbreviations: AgD, aggregate data; IPD, individual patient data; IQR, interquartile range; Med, median; RCT, randomized controlled trials; OS, observational study.

assess the effectiveness or safety of pharmacological interventions (96.3%, $n = 156$) on a binary (46.9%, $n = 76$) or time-to-event outcome (40.7%, $n = 66$). Regarding sponsorship, almost all eligible records (96.9%, $n = 157$) were conducted by, or received funding from pharmaceutical companies. Other general characteristics of the eligible records are also provided in Table 1.

3.2 | Methodological characteristics of the population adjustment analysis

The methodological characteristics of the PAIC analyses are provided in Table 2. Regarding adjustment method, 88.9% of the records used MAIC ($n = 144$), 6.8% used STC ($n = 11$), 3.7% used both MAIC and STC ($n = 6$), and 0.6% used ML-NMR ($n = 1$). An unanchored comparison was conducted in 64.8% of the records ($n = 105$). When there were multiple treatment arms (>2 in unanchored comparisons and >3 in anchored comparisons, 30.9%, $n = 50$), a separate PAIC analysis was often conducted for each pair of treatments (29.6%, $n = 48$). When there were more than two individual studies with IPD (25.9%, $n = 42$), the IPD from these studies were often merged before an adjustment was considered (25.3%, $n = 41$). In 44.5% of the records ($n = 72$), the eligibility criteria of the AgD study were applied to the IPD study/studies to reduce potential heterogeneity between their target populations.

Heterogeneity assessment was not considered in 11.1% of the records ($n = 18$). In 24.7% of the records ($n = 40$), the potential difference between individual studies was briefly mentioned prior to the PAIC analysis or in the discussion. In 64.2% of the records ($n = 107$),

TABLE 2 Methodological characteristics of the population adjustment analysis across studies.

Characteristics	Statistics
Population adjustment methods, *N* (%)	
Matching-Adjusted Indirect Comparison (MAIC)	144 (88.9)
Simulated Treatment Comparison (STC)	11 (6.8)
Both MAIC and STC	6 (3.7)
Multilevel Network Meta Regression (ML-NMR)	1 (0.6)
Type of comparison, *N* (%)	
Anchored	57 (35.2)
Unanchored	105 (64.8)
Handling multiple treatments (>2 for unanchored and >3 for anchored comparisons), *N* (%)	50 (30.9)
Separate PAIC analysis for each pair (or each group of three) of treatments	48 (29.6)
One common analysis for the entire treatment network (e.g., by using ML-NMR)	2 (1.2)
Handling multiple studies (>2) with IPD, *N* (%)	42 (25.9)
Studies with IPD merged	41 (25.3)
Studies with IPD kept apart (by using ML-NMR)	1 (0.6)
Handling multiple studies (>2) with AgD, *N* (%)	37 (22.8)
Studies with AgD pooled	34 (21.0)
Separate PAIC analysis for each AgD study	3 (1.8)
Before adjustment, the eligibility criteria of one study (i.e., the one with AgD) were used to refine the patient sample of other studies (with IPD), *N* (%)	
No	90 (55.5)
Partially	33 (20.4)
Fully	39 (24.1)
Bias/quality assessment of each study included in the PAIC analysis	
Yes	15 (9.3)
No	147 (90.7)
Heterogeneity assessment, *N* (%)	
No description/discussion about potential heterogeneity	18 (11.1)
No formal assessment, but the authors mentioned/discussed (informally) the potential difference between studies in:	40 (24.7)
Inclusion/exclusion criteria	26 (16.0)
Common comparator (only for anchored comparison)	2 (1.2)
Outcome definition/measurement	20 (12.3)
Follow-up time	19 (11.7)
The authors conducted a (partially) formal and systematic assessment of heterogeneity between studies in:	104 (64.2)
Inclusion/exclusion criteria	98 (60.5)
Common comparator (only for anchored comparison)	8 (4.9)
Outcome definition/measurement	81 (50.0)
Follow-up time	77 (47.5)
Covariates	
Number of adjusted covariates, Med (IQR)	7 (5–9)
Covariates selection based on, *N* (%):	
Availability of covariates across eligible studies	67 (41.4)
Experts' opinions	77 (47.5)
Statistical methods (e.g., running a series of univariate regression analyses to explore the association between each covariate and the outcome, then selecting covariates for the PAIC based on results of these analyses)	47 (29.0)
Literature reviews	45 (27.8)
Evaluation of case-mix overlap between populations (MAIC-specific), *N* (%)	
No evaluation	24 (16.0)
By effective sample size	117 (78.0)
By checking the presence of extreme values in the distribution of weights	15 (10.0)
Others (i.e., by testing the difference between studies in the distribution of each covariate before and after weighting)	20 (13.3)
Sensitivity analysis to assess the robustness of PAIC results	
No sensitivity analysis	77 (47.5)
Adjusting for different sets of covariates	55 (34.0)
Applying additional inclusion/exclusion criteria to the IPD study	19 (11.7)
Using different outcome definitions	7 (4.3)
Using different follow-up time	11 (6.8)
Other (e.g., using different approaches for handling missing data, implementing additional anchored/unanchored comparisons)	12 (7.4)

Abbreviations: AgD, aggregate data; IPD, individual patient data; IQR, interquartile range; MAIC, matching-adjusted indirect comparison; Med, median; PAIC, population-adjusted indirect comparison; STC, simulated treatment comparison.

TABLE 3 Reporting and discussions of population adjustment results.

Characteristics	Statistics
Covariate distribution in each study before adjustment reported, *N* (%):	
Not described	11 (6.8)
Partially described[a]	42 (25.9)
Fully described[b]	109 (67.3)
Covariate distribution in each study after adjustment reported (MAIC-specific), *N* (%): (*N* = 150)	
Not described	15 (10.0)
Partially described[a]	40 (26.7)
Fully described[b]	95 (63.3)
Results of the model fitting procedure reported (i.e., coefficient estimates and uncertainty measures, STC-specific), *N* (%): (*N* = 18)	
Not reported	15 (83.3)
Fully reported (both coefficient estimates and uncertainty measures)	3 (16.7)
Comparison of results before and after population adjustment	
No	48 (29.6)
Yes	114 (70.4)
Discussion about whether the change of results after population adjustment is clinically relevant (*N* = 114)	
No	88 (77.2)
Yes	26 (22.8)
Limitations acknowledged by authors:	
No acknowledgement	5 (3.1)
Unmeasured covariates	136 (84.0)
Unmeasured covariates explicitly mentioned	33 (20.4)
Important covariates not reported in one of the included studies	60 (37.0)
Limited sample size	31 (19.1)
Heterogeneity across studies	139 (85.8)
Small ESS/little overlap between populations	35 (21.6)
Lack of a common comparator	23 (14.2)
Others	7 (4.3)

Abbreviations: CI, confidence interval; ESS, effective sample size; MAIC, matching-adjusted indirect comparison; SE, standard error.
[a]Either (i) a measure of central tendency (e.g., mean, median) or (ii) a measure of dispersion (e.g., standard deviation, interquartile range) was not reported for each covariate.
[b]Both (i) and (ii) were reported for each covariate.

clinical heterogeneity was formally evaluated by describing the characteristics of each eligible study regarding inclusion/exclusion criteria (60.5%, $n = 98$), common comparator (4.9%, $n = 8$), outcome definition (50.0%, $n = 81$), and follow-up time (47.5%, $n = 77$). In 37.0% of the records ($n = 60$), all three components (inclusion/exclusion criteria, outcome, and follow-up time) were compared among eligible studies. Besides, only 9.3% of the records ($n = 15$) considered a risk of bias assessment prior to conducting the PAIC analysis.

Selecting baseline covariates into the weight model (MAIC) or the outcome model (STC and ML-NMR) is a critical step in PAIC analyses. We found that eligible records adjusted for a median of 7 covariates (IQR 5–9). In 47.5% of the records ($n = 77$), covariates were selected based on experts' opinions. In some other cases, this selection was also based on the availability of covariates in all individual studies (41.4%, $n = 67$), results of a statistical procedure (29.0%, $n = 47$), or results of a literature review (27.8%, $n = 45$). Among studies using MAIC ($n = 150$), the case-mix overlap between populations after adjustment was evaluated by effective sample size in 78.0% of the records ($n = 117$). In 47.5% of the records ($n = 77$), sensitivity analysis was not implemented to assess the robustness of the findings. When sensitivity analysis was considered, the most common practice was to assess the change of the results when adjusting for different sets of baseline covariates (34.0%, $n = 55$).

3.3 | Reporting and discussions of results of the population adjustment analysis

Among eligible records, 67.3% ($n = 109$) adequately described the covariate distribution before population adjustment. Besides, 63.3% of the MAIC records ($n = 95$) adequately described the covariate distribution after population adjustment. Remarkably, among 18 studies using STC and/or ML-NMR (11.1%), only three records reported results of the outcome model estimation. In 70.4% of the records ($n = 114$), results before and after population adjustment were directly compared, among which 22.8% ($n = 26$) explicitly stated that the change of the results after adjustment was clinically relevant. Finally, the most common limitations acknowledged by the authors included important heterogeneity between studies (85.8%, $n = 139$), unmeasured covariates (84.0%, $n = 136$), and unavailability of important covariates in one of the included studies (37.0%, $n = 60$). These findings are summarized in Table 3.

4 | DISCUSSION

Our study confirms the rising trend of PAICs in HTA over the last decade.[4,7] Among different PAIC methods to combine one IPD study and one AgD study, MAIC is the

most used. MAIC has been shown to be less powerful than STC and can be quite unstable in the presence of extreme weights.[3] STC, however, can hide the risk of unreasonable extrapolation due to the study populations being too different in terms of case mix (which makes the so-called positivity assumption violated). Furthermore, STC has only been recently extended to settings in which the outcome generating mechanism is complex, for example, due to treatment-covariates interactions and nonlinearities.[13,14] As a result, the use of this approach remains quite scarce in practice.

Although unanchored comparisons are more prone to bias than anchored comparisons, the former are more commonly implemented in HTA. This can be explained by the fact that many of the eligible records evaluated cancer and rare diseases. Due to ethical concerns, observational or single-arm studies might be the best available evidence one can have for cancer and rare disease treatment evaluation. This is acknowledged in the current drug approval process of the United States Food and Drug Administration and the European Medicines Agency.[15,16]

Overall, our review indicated that the conduct and reporting of PAICs remained suboptimal in practice. First, the percentage of PAIC analyses extensively assessing the clinical and methodological heterogeneity across studies remained low. As different studies are often conducted by different research teams in different settings, clinical and methodological heterogeneity between them is likely unavoidable. Similar to meta-analysis, overlooking potential heterogeneity in a PAIC analysis may lead to biased treatment effect estimates and hamper the interpretability of the findings. This should be further improved in future practice.

Second, covariates in PAIC are required to be effect modifiers (e.g., in anchored comparisons) and prognostic factors (e.g., in unanchored comparisons), which are differently distributed across studies. Besides, covariates that are not differently distributed may also need to be balanced in MAIC because balance pre-weighting does not guarantee balance post-weighting if the covariates are not accounted for. In this review, we found that a variety of approaches were used to select covariates in the analysis. However, in many cases, variable selection was affected by the availability of covariates data in individual studies. This is practically challenging as recent evidence continues to show that the collection and reporting of patient characteristics data among clinical studies in the same field remain very inconsistent.[17–19] In recent years, significant efforts have been made to improve this concern. Across many therapeutic areas, a so-called core patient characteristic set (CPCS) is specifically developed to identify all key prognostic factors that should be commonly collected and reported (among studies and databases evaluating a similar target condition), while keeping the additional burden on the implementation acceptable.[18,20,21] While CPCS should be further considered in the future, rationales for covariate selection should always be included when conducting a PAIC analysis.[3,22]

Third, we observed that a large number of studies did not assess the potential overlap of covariate distribution between populations before and after adjustment. In a PAIC analysis, although the target populations of different studies might be different, it is important that they are still sufficiently similar so that we can learn about one population via observing the other without erroneous extrapolations. Such an assumption is often referred to as positivity in the technical literature.[14] Visualizing the covariate distribution in each individual study is an important way to (informally) assess positivity. In contrast, calculating the effective sample size (ESS) or the percentage of extreme weights (i.e., in MAIC) allows one to roughly quantify the similarity level between populations. Also related to positivity, we found that many PAIC analyses applied the eligibility criteria of one study (AgD) to refine the patient samples of other studies (IPD). This approach may partially reduce the risk of positivity violation, as otherwise trials may have different inclusion/exclusion criteria and hence are not similar enough for a PAIC to be feasible.[23] As an example, consider two single-arm studies assessing the effectiveness of warfarin and a direct oral anticoagulants (DOAC) among patients with atrial fibrillation (AFib). The warfarin study includes AFib patients with severe renal impairment, who are nonetheless excluded from the DOAC study due to contraindication. To avoid structural positivity violation, one may refine the patient sample of the warfarin study by the eligibility criteria of the DOAC study, before conducting a PAIC to compare these treatments. Aligning eligibility criteria of different studies is thus useful and hence, should be advocated in future practice whenever possible.

Fourth, sensitivity analysis to assess the robustness of PAIC findings was not often considered, though PAIC methods often require many (untestable) statistical and clinical assumptions. For instance, when unmeasured effect modifiers (or unmeasured outcome prognostic factors in the case of unanchored comparisons) are suspected, one can perform sensitivity analyses to assess the extent to which it can impact results or to generate bounds on the treatment effect when only partial identification is possible.[24] Many sensitivity analysis approaches for population adjustment, however, can only be used when full IPD from all studies are available. This may also explain the limited use of sensitivity analysis in current practice.

Fifth, results of the model fitting procedure were barely reported across studies using STC or ML-NMR. Such a poor reporting practice implies a lack of transparency and reduces the reproducibility of the conducted

TABLE 4 Methodological recommendations for population-adjusted indirect comparisons (PAICs).

Conduct	• Decide carefully which variables to adjust for in the PAIC.
	  ○ Based on the type of comparison (anchored or unanchored), subject-matter knowledge and the availability of covariates across studies. For a more detailed guidance, see Reference 3.
	• Assess the risk of bias in each eligible study.
	  ○ For instance, by the RoB-2 tool for RCTs, by the ROBINS-I tool for observational controlled studies, by the NIH quality appraisal tool (or many other tools) for case-series/single-arm studies.[29]
	• Assess clinical and methodological heterogeneity across eligible studies.
	  ○ For instance, by comparing the characteristics of included studies (i.e., eligibility criteria, common control (in anchored comparisons), outcome definition and follow-up time) in a table.[12]
	  ○ PAICs should only be implemented when individual studies can be judged as sufficiently similar.
	• On many occasions, the eligibility criteria of different studies can be aligned to reduce heterogeneity and/or to avoid positivity violation.
	• Choose an appropriate statistical method to implement PAIC.
	  ○ Anchored comparisons should be prioritized whenever possible.[3]
	  ○ In the standard setting of PAIC where two treatments are indirectly compared (possibly via a common control), MAIC is still the primary method to use. However, extreme weights may arise and bias the effect estimate by MAIC when there is limited overlapping among study populations.[13,14]
	  ○ In contrast, STC only provides estimates for the conditional treatment effect, which does not coincide with the population-level treatment effect when the effect measure is non-collapsible (e.g., odds ratio or hazard ratio).[13] Because PAIC findings are often used for reimbursement decision making on the population level, STC may not be suitable for binary and time-to-event outcomes.
	  ○ Recently, enhancements to the standard versions of MAIC and STC have been proposed in the literature, with aims to improve statistical performance and to overcome the above-mentioned limitations.[13,26,28]
	  ○ In complex settings where a connected network of treatments is evaluated across multiple trials, new methods such as ML-NMR should be used.[8] However, ML-NMR, as currently conceptualized, is not readily extended to disconnected networks/unanchored comparisons of multiple treatments, or to time-to-event outcomes. This leaves room for future methodological research.
	  ○ When there are multiple studies with IPD, combining the IPD of different studies into one and then using a simple method such as MAIC or STC for population adjustment is suboptimal. This overlooks potential heterogeneity across IPD studies.
	• Sensitivity analysis to assess the robustness of PAIC findings should be considered.

Reporting and discussion	• Provide a clear description of the PAIC protocol, for example, what method is used for:
	  ○ Variable selection
	  ○ Bias and heterogeneity assessment
	  ○ Handing missing data
	  ○ Population adjustment (e.g., whether MAIC, STC, or any other method is used)
	  ○ Sensitivity analyses
	• Report the results of every step in the implemented PAIC, including:
	  ○ Which covariates are adjusted for in the analysis
	  ○ Covariate distribution in each study before adjustment
	  ○ Bias and heterogeneity assessment results
	  ○ Results of the outcome model estimation when using STC and ML-NMR; effective sample size, distribution of weights, percentage of extreme weights, distribution of covariates after adjustment by MAIC (or by any other weighting-based methods).
	  ○ Treatment effect estimates and corresponding uncertainty measures.
	  ○ Results of sensitivity analyses.
	• Discuss the clinical relevance of the obtained PAIC results by, for instance, comparing treatment effect estimates before versus after population adjustment.[3]
	• Discuss the limitations of the PAIC analysis, for example, whether there are important covariates that cannot be adjusted for (due to the unavailability of such covariates in some or all studies). Mention explicitly the unmeasured covariates and explain their potential impact on the validity of the findings.

Abbreviations: AgD, aggregate data; IPD, individual patient data; MAIC, matching-adjusted indirect comparison; NIH, National Institutes of Health; PAIC, population-adjusted indirect comparison; RoB-2, version 2 of the Cochrane risk-of-bias tool for randomized trials; ROBIN-I, risk of bias in non-randomized studies—of interventions; STC, simulated treatment comparison.

STC or ML-NMR analysis. As criticized by Holmes et al (2019), this practice also increases the uncertainty in the decision making process of regulatory committees when handling a new drug/indication application.[9] In addition, the clinical relevance of the PAIC findings (especially compared to before adjustment) was not often

considered. As PAIC methods are relatively new in HTA, it would be insightful to consider such a before-after comparison to assess the usefulness of PAIC in practice.

Finally, our review highlighted the urgent need of promoting novel PAIC methods in future practice. Indeed, classical approaches such as MAIC and STC can only be used to compare two treatments (or three in the case of anchored comparisons) across two different populations. In our review, many PAIC analyses included more than two studies and three treatments. Almost all analyses handled the problem of multiple studies by merging the IPD studies. This practice raises important concerns since it overlooks the risk of potential heterogeneity between different IPD studies. Likewise, conducting separate PAIC analyses when there are multiple treatments is suboptimal, since it does not allow the assessment of the entire treatment network. To remedy this, advanced methods such as ML-NMR should be more widely considered when multiple treatments and/or multiple studies are available.[8,25]

Recently, novel methods have been proposed to improve standard MAIC and STC, for example, by allowing the use of data-driven or machine learning techniques in the estimation of the weight model (in MAIC) and of the outcome model (in STC).[13] Alternative matching schemes are also suggested to enhance the performance of MAIC when the effective sample size is small,[14,26,27] or when an observational study is included in the analysis.[28] These new techniques should be more widely used for practical applications in the future. Also, to ameliorate the methodological quality of PAICs, it is important that hands-on guidance on the conduct and reporting of PAICs is soon developed. As a first step toward such endeavor, we propose in Table 4 some practical recommendations to assist applied researchers in avoiding the pitfalls encountered in this review.

Our review has some limitations. First, we did not assess PAIC analyses published in non-English language, which may introduce selection bias. Second, a large number of eligible PAIC analyses were not included in this review, as they were conference abstracts and were likely not published. This raise concerns over publication bias, as these analyses might have been unpublished due to not having statistically significant findings or not advocating the investigators' hypothesis. Finally, there might be some potential mistakes in our data extraction, even when this was cross-checked by an independent reviewer.

5 | CONCLUSIONS

Population-adjusted indirect comparison methods are increasingly used in HTA. However, the conduct and reporting of these methods are substantially heterogeneous and suboptimal in practice. More recommendations and guidelines on the methodological and reporting standards for PAIC are warranted, to enhance the quality of these analyses in the future.

AUTHOR CONTRIBUTIONS

Bang Truong: Data curation; formal analysis; investigation; methodology; writing – original draft. **Lan Anh Tran:** Data curation; formal analysis; methodology. **Tuan Anh Le:** Data curation; formal analysis; methodology. **Thu Pham:** Data curation. **Tat Thang Vo:** Conceptualization; methodology; supervision; writing – review and editing.

ACKNOWLEDGMENTS

We sincerely thank the associate editor and two anonymous peer-reviewers for their insightful comments and feedback on the previous version of the paper.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

All data is available in the Online Supplementary Materials.

ORCID

Tat-Thang Vo https://orcid.org/0000-0001-6485-9947

REFERENCES

1. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol*. 1997;50(6):683-691.
2. Dias S, Sutton AJ, Ades AE, Welton NJ. Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Med Decis Making*. 2013;33(5):607-617.
3. Phillippo DM, Ades T, Dias S, Palmer S, Abrams KR, Welton NJ. NICE DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in submissions to NICE2016.
4. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. Methods for population-adjusted indirect comparisons in health technology appraisal. *Med Decis Making*. 2017; 38(2):200-211.
5. Caro JJ, Ishak KJ. No head-to-head trial? Simulate the missing arms. *Pharmacoeconomics*. 2010;28(10):957-967.
6. Signorovitch JE, Wu EQ, Yu AP, et al. Comparative effectiveness without head-to-head trials: a method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept. *Pharmacoeconomics*. 2010; 28(10):935-945.
7. Remiro-Azócar A, Heath A, Baio G. Methods for population adjustment with limited access to individual patient data: a review and simulation study. *Res Synth Methods*. 2021;12(6): 750-775.

8. Phillippo DM, Dias S, Ades AE, et al. Multilevel network meta-regression for population-adjusted treatment comparisons. *J R Stat Soc Ser A Stat Soc*. 2020;183(3):1189-1210.
9. Holmes EM, Leahy J, Walsh CD, White A, Donnan PT, Lamrock F. Difficulties arising in reimbursement recommendations on new medicines due to inadequate reporting of population adjustment indirect comparison methods. *Res Synth Methods*. 2019;10(4):615-617.
10. Moher D, Liberati A, Tetzlaff J, Altman DG, The PG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*. 2009;6(7):e1000097.
11. National Institute for Health Research. International prospective register of systematic reviews (PROSPERO). About PROSPERO. 2022 Available from: https://www.crd.york.ac.uk/prospero/#aboutpage
12. Higgins JPT, Thomas J, Chandler J, et al. *Cochrane Handbook for Systematic Reviews of Interventions version 6.3 (updated February 2022)*. Cochrane; 2022.
13. Remiro-Azócar A, Heath A, Baio G. Parametric G-computation for compatible indirect treatment comparisons with limited individual patient data. *Res Synth Methods*. 2022;13:716-744.
14. Vo T-T. A cautionary note on the use of G-computation in population adjustment. *Res Synth Methods*. 2023;14(3):338-341.
15. Beaver JA, Howie LJ, Pelosof L, et al. A 25-year experience of US Food and Drug Administration accelerated approval of malignant hematology and oncology drugs and biologics: a review. *JAMA Oncol*. 2018;4(6):849-856.
16. Hatswell AJ, Baio G, Berlin JA, Irs A, Freemantle N. Regulatory approval of pharmaceuticals without a randomised controlled study: analysis of EMA and FDA approvals 1999-2014. *BMJ Open*. 2016;6(6):e011666.
17. Cahan A, Anand V. Second thoughts on the final rule: an analysis of baseline participant characteristics reports on ClinicalTrials.gov. *PLoS One*. 2017;12(11):e0185886.
18. Sorbye H, Köhne CH, Sargent DJ, Glimelius B. Patient characteristics and stratification in medical treatment studies for metastatic colorectal cancer: a proposal for standardization of patient characteristic reporting and stratification. *Ann Oncol*. 2007;18(10):1666-1672.
19. van Boven FE, de Jong NW, Braunstahl G-J, Gerth van Wijk R, Arends LR. A meta-analysis of baseline characteristics in trials on mite allergen avoidance in asthmatics: room for improvement. *Clin Transl Allergy*. 2020;10(1):2.
20. Khalil A, Gordijn SJ, Beune IM, et al. Essential variables for reporting research studies on fetal growth restriction: a Delphi consensus. *Ultrasound Obstet Gynecol*. 2019;53(5):609-614.
21. Meyer T, Selb M, Kiekens C, et al. Toward better reporting standards of patients' characteristics in rehabilitation trials: applying a new conceptual framework to current standards. *Am J Phys Med Rehabil*. 2020;99(3):216-223.
22. My Luong V, Pham Hien Trang T, Khanh Linh D, Tat-Thang V. Development of minimum reporting sets of patient characteristics in epidemiological research: a methodological systematic review. medRxiv. 2023:2023.02.07.23285508.
23. Cheng D, Ayyagari R, Signorovitch J. The statistical performance of matching-adjusted indirect comparisons. arXiv preprint arXiv:191006449. 2019.
24. Degtiar I, Rose S. A review of generalizability and transportability. arXiv preprint arXiv:210211904. 2021.
25. Vo T-T, Porcher R, Vansteelandt S. Assessing the impact of case-mix heterogeneity in individual participant data meta-analysis: novel use of I2 statistic and prediction interval. *Res Methods Med Health Sci*. 2020;2(1):12-30.
26. Jackson D, Rhodes K, Ouwens M. Alternative weighting schemes when performing matching-adjusted indirect comparisons. *Res Synth Methods*. 2021;12(3):333-346.
27. Van Lancker K, Vo T-T, Akacha M. Estimands in heath technology assessment: a causal inference perspective. *Stat Med*. 2022;41(28):5577-5585.
28. Remiro-Azócar A. Two-stage matching-adjusted indirect comparison. *BMC Med Res Methodol*. 2022;22(1):217.
29. Ma L-L, Wang Y-Y, Yang Z-H, Huang D, Weng H, Zeng X-T. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better? *Mil Med Res*. 2020;7(1):7.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Truong B, Tran L-AT, Le TA, Pham TT, Vo T-T. Population adjusted-indirect comparisons in health technology assessment: A methodological systematic review. *Res Syn Meth*. 2023;1-11. doi:10.1002/jrsm.1653