



Population-Adjusted Indirect Comparisons : methodological review and simulation-based assessment of robustness

Arnaud Serret-Larmande

► To cite this version:

Arnaud Serret-Larmande. Population-Adjusted Indirect Comparisons: methodological review and simulation-based assessment of robustness. Bioinformatics [q-bio.QM]. Sorbonne Université, 2025. English. <NNT : 2025SORUS431>. <tel-05525068>

HAL Id: tel-05525068

<https://theses.hal.science/tel-05525068v1>

Submitted on 24 Feb 2026

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization



SORBONNE UNIVERSITY

École doctorale 393 - Pierre Louis de Santé Publique
Institut Pierre Louis d'Epidémiologie et de Santé Publique

Population-Adjusted Indirect Comparisons: methodological review and simulation-based assessment of robustness

by ARNAUD SERRET-LARMANDE

PhD in BIOSTATISTICS

Supervised by DAVID HAJAGE

And by JÉRÔME LAMBERT

Publicly defended on 4 December 2025

Before a jury composed of:

CARINE BELLERA, PhD-HDR	Université de Bordeaux	Rapporteure — Présidente
FLORIAN NAUDET, PU-PH	Université de Rennes	Rapporteur
LAURA RICHERT, PU-PH	Université de Bordeaux	Examinatrice
TAT-THANG VO, JUNIOR PROFESSOR	Université Paris Est Créteil	Examineur
DAVID HAJAGE, PU-PH	Sorbonne Université	PhD advisor
JÉRÔME LAMBERT, PU-PH	Université Paris Cité	Co-PhD advisor

Résumé en français

Titre: Comparaisons Indirectes Ajustées sur la Population : revue méthodologique et évaluation de leur robustesse par simulation

Mots-clés: Comparaisons indirectes ajustées sur la population ; Matching-Adjusted Indirect Comparison ; Simulated Treatment Comparison ; Évaluation des Technologies de Santé ; Comparaison Indirecte de Traitement ; Méta-Épidémiologie

Résumé : Les comparaisons indirectes occupent une place croissante dans les demandes d'autorisation de mise sur le marché ou de remboursement. En comparant des traitements évalués dans des études et des populations différentes, elles sont toutefois exposées à un risque de biais de confusion. La difficulté d'accès aux données individuelles a ainsi conduit au développement et à l'usage croissant des comparaisons indirectes ajustées sur la population (Population-Adjusted Indirect Comparison, PAIC), permettant d'estimer un effet de traitement ajusté malgré l'absence des données individuelles pour l'un groupe de traitement. Depuis leur formalisation au début des années 2010, celles-ci ont fait l'objet de nombreux développements méthodologiques. Ce travail de thèse a abordé cette thématique sur deux plans : leur implémentation pratique et leurs performances comparativement aux estimateurs utilisant l'ensemble de données individuelles.

Le premier travail a consisté en une revue méthodologique de la littérature conduite sur les PAIC publiés jusqu'en avril 2021. Cette revue visait à décrire les caractéristiques de publication des études utilisant des PAIC, à évaluer la qualité de leur mise en œuvre au regard des recommandations théoriques, et à examiner les résultats rapportés. Celle-ci a mis en évidence plusieurs résultats importants : une augmenta-

tion rapide du nombre de publications au cours des dernières années, une hétérogénéité marquée de la qualité du reporting, ainsi qu'un biais de publication majeur. Ces résultats nous ont conduits à formuler des recommandations en matière de reporting.

La deuxième contribution s'est intéressée à des estimateurs de l'effet de traitement dans les PAIC, et en particulier des estimateurs pondérés. Nous avons mené une étude par simulation de Monte-Carlo afin de comparer les performances de différents estimateurs : la pondération inverse par score de propension (Propensity Score Weighting, PSW) et les Matching-Adjusted Indirect Comparison (MAIC). Différents scénarios ont été explorés, variant selon la distribution des covariables, le degré de recouvrement entre les populations comparées et l'adéquation du modèle d'ajustement. Les résultats ont montré que les MAIC basées sur les moments d'ordre 1 étaient plus robustes que le PSW en situation de faible recouvrement des distributions, y compris lorsque les covariables suivaient des distributions non normales. Ces résultats suggèrent que l'utilisation de données agrégées, en l'absence de données individuelles, se traduit avant tout par une limitation de la flexibilité des analyses plutôt que par une baisse substantielle des performances.

English summary

Title: Population-Adjusted Indirect Comparisons: Methodological Review and Simulation-Based Assessment of Robustness

Keywords: Population-Adjusted Indirect Comparison ; Matching-Adjusted Indirect Comparison ; Simulated Treatment Comparison ; Health Technology Assessment ; Indirect Treatment Comparison ; Meta-Epidemiology

Abstract: Indirect treatment comparisons are increasingly used in applications for market authorization and reimbursement. However, when comparing treatments evaluated in different studies and thus different populations, these methods are at risk of confounding bias. The statistical adjustment required to address this bias, combined with the difficulties in accessing Individual Patient Data (IPD), has led to the development and widespread adoption of Population-Adjusted Indirect Comparison (PAIC). These approaches enable estimation of adjusted treatment effects despite the absence of IPD in one of the treatment groups. Since their initial formalization in the early 2010s, PAICs have undergone substantial methodological development. Nevertheless, key questions remained, and this thesis focused on two of them. First, the quality of their practical implementation had been little investigated. Second, we explored the trade-offs that are made when relying on PAIC estimators instead of methods that exploit full IPD; as PAICs may be seen as a "default" option, that is, when complete IPD access is not possible.

The first part of this work consisted of a methodological review of the literature on PAICs published up to April 2021. The review aimed to describe the bibliographic characteristics of PAIC studies, to assess the quality of

their implementation against theoretical recommendations, and to examine reported results. It highlighted several important findings: a sharp increase in the number of publications in recent years, substantial heterogeneity in reporting quality, and strong evidence of publication bias. These findings motivated the formulation of reporting recommendations.

The second part of the thesis focused on the evaluation of weighting estimators within PAICs. We conducted a Monte Carlo simulation study to compare the performance of different approaches: Propensity Score Weighting (PSW) and Matching-Adjusted Indirect Comparison (MAIC), including adjustments based on first-order or both first and second-order moments. Data-generating mechanisms explored scenarios varying in covariate distribution, degree of overlap between populations, and adequacy of the adjustment model. The results showed that first-order MAIC estimators were more robust than PSW under conditions of limited overlap, even when covariates followed non-normal distributions. These findings suggest that the use of aggregated data, in the absence of IPD, primarily affects the results by limiting the flexibility of analyses rather than by substantially reducing the performance of MAIC estimators.

Remerciements

Tout d’abord, je tiens à remercier sincèrement les membres du jury qui m’ont fait l’honneur d’évaluer ce travail. Je remercie tout particulièrement le Docteur Carine Bellera et le Professeur Florian Naudet pour avoir assumé le rôle de rapporteur avec la charge que ça représente. Je suis également particulièrement reconnaissant au Docteur Tat-Thang Vo et Professeur Laura Richert pour avoir accepté de participer à ce jury de thèse.

Je remercie également les Professeures Sylvie Chevret et Florence Tubach pour m’avoir offert un poste hospitalo-universitaire respectivement à Saint-Louis puis à la Pitié-Salpêtrière, et qui m’auront permis de réaliser cette thèse. Conjuger ces postes à ce travail de thèse n’a pas été de tout repos mais s’est avéré extrêmement enrichissant professionnellement.

Et bien sûr ensuite à mes deux directeurs de thèse, Jérôme et David. Merci pour votre accompagnement au cours de ces quatre années, qui aura été à l’image de votre réactivité et soutien dans la dernière ligne droite de rédaction de ce manuscrit. Jérôme, ton expertise n’a d’égale que ton humour égaillant nos points visio “quasi” hebdomadaires. David, l’association d’autant de talent et de gentillesse chez une seule personne est probablement unique au monde ; mais je le savais déjà et sont les deux principales raisons qui m’ont donné envie de me lancer dans une deuxième thèse sous ta direction.

Cette aventure a donc commencé au SBIM de Saint-Louis, j’en profite pour saluer toute l’équipe et tout particulièrement mes co-AHUs, Guillaume, Sofia et Anaïs, ainsi que tous les doctorants et assimilés, Valentin, Luana, Florian, Loïc, Côme et Anouk. Que de bons moments passés ensemble, des déjeuners à l’hôpital jusqu’aux bars de Thessalonique et de Saint-Malo.

Et elle s’est prolongée dans le département de Santé Publique de la Pitié. Ces deux années passées en votre compagnie resteront un souvenir précieux. J’espère vous voir nombreux au pot de thèse. Mention spéciale aux anciens internes devenus amis Xavier, Camille, Titiane, Florian, Jeanne et Lisa.

Et donc aux co-AHU de la Pitié. A toi Bastien, comparse de bureau, aussi drôle qu’impressionnant travailleur. Je te souhaite le meilleur pour ton année de l’autre côté de la

Manche. A Samy, on te laisse reprendre le flambeau ; tu as déjà montré que tu avais toutes les qualités, premier degré.

À la bande d'Edouard Herriot, Jérémy, Alexis, Arthur, Gautier, Andrew, Pierre-Alexandre, Aubry, Lucas, Florent, Charles, Nicolas et Odile ; promis c'est la dernière fois que vous m'entendez me plaindre d'une thèse.

Aux copains de promo de Santé Publique qui font tous des parcours plus brillants les uns que les autres.

À Clément et Bruno, le trio inséparable qu'on forme depuis maintenant trois décennies s'est à nouveau montré décisif dans la dernière ligne droite.

Lita, thank you for cheering me up across the Atlantic. Here's to many more milestones to celebrate together, no matter the distance.

À Hecki, dont le regard ~~indifférent~~ bienveillant aura insufflé de la sérénité lors des séances de rédaction.

Des bisous à toute la famille, et plus particulièrement à Xavier, Maud et Caroline pour ces quelques jours passés ensemble en toute fin de rédaction. Xavier, je te laisse suggérer des pistes d'amélioration de ce travail.

À Papa et Grand-Mère qui ont suivi tout ça de là-haut.

Enfin à Maman et Gérard. Gérard, tu t'étais improvisé coach en course à pied l'an passé ; tu te serais reconverti coach en biostatistiques si tu en avais eu la possibilité. Ce travail aura pâti de ton absence, j'espère que tu le jugeras avec indulgence. A Maman, deux phrases ne suffiront pas à rendre compte de ton soutien indéfectible depuis toujours. Merci pour tout ton amour.

Contents

Résumé	i
Remerciements	iii
List of Figures	ix
List of Tables	xi
List of acronyms	xiii
1 Motivation & Context	1
2 The Challenge of Cross-Trial Population Differences in Indirect Treatment Comparisons	5
2.1 Notation and estimand	5
2.2 An overview of ITC methods	8
2.2.1 Naive indirect treatment comparison	9
2.2.2 Anchored indirect treatment comparison	10
2.2.3 Adjusted ITC	12
2.3 Population-Adjusted Indirect Comparisons	16
2.3.1 Matching-Adjusted Indirect Comparisons	16
2.3.2 Simulated Treatment Comparisons	19
2.3.3 Comparison of MAIC and STC	20
2.3.4 Distinguishing between prognostics factors and treatment-effect modifiers: a blurred distinction	23
2.3.5 Anchored versus unanchored PAIC assumptions	28
3 State of Practice	31
3.1 PAIC in the regulatory HTA environment	31
3.2 PAIC in peer-reviewed publications	36
3.3 Summary and comparative analysis of related reviews	62

3.3.1	Matching-Adjusted Indirect Comparison Studies in Oncology: A Scoping Review Focused on Reporting Quality. <i>Value in Health Regional Issues</i> . Farinasso CM, et al.	62
3.3.2	Population adjusted-indirect comparisons in health technology assessment: A methodological systematic review. <i>Research Synthesis Methods</i> . Truong B, et al.	62
3.3.3	A Comparison of Relative-Efficacy Estimate(S) Derived From Both Matching-Adjusted Indirect Comparisons and Standard Anchored Indirect Treatment Comparisons: A Review of Matching-Adjusted Indirect Comparisons. <i>Value in Health</i> , Cassidy O., et al.	63
4	The impact of limited access to IPD	65
5	Discussion and perspectives	97
5.1	Findings summary	97
5.2	Perspectives and avenues for future research	100
	Bibliography	103

List of Figures

2.1	Relative efficacy of A versus B , as evaluated in two different studies a and b	9
2.2	Anchored Indirect Comparison of A vs B , using C as an anchor	10
2.3	Summary of the notation used in this chapter: the objective is to estimate Δ_b (<i>i.e.</i> the A versus B treatment effect in the b population), the relative treatment effect of A versus B , while having access to IPD for the a study, and only AgD for the b study, using in the anchored case a common treatment arm C evaluated in both studies.	13
2.4	Adjusted unanchored Indirect Treatment Comparison of A versus B in the B population	14
2.5	Adjusted anchored Indirect Treatment Comparison of A versus B , using C as an anchor, in the B population	15
2.6	Interchanging status of sex as a PF, or PF and TEM, depending on contrast measure chosen.	25

List of Tables

2.1	Cardiovascular events depending on treatment and sex	24
2.2	Summary of pairwise Indirect Treatment Comparisons methods and of their assumptions	29

List of Acronyms

AgD Aggregated Data. 3, 16, 17, 19, 20, 22, 35, 62, 63, 65, 66, 99

CDA Canada's Drug Agency. 8

DGM Data Generating Mechanism. 4, 101, 102

EMA European Medicines Agency. 1, 99

ESS Effective Sample Size. 18, 100

EUnetHTA European Network for Health Technology Assessment. 31, 32, 34

FDA Food and Drug Administration. 1, 35

HAS Haute Autorité de Santé. 1, 3, 32

HTA Health Technology Assessment. ii, 1–3, 8, 11, 12, 17, 21, 29, 31–33, 35, 36, 65, 66, 97, 98, 100, 101

HTACG Member State Coordination Group on Health Technology Assessment. 33, 34, 101

IPD Individual Patient Data. ii, 2–4, 8, 14–17, 19, 20, 22, 27, 35, 62, 63, 65–67, 98–100

IQWiG Institute for Quality and Efficiency in Health Care. 32

ITC Indirect Treatment Comparison. ii, 2, 3, 5, 7–10, 12, 14–16, 20–22, 29, 31–36, 63, 99, 101

JCA Joint Clinical Assessment. 31, 33, 34, 99

MAIC Matching-Adjusted Indirect Comparison. i, ii, 2, 4, 16, 17, 20, 22, 23, 32, 33, 35, 62, 63, 65, 66, 99, 100, 102

ML-NMR Multilevel Network Meta-Regression. 2, 3, 16, 32, 62, 65, 66, 99

NICE National Institute for Health and Care Excellence. 3, 32, 33, 62, 100

NMA Network Meta-Analysis. 3, 15, 31, 32, 34, 35, 65, 98

PAIC Population-Adjusted Indirect Comparison. i, ii, 2–6, 12, 13, 16–18, 20–22, 26, 27, 29, 31–36, 62, 65–67, 97–102

PF Prognostic Factor. ix, 12, 15, 22–25, 27, 28, 100, 101

PSW Propensity Score Weighting. i, ii, 4, 65, 66, 102

RCT Randomized Controlled Trial. 2, 7, 9, 10, 12, 66

STC Simulated Treatment Comparison. i, ii, 2, 16, 19, 20, 22, 23, 32, 33, 35, 62, 65–67, 99, 102

TEM Treatment Effect Modifier. ix, 12, 13, 15, 20–25, 27, 28, 34, 100, 101

Contributions

Original Articles

1. Serret-Larmande A, Zenati B, Dechartres A, Lambert J, Hajage D. A methodological review of population-adjusted indirect comparisons reveals inconsistent reporting and suggests publication bias. *Journal of Clinical Epidemiology*. 2023;163:1-10. doi:10.1016/j.jclinepi.2023.09.004
2. Serret-Larmande A, Lambert J, Gaudry S, Hajage D. Assessing the Impact of Covariate Distribution and Positivity Violation on Weighting-Based Indirect Comparisons: a Simulation Study. *arXiv*. Preprint posted online July 20, 2025. doi:10.48550/arXiv.2507.12241. Under review in *Statistics in Medicine* at the time of writing.

Oral presentation

1. *EPICLIN 2024—Dijon* Serret-Larmande A, Lambert J, Hajage D. CO1. 1-Comparaison indirectes ajustées sur la population: une étude de simulation. *Journal of Epidemiology and Population Health*. 2024;72:202400.
2. *Workshop Laboratoires Pierre Fabre, 2023—Toulouse*: Comparaisons Indirectes Ajustées sur la Population: Théorie et Pratique

Poster presentation

1. *ISCB 2024—Thessaloniki* Serret-Larmande A, Lambert J, Hajage D. Population-Adjusted Indirect Comparisons: a simulation study.
2. *EPICLIN 2023—Nancy* Serret-Larmande A, Lambert J, Hajage D. Revue méthodologique des Comparaisons Indirectes Ajustées sur la Population. *Best poster award*
3. *EUROCIM 2023—Oslo* Serret-Larmande A, Lambert J, Hajage D. Population-Adjusted Indirect Comparisons: a methodological study.

Chapter 1

Motivation & Context

When regulators or payers evaluate a new therapy for market authorization or reimbursement, they require evidence that the therapy offers improved outcomes over the current standard of care, demonstrates a superior benefit-risk balance, or brings substantial economic advantages while maintaining comparable efficacy.

This evaluation is carried out by regulatory agencies and Health Technology Assessment (HTA) bodies. Regulatory agencies, such as the Food and Drug Administration (FDA) in the United States or the European Medicines Agency (EMA) in Europe, focus on establishing that a therapy's benefit-risk balance is favorable for a specific indication. HTA is a multi-disciplinary evaluation that integrates both clinical evidence and economic considerations to determine the cost-effectiveness based on the clinical, economic, and social impact of a health technology, leading to its eventual reimbursement by public or private health insurance [1].

A concrete example of HTA's role can be seen with the functioning of Haute Autorité de Santé (HAS) in France: after a drug is authorized for market use, the French HTA body assesses its medical benefit (*Service Médical Rendu*), and its medical improvement compared to existing therapies (*Amélioration du Service Médical Rendu*). The *Service Médical Rendu* rating determines if the therapy should be reimbursed by the healthcare system (and at what level), whereas the *Amélioration du Service Médical*, which ranges from I (major improvement) to V (no improvement), guides the price that can be negotiated [2].

Decisions carry tremendous consequences for patients whose outcomes may hinge on the efficacy of innovative treatments, and for society, which will eventually bears the cost of the negotiated price. Because HTA bodies must judge a therapy's relative effectiveness (or added benefit) against the current standard of care, they require robust comparative evidence.

Such evidence is primarily provided by clinical research. In this context, appropriate statistical analysis and careful interpretation of trial outcomes is paramount to drawing reliable conclusions. Numerous guidelines coming from both regulatory bodies and academic consortia aimed to standardize clinical research conduct, reporting and results interpretation [3, 4, 5]. Randomized Controlled Trials (RCTs) remain the gold standard for evaluating new treatments, in that they bring the most convincing evidence efficacy of a new health technology. However, head-to-head RCTs against all relevant comparators are often impractical, if not impossible. Numerous reasons may explain this situation, from temporal problems with alternative treatments that were not yet on the market when the pivotal RCT was designed, dimensional problems when the profusion of active comparators prevents from comparing each pair head-to-head, which is particularly true in the field of oncology [6, 7], to ethical or pragmatic reasons, such as rare diseases or extremely severe conditions (*e.g.* late-stage cancer refractory to approved therapies), where a single-arm trial may be deemed more appropriate [8, 9]. As a result, decision-making may rely on evidence synthesis of information fragmented across separate trials. Under these circumstances, Indirect Treatment Comparison (ITC) may be required to estimate a relative treatment effect when competing interventions of interest have not been compared head-to-head.

ITCs have a long-standing role in clinical research and are nowadays a recognized framework for evidence synthesis in the context of HTA. It is an umbrella term encompassing a variety of different methods, whose choice may depend on the presence of a common comparator, the size of the network (*i.e.* the number of treatment arms involved in the comparison), access to Individual Patient Data (IPD), or the target estimand of interest. Fundamentally, all these methods aim to address the same challenge: how to account for systematic differences in patient characteristics across the populations that may confound the estimated relative treatment effect.

Among ITC methods, Population-Adjusted Indirect Comparisons (PAICs) allow to adjust for cross-trial population heterogeneity in situations where IPD is only partially available. Such situations commonly occur when manufacturers are required, as part of HTA submissions, to provide comparative evidence against competing treatments, while lacking access to the competitors' IPD. They encompass mostly Matching-Adjusted Indirect Comparison (MAIC) and Simulated Treatment Comparison (STC) that were first mentioned in the literature in 2010 and allow for pairwise comparisons [10, 11, 12]. Very briefly, MAICs are an analogue to propensity score weighting, while STCs are based on outcome regression models. Both methods allow for estimating an adjusted effect while having access to the IPD for only one of the two treatments compared. More recently, Multilevel Network Meta-Regression (ML-NMR) has been introduced and allows extending PAIC beyond simple

pairwise comparisons to a larger network of treatments, while still integrating evidence coming both from studies with IPD and with Aggregated Data (AgD) [13, 14].

Research Questions and Objectives

Because they offer a pragmatic solution to a frequent practical problem (*i.e.* a pharmaceutical company that needs to conduct an ITC between a drug it developed and one developed by a competitor, with no access to IPD for the latter), PAICs have seen considerable practical adoption and methodological development since their initial description.

Some milestones can be identified among methodological developments in PAIC. The first and to date arguably most influential formalization of the theoretical assumptions and their practical use is the Technical Support Document 18 issued by the National Institute for Health and Care Excellence (NICE), conducted by Phillippo et al. (2016) [15]. Remiro-Azócar's PhD dissertation can also be cited among the in-depth rigorous specifications that helped strengthen the field [16]. It focused specifically on PAIC estimators that use a common treatment arm and contributed to advancing the field by providing rigorous theoretical developments regarding the nature of the estimand targeted by these estimators and the conditions for effect identification followed by extensive simulation work. Several works also unified in various ways the concepts of regression-based and weighting-based estimators, including Cheng, Tchetgen, and Signorovitch (2022), Park et al. (2024), and Remiro-Azócar (2022) [17, 18, 19]. Another instrumental development is the introduction of ML-NMR by Phillippo et al. et al. in 2020, which generalizes the concept of Network Meta-Analysis (NMA) to integrate both studies with IPD or only AgD to produce estimates that are transportable (that is adapted to different population distributions) across multiple treatment comparisons [13]. A more in-depth analysis of these contributions will be developed later in this document.

Alongside these theoretical advances, the use of PAICs in applied settings has grown rapidly. HTA agencies, notably NICE in the UK, began receiving submissions including PAICs in the early 2010s [20]. Similar trends have been observed in France, where the Haute Autorité de Santé (HAS; French HTA body) increasingly encounters PAIC-based evidence in reimbursement submissions. In parallel to the regulatory context, a similar trend in the literature were also noted by Remiro-Azócar, mentioning a rapid PubMed search that yielded a significant increase in the number of publications[16]. This practical uptake raised critical questions that shaped the objectives this work, and the two research articles featured in this thesis aimed to address knowledge gaps in this domain.

The first work focused on the practical aspect of the problem, that is the implementation of PAIC in the peer-reviewed literature, therefore aiming to bridge the gap in information between existing methodological developments and rather limited knowledge on practical implementations. It sought to address broad yet unanswered bibliographical questions regarding the trend and evolution of the PAIC literature over time in terms of volume and authorship. Additionally, it was designed to assess the implementation quality of PAIC relative to methodological guidance references, with the aspiration of providing insight into the consequences there may have on the credibility and acceptability of submitted evidence. This work resulted in the first peer-reviewed article, a methodological review that searched PubMed for articles featuring PAIC until 2nd April 2023, analyzed the methodology employed, the reporting quality and the reported results, and additionally provided reporting recommendations. It has been published in the *Journal of Clinical Epidemiology* in November 2023 “A Methodological Review of Population-Adjusted Indirect Comparisons Reveals Inconsistent Reporting and Suggests Publication Bias” [21].

The second component of this thesis was rooted in a central question: what are the consequences of using PAICs in place of methods that use the full IPD on the accuracy and precision of treatment effect estimates? Following this idea, we compared different weighting estimators based on the availability of IPD. We focused on weighting estimators mostly because they were revealed by the methodological review to be overwhelmingly the most prevalent implementation of PAIC. This led to the second original article featured in this doctoral thesis: Assessing the Impact of Covariate Distribution and Positivity Violation on Weighting-Based Indirect Comparisons [22]. At the time of writing this manuscript, the article is under review in *Statistics in Medicine* and a first version is available on a pre-print server. It compared the performance of weighting-based estimators using full-IPD, that is Propensity Score Weighting (PSW), to two different implementations of MAIC. The performance was evaluated through Monte Carlo simulations in terms of accuracy and precision of estimations, depending on various Data Generating Mechanisms (DGMs) that included non-normally distributed covariates and positivity assumption issues. Results from this simulation study contributed to a better understanding the actual trade-off that is being made when defaulting to using PAIC due to the partial lack of IPD.

Chapter 2

The Challenge of Cross-Trial Population Differences in Indirect Treatment Comparisons

The following section formalizes the identification and estimation problem underlying ITC. For clarity, the exposition is limited to the case of pairwise comparisons, which remains the most frequently encountered in the context of PAIC.

2.1 Notation and estimand

Causal framework and potential outcomes

Hereafter, we are interested in comparing the efficacy of a given treatment, noted treatment A , with another treatment B . For each individual, we define two potential outcomes:

- $Y(A)$: the outcome that would be observed if the individual were to receive treatment A .
- $Y(B)$: the outcome that would be observed if the individual were to receive treatment B .

In practice, each individual receives only one treatment, denoted by the variable $Z \in \{A, B\}$. Thus the observed values of the outcome Y under treatment A and B are written $y(Z = A)$ and $y(Z = B)$ respectively, simplified as $y(A)$ and $y(B)$.

Target estimand

A contrast effect is marginal if it represents the effect averaged over an entire population. It is conditional if it represents the average effect for a specific subgroup of the population, where the subgroup is defined by a particular set of covariate values. A debate exists regarding which of these estimands is most appropriate for policymakers and consequently for PAIC [23, 24, 25]. We will summarize the main arguments in a later section and here introduce the notation for both marginal and conditional estimands.

The marginal average treatment effect is a common policy-relevant estimand. It is defined as

$$\Delta_M = g(\mathbb{E}[Y(A)]) - g(\mathbb{E}[Y(B)]) \quad (2.1)$$

Here $\mathbb{E}$ represents the expectation over the full distribution of individuals of a target population, and g is a link function which transforms the outcome to a linear scale (meaning that the effect of the different underlying variables predictive of the outcome is additive on that scale). A marginal estimand answers the question: “What is the effect, as measured by Y , of switching the entire population of interest from treatment B to A ?”.

The conditional average treatment effect can be especially of interest in the context of individualized-treatment framework. It is defined as

$$\Delta_C = g(\mathbb{E}[Y(A)|\mathbf{X} = x]) - g(\mathbb{E}[Y(B)|\mathbf{X} = x]) \quad (2.2)$$

where $\mathbf{X}$ is a matrix of covariates associated with the outcome Y , and x are specific values for these covariates, defining a subpopulation under consideration. A typical conditional estimand of interest is the conditional effect at the average values of $\mathbf{X}$, that is $\mathbf{X} = \bar{x}$.

The g link function specifies the scale for either estimand.

- If Y is continuous, g is usually the identity function, and $\Delta_\bullet$ is the mean difference between treatment groups ($g(x) = x$).
- If Y is binary, g can be the identity function ($\Delta_\bullet$ is a difference of proportions), log function ($g(x) = \log(x)$) when $\Delta_\bullet$ is a log relative risk, or logit function ($g(x) = \log(\frac{x}{1-x})$) when $\Delta_\bullet$ is a log odds ratio.
- If Y is a time-to-event outcome, the log function yields a hazard ratio.

Hereafter, for the sake of simplicity, we will refer to the estimand of interest as simply Δ

when the description is equally suited to both a marginal and conditional estimand.

Estimating a treatment effect in a cross-population setting

If we consider $\bar{y}(A)$ and $\bar{y}(B)$ to be sample means of Y under A and B respectively, we have an estimate of Δ with $\hat{\Delta} = g(\bar{y}(A)) - g(\bar{y}(B))$. In the context of a Randomized Controlled Trial (RCT), this would be an unbiased estimate of Δ since, by definition, the only systematic difference in outcomes between treatment groups would be attributable to the treatments themselves. However, in the absence of randomization, treatment groups are coming from different populations and thus not only differing by the treatment that they are receiving but also in characteristics susceptible to influence the outcome. The estimation problem is thus to isolate the treatment effect from these cross-trial population differences. In order to adequately estimate the treatment effect, the estimation process has to identify and account for the differences between populations that influence the outcome. Moreover, if a marginal estimand is of interest, the problem becomes a transportation problem requiring the estimation of a treatment effect in a given target population. In the context of ITC, the two treatments A and B have been evaluated in two separate studies, which will be noted a and b . Participants in these studies have thus been sampled from two different populations. We can define a set $\mathbf{X}$ of baseline characteristics that may influence the outcome. We note x_a and x_b the actual values of $\mathbf{X}$ within a and b respectively. The observed sample mean of Y in each of these populations are noted $\bar{y}_a(A)$ and $\bar{y}_b(B)$.

The practical estimand-estimation dependency

The cross-trial population baseline imbalance leads to different challenges:

- **Defining the target estimand.** It entails specifying the population of interest, the groups being compared, the outcome which will be measured, and the contrast of interest. This step is guided by scientific rationale and the research objective.
- **Assessing identifiability.** Once the target estimand has been specified, the next question is whether it is possible to estimate it without bias from the available data. This identifiability assessment requires listing the assumptions that link the observed data to the counterfactual contrast of interest (e.g., exchangeability between treatment groups, either unconditionally or conditionally on all confounders, consistency of interventions, positivity between treatment groups) [26]. Determining plausibility of these assumptions is a prerequisite before attempting estimation.

- **Estimating the effect.** Provided identifiability is judged plausible, the final step is to choose an estimation strategy adapted to the available data.

In theory, these three steps are conceptually distinct: the identification of the causal estimand should be determined by the scientific rationale and the research objective, while the estimation procedure corresponds to the practical tools yielding an estimation. However, in the context of ITC, they may become intertwined: the definition of the population of interest, the choice of the causal contrast, the hypotheses for identifiability, the adjustment model and the estimator choice may be dependent because dictated by the same practical constraints. This is especially true for the choice of target population and estimand is often constrained by the availability of data, in particular by whether Individual Patient Data (IPD) are accessible. For instance, while one might wish to define the effect in the overall population, this estimand may not be identifiable without IPD in all relevant trials. In such cases, the target estimand is often chosen conditional on what can plausibly be identified. This dependency between scientific aims, identifiability, and practical estimation tools is especially evident in PAIC, where the lack of IPD typically forces a redefinition of the target population and the estimand itself. This problem is being mentioned by HTA bodies, which call to explicitly define a target population, as for instance *Methods Guide For Health Technology Assessment* issued by the Canada's Drug Agency (CDA): “*The ability to draw high-certainty conclusion can be equally reliant on the rigour of the ITC ...and the alignment between trial populations and target population.*” [27]. Consequences on the modification of the target population by the estimation procedure are also mentioned in the recently issued *Methodological Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons* by the Member State Coordination Group on Health Technology Assessment: “*... the final target population must be described in detail, especially if trimming or truncation methods are applied. An investigation of whether this target population sufficiently represents the population selected for the original research question is required.*” [28].

2.2 An overview of ITC methods

The ITC framework encompasses different methods, which choice depends on the estimand of interest, the nature of the cross-population difference, and practical data constraints. We will review the different possibilities, alongside their use-case and assumptions.

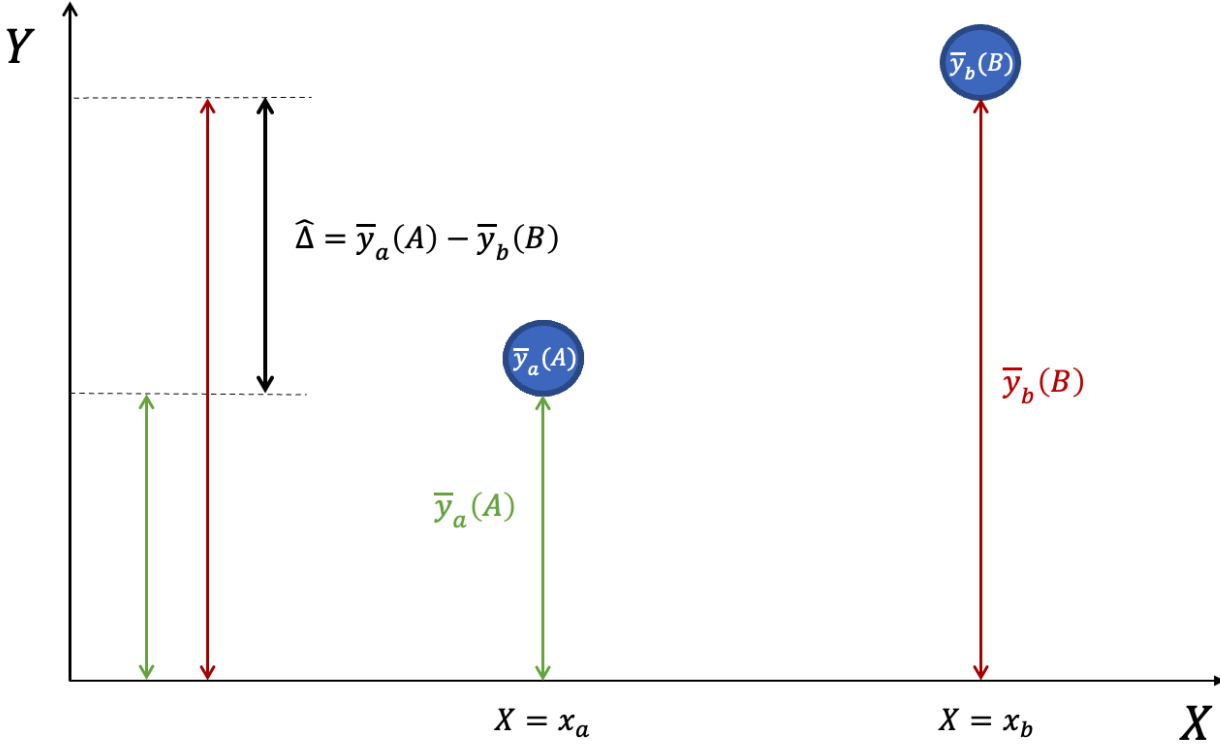


Figure 2.1: Relative efficacy of A versus B , as evaluated in two different studies a and b

2.2.1 Naive indirect treatment comparison

Keeping with previous notations, we are interested in estimating the relative treatment effects between two treatments A and B that have been evaluated in two different populations a and b . The estimate of the treatment effect is simply obtained as

$$\hat{\Delta} = g(\bar{y}_a(A)) - g(\bar{y}_b(B)) \quad (2.3)$$

with $\bar{y}_a(A)$ and $\bar{y}_b(B)$ being the sample means of the outcome in both treatment arms. Such an analysis is illustrated in Figure 2.1: the treatment effect estimation would simply be the distance on the y -axis between the two $y_a(A)$ and $y_b(B)$ outcome point estimates (Figure omits link function g for conciseness). This is a simple unadjusted indirect comparison. $\hat{\Delta}$ obtained with such a naive ITC will be biased as soon as there are systematic differences between $\mathbf{X}_a$ and $\mathbf{X}_b$ in terms of characteristics that may influence the outcome. In practice, the situation of absence of confounders (also known as the exchangeability assumption) would be especially rare and remains a very strong assumption requiring statistical evidence [29, 30, 31].

Hence, a naive comparison based on the raw difference of observed outcomes using data from non-randomized studies will produce a biased estimate. The defining feature of an RCT

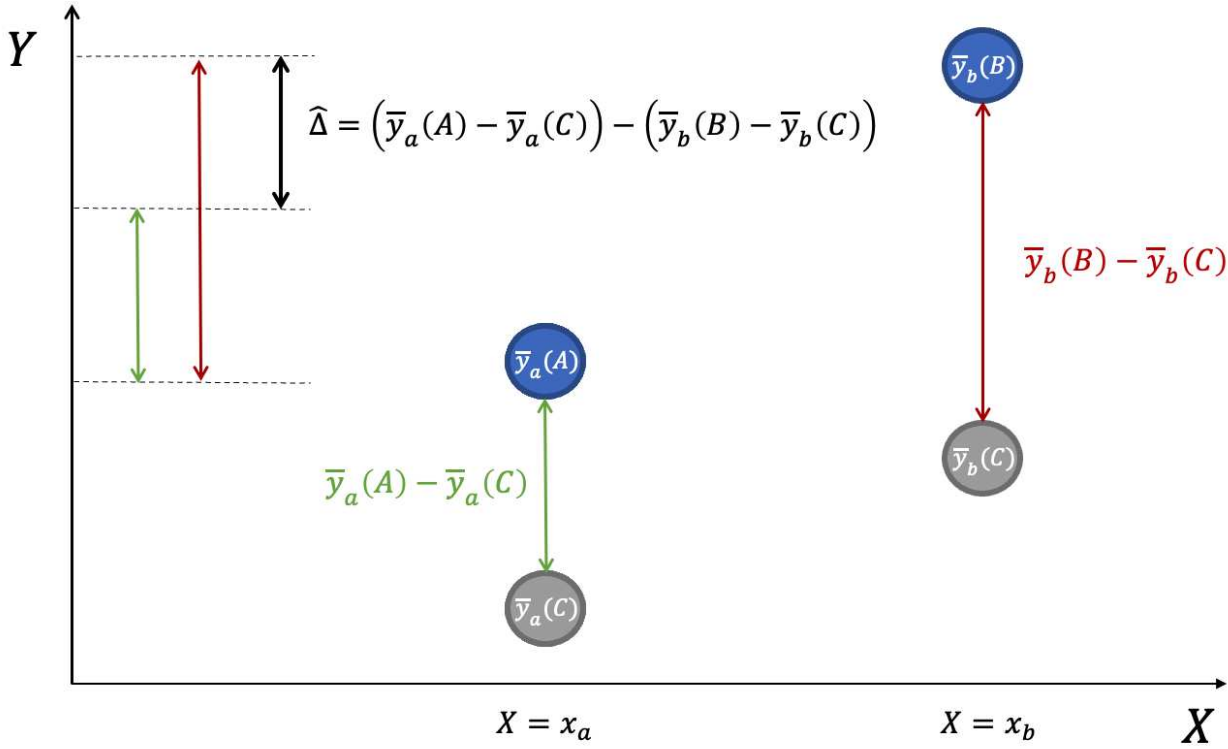


Figure 2.2: Anchored Indirect Comparison of A vs B , using C as an anchor

is that randomization balances confounders (both measured and unmeasured) between the treatment arms. Consequently, and provided that the RCT has been correctly conducted, any difference observed between an intervention of interest and its comparator can be causally attributed to the different treatment evaluated (up to some level of statistical uncertainty). Conversely, in the absence of randomization, baseline characteristics of compared groups may differ and an investigator willing to compare these groups must consequently adjust for confounders of the association between the treatment and the outcome.

2.2.2 Anchored indirect treatment comparison

When two treatments of interest have not been directly compared in an RCT, it may still be possible to use existing randomized evidence to compare these treatments, provided each treatment has been evaluated in a separate randomized trial against the same control. The common comparator allows to exploit the within-trial randomization to yield robust ITCs that relaxes the exchangeability assumption (*i.e.* partly takes into account confounding bias) as compared to a naive ITC.

Considering the same two studies a and b , that evaluated respectively A and B . Let's assume that each one of these studies additionally compared an identical third treatment

C. Moreover, each study was randomized, and thus the within-study estimated relative treatment effect is unbiased. Assuming that the treatment *C* is indeed identical across both studies, it is possible to leverage the two within-studies relative treatment effect (comparing *A* to *C* within *a*, and *B* to *C* within *b*) to obtain an unbiased estimate $\hat{\Delta}$. The common treatment arm *C* is used as an "anchor" in the *A* versus *B* comparison, in the sense that the outcome value under *A* and *B* is compared through their respective differences to *C*. In other words, in an anchored comparison, the relative treatment effect Δ is estimated using a difference-in-difference, as shown below:

$$\hat{\Delta} = (g(\bar{y}_a(A)) - (g(\bar{y}_a(C))) - (g(\bar{y}_b(B)) - g(\bar{y}_b(C)))) \quad (2.4)$$

For an unbiased estimate of Δ to be obtained, it is only required that either i. the within-trial relative differences (*i.e.* $g(\bar{y}_a(A) - \bar{y}_a(C))$ and $g(\bar{y}_b(B) - \bar{y}_b(C))$) be constant across every level of **X** (*i.e.* there is no heterogeneity of the treatment effect), or ii. that the baseline characteristics defining heterogeneity strata of the treatment effect are balanced across the two compared populations (*i.e.* in other words, the variables modifying the treatment effect are balanced across the populations *a* and *b*). This assumption is called the constancy of relative effects. The principle of an anchored comparison is illustrated in Figure 2.2 (Figure omits link function *g* for conciseness). A practical prerequisite is that *C* is truly the same intervention across trials, regarding dose and regimen (or practical implementation in the case of an intervention), and that the outcome definition and measurement is identical, so that the within-study differences are actually comparable.

The idea to leverage a common comparator to address cross-trials comparison issues in the context of HTA can be traced back to 1988, with the work of Eddy (1989) [32]. This work mentions a situation where "*two technologies in a family*" have been compared to "*a common third technology, and a policymaker wants to compare the first two technologies to each other*". It then proceeds to illustrate using a Bayesian framework how one can simply incorporate the within-trial relative differences with the common comparator to yield an estimate of the relative efficacy of the pair of treatments of interest. However, it does not give an explicit statistical explanation of the advantage of comparing relative treatment effect estimates over absolute outcome values. Similarly, Higgins and Whitehead (1996) mentions the possibility of estimating pairwise indirect comparison from a meta-analysis comparing three treatments [33]. However their emphasis is on proper variability estimation and independence of trials included in the meta-analysis, rather than on bias reduction via difference-in-differences estimation.

Finally, the pivotal work came with Bucher et al. (1997) who formalized the argument

for anchored indirect comparisons [34]. They explicitly mentioned the superiority of using a common treatment arm when estimating a relative efficacy of two treatments that have been compared in two different trials, as compared to unanchored indirect comparison, in terms of bias reduction. The article is constructed around an example comparing odds ratio, but the principle extends to any outcome and contrast measure for which a linearizing link function exists. The estimation of a treatment effect using an unadjusted pairwise comparison of relative treatment effects coming from two different RCTs with a common treatment arm is nowadays commonly referred to as a Bucher ITC in the literature.

Prognostic Factors and Treatment-Effect Modifiers To clarify the benefits and limitations of using an anchored ITC, it is necessary to introduce the distinction between two types of confounders: PFs (PFs) and TEMs (TEMs). A PF is a baseline characteristic whose association with the outcome is constant regardless of the treatment received. A TEM is a baseline covariate that modifies the effect of treatment on the outcome, that is, there is a treatment–covariate interaction on the chosen effect scale (e.g., risk difference, log-odds, or log-hazard). Associated with a given treatment, a TEM will lead to a synergistic, or antagonistic effect on the outcome. Real-life examples of such interaction relationship include the synergistic effects observed between the wild-type KRAS protein genotype and panitumumab on survival in metastatic colon cancer [35], or the antagonistic effect of rifampicin on oral contraceptive pills when both medications are taken in conjunction [36]. Of note, the concept of statistical interaction is referred to using different labels depending on the domain, hinting at the perspective of interest: statistical heterogeneity or interaction factor (classically in the clinical trials literature), predictive factor (especially in the industry) [37, 38], or Treatment Effect Modifier in the domain of HTA and PAICs. We will thereafter use the term TEMs. Of note, the denomination “Treatment Effect Modifier” itself reflects the perspective of the actual effect studied in a causal analysis, but is a matter of interpretation as to which variable modifies which. From a statistical perspective, the effect of a treatment is not necessarily more modified by the TEM than the other way around. A follow-up discussion of the links between TEM and PF and the implication for anchored and unanchored comparisons is presented in section 2.3.4.

2.2.3 Adjusted ITC

Imbalance in confounders across populations naturally raises the question of statistical adjustment. An additional constraint that has not been mentioned so far is the population that is being targeted. Following the notation introduced earlier, we denote the two populations

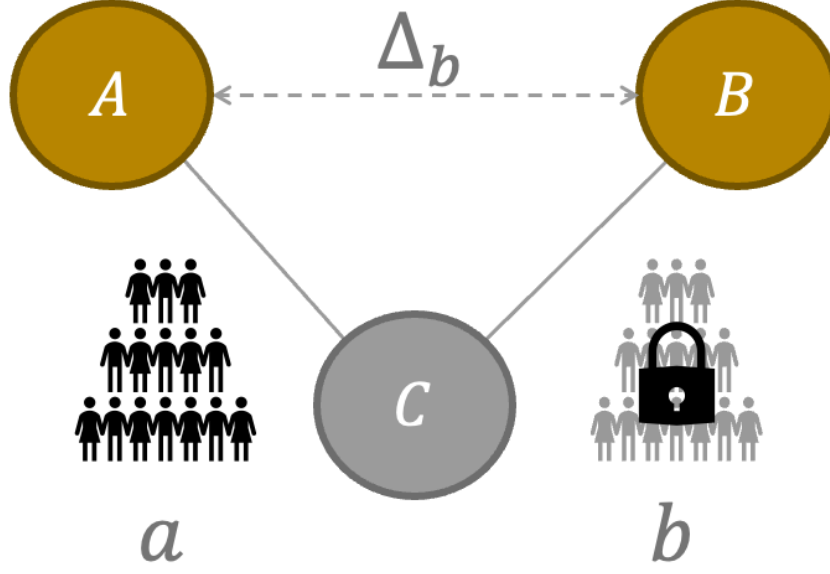


Figure 2.3: Summary of the notation used in this chapter: the objective is to estimate Δ_b (*i.e.* the A versus B treatment effect in the b population), the relative treatment effect of A versus B , while having access to IPD for the a study, and only AgD for the b study, using in the anchored case a common treatment arm C evaluated in both studies.

as the “source” a and “target” b populations. In the framework of PAIC, we will consider that the population of interest for the difference in efficacy between A and B is the b population, for reasons that will be explained shortly after. We will note the b -specific treatment effect estimand Δ_b . Figure 2.3 summarizes the notations introduced so far.

Importantly, this precision matters not only for marginal estimands, that are intrinsically tied to a specific population distribution, but also for conditional estimands. Indeed, in the presence of TEM, any conditional estimand will also be inherently population-specific [23]. As a reminder, the conditional effect is defined as $\Delta = g(\mathbb{E}[Y(A)|\mathbf{X} = x]) - g(\mathbb{E}[Y(B)|\mathbf{X} = x])$, thus the conditional effect of A versus B will not be constant if $\mathbf{X}$ influences this relative difference, that is if the vector $\mathbf{X}$ contains TEMs. In the case of PAIC, the population-specific conditional estimand is usually defined at the average level of the combination of TEMs of the target population (*i.e.* in the b population, thus at $X = \bar{x}_b$).

Overall, in the framework of PAIC, the objective of the adjustment procedure is to estimate the treatment effect of A versus B that would have been observed in b , using available data. We have observed the mean outcome values of y under treatments B , and C in the case of an anchored comparison ($\bar{y}_b(B)$ and $\bar{y}_b(C)$), but not under treatment A , *i.e.* $\bar{y}_b(A)$. Thus, the task of the adjustment procedure is to obtain an unbiased estimate of the value of the outcome y under A in the population b , *i.e.* $\bar{y}_b(A)$ in an unanchored comparison,

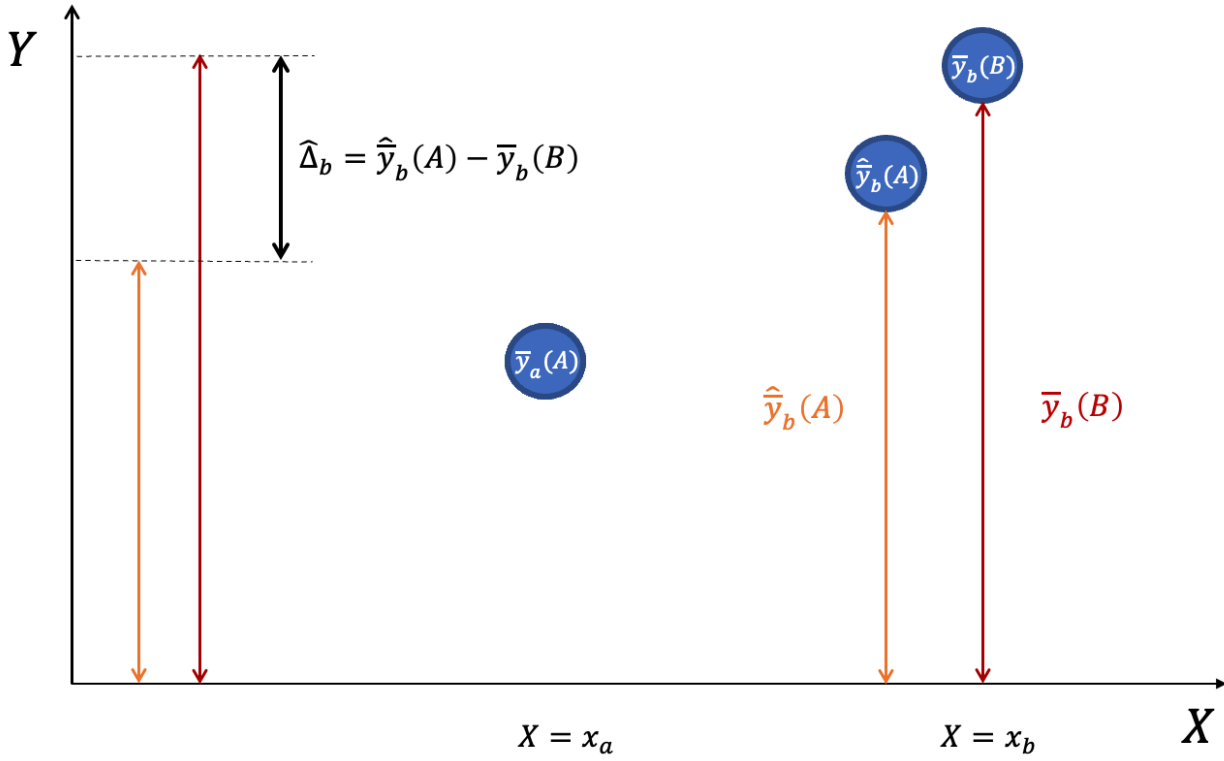


Figure 2.4: Adjusted unanchored Indirect Treatment Comparison of A versus B in the B population

and of the relative difference between A and C as measured by y in the b population in an anchored comparison: $\bar{y}_b(A) - \bar{y}_b(C)$. The unobserved component $\bar{y}_b(A)$ (and $\bar{y}_b(C)$ if anchored) is replaced by its estimate $\hat{\bar{y}}_b(A)$ (and $\hat{\bar{y}}_b(C)$) in equations 2.3 and 2.4.

$$\hat{\Delta}_b = g(\hat{\bar{y}}_b(A)) - g(\bar{y}_b(B)), \quad (2.5)$$

$$\hat{\Delta}_b = (g(\hat{\bar{y}}_b(A)) - g(\hat{\bar{y}}_b(C))) - (g(\bar{y}_b(B)) - g(\bar{y}_b(C))) \quad (2.6)$$

A schematic illustration of adjusted unanchored and anchored ITCs is shown in Figures 2.4 and 2.5 respectively (Figures omit the g link function for conciseness).

Different methods exist to obtain $\hat{\bar{y}}_b(A)$ (and $\hat{\bar{y}}_b(C)$). Broadly, they fall into two main categories, with either outcome modeling (e.g., regression, g-computation), or exposure modeling (propensity score methods).

In the context of ITC, the choice of the adjustment method depends on the availability of IPD. If no IPD are available, only unadjusted unanchored or anchored comparisons would be technically feasible. Unadjusted unanchored comparisons make the assumption of constancy

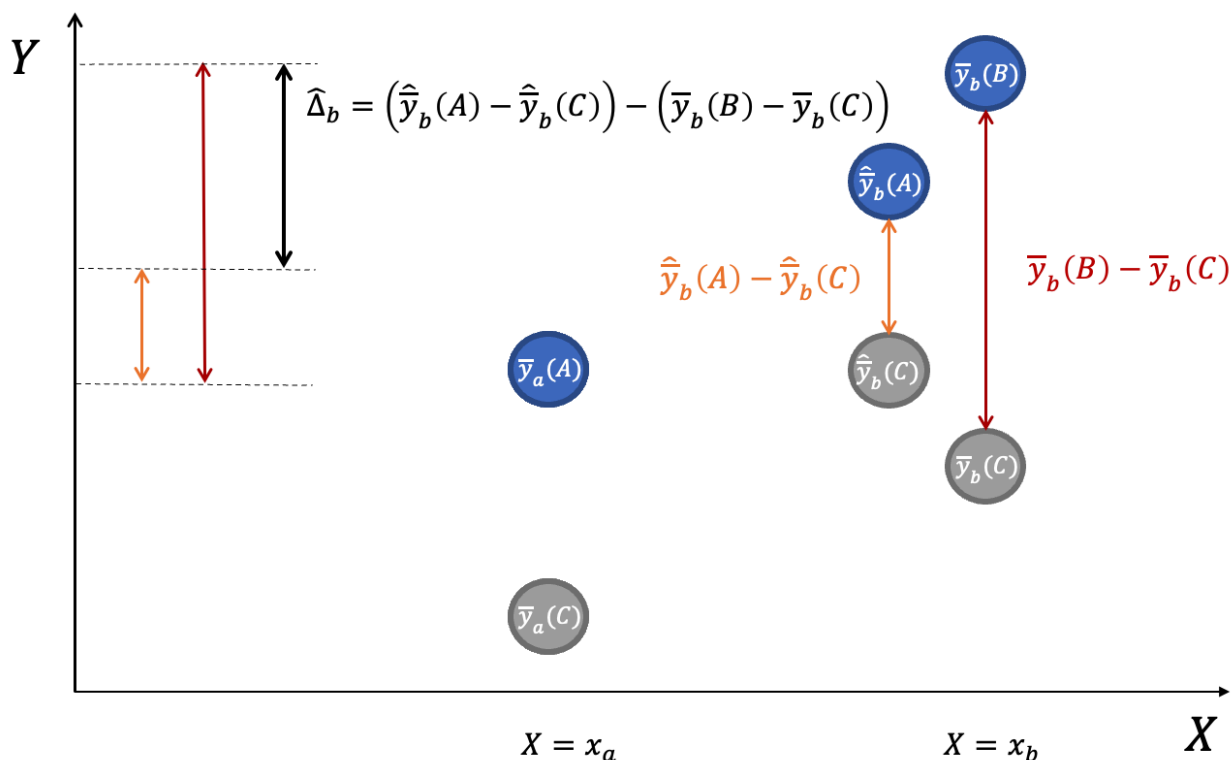


Figure 2.5: Adjusted anchored Indirect Treatment Comparison of A versus B , using C as an anchor, in the B population

of absolute effects, which implies that the distribution of all prognostic factors and effect modifiers is sufficiently similar across the two trial populations for the observed absolute effect in one population to be directly generalized to the other. In other words, it makes the assumptions of the absence of confounders.

As already stated, unadjusted anchored comparisons, also known as Bucher ITC, rely on the less strong assumption of constancy of relative effects, also known as the homogeneity assumption in the Network Meta-Analysis (NMA) literature. This requires that prognostic factors may vary between populations, but that effect modifiers must be either absent or balanced across studies involved in the indirect comparison, allowing valid estimation of the relative effect between the treatments of interest.

In practice, the assumption of no imbalance in PFs and/or TEMs across populations is often difficult to prove or implausible, which limits the validity of unadjusted comparisons. In such cases, adjustment methods can be used, but these require access to IPD. When IPD are available for both studies, methods for pairwise ITC converge to a specific case of IPD NMA. Interestingly, propensity score methods are rarely discussed in the context of IPD NMA, even though they are theoretically capable of addressing latent population differences. This topic will be examined in more details in Chapter 4. Finally, when IPD are available

for only one study, PAIC methods are required. Two main methods exist, and are based on either outcome modeling, which forms the basis of Simulated Treatment Comparison (STC), or trial assignment modeling, which underlies Matching-Adjusted Indirect Comparison (MAIC). Doubly robust approaches have also been suggested [18, 39]. More recent methodological developments extended these methods beyond simple pairwise comparison, with the introduction of Multilevel Network Meta-Regression (ML-NMR) by Phillippo et al. (2020) [13]. The following sections provide explanations of these different methods.

2.3 Population-Adjusted Indirect Comparisons

2.3.1 Matching-Adjusted Indirect Comparisons

Signorovitch et al. (2010) introduced the concept of MAICs in a seminal paper, “Comparative effectiveness without head-to-head trials: a method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept” [10]. In the absence of IPD for one of the studies in the ITC, conventional adjustment methods cannot be applied. Propensity score methods involve fitting a logistic regression model via maximum likelihood, inferring individual propensity scores using prediction from that model, and then resampling accordingly based on the estimated propensity scores (which is either done through matching, weighting or stratification). When only aggregated data are available for one of the studies, none of these steps are feasible: the logistic model cannot be fitted, individual propensity scores cannot be estimated, and the full population cannot be resampled. PAICs circumvent these limitations with two key characteristics: 1. they target an estimand which requires only resampling the IPD population, and 2. they make use of the method of moments, rather than the maximum likelihood estimation, to estimate the conditional individual probability of belonging to the IPD trial.

MAIC estimand

MAIC targets the treatment effect in the population for which only Aggregated Data (AgD) are available, by reweighting the IPD to match the AgD population covariate distribution. This estimand is a marginal effect, analogous to the Average Treatment effect in the Control estimand in propensity score weighting (considering that the “control” population is the b population from which the AgD come from). Hence, in other words, MAIC will estimate the treatment effect in the population in which the investigator does not have access to IPD. Hence, in other words, MAIC will estimate the treatment effect in the population for which

the investigator does not have access to IPD. This feature has a somewhat counterintuitive implication in the context of PAICs conducted for HTA dossiers: MAICs allow the relaxation of the constancy assumption by accounting for variables that are imbalanced between the two studies (through an adjustment procedure that will be described later) yet the resulting estimand corresponds to the treatment effect in the competitor's trial population, rather than in the population from which the IPD come from (*i.e.* the MAIC's study sponsor).

Method of moments

Partial unavailability of IPD has two main implications. First only some summary measures of baseline characteristics are typically available, *e.g.* means, medians, proportions, and variability measures such as standard deviations or interquartile ranges. Second, information on the joint distribution of covariates is almost systematically missing. Consequently, conventional multivariable logistic regression cannot be fitted. To address this limitation, Signorovitch et al. leveraged the method of moments to estimate balancing weights enabling to resample IPD to match the moments of the AgD population [10].

The method of moments traces back to the 19th century, namely with the work of Pafnuty Chebyshev in 1887 on large numbers [40].

The method aims at equating the k -order moments of two multivariate distributions. In the context of PAIC, consider two studies sampled from two different populations a and b , from which we observed sample distributions of $\mathbf{X}$, a multivariable matrix of p covariates:

- In the index study a (study using a sample of population a), we had access to the individual-level data for each covariates $\mathbf{X}_{(a)i,j}$, with $i = 1, 2, \dots, n_a$ and $j = 1, 2, \dots, p$, with n_a being the number of subjects and p the number of covariates, which allowed to estimate the entire distribution of $\mathbf{X}$.
- In the target study b , we only estimated the first k -order moments of each covariate of $\mathbf{X}$, written $\mu_k(\mathbf{X}_{(b),j})$, with μ_k denoting the k -th raw moment, that is $\mu_k(\mathbf{X}_{(b),j}) = \mathbb{E}[\mathbf{X}_{(b),j}^k] = \sum_i^{n_b} \frac{\mathbf{X}_{(b)i,j}^k}{n_b}$.

Thus the objective is to find a vector of weights w of length n_a such that the k -moments of the weighted distribution of $\mathbf{X}_{(a)}$ equates the sample moments of $\mathbf{X}_{(b)}$, that is:

$$\mu_k(w\mathbf{X}_{(a)}) = \mu_k(\mathbf{X}_{(b)})$$

for all k .

An exponential parametric form is specified for $\hat{w}$ such as $\hat{w} = \exp(\beta_0 + \beta \mathbf{X}_{(a)})$, with β a vector of coefficients. Other parametric forms are theoretically possible, though the exponential is preferred because i. weights are always positive, and ii., it is proven to be equivalent to entropy balancing thus ensuring that the weights are as close as possible to 1 (that is as close as possible to no weight at all) [41].

Finally the vector β , and the corresponding weights $\hat{w}$, is estimated by solving the following system of equations for all j and for $k = 1, 2, \dots, K$:

$$0 = \frac{1}{n_a} \sum_i^{n_a} \hat{w} \mathbf{X}_{(a)i,j}^k - \mu_k(\mathbf{X}_{(b)j}^k)$$

In the context of PAIC, $K = 1$ almost systematically, and potentially 2 at most [21]. Balancing on the first-order moment equalizes the means of the covariate distributions, and first- and second-order equalizes their means and variances.

In practice, reweighting implies that some individuals contribute more to the analysis (over-represented) while others contribute less (under-represented), in order to match the target population. In the context of PAIC, this inevitably reduces the amount of independent information carried by the data: although the nominal sample size remains unchanged, the Effective Sample Size (ESS) is smaller. Estimating the ESS is therefore crucial, as it quantifies the loss of precision due to reweighting and provides a measure of the sample size reduction. An estimation of the sample size reduction can be estimated using the ESS, as

$$\text{ESS} = \frac{(\sum_{i=1}^{n_a} \hat{w}_i)^2}{\sum_{i=1}^{n_a} \hat{w}_i^2}$$

The larger the difference between x_a and x_b (*i.e.* the more different the initial populations are), the smaller the resulting post-weighting sample size will be.

MAIC estimator

The estimation of the average outcome $\hat{y}$ under treatment $Z \in \{A, C\}$ in trial b , using observation from trial a is the average outcome weighted with $\hat{w}$, *i.e.*

$$\hat{y}_{(b)}(Z = z) = \frac{1}{\sum_i \hat{w}_i} \sum_i^{n_a} \hat{w}_i y_{(a)i}(Z = z)$$

The anchored and unanchored treatment effect estimators are then obtained by plugging these estimated quantities in equations 2.5 and 2.6. The resulting variance of the treatment

effect can be estimated by summing the variances of the outcome under each treatment arms involved in the indirect comparison.

2.3.2 Simulated Treatment Comparisons

The concept of adjusted indirect comparisons using a conditional outcome was formalized by Ishak in 2015 [12], building on some previous work by Caro and Ishak [11]. They coined the method Simulated Treatment Comparison, since originally it entailed simulating the missing IPD for the index study. However, STC do not necessarily require simulations as two main implementations are described.

Both approaches setup a conditional outcome model that estimates the outcome that would have been observed under treatment Z in a population with characteristics of the AgD population. Initially, the following outcome models are fitted in the IPD sample, for respectively the unanchored (Equation 2.7) and anchored cases (Equation 2.8):

$$\hat{y}_{(b)} = g^{-1}(\beta_0 + \beta_{PF}\mathbf{X}_{(b)PF} + \beta_{TEM}\mathbf{X}_{(b)TEM}) \quad (2.7)$$

$$\hat{y}_{(b)}(Z) = g^{-1}(\beta_0 + \beta_{PF}\mathbf{X}_{(b)PF} + \beta_Z Z + \beta_{TEM}\mathbf{X}_{(b)TEM} + Z\theta_{TEM}\mathbf{X}_{(b)TEM}) \quad (2.8)$$

The subsequent step of estimating the outcome $y_{(b)}(Z)$ is dependent on the version of the STC implementation.

Plug-in implementation

In practice, the most common form of STC implementation can be referred to as the “plug-in” method. It was first described succinctly by Ishak, Proskorovsky, and Benedict (2015), and later in more detail by Phillippo et al. (2016) [12, 15]. As a straightforward two-step process, its simplicity likely explains its appeal in applications. First, a regression model is fitted in the IPD cohort (population a), including all the relevant confounders, as laid out in 2.7 and in 2.8. Second, a prediction of the outcome under each treatment modality is being made (A for unanchored comparison, and A and C for anchored comparison). If the regression model includes treatment-effect modifiers, the prediction needs to be made at the average level of these variables. Following description by Phillippo et al. (2016), this “average-level” conditional prediction can be easily extracted from a regression model by

first centering the covariates of the IPD cohort included in the model by the average value of these same variables in the AgD population. Consequently the average effect of the TEMs in the centered AgD population would be null, and the treatment effect is simply either the intercept of the model in an unanchored comparison (the conditional outcome under treatment A in the b population), or the treatment effect coefficient in an anchored comparison (the conditional effect of A versus C in the b population). Similarly to MAIC, the treatment effect variance can be estimated by summing the variances of the outcome under each treatment arms involved in the indirect comparison.

Simulation implementation

The STC method introduced in the original article, and less frequently used, is akin to g-computation as it requires i. simulating the missing arm, ii. predicting individual outcome values in the simulated arms, and iii. obtaining a marginal outcome by averaging the individual predictions. Missing information for adequate target trial covariate simulations (such as covariate correlations and covariate parametric distributions) may be taken from external sources as argued by Caro and Ishak, but is usually borrowed from the index study (IPD trial) used in the ITC [11]. The original specification for the simulation entailed simulating a single trial with the same size as the original AgD one. More recent developments suggested potential improvements over this initial methodology regarding the sample size of the simulated data [42]. They recommended either simulating multiple samples while keeping the same sample size in order to mitigate the arbitrary impact of the chosen seed on the simulated dataset, or even simulating an “infinite” population (in practice very large, that is encompassing $> 100,000$ individuals). Individual outcome predictions are subsequently made using this newly simulated data, and averaged over the entire population to obtain a marginal treatment effect estimate. An adequate variance of the estimate in the simulated dataset can then be obtained through non-parametric bootstrapping.

2.3.3 Comparison of MAIC and STC

Among PAIC, MAICs are by far the most common over STCs in applied settings [43, 44, 21]. We will present the main distinctions between the two methods below.

Conditional versus marginal estimand

As mentioned before, MAIC and STC “simulation implementation” target marginal estimands, while STC “plug-in implementation” target conditional estimands. The choice between them

may depend on the estimand most relevant to stakeholders, as well as on which is most appropriate for constructing the indirect comparison (this later point is being discussed in the following subsection).

As a reminder, a conditional effect is the individual-level effect and quantifies the treatment effect for an average individual given a specific set of covariates. The marginal effect is the population-level effect and is the contrast of two average effects, obtained by integrating a conditional model over the multivariate distribution of a given population. Thus marginal and conditional effects differ in definition, and will coincide numerically only when the contrast is additive on the linear scale (and in the presence of treatment-effect heterogeneity, when the conditional effect is the one for an average individual, as explained below). Such measures are said to be collapsible and include risk difference, means difference and risk ratio. Frequent non-collapsible measures are odds ratio and hazard ratios. They are defined on a log scale, and averaging conditional effects across strata and then back-transforming does not yield the same result as computing the marginal effect directly. Since the hazard and odds ratio are among the most prevalent treatment effect contrast measures encountered in medical research, non-collapsibility and consequently the targeting of conditional or marginal effects bears non-trivial importance in practice. Moreover, in situations of heterogeneity (*i.e.* in the presence of TEM), multiple conditional effects exist as there is one conditional effect for each stratum defined by the combination of the TEM variables. A sensible choice when needed to report a single conditional effect in situation of heterogeneity is to take the one at the average level (arithmetic mean) of the covariates defining the treatment-effect heterogeneity [23].

In the context of PAIC, which of the conditional or marginal estimand should be targeted has been debated in a series of articles [23, 24, 25, 45, 46]. The arguments revolved around differentiating between the qualities of the estimator and the estimand. PAIC results are mostly used in the context of market authorization and reimbursement dossiers, where the treatment effect reported by ITC is added to the synthesis to provide an estimation of the relative benefit. This information is usually interpreted in light of the total expected cost of the treatment to inform reimbursement decision by HTA bodies. The idea that a population-oriented question should expect a population-based answer plead to favor marginal over conditional estimands, and thus weighted-based estimators over regression-based models when the contrast effect used is non-collapsible [47].

Compatibility of estimands forming the indirect comparison

Treatment effects reported in publications from clinical trials are traditionally marginal estimates, although conditional estimands are increasingly required [48, 49]. In practice, the “plug-in” method is by far the most frequent implementation. Consequently, in practice, the case of constructing an ITC by comparing a marginal estimate (AgD *b* study treatment effect, as marginal effects are often reported in clinical trials publications) to a conditional estimate (estimated using the IPD from study *a* with an “plug-in” implementation STC) is not rare, and leads to comparing quantities that are not “compatible”. The estimand is difficult to define, and it is unlikely to match the originally intended target. Moreover, even if a conditional outcome was actually reported in the AgD study *b*, it would likely be coming from a slightly different outcome model than the one fitted in the STC procedure (*i.e.* namely different variables and interactions included in the regression model). In such a case, the two estimates used in the ITC would still target different effects, leading to comparing quantities that are not directly comparable.

Relative performance of PAIC methods: MAIC vs STC

Broadly speaking, regression-based estimators are considered more efficient than weighting-based estimators [50, 51]. The actual comparison of the two is usually limited to collapsible outcomes for reasons explained before in this section. In the context of PAIC, this specific aspect has been studied by Phillippo et al. (2020) who found that STCs were providing better estimation performance as compared to MAICs, both in terms of accuracy and precision [52]. A debate ensued regarding the fact that this study targeted a marginal estimand with a non-collapsible effect measure to evaluate performance of both MAIC and STC; however, the difference in performance observed was deemed unlikely to be explained only by this aspect [24, 23].

Choice of the covariates to be included in the model

As explained in 2.2.2, anchored comparisons only need to include TEM to be unbiased, while unanchored comparisons need to include both PF as well as TEM, that is all confounders of the association between the exposure and the outcome. This simple rule is actually complicated depending on the estimator used, *i.e.* MAIC or STC. Consequently, whether an MAIC or an STC is being used, the precision of the estimation will vary in opposite directions when including PFs in an anchored comparison. Provided usual assumptions when fitting a regression model hold (no collinear regressors, number of observations sufficient in regards

to the number of covariates included), an STC model will typically see an increase in the precision of its estimation when including additionally relevant PF, even though it will not impact the accuracy of the estimate. Conversely, the MAIC weights will increasingly depart from unity as more covariates are included in the trial selection model, leading to a reduction in the sample size of the resulting weighted cohort. Consequently, the estimation will suffer from increased variability from the smaller sample size.

Hence, including PF that are not required to deconfound the trial-outcome association worsen precision of the estimation of anchored MAIC, and may be beneficial in anchored STC.

As stated before, anchored comparisons need only to adjust for TEMs to yield an unbiased treatment effect, while unanchored must include both pure PFs and TEMs, in order to resolve any imbalance in confounders distributions between studies. This distinction points out that confounders which distributions happen to be balanced may not need to be included in the model to yield an accurate treatment effect estimate. However, reweighting a cohort, as performed by MAICs, is intrinsically a resampling procedure. Consequently, variables that may have been balanced in the initial unweighted populations may become imbalanced after weighting. If these variables happen to also be associated with the outcome, they will confuse the association between treatments and outcome in the weighted cohort, leading to biased estimates. Therefore, it is argued that in the case of MAICs, all relevant confounders should be included, regardless of the balance before reweighting [16]. If one or several such confounders have been identified but purposefully not included in the model (probably because of sample size limitations), cross-study balance between these confounders should be checked in the reweighted cohort.

Finally, another important aspect to consider is that STC requires explicit specification of the outcome model, that is distinguishing between TEM and PF as well as the actual form of the treatment effect interaction. Conversely, MAIC does not require this level of precision in the trial assignment model.

2.3.4 Distinguishing between prognostics factors and treatment-effect modifiers: a blurred distinction

As stated previously, an anchored comparison will be unbiased only if the relative difference between evaluated treatments is constant across any strata of a studied population, that is if there are no TEM imbalance between populations susceptible to influence any of the treatment association with the outcome.

However, the distinction between TEM and PF is not straightforward. First, it depends on the scale upon which the treatment effect is defined [53, 15, 54]. A variable with a (non-null) constant effect on the outcome on a given scale, across all strata defined by the unique values of a PF, may have on another scale a non-constant effect across these same strata. In other words, a variable identified as a pure PF on a given scale may become a TEM on another scale.

An illustration of this phenomenon is shown in Table 2.1, inspired by Laupacis, Sackett, and Roberts (1988) [55]. This contingency table is representing the proportion of cardiovascular event depending on the treatment received, stratified by sex. In men and in women, the absolute risk reduction of experiencing a cardiovascular event when being given treatment A as compared to treatment B is 0.04. However, the risk ratio of A vs B is 0.6 in women, and 0.2 in men (the corresponding odds ratio is 0.57 and 0.19 respectively). The marginal risk ratio is 0.47 (and marginal odds ratio is 0.45). Hence, when interested in risk or odds ratio, the sex is a TEM of the relative efficacy of A versus B; but is not anymore when one is interested in the absolute risk reduction, which is constant across sex-strata.

Table 2.1: Cardiovascular events depending on treatment and sex

Sexe	Treatment	Event	No event	Odds ratio	Risk ratio	Risk difference
Female	A	1	99	<i>0.19</i>	<i>0.2</i>	<i>0.04</i>
	B	5	95			
Male	A	6	94	<i>0.57</i>	<i>0.6</i>	<i>0.04</i>
	B	10	90			
Total	A	7	193	<i>0.45</i>	<i>0.47</i>	<i>0.04</i>
	B	15	185			

Figure 2.6, inspired by Vickers and Steineck (2018) [53] gives a broader outlook on this problem. Keeping with the previous example, we are still interested in the relationships between two variables (sex and treatment) and a binary outcome. The effect of sex on the outcome is constant on a given scale (for any baseline event rate, that is for any combination of other PF). That scale is indicated by the annotation as the title of each subplot; Risk Ratio (RR), Odds Ratio (OR), Risk Difference (RD) for respectively panels A, B and C. Focusing on the panel A for instance (reasoning is identical for panels B and C) the treatment effect and the sex “effect” on the outcome measured through RR are constant for any baseline event rate, and the treatment effect is homogeneous for both sexes, that is the variable sex is not a TEM when the contrast measure is a RR. However, when quantifying the treatment effect using an alternative scale while keeping the exact same data generating mechanism, the treatment effect “becomes” heterogeneous regarding sex. Indeed, a variable that changes baseline risk inevitably changes at least one of the three usual contrast measures; whichever

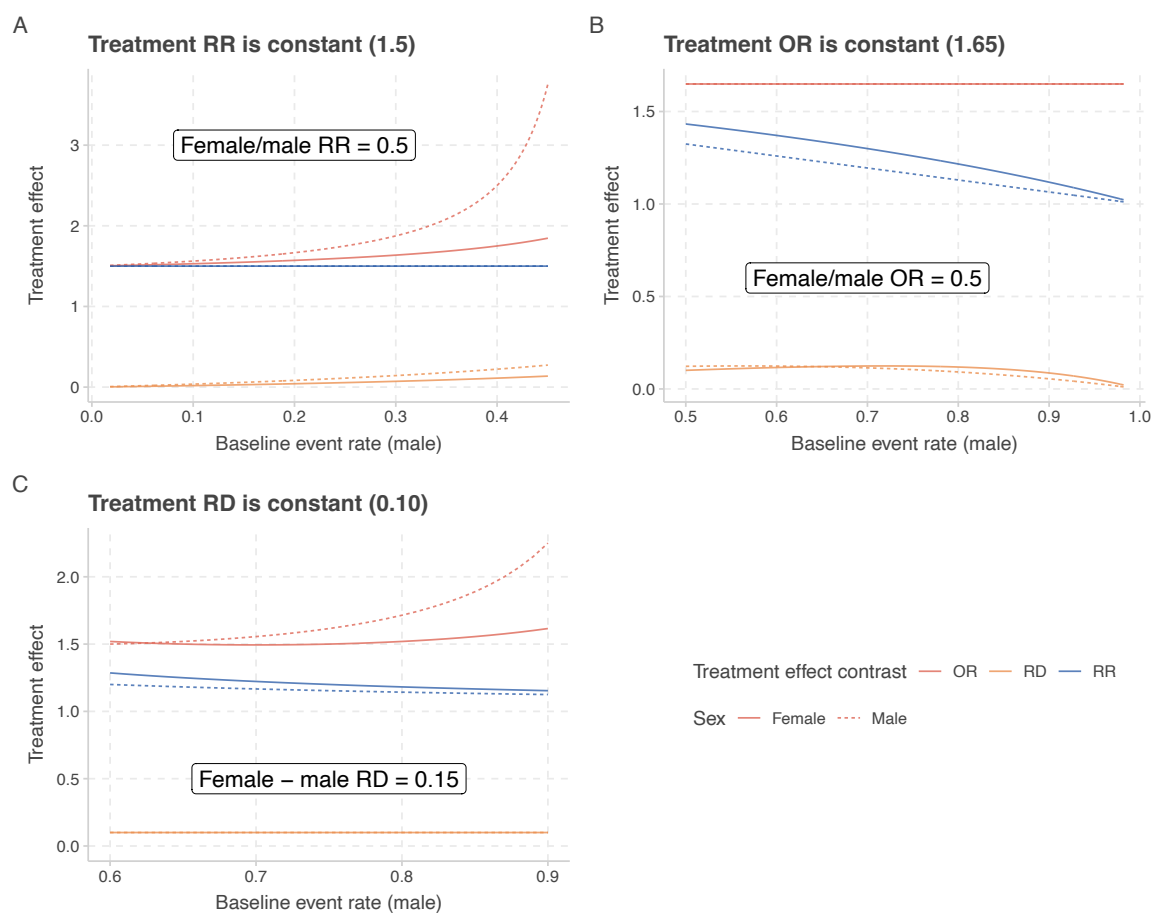


Figure 2.6: Interchanging status of sex as a PF, or PF and TEM, depending on contrast measure chosen.

measure stays constant defines the scale on which that variable is not an effect modifier.

Indeed, keeping with this example, we have:

$$\left\{ \begin{array}{l} T \in \{0, 1\} \text{ denoting treatment} \\ S \in \{0, 1\} \text{ denoting sex} \\ Y \in \{0, 1\} \text{ denoting a binary outcome} \end{array} \right.$$

Let's write the expected success under each treatment modality, conditional to sex with

$$p_A(S) = \Pr(Y = 1 \mid T = 0, S), p_B(S) = \Pr(Y = 1 \mid T = 1, S)$$

Thus, the treatment effect can be quantified as

$$\left\{ \begin{array}{l} RD = p_A(S) - p_B(S) \\ RR = \frac{p_A(S)}{p_B(S)} \\ OR = \frac{p_A(S)(1-p_B(S))}{(1-p_A(S))p_B(S)} \end{array} \right.$$

Consequently, if the RD is constant for all strata defined by S, we have: $\theta = p_A(S) - p_B(S)$, and $p_B(S) = p_A(S) - \theta$; we can rewrite the other contrast effect measure as:

$$\left\{ \begin{array}{l} RR = \frac{p_A(S)}{p_A(S) - \theta} = \frac{1}{1 - \frac{\theta}{p_A(S)}} \\ OR = \frac{p_A(S)(1-p_A(S) + \theta)}{(1-p_A(S))(p_A(S) - \theta)} = \frac{p_A(S)(1-p_A(S)) + \frac{\theta p_A(S)(1-p_A(S))}{1-p_A(S)}}{p_A(S)(1-p_A(S)) - \frac{\theta(1-p_A(S))p_A(S)}{p_A(S)}} = \frac{1 + \frac{\theta}{1-p_A(S)}}{1 - \frac{\theta}{p_A(S)}} \end{array} \right.$$

which shows that these quantities are only dependent on the values of $p_A(S)$ when θ is constant, and that they will consequently vary when the baseline risk of event in the treatment A arm (defined as $p_A(S)$) changes.

Identification of covariates to be included in a PAIC model

Previous sections discussed which variables should be included in a PAIC model to obtain accurate and precise estimates; here we outline practical approaches to identify them.

The proper identification and handling of confounders may represent the foundational purpose of epidemiology as a scientific discipline. It led to the rise of causal inference as a

standalone research field, highlighting both the centrality as well as the complexity of the task.

In practice, different ways can be taken to build this knowledge. We argue that ideally the adjustment model should be constructed based on prior, domain-specific knowledge, as opposed to relying primarily on empirical statistical exploration of a given dataset, which will in most cases be the IPD trial dataset. While expedient, proceeding this way relies entirely on recorded variables available in that data, potentially overlooking unrecorded key variables, and exposes itself to type II errors, *i.e.* associations deemed “non-significant” because of sampling variation. Hernán et al. provided a worked example of how conclusions regarding a given research question (association between folic acid supplementation and risk of neural tube defects) answer can depart depending on the method used to identify confounders, namely empirical statistics as opposed to domain-knowledge based. [56]

Methods for identifying confounders and defining their relationships has been vastly explored in the causal inference field. While developing a Directed Acyclic Graph (DAG), the workhorse of causal inference, is outside of the scope of this work, the vast body of work of this field regarding how to identify confounders is of high interest in the context of PAIC [57, 58]. Given their non-randomized nature, adjusted indirect comparisons should rely on a robust methodology to identify and take into account confounders in order to provide trustworthy results. We argue that systematic review should be carried out to identify PF and TEM [21]. Ferguson et al. offer guidance on how to translate systematic reviews on a specific research question into DAGs [59].

When literature is conflicting or sparse, which may often be the case when studying newly approved drugs (a main use case of PAIC), domain-experts should be called up to provide additional guidance. Rodrigues et al. provide an example on how to reach a consensus on a causal model in such a situation [60]. An example combining literature-based background knowledge with expert opinion is shown in Hasani et al. (2024) [61]. Finally, selecting an appropriate set of covariates to adjust for in situation where uncertainty remains regarding the true causal graph was already discussed by Sauer et al. (2013) [62].

In the context of PAIC, a special emphasis is put on differentiating between pure PFs vs. TEMs, as the existence of TEMs (as opposed to pure prognostic factors) is the assumption justifying conducting an anchored PAIC in the first place, rather than a more straightforward Bucher ITC. However, information on factors predicting the efficacy of a treatment may be even more scarce than on PFs, especially if that treatment has yet to be approved. Since heterogeneity of effect may depend on underlying biological factors, it may be possible to base this assessment from information from other drugs pertaining to the same class [63].

A further complication is the scale dependence of treatment-effect heterogeneity. Heterogeneity is not purely a pharmacological property but depends on the contrast used to quantify the effect as shown in section 2.3.4. While it has been described for binary outcomes, the same applies for continuous outcomes: if the baseline outcome differs across strata, a treatment with a constant effect across these strata as measured by an absolute contrast measure (e.g. absolute difference) will necessarily show a heterogeneous effect across these same strata when switching to a relative contrast effect (e.g. relative difference). The magnitude of the treatment effect heterogeneity will be proportional to the difference in baseline risk across strata, and by the treatment efficacy, besides any underlying explanation regarding biological efficacy tied to certain characteristics. Hence the distinction regarding the status of covariates as purely prognosis or TEM may be more quantitative than qualitative: is the magnitude of treatment-effect heterogeneity across a PF large enough to be relevant at all, and is that scale of any interest? More rarely, a variable that is not prognostic may be a TEM. In other words, the treatment will show some efficacy only on a given strata of the population [53]. This kind of specific biological pattern may be especially seen in oncology, with therapies selectively efficient in tumors expressing specific characteristics, e.g. pemetrexed showing an efficacy on small cell lung cancer specifically for the non-squamous histological type [64], while the histological type has not been shown to influence prognostic overall [65], or the efficacy of pembrolizumab being dependent on the mutational signature of a tumor [66]. While we recommend that the adjustment model be primarily identified through medical expertise and literature review, it is fair to acknowledge that this requirement may be unrealistic in some cases, particularly for innovative treatments for which such a broader contextual knowledge is scarce.

Hence, in practice, care should be taken when conducting adjusted indirect comparison, as the rationale for conducting an adjusted comparison and the validity of the modelization carried-out depends on the adequate identification of the relationship between covariates.

2.3.5 Anchored versus unanchored PAIC assumptions

As mentioned before, unanchored comparisons need to adjust for all confounders to yield a reliable estimate. Thus, an indirect comparison relying on an anchor treatment arm makes the assumption of constancy of the relative effects across compared populations, while one without anchor make the assumption of constancy of the absolute treatment effects across populations. This crucial difference in assumptions led some authors to question the validity and usefulness of any unanchored comparisons altogether. For instance, Phillippo et al. (2016) mention that unanchored ITC assumption “*is largely deemed unreasonable (if it were, there would be no reason to undertake randomised controlled trials).*” [15]. Remiro

Azócar, whom PhD thesis is focused on anchored comparisons, justify the choice of not including unanchored comparisons: “*it does not respect within-study randomization and is not protected from imbalances in any covariates that are prognostic of outcome (in essence implying that absolute outcomes can be predicted from the covariates, a heroic assumption)*” [16]. While we acknowledge these objections, studying unanchored ITC remains unavoidable when undertaking a comprehensive work on PAIC as reviews found that they represented approximately 65% of published PAICs [44, 21]. Underlying reasons for this observation as well as position of HTA bodies on unanchored comparisons are discussed in Chapter 3. Moreover, relative performance of anchored and unanchored methods with respect to the exactitude of the adjustment model is explored in Chapter 4. Table 2.2 summarizes the different methods available and their assumptions in terms of confounders identification in the context of pairwise indirect comparisons indirect comparisons. Unadjusted unanchored ITC are not represented because of their unrealistic assumptions that prevent them from being actually used in practice.

Method	Estimator	Estimand targeted	Adjustment requirement	Hypothesis
Anchored	Bucher ITC	Marginal	—	Constancy of relative effects
	MAIC	Marginal	TEMs	Conditional constancy of relative effects
	STC	Conditional (plug-in) or marginal (simulation)	TEMs ($\pm$ PFs for precision)	
	IPD NMA	Conditional or marginal		
Unanchored	MAIC	Marginal	All confounders	Conditional constancy of absolute effects
	STC	Conditional (plug-in) or marginal (simulation)		
	IPD NMA	Conditional or marginal		

Table 2.2: Summary of pairwise Indirect Treatment Comparisons methods and of their assumptions

Chapter 3

State of Practice

This chapter addresses practical aspects of PAIC. It will first provide an up-to-date review and breakdown of different guidelines and positions of major HTA and regulatory bodies, before presenting the results of the first original article featured in this work, focusing on PAIC implementations in the literature [21]. Together, these perspectives shed light on how PAIC methods are expected to be reported and justified in the context of market access and regulatory approvals, and how they are being implemented in practice.

3.1 PAIC in the regulatory HTA environment

This part will compile evidence from peer-reviewed publications and current HTA bodies' guidelines in conducting and reporting PAIC studies in the context of market access and regulatory approvals.

In a recent review van Beekhuizen et al. (2025) report on the European Network for Health Technology Assessment (EUnetHTA) relative effectiveness assessments, submitted from 2010 up to 2021 [43]. The European Union Regulation 2021/2282 on HTA aims to harmonize the HTA process across the European Union through Joint Clinical Assessment. This program will be rolled out in phases and is set to become the unique pathway across European Union for evaluating new medicinal products (starting January 2025 with oncology and advanced therapy medicinal products). It was first evaluated as a pilot programme, the EUnetHTA Joint Action project, which van Beekhuizen et al. review is focused on. They counted twelve relative effect assessments that featured at least one ITC. These twelve dossiers represented twenty-five ITC in total, with 8 being Bucher ITCs, 7 used PAICs, and 7 NMAs (a single NMA accounting for more than a single ITC). They categorized opinions of the Joint Action project on the ITC as either “appropriate”, “unsuitable”, or “unclear”. Out

of the 25 ITCs, 24 (96%) were labeled as “unclear”, and only one as “appropriate”. Still, they note that in the 12 relative efficacy assessment dossiers from which these 25 ITCs come from, 10 (83%) included the ITC result in the final assessment report. The most frequent critique found was related to “data uncertainty”, which often related to differences in the compared populations in potential confounders that could not be adjusted on, but also small sample sizes. Finally, it is worth noting that none of these ITCs has been labeled as “unsuitable”. Hence, in all but one case, the commission was unable to determine whether the ITC provided convincing evidence, or conversely, whether its methodology was sufficiently flawed to be disregarded in the proceedings. This suggests that the commission stakeholders might not have considered themselves in a position to take a decision based on ITCs. This could either reflect limited understanding of the assumptions and limits of these methods, or insufficient transparency and consistency in the reporting of the ITC by the industry, with either of which would prevent them from taking a hard stance on the relevance of indirect evidence. Additionally, the EUnetHTA being a joint effort from different national HTA bodies, this “indecisiveness” may also reflect different national HTA cultures.

Macabeo et al. (2024) conducted a comparative evaluation of the appreciation of ITC in HTA dossiers across five European agencies (England, France, Germany, Italy, and Spain) [67]. ITC methods included PAIC (STC, MAIC, ML-NMR), Bucher ITC, NMA as well as “undetermined”. Among the 120 evaluation proceedings mentioning an ITC in their report, the acceptance rate of ITC evidence was heterogeneous, with the highest for the NICE (English HTA body) at 47% (18/47), and the lowest for the AEMPS (Spain HTA body) and the HAS (France HTA body) with respectively 5% (1/21) and 0% (0/11) acceptance rates. Heterogeneity and potential risk of bias was the most often cited complaint, and it appeared roughly as frequent whether the ITC was adjusted or unadjusted (around two thirds of cases for PAIC, and 57% for Bucher ITC). Moreover, this difference in opinion around the general relevance of ITC is reflected in each HTA body guidelines: as quoted by Macabeo et al., NICE recommendations state that “data from trials that compare the technology with non-relevant comparators may be needed to enable the technology and the comparators to be linked in an indirect or mixed treatment comparison”. On the other hand, the Institute for Quality and Efficiency in Health Care (IQWiG), the German HTA authority, mentions that “there are still several unsolved methodological problems, so that currently the routine application of indirect comparison methods within the framework of benefit assessments is not advisable” [68]. How these varying assessments of indirect comparisons will shape the upcoming European-level joint assessments remains an open question.

In line with these observations, Katsoulis et al. (2024) conducted a study on HTA stakeholders’ perceptions of the relevance and acceptability of ITC in HTA dossiers, specifically

in the oncology field [69]. They surveyed 30 current and former members of national HTA bodies in Australia, Germany, France, the UK and the USA. Consistent with the findings from Macabeo et al., it emerges from this work a heterogeneity in practices across countries, with Australian and British respondents reporting an overall favorable perspective of their respective agencies on these methods, as compared to the US, Germany and France respondents who overall reported lower acceptance rates. German respondents were adopting the most negative views on these methods overall. Beyond countries heterogeneity, differences in acceptance emerged regarding context and actual methodologies used. Opinion were divided regarding the acceptance of unanchored methods: most respondents stated that an unanchored ITC could be accepted provided a strong justification and methodology, unless for German respondents who stated that an unanchored comparison is usually simply not accepted as evidence for relative efficacy assessment. Regarding the narrower set of PAIC, survey respondents reported lower likelihood of acceptance of STC as compared to MAIC. Given the still rather low acceptance of PAIC by HTA bodies and the novelty of these methods, this observation might either reflect genuine concerns with the specificity of the method (*e.g.* conditional estimand, extrapolation beyond observed covariate space), or lack of familiarity as compared to MAICs. Finally, when asked what were the criteria most important for an ITC to gain credibility in the eyes of the payers, several aspects stood out, with data quality of the source studies, justification for the conducting of an indirect comparison (as compared to providing direct evidence), and justification regarding the chosen ITC methodology being the most important aspects. While this survey does not reflect the official position of the different HTAs bodies, it allows interesting insights from individuals actually involved in these decisions.

Several regulatory agencies and HTA bodies issued guidelines on PAIC methodologies. Reflecting early adoption of these methodologies, NICE published as early as 2016 such a guidance in the form of the Technical Support Document 18: “Methods for population-adjusted indirect comparisons in submissions to NICE” [15]. By providing a comprehensive review of the methods, from theoretical aspects to software implementations, this document has been considered a reference point for conducting and reviewing PAICs beyond the UK, with more than 300 citations as of mid-2025. In 2024, the European Union, through the Member State Coordination Group on Health Technology Assessment (HTACG) issued two documents, *Methodological Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons* and *Practical Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons* which provided dedicated chapters to PAICs [28, 70]. This documents’ release is timely in the context of the incoming generalization of the Joint Clinical Assessments. Indeed, ITCs may become increasingly important with the new regulatory pipeline emphasis on multiple PICO per dossier (PICO refers to Population, Intervention, Comparison, Outcome,

i.e. a standard framework used for structuring clinical research questions). The proliferation of PICO's reflects the diversity of national perspectives across the EU, each placing different emphases on PICO's components. Beyond general reporting recommendations of PAIC instances in submission to Joint Clinical Assessments, it emphasizes wariness regarding unanchored comparisons: "...we are not endorsing them (*i.e.* unanchored ITC), and once again reiterate that estimates arising from using population based adjustment methods when performing unanchored ITCs are unreliable.". Interestingly it also discusses the intrinsic limitations of standard PAIC regarding the targeted population, and mentions considering the shared effect-modifier assumption, introduced by Phillippo et al. in NICE TSD 18, that allows transporting an effect into a different population, provided the underlying biological mechanisms of indirectly compared treatments are similar enough to assume a quantitatively similar interaction with TEMs. A typical concern of non-randomized studies is the amount of remaining bias in the analysis. The report put forth the "shifted null hypotheses" approach to address this limitation, stating that *"that a conclusion for an existing treatment effect can only be drawn if the confidence interval lies above or below a certain threshold shifted away from the zero effect"*. A closely related concept is the E-value, which quantifies the minimum strength of association with the outcome that an unmeasured or a group of unmeasured confounders should have in order to explain away a given statistical association [71].

Aballéa et al. (2025) discusses the merits and purportedly shortcomings of these EU-netHTA guidelines in a preprint [72]. This group of academics and industry experts pointed out several criticisms of these recommendations, namely an overly restrictive approach to unanchored comparison and an unjustified higher conservatism on PAIC as compared to NMA and Bucher ITC. While the matter of the relative acceptance of unanchored comparisons by different regulatory bodies across the world is a complicated and context-specific debate, we definitely agree with the observation regarding the relatively more cautious approach on anchored PAIC as compared to Bucher ITC and NMA. The argument is that Bucher ITC and NMA make the assumption of absence of heterogeneity, that is similar populations in terms of TEMs, with thus no impact on the pooled estimate; on the other hand the *raison d'être* of anchored PAIC is the assumption of imbalance in TEMs distribution between compared populations. The HTACG rightfully mentions that one cannot be certain that an adjusted method will have correctly adjusted for all interaction factors: *"In practice, however, one can never be sure that all the relevant effect modifiers are included. Therefore, population-adjusted methods have to be applied with the utmost care."* Moreover, the HTACG rightfully points out that the anchored PAIC "conditional constancy of relative effects" may not be upheld. Consequently, it recommends that extra analyses be conducted in order to assess robustness of the estimate, with for instance shifted null hypothesis testing. In the conclusion, the report explicitly acknowledges the validity of NMA as an indirect com-

parison method and reiterates warnings for anchored PAICs: “*relies on accounting for all relevant effect modifiers in the model, which is difficult to ensure, and therefore, in general not an adequate analysis.*”. This seems to overlook that Bucher ITC or NMA assume (absolute) constancy of relative effects, a necessarily stronger assumption than anchored PAIC’s one. Paradoxically, an explicit, albeit potentially imperfect attempt to address acknowledged heterogeneity seems to attract more scrutiny than the implicit assumption that no such heterogeneity exists. This extra carefulness with PAIC results may be rooted in the still relative degree of novelty of these methods or the fact that they partially rely on AgD and thus that the adjustment provided is somehow inadequate. This latter aspect will be studied in more details in the Chapter 4, where performance of PAIC methods is compared with their full-IPD counterparts. A more important concern in our opinion with using these methods, is the large “degrees-of-freedom” it allows for the analyses, which may be conducive to selective results reporting. We agree with this concern, and results of the review we conducted as the first part of this doctoral work lends credibility to this observation [21]. However, again it seems that these concerns are mostly brought up for PAIC, as compared to IPD NMA for instance. While having full access to individual data addresses many shortcomings of PAICs, especially regarding the target population, the problem of the degrees of freedom of the analysis is identical whether using full IPD or relying partly on AgD. However, addressing selective reporting of adjusted ITC is paramount. One approach would be to systematically report the unadjusted estimate alongside the adjusted one. Adopting a conservative standpoint, this adjusted analysis could be considered a confirmatory result, strengthening an anchored ITC. Any discrepancy between the two estimates should be analyzed in light of the sample sizes shrinkage before and after adjustment (for MAIC and for methods based on weights derived from a treatment assignment model), realism of the assumption of extrapolation beyond the observed data space (for STC and methods based on outcome regression model), justification for the confounders included in the adjustment models, and heterogeneity of the two target populations of the unadjusted and adjusted models.

It is worth noting that no such guidelines on indirect comparisons have been issued by North American regulatory bodies. In its report “Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products”, the FDA does not mention any specific methodologies to carry anchored or adjusted indirect comparisons [73], and nor in “Meta Analyses of Randomized Controlled Clinical Trials to Evaluate the Safety of Human Drugs or Biological Products” [74]. Tanaka et al. (2024) conducted a targeted review of ITC guidelines issued by various HTA bodies worldwide until August 2023 and confirmed this observation [75]. The Canadian Agency for Drugs and Technologies in Health did not issue any specific guidelines on Indirect Treatment Comparisons. Among the multiple healthcare

payers in the American health insurance landscape, no detailed guidelines on the topic can be found. The Institute for Clinical and Economic Review, one of the most important non-governmental and non-profit body, mentions ITC in its Value Assessment Framework, but without actual recommendations regarding their implementation and evaluation [76]. Moreover, data is relatively scarce regarding prevalence of use and acceptance of PAIC. While it is clear that the number of HTA filings relying on single-arm trials has seen a massive uptick in recent decades, especially in the United States, no review has quantified the importance taken by indirect comparisons in this context [77]. A review from 2020 focusing on “Institute for Clinical and Economic Review reports” concluded that for the 28 reports eligible for indirect comparisons issued during the period of interest, 14 of them deemed indirect comparisons to be infeasible [78].

3.2 PAIC in peer-reviewed publications

One of the objectives of this PhD thesis was to assess the current use of PAIC worldwide. A systematic review of real-world implementation of PAIC appeared necessary, as the only existing assessment at the time was conducted by Phillippo et al. (2016), in the context of NICE evaluations [15]. At that time, more than 10 years after their formal introduction in the literature, no review of publications featuring PAIC existed. Moreover it appeared clear that between 2016 and 2022 (the year which we initiated the systematic review), their use appeared to have increased markedly, as suggested by an informal quantitative analysis featured in Remiro-Azócar’s PhD thesis [16]. Additionally, no dedicated reporting guidelines were available when we initiated this work (which is still the case today), raising concerns about the coherence and quality of reporting in published studies.

Consequently, we undertook a methodological review of PAIC published in peer-reviewed journals, with two main objectives: first, to provide a clear picture of the current quantitative trends and adoption of these methodologies by researchers, and second, to identify objective information regarding the quality of their implementation.

This review materialized in a peer-reviewed article published in the *Journal of Clinical Epidemiology*, on September 2023. It is reproduced below in its entirety, followed by the supplementary online materials [21].

A subsequent section compare and contrasts these results to the ones from three other reviews conducted concurrently and published in 2023 and 2025, by Truong et al. (2023), Farinasso et al. (2025), and Cassidy et al. (2023) [44, 79, 80].

Published article I (contributed work)

The first peer-reviewed article of this thesis is Serret-Larmande et al. (2023). [21] It is reproduced here in its published form under the Creative Commons CC-BY license, together with the online Supplementary Materials.

ORIGINAL ARTICLE

A methodological review of population-adjusted indirect comparisons reveals inconsistent reporting and suggests publication bias

Arnaud Serret-Larmande^{a,b,*}, Belkacem Zenati^a, Agnès Dechartres^a, Jérôme Lambert^b,
David Hajage^a

^aINSERM, Institut Pierre Louis d'Epidémiologie et de Santé Publique, AP-HP, Hôpital Pitié-Salpêtrière, Département de Santé Publique, Centre de Pharmacopépidémiologie, Sorbonne Université, Paris, France

^bECSTRA Team UMR-1153 INSERM, AP-HP Saint Louis Hospital, Université Paris Cité, Paris, France

Accepted 11 September 2023; Published online 16 September 2023

Abstract

Objectives: Population-adjusted indirect comparisons (PAICs) were developed in the 2010s to allow for comparisons between two treatments evaluated in different trials while accounting for differences in patient characteristics if individual patient data (IPD) are available for only one trial. Such comparisons are increasingly used in market access applications when a pharmaceutical company compares its new treatment (with IPD available) to another treatment developed by a competitor (with only aggregated data available). This study aimed to describe the characteristics of these PAICs, assess their methodology, and describe the reported results.

Study Design and Setting: Original articles reporting the use of at least one PAIC were searched on PubMed between January 1, 2010 and April 2, 2022. Two reviewers independently selected articles and extracted data.

Results: We included 133 publications reporting the results of 288 PAICs. Half of the articles were published on or after May 7, 2020, and 71 (53%) pertained to onco-hematology. The pharmaceutical industry was involved in 130 (98%) articles. Key methodological aspects were reported inconsistently, with only three articles adequately reporting all aspects. A total of 161 (56%) articles reported a statistically significant benefit for the treatment evaluated on IPD. Conversely, only one PAIC significantly favored the treatment evaluated on aggregated data.

Conclusion: Although the number of published PAICs is increasing, the methodology and transparency need to be improved. Moreover, our study strongly suggests a reporting bias. This situation calls for strengthening guidelines to improve trust in PAIC results and thus their reliability in market access applications. © 2023 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Keywords: Population-adjusted indirect comparisons; Matching-adjusted indirect comparisons; Simulated treatment comparisons; Methodological review; Health technology assessment; Clinical research

Data availability: Articles that support our findings are publicly available. Template data collection forms, detailed data extraction on all included studies, and analytical code are available on a GitHub repository: https://github.com/Nworno/paite_literature_review.

Prospero registration number: CRD42022326154 (https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=326154).

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Registration and protocol: The protocol of the review was registered in PROSPERO (CRD42022326154).

Ethical approval: Not required for this work.

* Corresponding author. Département de Santé Publique - Hôpital de la Pitié-Salpêtrière, 47-83 boulevard de l'Hôpital, 75013 Paris, France. Tel.: +33-6-50-57-49-30; fax: +33-1-42-16-00-00.

E-mail address: Amaud.serret-larmande@u-paris.fr (A. Serret-Larmande).

1. Introduction

Health technology assessment (HTA) bodies grant approvals based on the premise that a given new treatment provides a benefit as compared with the current standard of care (based on efficacy, safety, and economic considerations). Ideally, the decision is informed by direct evidence obtained from randomized controlled trials. However, various practical reasons prevent this ideal scenario from being the norm in the current HTA context, from ethics concerns to the desire for innovative therapies to be quickly accessible [1]. In such cases, indirect treatment comparisons (ITCs) are increasingly being used to estimate the relative benefits of treatments evaluated across different studies [2–5]. However, in practice, individual patient data (IPD) are rarely available for all studies, especially in the context

What is new?

Key findings

- The number of published population-adjusted indirect comparisons (PAICs) has significantly increased in recent years. Population-adjusted indirect comparisons (PAICs) focus mostly on oncologic and hematologic conditions and are almost systematically conducted in collaboration with the pharmaceutical industry.
- The methodology and reporting of these studies were heterogeneous and overall insufficient regarding key criteria.
- Study results for the primary outcome suggest a major publication and reporting bias.

What this adds to what was known?

- Population-adjusted indirect comparisons (PAICs) can be effective tools to estimate relative treatment effects but rely on assumptions that may be difficult to comply with. This review unveiled insufficiency in PAIC implementation methodology and reporting, which undermines confidence in the results.

What is the implication and what should change now?

- Improving the methodology and reporting standards, and preregistering study protocols seem paramount to increasing trust in PAIC results and improving their utility in market access applications.

of HTA. Generally, sponsors have access to IPD for the treatment that they are developing, but they can only access published aggregated data for treatments developed by competitors. Methods allowing for conducting ITCs combining IPD and aggregated data while adjusting for imbalance between compared populations are called population-adjusted indirect comparisons (PAICs). Matching-adjusted indirect comparisons (MAICs) and simulated treatment comparisons (STCs) are the most established PAICs [6–9].

Since their introduction in the 2010s, the methodological aspects of these methods have been discussed in depth in a significant body of work [10–13]. One review focused on their use in the context of the HTA submissions to NICE (UK HTA body) until 2018 [14]. However, no systematic literature review has evaluated PAICs published in peer-reviewed journals, and no reporting guidelines have been published in the literature (beside guidelines issued by HTA bodies) [15,16].

This need seems even more pressing because market access bodies are increasingly relying on these approaches to inform their decisions [11,17], sometimes solely based on single-arm trials [18–20].

We conducted a methodological literature review of PAICs published in peer-reviewed journals. The objectives were to describe the characteristics of these publications, evaluate their methodology, and describe the reported results. We conclude this work with several recommendations aiming at improving reporting consistency.

1.1. A brief introduction to PAIC methods

This section briefly introduces the key theoretical concepts of PAIC methods. Comprehensive methodological descriptions are available elsewhere [10,11,21,22]. Thereafter, the term source studies will refer to the original studies (i.e., usually clinical trials, even if data from noninterventional studies could be used) involved in the indirect comparison. An ITC is said to be anchored if it relies on a common treatment arm and unanchored in the absence of common a comparator.

The aim of an ITC is to estimate the relative effect of two treatments a and b on an outcome Y , using evidence from two different study samples coming from two distinct populations A and B . The marginal relative effect of a vs. b , as estimated by an unadjusted indirect comparison, is simply the difference between the observed average outcomes $\bar{Y}_a(A)$ and $\bar{Y}_b(B)$. This unanchored estimate would be unbiased only in the absence of an imbalance between populations in terms of both prognostic factors and treatment-effect modifiers, that is, if the populations A and B were identically distributed for any relevant baseline characteristic.

If a common comparator arm exists (e.g., if both a and b were evaluated by a randomized controlled trial against the same standard of care or placebo), the relative treatment effect of a vs. b can be estimated by the difference between the relative effects a vs. c and b vs. c . Due to the within-studies randomization, this *anchored* estimate cancels out the consequences of prognostic factors imbalance (but not treatment-effect modifiers imbalance) across populations. This approach is sometimes called Bucher ITC [23].

When there is an imbalance of prognostic factors or treatment-effect modifier between populations, unanchored or anchored *adjusted* ITCs can be used. To obtain unbiased treatment effect estimates, anchored ITCs should adjust for all imbalanced treatment-effect modifiers and unanchored ITCs for both imbalanced treatment-effect modifiers and prognostic factors. Ideally, adjustment for imbalance in relevant baseline characteristics would require IPD to be available for both studies [24,25]. Conversely, PAICs can be used when IPD are available for one study (e.g., A study) but only aggregated data for the other (e.g., B study). For the two most commonly used methods (MAIC and STC), the principle is the same: to estimate $\hat{Y}_a(B)$ (i.e., the

outcome that would have been observed under treatment A in the B study) and then compare it with the observed $\overline{Y}_b(B)$.

Matching-adjusted indirect comparison (MAIC) carries out the adjustment by reweighting the A sample with the inverse of the odds of being enrolled in the B study. The weights are estimated from the baseline covariates by using a logistic model fitted with the method of moments [6]. The estimated effect is the marginal treatment effect in the aggregated data (population). The post-weighting sample size, called the effective sample size, is frequently reported. The method of moments guarantees mathematically that the mean (and potentially standard deviation if included) of the characteristics included in the weighting model will match exactly the ones of the aggregated data study after weighting. The resulting effective sample size represents the “cost” of this exact matching: a small effective sample size reflects a lack of population overlap [6].

In contrast, STC estimates the relative treatment effect by fitting a regression model in the A sample data with adjusted covariates centered on their average value in the B aggregated data. This method estimates a conditional effect in a population with the average covariates distribution of the B population [10].

As previously stated, unanchored PAIC requires that all imbalanced prognostic factors and treatment-effect modifiers be included in the analysis. In this setting, the assumption of absence of unobserved confounders is widely considered difficult or even impossible to achieve [10,14,21]. In the anchored setting, adjusting for treatment-effect modifier only is sufficient to obtain an unbiased treatment effect estimate. With anchored MAIC, including prognostic factors in addition to treatment-effect modifiers in the weight estimation model reduces the precision without further bias reduction [10]. Conversely, with anchored STC, adding prognostic factors in the adjustment model increases the precision and is thus encouraged [10].

2. Methods

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analysis 2020 for reporting this methodological review [26].

2.1. Search and selection process

On April 2, 2022, we searched MEDLINE via PubMed for PAICs published between January 1, 2010 (the year when PAIC methodologies were first introduced) and April 2, 2022. The search algorithm combined terms aiming at retrieving STCs, MAICs or other types of PAICs. We also retrieved articles citing the four original articles establishing the STC and MAIC methodologies [6–9]. The search equation is reported in [supplementary materials](#).

We included all studies reporting the results of a PAIC in the main text as a primary or secondary analysis. We defined PAIC as a methodology that 1) performs a pairwise ITC, 2) uses a statistical method aimed at reducing the confounding bias in estimating the treatment effect, and 3) relies on both IPD and aggregated data to do so [11,12]. Only original research articles were included. We excluded nonhuman or preclinical research and methodological articles (i.e., cases in which the PAIC is conducted to illustrate a methodological work).

Using Rayyan software [27], two reviewers (ASL, BZ) independently selected references by reading the title and abstract and the full text if necessary. Discrepancies between the two reviewers were adjudicated by a third reviewer (DH).

2.2. Data collection

For each included article, two reviewers (ASL, BZ) independently collected the information detailed below. Disagreements were resolved by discussion, referring to the third reviewer (DH) when necessary. Data were collected to inform three different dimensions: the main characteristics of the article, the source studies, and the methodology and results of the PAIC (for the primary outcome if defined or the first outcome reported otherwise). The detailed description of the collected data can be found in [Supplementary material](#) (see “Data collection details”).

Based on the methodological literature [10,14,21], we defined several key elements that should be reported to allow readers to adequately judge the validity of the methods used and the relevance of the results.

1. Completion of a systematic review to identify direct comparisons. The aim is twofold: asserting that no direct comparison exists and that the whole body of available evidence to inform the indirect comparison has been considered.
2. Defining a primary outcome. Despite their non-randomized nature, PAICs often serve as substitutes for missing randomized control trials. The requirement for a primary outcome is rooted in the framework of randomized control trials, where it aims to prevent type I error inflation and the possibility of selective reporting and cherry-picking [28].
3. Sample size reporting. Information on the sample size at the different steps of a PAIC informs the magnitude of the selection required in the IPD treatment arm. In a first step, it may be required to align source studies inclusion criteria, by taking a subset of the IPD population that matches the inclusion criteria of the aggregated data population. The unadjusted sample size has been defined as the sample size after subsetting the IPD population. The adjusted sample size is the sample size used in the adjustment model for STC and the effective sample size for MAIC [10].

4. Rationale for covariates included in the model. Adequate choice of covariates to be included in the adjustment/weighting model requires considering two aspects of the problem. One has to do with the chosen PAIC method, that is, appropriately accounting for treatment-effect modifiers only, prognostic factors only, or both prognostic factors and treatment-effect modifiers (see “A brief introduction to PAIC methods”). The other one is the authors’ justification for the status of each variable (prognostic factors or treatment-effect modifiers), which depends on each application [29]. Covariates included in the adjustment/weighting model should be based on clinical relevance using a priori medical knowledge and not on data-driven statistical considerations [10,11]. Medical knowledge can refer to medical experts’ opinions or a literature review [30].

2.3. Data analysis

The analysis was mostly descriptive. Median and Q1-Q3 were reported for continuous variables, and frequencies and percentages for categorical variables. The analysis was performed with R 4.2.2 [31].

The collected data as well as the source code are freely accessible on a GitHub repository: https://github.com/Nworno/paict_literature_review.

3. Results

3.1. Selection process and article characteristics

Our search retrieved 799 unique articles. We excluded 586 records based on the titles and abstracts; 80 were subsequently excluded based on the full text, which left 133

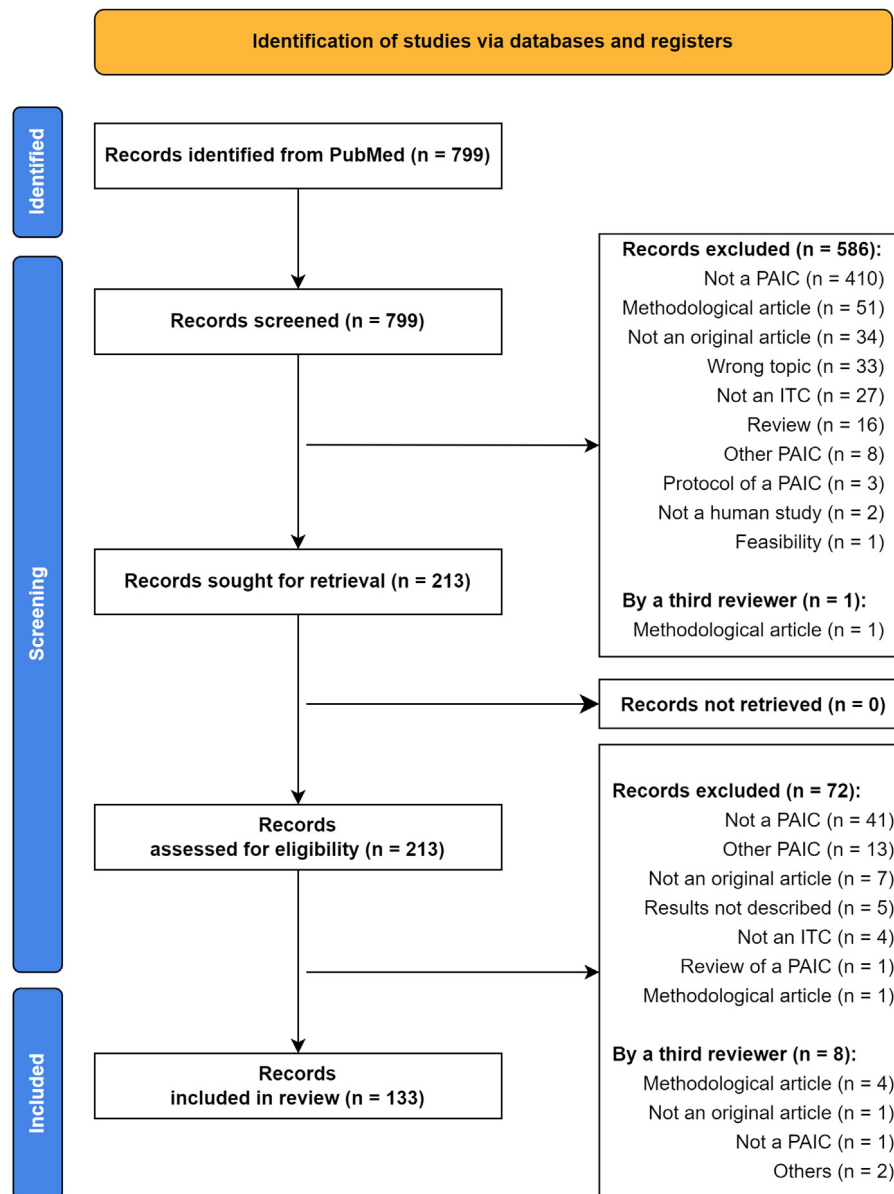


Fig. 1. Flow diagram of selection of articles. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

articles meeting inclusion criteria. Details of this selection process are in Fig. 1 (see complete list of included articles in Supplementary file included_articles.xlsx).

Table 1 summarizes the most important attributes of articles. The first article was published on April 29, 2011 [32]. Half of the articles were published on or after May 7, 2020 (Fig. 2). The pharmaceutical industry was directly involved in 130 (98%) articles, either through authorship for 125 (94%) and/or by funding the research for 129 (97%). Many articles involved collaboration between academia and the pharmaceutical industry (68%, 91/133). Methodological experts were involved in 121 (91%) articles. Most (53.4%) studies focused on malignant diseases, and approximately one-third involved at least one monoclonal antibody (see Supplementary Table 1). A comparison of characteristics presented in Table 1 based on the presence of at least one academic coauthor was performed and is available as Supplementary Table 2. We did not perform the preplanned comparison according to the support of pharmaceutical industry, since we only found four academic-only articles.

Overall, we identified 470 unique source studies (Supplementary Table 3). Most were clinical trials (93.4%, 439/470), of which 115 (26.2%) were single-arm. Eighty-nine (66.9%) studies reported the inclusion criteria of the source studies, either with a table (32.3%), or through a detailed comparison of the inclusion criteria in the text (34.6%). Sixty-six studies (49.6%) took a subset of the IPD population that matched the inclusion criteria of the aggregated data population, and 18 (13.5%) stated that inclusion criteria were similar enough to include the entire IPD population. Forty-nine (36.8%) did not report any information about the alignment of inclusion criteria.

3.2. Main methodological and reporting characteristics

A total of 288 PAICs were found in the selected articles: 257 MAICs (89%), 29 STCs (10%), and two multilevel network-meta regression (from a single article), a recently developed method allowing to extend PAIC beyond pairwise comparisons [33]. These indirect comparisons were mostly unanchored (73%, see Supplementary Figure 1). Table 2 presents the list of key methodological and reporting aspects of PAICs, as defined in the Methods section.

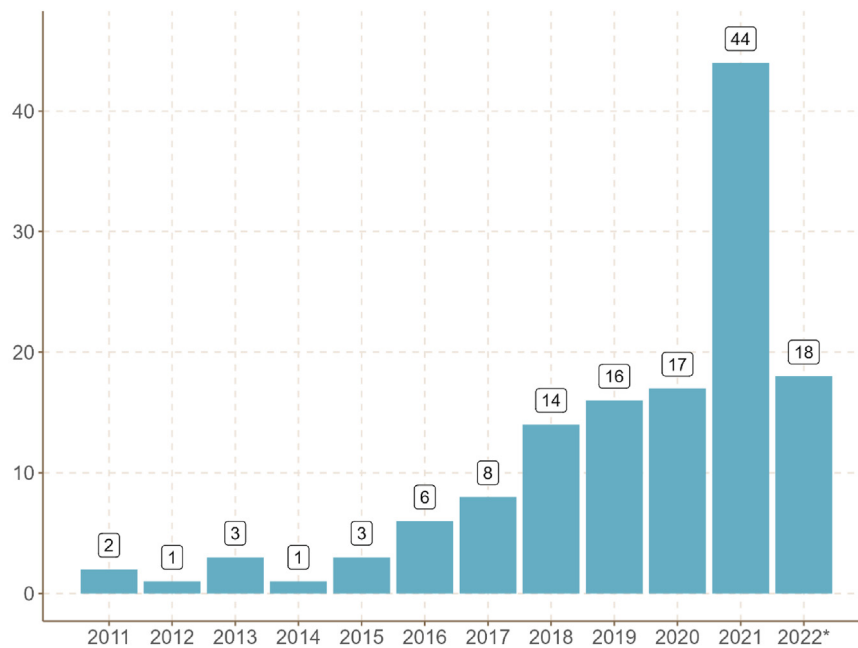
Only 53% (71/133) articles (61%, 175/288 at the PAIC level) reported having conducted a systematic review to find source studies suitable for the indirect comparison. Among the 12 articles with no collaboration with methodological experts, the percentage of systematic reviews fell to 17% (2/12 studies). Only 72 (24.7%) comparisons mentioned having defined a primary outcome.

Almost all PAICs reported information about the initial sample size before adjustment. Overall, 77% (199/257) of the MAICs reported the effective sample size after adjustment. One-quarter of the MAICs reported an effective

Table 1. Characteristics of the articles included in the review ($n = 133$)

Characteristics	n (%) / median [Q1, Q3]
Publication period	
Median [Q1, Q3]	7 May, 2020 [9 August 2018, 3 June 2021]
Countries of first author affiliations	
United States	68 (51.1)
United Kingdom	19 (14.3)
Canada	9 (6.8)
Other	37 (27.8)
Countries of last author affiliations	
United States	68 (51.1)
United Kingdom	15 (11.3)
Germany	13 (9.8)
Other	37 (27.8)
Methodological department or Contract Research Organization	
Yes	121 (91)
Not mentioned	12 (9)
Positions of investigators	
Pharmaceutical company	125 (94)
Academia	96 (72.2)
CRO	109 (82)
Condition (ICD-10 chapter)	
Neoplasms	71 (53.4)
Diseases of the nervous system	15 (11.3)
Diseases of the skin and subcutaneous tissue	11 (8.3)
Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	8 (6)
Certain infectious and parasitic diseases	6 (4.5)
Endocrine, nutritional and metabolic diseases	5 (3.8)
Other	17 (12.8)
Authorship or financial support from pharmaceutical industry	
Yes	129 (97)
No	4 (3)
Number of PAIC(s) per article	
Median [Q1, Q3]	1 [1, 3]
Type of PAIC conducted	
MAIC	118 (88.7)
STC	10 (7.5)
MAIC and STC	4 (3)
ML-NMR	1 (0.8)
Anchored vs. unanchored PAIC	
Anchored	38 (28.6)
Unanchored	86 (64.7)
Both	9 (6.8)

Abbreviations: PAIC, population-adjusted indirect comparison; MAIC, matching-adjusted indirect comparison; STC, simulated treatment comparison; ML-NMR, multilevel network meta-regression; ICD-10, international classification of diseases, revision 10.



* For the year 2022, data collected up to April 2 only

Fig. 2. Annual publication count of PAIC application. PAIC: population-adjusted indirect comparison.

sample size ≤ 50 . The median sample size reduction after adjustment was 39.3% ([Supplementary Table 4](#)).

Detailed information about covariates is in [Supplementary Tables 5 and 6](#). Choice of the covariates used in the analysis was based solely on the literature or a priori medical knowledge for 33% (96/288) of the PAICs and on statistical considerations for 31% (88/288). For 36.1% (104/288) of the PAICs, this choice was not explained. Many PAICs (56%, 127/288) did not discuss the status of the covariates (i.e., whether the covariates were prognostic factors or treatment-effect modifiers). Among the 127 PAICs that did report this information, 12 (9%) unanchored PAICs did not include prognostic factors in the adjustment model and 19 (15%) anchored MAICs incorrectly included prognostic factors.

Overall, only 3 (1%) comparisons from three different studies reported all key defined criteria.

3.3. Reported results

More than half (167/288, 58%) of the PAICs reported results for both unadjusted and adjusted comparisons, and adjusted results were reported for 96.9% of PAICs ([Supplementary Table 7](#)). Time-to-event was the most frequent type of outcome ([Supplementary Table 8](#)). The median absolute change between unadjusted and adjusted effect size was 14.9%.

Among the 279 PAICs for which an adjusted effect was reported, 58% (162/279) were statistically significant ([Table 3](#)). The treatment effect was beneficial for the IPD

treatment arm for all but one comparison (161/162, 99.6%). The only PAIC in favor of the aggregated data treatment arm was one of the three articles without any involvement of the pharmaceutical industry. [Fig. 3](#) represents the treatment effects reported along with the adjusted sample size (for relative measures of association only).

We identified six pairs of articles comparing the same two treatments with the same data sources but with the IPD and aggregated data swapped. Their main differences in terms of methodological aspects and conclusions are summarized in [Supplementary Table 9](#). In brief, in four of these six pairs, one study reported a significant effect in favor of the IPD treatment, and the other study yielded nonsignificant findings. In one pair, both studies reported a significant effect in favor of the IPD treatment (thus arriving at opposite conclusions on the most effective treatment). Finally, the last pair showed point estimates favoring the IPD treatment but without any significant results.

4. Discussion

This review found a substantial increase in the number of PAICs published in recent years. About three-quarters of PAICs were unanchored (i.e., without common treatment arm), and MAICs were significantly more prevalent than STCs. We found significant heterogeneity in the granularity and exhaustiveness of the reported information, which hampers the reader's ability to correctly assess the

Table 2. Key methodological and reporting aspects of the PAICs included in the review ($n = 288$)

Key methodological and reporting aspects	n (%)
Systematic review to identify source articles	
Yes	175 (60.8)
Primary outcome defined	
Yes	71 (24.7)
Sample size reporting	
IPD	
Before adjustment	281 (97.6)
Adjusted (ESS for MAIC)	228 (79.2)
Aggregated data	228 (79.2)
Adequate justification of the covariates used in the analysis	
Yes: <i>a priori</i> medical-based only	96 (33.3)
Partly statistical-based	67 (23.3)
Exclusively statistical-based	21 (7.3)
No justification	104 (36.1)
Adequate covariates inclusion in the weighting/ outcome model	
Yes	96 (33.3)
Anchored MAIC ($N = 69$): inclusion of prognostic factors	19 (6.6)
Unanchored PAIC ($N = 209$): no inclusion of prognostic factors	12 (4.2)
No mention of covariates' effects	161 (55.9)
Satisfying each of the five listed criteria	
Yes	3 (1.0)

Abbreviations: PAIC, population-adjusted indirect comparison; IPD, individual patient data; MAIC, matching-adjusted indirect comparison; ESS, effective sample size.

relevance of the results. Reported results almost never favored the treatment evaluated with aggregated data.

Phillippo et al. [14] performed a review focusing on PAICs submitted to the NICE HTA until April 2018. Contrary to the prediction of these authors, we did not observe a significant increase in the proportion of unanchored vs. anchored PAICs conducted before and after 2019; this proportion remained approximately stable [14,21]. Another recent conference abstract also reviewed PAICs submitted to NICE and found overall consistent results with our study [34].

Studies involving PAIC were almost exclusively produced in collaboration with the pharmaceutical industry.

Population-adjusted indirect comparisons (PAICs) seem especially appealing in the context of innovative treatments (particularly in oncology and hematology), an area in which single-arm trials and accelerated approval requests are common [19,20].

Reporting for key methodological aspects in PAICs was heterogeneous. In many cases, a systematic review was missing, making it difficult to determine whether all available information had been found before performing the comparisons. Most PAICs lacked a clearly defined primary outcome. Because published PAICs aim at filling the evidence gap due to the absence of direct comparisons, defining a primary outcome is not less crucial than it is for randomized controlled trials [35]. However, PAICs are often conducted with the direct intent to inform HTA files and economics modeling, wherein payers may establish a different hierarchy of outcomes. Owing to this specific context, it is understandable that research using PAIC may report on a broad spectrum of secondary outcomes. Sample size reporting was unclear and inconsistent: we believe that sample sizes should be systematically reported in a flowchart, starting from the sample size of each source study up to the number of individuals included for the treatment effect estimation and reasons for exclusion at each stage.

Regarding covariates adjustment, a discussion about the landscape of confounders, and among them which ones were adjusted for and which ones could not, was lacking most of the time. The depth of information provided varied greatly among articles, from a mere mention that medical expertise was sought after to identify confounders, to a specific description of the confounder status of each variable (e.g., Ribero et al., [36] Bhak et al. [37]). Providing evidence regarding the availability of treatment-effect modifier data (and prognostic factor data for unanchored comparisons) is crucial: without them, there is no reason to think that an adjusted treatment effect will be more precise than an unadjusted one [17]. Relying solely on statistics to identify covariates exposes to the risk of under or overadjustment (the latter may lead to unnecessary small effective sample size when MAIC is used). More importantly, overlooking medical knowledge will prevent identifying unobserved confounders that could not be accounted for, leading to unacknowledged residual bias.

When a pharmaceutical company is involved in a PAIC, the IPD arm represents the firm's promoted treatment, and the aggregated data arm represents a competitor's

Table 3. Statistical significance of the adjusted treatment effect (primary outcome, or first one reported)

	Unknown ($N = 7$)	Not statistically significant ($N = 110$)	Statistically significant ($N = 162$)	Total ($N = 279$)
Significance (5% alpha risk)	n (%)	n (%)	n (%)	n (%)
Favors IPD study intervention				
No	1 (14.3)	33 (30)	1 (0.6)	35 (12.5)
Yes	6 (85.7)	77 (70)	161 (99.4)	244 (87.5)

Abbreviations: IPD, individual patient data.

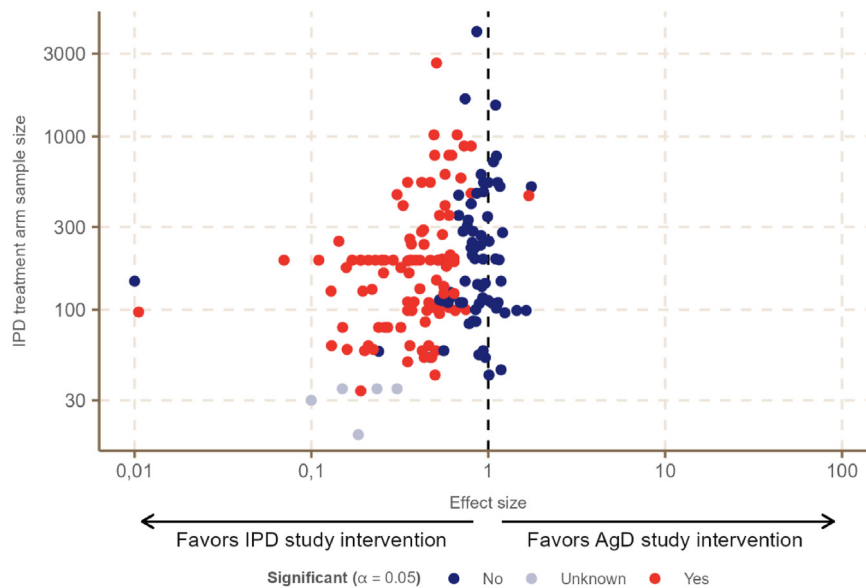


Fig. 3. Reported adjusted treatment effects (relative measures of association only: OR, RR, and HR). OR: odds ratio; RR: relative risk; HR: hazard ratio; IPD: individual patient data; AgD Aggregated Data. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

treatment. Only one PAIC reported a statistically significant benefit for the aggregated-data treatment arm overall and none when the pharmaceutical industry was involved. This finding suggests publication and reporting bias, as studies may only be submitted if the results favor the sponsored treatment, and favorable results may be “cherry-picked” for reporting. Reporting bias is particularly problematic in the PAIC context because of the numerous degrees of freedom offered by the methodology and the lack of reporting recommendations and protocol preregistration [38]. The hypothesis of reporting bias is reinforced by the existence of several pairs of articles interested in the exact same treatment pair, with the same data sources but with the IPD and aggregated data swapped. In such cases, the study published later usually refutes some results of the earlier study and draws at least partially diverging conclusions.

The five criteria presented in Table 2 were deemed the most important to enhance PAICs’ utility, reproducibility, and robustness. Nonetheless they do not represent an exhaustive list and further work could focus on other aspects such as model details used for weight estimation in MAICs and outcome modeling in STCs [11], or the choice of a relative or absolute scale for the indirect comparison [10,14].

Even though we found several studies with noteworthy methodology and reporting completeness [36,39], our review suggests room for improvement. Therefore, we give recommendations for future studies (Box 1). The recommendations are based on the insights of this review, the methodological literature on PAIC methods [12], as well as the established guidelines to report results from observational studies (STROBE [40]) and systematic reviews (PRISMA [26]).

4.1. Limitations

To simplify data collection with numerous reported outcomes, we collected only one outcome per comparison, typically the primary outcome, or the first reported outcome if not defined. Although the results could have differed if all outcomes had been collected, the reporting bias persisted even when considering only articles with defined primary outcomes.

This reporting bias should be evaluated in light of the actual information submitted to HTA agencies, as a previous study revealed discrepancies between article publications and HTA submission, with the former reporting more favorable outcomes [43].

The search was restricted to PubMed, and conference abstracts were excluded. Although some PAIC literature may have been missed, conference abstracts lacked the detailed information required for the review.

Our conclusions primarily centered around the comprehensiveness and the level of detail of reporting rather than the methodology itself. Particularly, we could not appraise the assertion regarding the prognostic factor or treatment-effect modifier status of the covariates, which would have required domain-specific knowledge.

5. Conclusions

Population-adjusted indirect comparisons (PAICs) can provide valuable insights for HTA decisions, yet this review highlighted inconsistencies in reporting and flaws in the methodology implementation that warrant caution. Clear PAIC reporting guidelines need to be developed and

Box 1 Reporting Recommendations [41,42].

1. Register the protocol of the planned PAIC on any platform accepting observational studies protocol (e.g., [ClinicalTrials.gov](https://www.clinicaltrials.gov) [41], Open Sciences Foundation [42]) before the start of the work. Critically, the protocol should specifically mention a primary outcome (and justify its choice) and a primary analysis methodology (i.e., regarding the set of covariates adjusted for, the type of PAIC carried-out, the chosen population subset ...). Secondary and sensitivity analyses should also ideally be planned and mentioned beforehand. Post hoc analyses should be reported as such.
2. Conduct a literature review to 1) search for potential existing direct comparisons, and 2) gather an exhaustive list of studies eligible for the PAIC. Incidentally, a literature review can also identify ongoing trials directly comparing the treatments of interest, thus providing an opportunity to indicate that the PAIC result will be "temporary" until the completion of the trial providing the direct comparison. Details of the review should be reported (including at least the search equation and the article selection flowchart).
3. Cite inclusion criteria of source studies as well as their eventual differences and how they have been considered in the PAIC.
4. Report sample sizes at different steps: before and after matching (i.e., inclusion/exclusion criteria harmonization), and after reweighting in MAICs (ESS). Sample sizes should also be reported separately for both treatment arms for an anchored comparison. Additionally, reporting the weight distribution in MAICs allows for assessing the degree of similarity between compared populations.
5. Mention known prognostic factors/treatment-effect modifiers based on references from a literature search or justified medical expertise. Specifically indicate which have been adjusted for and which have not. Furthermore, report details of the adjustment/weighting model used (e.g., first-order moments, second-order moments, potential interaction terms, eventual splines, etc.).
6. Reporting unadjusted in addition to adjusted treatment effect (with statistical testing significance or confidence/credible intervals).

endorsed by relevant organizations to improve PAIC reporting, which will allow readers to adequately assess the quality of the PAICs. The recommendations provided in this article should serve as a starting point to help addressing inconsistencies in PAIC reporting. However, even with

enhanced reporting, reviewers should be aware of potential publication bias, as industrials are unlikely to publish results that are not in favor of their product.

CRedit authorship contribution statement

Arnaud Serret-Larmande: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing — original draft. **Belkacem Zenati:** Investigation, Formal analysis, Writing — review & editing. **Agnès Dechartres:** Methodology, Formal analysis, Writing — review & editing. **Jérôme Lambert:** Conceptualization, Methodology, Formal analysis, Writing — review & editing, Supervision. **David Hajage:** Conceptualization, Methodology, Investigation, Formal analysis, Writing — review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no competing interest.

Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jclinepi.2023.09.004>.

References

- [1] Vanier A, Fernandez J, Kelley S, Alter L, Semenzato P, Alberti C, et al. Rapid access to innovative medicinal products while ensuring relevant health technology assessment. Position of the French National Authority for Health. *BMJ Evid-Based Med* 2023. <https://doi.org/10.1136/bmjebm-2022-112091>.
- [2] Comparators and comparisons: direct and indirect comparisons. European Network for Health Technology Assessment. Available at https://www.eunetha.eu/wp-content/uploads/2018/01/Comparators-Comparisons-Direct-and-indirect-comparisons_Amended-JA1-Guide-line_Final-Nov-2015.pdf. Accessed January 5, 2023.
- [3] Patel D, Grimson F, Mihaylova E, Wagner P, Semenzato P, Van Engen A, et al. Use of external comparators for health technology assessment submissions based on single-arm trials. *Value Health* 2021;24(8): 1118–25.
- [4] Cucherat M, Izard V. Indirect comparisons methods and validity. Available at. Haute Autorité de Santé; 2009:66. Available at https://www.has-sante.fr/upload/docs/application/pdf/2011-02/summary_report_indirect_comparisons_methods_and_validity_january_2011_2.pdf. Accessed February 10, 2022.
- [5] Lambert J, Lengliné E, Porcher R, Thiébaud R, Zohar S, Chevret S. Enriching single-arm clinical trials with external controls: possibilities and pitfalls. *Blood Advances* 2023;7:5680–90.
- [6] Signorovitch JE, Wu EQ, Yu AP, Gerrits CM, Kantor E, Bao Y, 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: 935–45.
- [7] Signorovitch JE, Sikirica V, Erder MH, Xie J, Lu M, Hodgkins PS, et al. Matching-adjusted indirect comparisons: a new tool for timely comparative effectiveness research. *Value Health* 2012;15(6):940–7.

- [8] Caro JJ, Ishak KJ. No head-to-head trial? Simulate the missing arms. *Pharmacoeconomics* 2010;28:957–67.
- [9] Ishak KJ, Proskorovsky I, Benedict A. Simulation and matching-based approaches for indirect comparison of treatments. *Pharmacoeconomics* 2015;33:537–49.
- [10] Phillippo D, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. Methods for population-adjusted indirect comparisons in submission to NICE. Decision support unit. Available at: National Institute for Health and Care Excellence; 2016:82. Available at <http://nicedsu.org.uk/wp-content/uploads/2018/08/Population-adjustment-TSD-FINAL-ref-rerun.pdf>. Accessed February 10, 2022.
- [11] Remiro Azócar A. Population-adjusted indirect treatment comparisons with limited access to patient-level data. PhD thesis. London: University College London; UCL Discovery; 2021.
- [12] 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* 2018;38:200–11.
- [13] Remiro-Azócar A. Target estimands for population-adjusted indirect comparisons. *Stat Med* 2022;41:5558–69.
- [14] Phillippo DM, Dias S, Elsad A, Ades AE, Welton NJ. Population adjustment methods for indirect comparisons: a review of national institute for health and care excellence technology appraisals. *Int J Technol Assess Health Care* 2019;35(03):221–8.
- [15] D4.3.1. Direct and indirect comparisons - individual practical guideline document. EUnetHTA 21. 2022. Available at <https://www.eunetha.eu/wp-content/uploads/2022/12/EUnetHTA-21-D4.3.1-Direct-and-indirect-comparisons-v1.0.pdf>. Accessed August 1, 2023.
- [16] D4.3.2 direct and indirect comparisons. EUnetHTA 21. 2022. Available at https://www.eunetha.eu/wp-content/uploads/2022/05/D4.3.2_GL_Comparisons_for-public-consultation.pdf?x69613=. Accessed August 1, 2023.
- [17] Pooley N, Kisomi M, Embleton N, Langham S. CO163 the increasing use of population-adjusted indirect comparisons in the NICE health technology assessment (HTA) submission process and the response to these methods. *Value Health* 2022;25(12):S49.
- [18] 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.
- [19] Beaver JA, Howie LJ, Pelosof L, Kim T, Liu J, Goldberg K, 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):849–56.
- [20] Agrawal S, Arora S, Amiri-Kordestani L, de Claro RA, Fashoyin-Aje L, Gormley N, et al. Use of single-arm trials for US food and drug administration drug approval in oncology, 2002–2021. *JAMA Oncol* 2022;9:266–72.
- [21] 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):750–75.
- [22] Phillippo DM, Dias S, Ades AE, Welton NJ. Assessing the performance of population adjustment methods for anchored indirect comparisons: a simulation study. *Stat Med* 2020;39:4885–911.
- [23] 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;50:683–91.
- [24] Lambert PC, Sutton AJ, Abrams KR, Jones DR. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *J Clin Epidemiol* 2002;55:86–94.
- [25] Stewart LA, Clarke MJ. Practical methodology of meta-analyses (overviews) using updated individual patient data. *cochrane working group. Stat Med* 1995;14:2057–79.
- [26] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann T, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Syst Rev* 2021;10(1):89.
- [27] Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. *Syst Rev* 2016;5(1):210.
- [28] Guidelines for reporting outcomes in trial reports: the CONSORT-outcomes 2022 extension, research, methods, statistics, JAMA, JAMA network. Available at <https://jamanetwork.com/journals/jama/fullarticle/2799401>. Accessed August 4, 2023.
- [29] Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med* 1988;318:1728–33.
- [30] 5 The reference case, Guide to the methods of technology appraisal 2013, Guidance, NICE. Available at <https://www.nice.org.uk/process/pmg9/chapter/the-reference-case#synthesis-of-evidence-on-health-effects>. Accessed March 6, 2023.
- [31] R Core Team. R: a language and environment for statistical computing 2022. Available at <https://www.R-project.org/>.
- [32] Signorovitch JE, Wu EQ, Swallow E, Kantor E, Fan L, Gruenberger JB. Comparative efficacy of vildagliptin and sitagliptin in Japanese patients with type 2 diabetes mellitus: a matching-adjusted indirect comparison of randomized trials. *Clin Drug Investig* 2011;31(9):665–74.
- [33] Phillippo DM, Dias S, Ades AE, Belger M, Brnabic A, Schacht A, et al. Multilevel network meta-regression for population-adjusted treatment comparisons. *J R Stat Soc Ser A Stat Soc* 2020;183(3):1189–210.
- [34] Kroi F, Hu Y, Beekhuizen S van, Heeg B. P43 navigating the maze of standard and novel population-adjusted indirect comparison (PAIC) methods. *Value Health* 2022;25(12):S10.
- [35] Page MJ, McKenzie JE, Kirkham J, Dwan K, Kramer S, Green S, et al. Bias due to selective inclusion and reporting of outcomes and analyses in systematic reviews of randomised trials of health-care interventions. *Cochrane Database Syst Rev* 2014;2014:MR000035.
- [36] Ribero VA, Daigl M, Martí Y, Gorni K, Evans R, Scott DA, et al. How does risdiplam compare with other treatments for Types 1-3 spinal muscular atrophy: a systematic literature review and indirect treatment comparison. *J Comp Eff Res* 2022;11(5):347–70.
- [37] Bhak RH, Mody-Patel N, Baver SB, Kunzweiler C, Yee CW, Sundaresan S, et al. Comparative effectiveness of pegcetacoplan versus ravulizumab in patients with paroxysmal nocturnal hemoglobinuria previously treated with eculizumab: a matching-adjusted indirect comparison. *Curr Med Res Opin* 2021;37(11):1913–23.
- [38] Patel CJ, Burford B, Ioannidis JPA. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J Clin Epidemiol* 2015;68:1046–58.
- [39] Rugo HS, Haltner A, Zhan L, Tran A, Bananis E, Hooper B, et al. Matching-adjusted indirect comparison of palbociclib versus ribociclib and abemaciclib in hormone receptor-positive/HER2-negative advanced breast cancer. *J Comp Eff Res* 2021;10(6):457–67.
- [40] Elm E von, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 2007;335:806–8.
- [41] Create a preregistration - OSF support. Available at <https://help.osf.io/article/158-create-a-preregistration>. Accessed February 28, 2023.
- [42] ClinicalTrials.gov. Protocol registration data element definitions for interventional and observational studies. Available at <https://prsinfo.clinicaltrials.gov/definitions.html#ObsStudyDesign>. Accessed February 28, 2023.
- [43] Rising K, Bacchetti P, Bero L. Reporting bias in drug trials submitted to the food and drug administration: review of publication and presentation. *PLoS Med* 2008;5(11):e217.

Supplementary materials to “*A methodological review of Population-Adjusted Indirect Comparisons reveals inconsistent reporting and suggests publication bias*”

This document contains the search equation, supplementary information on how the data was collected, supplementary figures and tables.

PubMed search equation

```
(
  (
    ("indirect treatment comparison*")
    OR ("simulated treatment comparison*")
    OR ("adjusted indirect comparison*")
    OR ("maic"[Title/Abstract])
    OR ( ("treatment comparison*") OR ("indirect comparison*"))
      AND (("individual patient data") OR ("aggregate data") OR ("aggregated data") OR
("population level data") OR ("patient level data"))
  )
  OR (("indirect comparison*") AND ("propensity score"))
) AND (2010/1/1:2022/4/2[pdat])
) OR
(
  (28823204,35267593,34912478,34877623,34822128,34754238,34616492,34493319,
34368918,34292540,34250585,34250584,34074239,34018204,33970454,33847901,3376812
3,33675139,33556898,33488773,33292314,33215981,33155118,33145711,33015906,329935
19,32684669,32524498,32395870,32007196,31930460,31838711,31390809,31364979,31218
655,31190671,31021449,30388073,30312378,30304407,29605841[UID])
  OR
  (20831304,34074239,33598857,33448252,33015906,32957804,32684669,32524498,324753
40,31564931,31190671,31140123,28823204,27529762,27116943,26979173,25795232,25414048,246
43323,22788256,20831287[UID])
  OR (20831302,34765537,34383273,34292540,34280545,34277983,34074239,34018204,338
71055,33847901,33830214,33822328,33448252,33291810,33244703,33015906,32684669,32524498,
32499262,32475340,32395870,32371431,32301116,31943429,31917873,31838710,31417326,31364
979,31190671,30758745,30472749,30388073,30312378,30309978,29961001,29954322,29605841,29
271481,29192016,28823204,28762213,28607774,28573505,27790010,27445511,27116943,2682414
5,26586608,26287812,25795232,25414048,24653837,24314093,23975660,22924967,21819162,2083
1287[UID])
  OR (22999145,35044552,34877623,34797506,34765537,34616492,34292540,34285621,342
77983,34250585,34080140,34067521,34067288,34031743,34021585,33970454,33847901,33822328,
33745079,33674928,33664606,33598857,33556898,33488773,33472427,33448252,33435879,33377
987,33244703,33062927,32993519,32957804,32804408,32367407,31943429,31838710,31813086,31
743962,31646068,31642037,31564931,31417326,31364979,31218655,31140123,31021449,3075874
5,30610657,30567533,30532569,30472749,30388073,30312378,30309978,29954322,29861642,2960
5841,29192016,28856580,28823204,28762213,28680955,28607774,28573505,27778240,27445511,2
7366049,26979173,26287812,25795232,25533141,23975660[UID])
)
```

Data collection details

Data were collected to inform three dimensions: the main characteristics of the article, the source studies, and the methodology and results of the PAIC (Population-Adjusted Indirect Comparison).

Article's main characteristics

The following data were collected: journal, publication year, country(ies) of first and last author affiliations, positions of study investigators (academic, pharmaceutical company, Contract Research Organization), mentioned source(s) of funding (public grant, industrial funding), and medical condition studied.

Information for the source studies used in the indirect comparison

For each source study, we collected whether the study was observational or interventional, the country(ies) in which the study was carried out, and for clinical trials, the administrative unit identifying information (NCT or EUDRA-CT), the phase(s), and the number of treatment arms.

Methodology and results of PAIC

For each comparison, the following data were collected: performance of a systematic review to find eligible source studies to include in the research, type of PAIC conducted, use of an anchored PAIC, definition of a primary outcome, outcome type (binary, categorical, continuous, time-to-event) and treatment contrast effect measure (e.g., mean difference, HR, RR etc.), rationale for covariates to be included in the adjustment model and their (expected) relation to the outcome (prognostic factors, treatment-effect modifiers), specification of the exposure/outcome model (e.g., inclusion of interaction terms, nonlinear relation between covariates, etc.), sample sizes (before and after adjustment: after adjustment corresponding to the Effective Sample Size for MAICs, and to the sample size used in the adjustment model for STCs which may differ from the original sample size in case of missing data for the variables included in the model), reporting of a weight distribution evaluation (matching-adjusted indirect comparison [MAIC] only), and relative treatment-effect point estimate and its significance level (adjusted and unadjusted).

Each article could report more than one PAIC result by comparing several pairs of treatments or conducting different types of PAICs (i.e., MAIC or simulated treatment comparison [STC]). We collected methods and results for each pair of treatment comparisons (a single treatment defined by a molecule and its regimen), and for each type of PAIC presented (e.g., collecting methodology information and results of both MAIC and STC if both were reported in a given article). For cases for which one given PAIC type was performed several times (e.g., several MAICs for the same compared pair of treatments with variations in the population selection process or in the variables considered), methods/results were collected for the primary analysis only (or the first one mentioned if no primary analysis was defined). Similarly, we collected information for only one outcome (the primary outcome, or the first one reported in the text of the “Results” part if no primary outcome was defined) for each comparison. Statistical significance of the reported outcomes was assessed with a 5% alpha risk.

Supplementary Tables

Supplementary Table 1: **Characteristics of treatments compared in the population-adjusted indirect comparisons (PAICs).**

		IPD	AgD	Total
		(N = 382)	(N = 377)	(N= 759)
		n (%)	n (%)	n (%)
Treatment group (ATC 3rd level[†])				
	Other antineoplastic agents	99 (25.9)	70 (18.6)	169 (22.3)
	Protein kinase inhibitors	58 (15.2)	66 (17.5)	124 (16.3)
	Immunosuppressants	61 (16)	54 (14.3)	115 (15.2)
	Direct acting antivirals	36 (9.4)	28 (7.4)	64 (8.4)
	Vitamin K and other hemostatics	20 (5.2)	20 (5.3)	40 (5.3)
	Corticosteroids for systemic use, plain	14 (3.7)	17 (4.5)	31 (4.1)
	Other ATC 3 rd -level categories	94 (24.6)	122 (32.4)	216 (28.5)
CAR T-cell therapy				
	No	374 (97.9)	374 (99.2)	748 (98.6)
	Yes	8 (2.1)	3 (0.8)	11 (1.4)
Monoclonal antibodies				
	No	290 (75.9)	302 (80.1)	592 (78)
	Yes	92 (24.1)	75 (19.9)	167 (22)

IPD: individual patient data; AgD: aggregated data; CAR: chimeric antigen receptor

[†] Treatments were mapped to their code in the 2023 Anatomical Therapeutic Chemical (ATC) classification and grouped to the third level (see *WHO Collaborating Centre for Drug Statistics Methodology, Guidelines for ATC classification and DDD assignment, 2023. Oslo, 2022.*)

Supplementary Table 2: **Characteristics of the articles included in the review, depending on whether academics were among the authors (n=133)**

		n (%) / median [Q1, Q3]			
		Academics among authors			
		Yes (N=96)	No (N=37)	Total (N = 133)	<i>p</i> -value
Publication period	Median [Q1; Q3]	August 31, 2020 [January 31 2019; June 3 2021]	February 23, 2019 [August 24, 2016; June 4, 2021]	May 7, 2020 [August 9, 2018; June 3, 2021]	0.071 [‡]
Countries of first author affiliations					0.825 [†]
	USA	47 (49)	21 (56.8)	68 (51.1)	
	UK	15 (15.6)	4 (10.8)	19 (14.3)	
	Canada	7 (7.3)	2 (5.4)	9 (6.8)	
	Other	27 (28.1)	10 (27)	37 (27.8)	
Countries of last author affiliations					0.032 [†]
	USA	42 (43.8)	26 (70.3)	68 (51.1)	
	UK	11 (11.5)	4 (10.8)	15 (11.3)	
	Germany	12 (12.5)	1 (2.7)	13 (9.8)	
	Other	31 (32.3)	6 (16.2)	37 (27.8)	
Methodological department or Contract Research Organization					0.114 [†]
	Yes	85 (88.5)	36 (97.3)	121 (91)	
	Not mentioned	11 (11.5)	1 (2.7)	12 (9)	
Positions of co-authors					
	Pharmaceutical Industry	88 (91.7)	37 (100)	125 (94)	0.070 [†]
	Private Data Analysis Company	75 (78.1)	34 (91.9)	109 (82)	0.064 [†]
Condition (ICD-10 chapter)					0.003 [†]
	Neoplasms	55 (57.3)	16 (43.2)	71 (53.4)	
	Diseases of the nervous system	9 (9.4)	6 (16.2)	15 (11.3)	
	Diseases of the skin and subcutaneous tissue	10 (10.4)	1 (2.7)	11 (8.3)	

		n (%) / median [Q1, Q3]			
		Academics among authors			
		Yes (N=96)	No (N=37)	Total (N = 133)	<i>p-value</i>
	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	6 (6.2)	2 (5.4)	8 (6)	
	Certain infectious and parasitic diseases	0 (0)	6 (16.2)	6 (4.5)	
	Endocrine, nutritional and metabolic diseases	4 (4.2)	1 (2.7)	5 (3.8)	
	Other	12 (12.5)	5 (13.5)	17 (12.8)	
Authorship or financial support from pharmaceutical industry					0.207 [†]
	Yes	92 (95.8)	37 (100)	129 (97)	
Number of PAIC(s) per article					0.921 [¥]
	Median [Q1, Q3]	1 [1, 2]	1 [1, 3]	1 [1, 3]	
Type of PAIC conducted					0.523 [†]
	MAIC	87 (90.6)	31 (83.8)	118 (88.7)	
	MAIC and STC	2 (2.1)	2 (5.4)	4 (3)	
	STC	6 (6.2)	4 (10.8)	10 (7.5)	
	ML-NMR	1 (1)	0 (0)	1 (0.8)	
Anchored vs unanchored					0.882 [†]
	Anchored	28 (29.2)	10 (27)	38 (28.6)	
	Unanchored	61 (63.5)	25 (67.6)	86 (64.7)	
	Both	7 (7.3)	2 (5.4)	9 (6.8)	

PAIC: population-adjusted indirect comparison; MAIC: matching-adjusted indirect comparison; STC: simulated treatment comparison; ML-NMR: multilevel network meta-regression; ICD-10, International Classification of Diseases, Revision 10

[†]: Fisher's exact test

[¥]: Wilcoxon test

Supplementary Table 3: **PAIC** source study characteristics

		IPD (N = 185)	AgD (N = 285)	Total (N = 470)
		n (%)	n (%)	n (%)
Study type				
	Observational study	9 (4.9)	22 (7.7)	31 (7.6)
	Clinical trial	176 (95.1)	263 (92.3)	439 (93.4)
Country where the study was conducted (international if more than one)				
	Missing	12	21	33
	International	131 (75.7)	199 (75.4)	330 (75.5)
	United States	25 (14.5)	29 (11)	54 (12.4)
	France	3 (1.7)	9 (3.4)	12 (2.7)
	Italy	4 (2.3)	8 (3)	12 (2.7)
	Japan	8 (4.6)	4 (1.5)	12 (2.7)
	Others	2 (1.2)	15 (5.7)	17 (3.9)
Number of arms (clinical trials only)				
	Single-arm trial	45 (25.6)	70 (26.6)	115 (26.2)
	Multiple-arm trial	131 (74.4)	191 (72.6)	322 (73.3)
	Unknown	0 (0)	2 (0.8)	2 (0.5)
Clinical trial phases[†]				
	Phase 1	20 (10.8)	16 (5.6)	36 (7.7)
	Phase 2	42 (22.7)	60 (21.1)	102 (21.7)
	Phase 3	125 (67.6)	171 (60)	296 (63)
	Phase 4	1 (0.5)	9 (3.2)	10 (2.1)
	Unknown	4 (2.2)	26 (9.1)	30 (6.4)

PAIC: population-adjusted indirect comparison; IPD: individual patient data; AgD: aggregated data

[†] multiple phases possible for each trial, so percentages may add up to more than 100%

Supplementary Table 4: **PAIC sample sizes**

		MAIC	STC	ML-NMR	Total
		(N = 257)	(N = 29)	(N = 2)	(N = 288)
IPD sample size	Missing	4	3	0	7
	Median	193 [110, 394]	193 [72.5, 193]	961 [961, 961]	193 [110, 380]
	[Q1, Q3]				
Aggregate data sample size	Missing	8	3	0	11
	Median	208 [107, 470]	53.5 [36.5, 352]	1227 [1226.5, 1227.5]	205 [99, 463]
	[Q1, Q3]				
Effective sample size (ESS)	Missing	58			
	Median	97 [49.5, 248]			
	[Q1, Q3]				
Reduction percentage ESS	Missing	59			
	Median	39.3 [23.4, 60.9]			
	[Q1, Q3]				

PAIC: population-adjusted indirect comparison; MAIC: matching-adjusted indirect comparison; STC: simulated treatment comparison; ML-NMR: multilevel network meta-regression; ESS: effective sample size

Supplementary Table 5: Authors' explanations for covariates adjusted for in PAICs (n=288 PAICs)

		n (%)
No justification		
	No	184 (63.9)
	Yes	104 (36.1)
Expertise		
	No	162 (56.2)
	Yes	126 (43.8)
Literature review		
	No	201 (69.8)
	Yes	87 (30.2)
Statistical effect in the IPD dataset		
	No	222 (77.1)
	Yes	66 (22.9)
Imbalance		
	No	256 (88.9)
	Yes	32 (11.1)
Covariates justification		
	<i>A priori</i> medical-based only	96 (33.3)
	Partly statistical-based	67 (23.3)
	Only statistical-based	21 (7.3)
	No justification	104 (36.1)

IPD: individual patient data

Supplementary Table 6: **Adjustment model details: number of covariates, and characteristics of covariates adjusted for (n=288 PAICs)**

		n (%) / median [Q1, Q3]
Number of covariates adjusted for/matched on	Missing Median [Q1, Q3]	13 7 [4, 8]
Type of covariates adjusted for		
	Age	231 (84)
	Sex	174 (63.3)
	Medical history	249 (90.5)
	Comorbidities	99 (36)
	Performance score	119 (43.3)
	Past treatments	155 (56.4)
	Race/ethnicity	86 (31.3)
	Others	84 (30.5)

Supplementary Table 7: Characteristics of reported results (primary outcome, or first one reported) (n=288 PAICs)

	n (%) / median [Q1, Q3]
Reporting an unadjusted effect size	
Yes	167 (58)
Reporting an adjusted effect size	
Yes	279 (96.9)
Reporting of a weight distribution evaluation (MAIC)	
Missing	28
Not mentioned	195 (75)
Yes, mentions that the weight distribution has been studied, and weight distribution reported in main article or Supplementary materials	30 (11.5)
Yes, mentions that the weight distribution has been studied, but weight distribution not reported	35 (13.5)
Primary outcome: variable type	
Binary (e.g., rates)	106 (36.8)
Continuous (count, mean, etc.)	54 (18.8)
Time-to-event	128 (44.4)
Primary outcome: treatment contrast effect	
Missing	5
Means/median difference	47 (16.6)
Rates/ratio	184 (65)
Rates/ratio difference	52 (18.4)
Effect size change between unadjusted and adjusted treatment effect (in percentage) *	
Missing	121
Median [Q1, Q3]	14.9 [6.6, 33.9]

MAIC: matching-adjusted indirect comparison

* Unadjusted effect sizes are either anchored in the case of an anchored PAIC, and unanchored if not

Supplementary Table 8: **Type of association measures used (absolute or relative), according to outcome types and PAIC method used (primary outcome, or first one reported) (n=288 PAICs)**

PAIC type		Binary	Continuous	Time-to- event	Total
		n (%)	n (%)	n (%)	n (%)
MAIC					
	Missing	4	0	1	5
	Absolute difference	52 (54.2)	37 (74)	4 (3.8)	93 (36.9)
	Relative difference	44 (45.8)	13 (26)	102 (96.2)	159 (63.1)
STC					
	Absolute difference	0 (0)	2 (100)	2 (9.5)	4 (13.8)
	Relative difference	6 (100)	0 (0)	19 (90.5)	25 (86.2)
ML-NMR					
	Absolute difference	0	2 (100)	0	2 (100)
	Relative difference	0	0 (0)	0	0 (0)

PAIC: population-adjusted indirect comparison; MAIC: matching-adjusted indirect comparison; STC: simulated treatment comparison; ML-NMR: multilevel network meta-regression

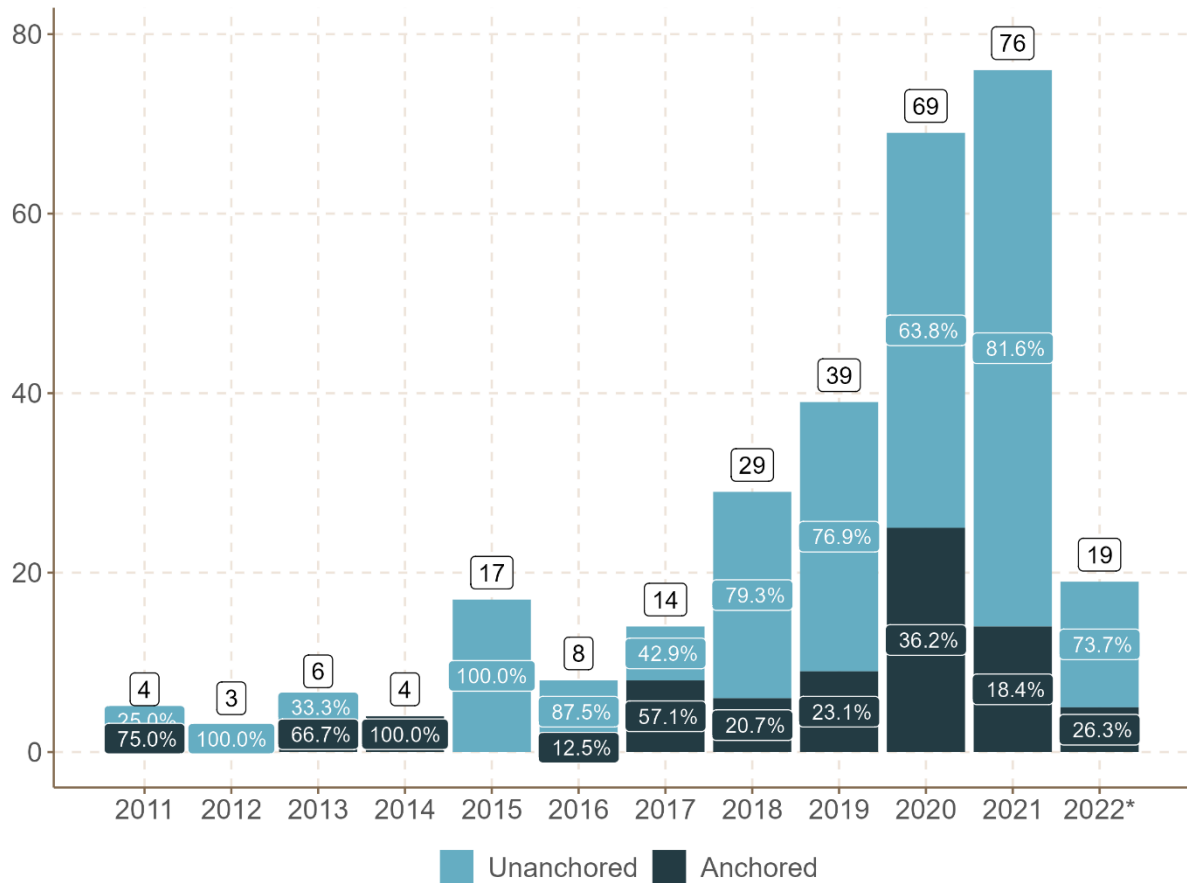
Supplementary Table 9: **Main findings of articles comparing the exact same pair of treatments, using the same source studies, but with IPD and aggregated data swapped.**

	Main findings reported in the article	Comments and main methodological differences
Anchored comparison of Dimethyl Fumarate (DMF) vs Fingolimod in Multiple Sclerosis		
<i>Nixon et al., 2014</i> 10.1007/s12325-014-0167-z	IPD: Fingolimod Fingolimod is superior to DMF in terms of No-Evidence of Disease Activity (NEDA)	More covariates taken into account in Fox et al. 2017 Different adjustment method used: STC for Nixon et al. 2014; MAIC for Fox et al. 2017
<i>Fox et al., 2017</i> 10.1080/03007995.2016.1248380	IPD: DMF Fingolimod and DMF are performing similarly in terms of efficacy (Relapse Rate, NEDA, etc.). Patients reported outcomes regarding quality of life are better for DMF as compared to fingolimod	
Anchored MAIC of Inotuzumab Ozogamicin (InO) VS Blinatumomab in Relapsed or Refractory Acute Lymphoblastic Leukemia (ALL)		
<i>Song et al., 2019</i> 10.1007/s12325-019-0873-7	IPD: Blinatumomab Blinatumomab superior to InO for means difference of Restricted Mean Survival Time (statistically significant), and Complete Remission differences (non-statistically significant)	Proskorovsky published MAIC/STC in 2019, in response to the MAIC published by Song et al in 2018 - Proskorovsky used a subset of the followup period, and included both Philadelphia positive and negative patients, and with <= 1 salvage therapies - Song included Philadelphia-negative patients only, and patients with possibly more than 1 salvage therapy - Different outcomes tested
<i>Proskorovsky et al., 2019</i> 10.1007/s12325-019-00991-w	IPD: InO Remission rate favors InO over Blinatumomab (statistically significant)	
Unanchored MAIC of BAY 94-9027 vs recombinant Factor VIII Fc (RFVIII Fc) in severe Hemophilia A		
<i>Batt et al., 2019</i> 10.2147/jbm.s206806	IPD: BAY 94-9027 Find similar Annualized Bleeding Rate (ABR) between compared treatments	Differences in the inclusion criteria of the compared populations
<i>Hakimi et al., 2021</i> 10.1007/s12325-020-01599-1	IPD: rFVIII Fc Find significantly lower ABR with Factor VIII Fc as compared to BAY	

Anchored MAIC of apalutamide VS darolutamide in prostate cancer		
<i>Halabi et al., 2021</i> 10.1097/ju.0000000000001767	IPD: darolutamide No statistically significant differences in metastatic-free survival, and darolutamide superior in terms of safety	Differences in the population inclusion criteria, in the set of covariates used, and in the statistical framework (frequentist versus Bayesian).
<i>Chowdhury et al., 2022</i> 10.1007/s12325-021-01885-6	IPD: apalutamide Apalutamide superior to darolutamide in metastatic-free survival, progression-free survival, and no statistically significant differences in adverse events	
MAIC of everolimus and sunitinib in pancreatic neuroendocrine tumor		
<i>Signorovitch et al., 2013</i> 10.1186/2162-3619-2-32	IPD: everolimus MAIC finds overall similar efficacy between everolimus and sunitinib in terms of OS and PFS (<i>i.e.</i> non-statistically significant), but point estimate favors of everolimus	Differences in point estimates direction are ascribed to updated survival data in Ishak et al. 2018, and a different set of covariates.
<i>Ishak et al., 2018</i> 10.2217/cer-2018-0020	IPD: sunitinib MAIC finds similar PFS and OS for sunitinib and everolimus, but point estimates favors sunitinib	
Unanchored indirect comparison of crizotinib versus entrectinib in non-small-cell lung cancer		
<i>Chu et al., 2020</i> 10.2217/cer-2020-0063	IPD: entrectinib MAIC finds statistically significant improved ORR with entrectinib	Differences purported to differences in the adjustment method used, and in longer follow-up in Tremblay et al. 2022
<i>Tremblay et al., 2022</i> 10.2217/fon-2021-1102	IPD: crizotinib STC finds no statistically significant difference in ORR between compared treatments	

IPD: individual patient data

Supplementary Figure



* For the year 2022, data collected up to April 2 only

Supplementary Figure 1: **Conducted PAIC annual count: anchored or unanchored**

PAIC: population-adjusted indirect comparison

3.3 Summary and comparative analysis of related reviews

3.3.1 Matching-Adjusted Indirect Comparison Studies in Oncology: A Scoping Review Focused on Reporting Quality. *Value in Health Regional Issues*. Farinasso CM, et al.

This review was published in 2025, focusing specifically on PAICs conducted in oncology [79]. It focused on MAIC exclusively, and emphasized the evaluation of MAIC quality according to NICE guidelines. Their review found 115 MAICs, with 72% of them being unanchored. An important feature is that 79 studies needed to reconstruct IPD, usually to be able to use an HR as a contrast measure. They also noted that on average, MAIC led to an average 44.9% sample size reduction. Finally, an interesting point they raised was that anchored and unanchored MAICs were likely to use expressions to convey positive results in their closing statements (43% of anchored MAICs and 31% of unanchored MAICs). Finally, they devised a list of NICE guidance criteria, and reached the conclusion that 3% of included MAICs met all the NICE guidance criteria, while 7% met at least 7 out of 8 recommendations.

3.3.2 Population adjusted-indirect comparisons in health technology assessment: A methodological systematic review. *Research Synthesis Methods*. Truong B, et al.

The overall goal of the review conducted by Truong et al. (2023) was very similar to ours, with the aim of assessing the use of PAIC as published in the literature, encompassing both MAICs and STCs, from the inception of these methods up to February 2023 [44]. They found 162 articles meeting the inclusion criteria, with MAIC being used as the adjustment method in 150 cases (92.6%), STC in 17 cases (10.5%), and ML-NMR in 1 case (6 articles used both MAIC and STC). Overall general results were consistent with Serret-Larmande et al. (2023), as they too noted an uptick in publication rate recently, a predominance of unanchored (64.8%) as compared to anchored studies (35.2%), of MAICs (92.6%) as compared to STC (10.5%)—some studies used both, and oncology/hematology as the main domain of interest of these comparative efficacy studies. Additionally, they noted that a formal assessment of heterogeneity between studies was carried out in 89.9% of the records, and that 44.5% of the studies first trimmed the IPD population to match inclusion criteria of the AgD one, before adjusting for confounders. Finally, they reported about the limitations acknowledged by the authors, with 84.0% of the studies mentioning the potential for unmeasured covariates, and

37.0% mentioning explicitly unmeasured covariates.

3.3.3 A Comparison of Relative-Efficacy Estimate(S) Derived From Both Matching-Adjusted Indirect Comparisons and Standard Anchored Indirect Treatment Comparisons: A Review of Matching-Adjusted Indirect Comparisons. *Value in Health*, Cassidy O., et al.

Also in 2023, Cassidy et al. (2023) focused on comparing the results of anchored unadjusted ITC (Bucher ITC) with anchored MAICs for the same pairwise ITCs (*i.e.* same pair of treatment compared, same data source, same definition of the outcome) [80]. By comparing the point estimates of Bucher ITCs to MAICs, and their confidence/credible intervals, it was intended as a way to quantify the difference in magnitude of adjusted estimates as compared to unadjusted ones. Eligible studies were selected from peer-reviewed publications (the authors browsed Embase, Medline, and the Cochrane Library databases) as well as the grey literature. The treatment effect estimated through the ITC could be said to either favor the treatment initially evaluated in the study with IPD available, or conversely favor the one evaluated in the study with only AgD available.

They found a total of 168 instances of the exact same pairwise treatment comparison. Among these 168 pairwise ITC, in 53 cases (31.5%) the point estimate of the Bucher ITC was more favorable to the IPD treatment (*i.e.* the treatment evaluated in the study for which IPD is available) as compared to the point estimate yielded by the MAIC, conversely the MAIC estimate was more favorable to the IPD treatment in 109 cases (64.8%); and point estimates were equal in 6 cases. Thus, the adjusted comparison turns out to favor the treatment coming from IPD more often than the Bucher ITC. Moreover, they found that among the 80 pairs of MAIC/unadjusted ITC reporting a confidence or credible interval for the unadjusted ITC, only six times the MAIC point estimates fell outside the boundaries of the reported unadjusted intervals, and every six times it was favoring the IPD treatment. These observations made them come to the conclusion that peer-reviewed MAICs exhibited a publication bias, namely that the adjusted analysis would be more likely to be reported if it favored the IPD treatment. Using a different approach, this study thus confirms the publication and reporting bias exhibited in our methodological review.

Chapter 4

The impact of limited access to IPD

PAICs became a popular tool thanks to their usefulness in a relatively frequent scenario in the realm of HTA, that is partially unavailable IPD. Consequently, the property that allows to perform cross-population adjustment without complete access to IPD raises a question: what are the compromises being made with such estimators? In a situation with fully-accessible IPD, one would naturally use IPD Network Meta-Analysis (NMA) and Propensity Score Weighting (PSW) in place of STC and MAIC respectively. It follows that the use of PAIC estimators should entail drawbacks. One of the most important constraints is the restriction of the target population of interest, which is bound to be the AgD (if one does not make some key additional assumptions about transportability of effects across populations). This specific issue has been described in the second chapter of this thesis, and has been discussed at large in the literature [24, 23]. Additionally, the smaller sample size may lead to challenges when adjusting for confounders. Outcome or trial assignment regression model will be fitted with, on average, twice as little observations. This bears direct consequences regarding the precision of estimated coefficients and potential limitations in the number of covariates included in the models. Moreover, a limited sample size may additionally restrict the ability to select confounders based on their statistical association with the outcome, irrespective of the appropriateness of this approach). A handful of studies studied the direct consequence of swapping a full-IPD estimator for one partly relying on AgD. Phillippo et al. (2020) compared performance of ML-NMR, STC, MAIC in a large range of scenarios, varying covariates distributions, adequate adjustment model specification, cross-study covariate distributions overlap, covariates correlation structures, shared effect modification, sample size and strength of effect modification [52]. Several key results from the study included the greater sensitivity of MAIC to small sample size and lack of covariate overlap across studies, as compared to ML-NMR and STC, and showed that MAIC yielded overall consistently larger standard error estimates as compared to the regression based models. A discussion ensued

on the different estimands targeted by MAIC, STC and ML-NMR (namely marginal versus conditional), and the amount of bias that could be due to a discrepancy between the numerical values of the different estimands [24, 81, 25]. This discussion also pointed out that ML-NMR had the ability to resolve shortcomings of STC and MAIC, namely because they allow to estimate an effect in any population of interest, provided assumptions about transportability of effect-modification hold. Nonetheless, ML-NMR requires a network of comparisons to be able to leverage additional information to incorporate into pairwise comparison. Other works focused on comparing the benefit from IPD over AgD in large network of treatments [82, 83]. Nonetheless, in the absence of such larger network, MAIC and STC remain tools of choice to conduct cross-populations adjustments in pairwise comparison settings.

Several empirical comparisons of MAIC to a weighting model that use full-IPD to estimate the treatment effect have been carried out. Wong et al. (2023) used alternatively anchored PSW and MAIC with data from the GEMINI-1 and ULTRA-1 RCTs in order to indirectly compare vedolizumab and adalimumab in the treatment of inflammatory bowel disease [84]. Park et al. (2024) conducted a case study on the treatment of third-line small cell lung cancer (comparing nivolumab to standard-of-care) to compare results obtained with PSW, regression adjustment, and a doubly robust method, to respectively the ones coming from MAIC, STC and a doubly robust PAIC estimator combining MAIC and STC [18]. In this specific empirical evaluation, estimates yielded by the different methods tested were very close, ranging from 0.63 (95% CI [0.44-0.90]) with the unadjusted estimates, to 0.69 (95% CI [0.44-0.98]) using the full-IPD outcome regression model, in favor of nivolumab. Regression-based and doubly robust estimates yielded larger confidence intervals as compared to weighting-based estimators.

A large body of literature has explored different aspects and properties of MAIC and STC, studying their respective performance in different situations, but it seems that none of them has been focused explicitly on comparing a PAIC estimator to its IPD equivalent. Since the lack of IPD is most often a pragmatic limitation, typically related to confidentiality, such findings could potentially contribute to calls for broader access to full-IPD in analyses conducted within regulatory and HTA settings.

Therefore, we aimed to compare PAIC with their full-IPD counterparts in terms of accuracy and precision of the estimation. The objective was to evaluate these estimators under a variety of scenarios that may represent the challenges of real-world implementation, with for instance non-normal covariate distributions. The question of robustness of generalized linear model to non-normally distributed data has been studied extensively [85, 86, 87], as well as the presence of non-normally distributed data in real-life datasets [88, 89, 90].

However, as of today, one of the unanswered questions in the literature remains: how do these estimators behave under less-than ideal conditions? While robustness of PAIC regarding non-normality of covariate distribution could be extrapolated from knowledge of generalized linear models, no simulation study has been conducted as an attempt to quantify this aspect. Their performance in the presence of unobserved covariates or non-normal covariate distributions, relative to their IPD counterparts, also remains uncertain. Depending on the results, this can potentially shed light for regulatory agencies to push for more liberal data sharing policies, or to try to find solutions to obstacles to data sharing.

The article that follows has been planned with the idea of studying both weighted estimators and regression-based estimators [22]. However, in the final version of the article, the focus has been narrowed down to weighting estimators only, for two main reasons: first, the wealth of results and cross-comparisons made the distillation of the results in an intelligible picture challenging. Indeed, estimators' performance have been compared across eight data generating mechanisms, presence of hidden confounders, with their anchored and unanchored versions. Second, the most frequent implementation of STC, that is the "plug-in" version (see STC section, chapter 2), is extremely close to its IPD counterpart, only fitted on half of the data. Consequently, their respective performance is extremely similar across the different scenarios and comparisons that have been evaluated. Therefore, the simulation results for STC and linear models have been omitted from this work.

Published article II (pre-print)

The second published article of this thesis is Serret-Larmande et al., 2025 (posted as preprint on the arXiv platform). It is reproduced here in its original form. It is currently under review in *Statistics in Medicine* at the time of writing of this thesis.

ASSESSING THE IMPACT OF COVARIATE DISTRIBUTION AND POSITIVITY VIOLATION ON WEIGHTING-BASED INDIRECT COMPARISONS: A SIMULATION STUDY

Arnaud Serret-Larmande

Sorbonne University, INSERM,
Pierre Louis Epidemiology and Public Health Institute
AP-HP, Pitié-Salpêtrière Hospital, Public Health Department
PSL-CFX Clinical Research Unit, CIC-1901
Paris, France
arnaud.serretlarmande@gmail.com

Jérôme Lambert

ECSTRRA Team IRSL
INSERM UMR 1342, Université Paris Cité
Hôpital Saint-Louis
Paris, France
jerome.lambert@aphp.fr

Stéphane Gaudry

AP-HP, Hôpital Avicenne,
Service de Réanimation Médico-Chirurgicale
INSERM, UMR-S 1155 CORAKID
Paris, France
stephane.gaudry@aphp.fr

David Hajage

Sorbonne University, INSERM,
Pierre Louis Epidemiology and Public Health Institute
AP-HP, Pitié-Salpêtrière Hospital, Public Health Department
PSL-CFX Clinical Research Unit, CIC-1901
Paris, France
david.hajage@aphp.fr

July 22, 2025

ABSTRACT

Population-Adjusted Indirect Comparisons (PAICs) are used to estimate treatment effects when direct comparisons are infeasible and individual patient data (IPD) are only available for one trial. Among PAIC methods, Matching-Adjusted Indirect Comparison (MAIC) is the most widely used. However, little is known about how MAIC performs under challenging conditions such as limited covariate overlap or markedly non-normal covariate distributions.

We conducted a Monte Carlo simulation study comparing three estimators: (i) MAIC matching first moment (MAIC-1), (ii) MAIC matching first and second moments (MAIC-2), and (iii) a benchmark method leveraging full IPD—Propensity Score Weighting (PSW). We examined eight scenarios ranging from ideal conditions to situations with positivity violations and non-normal (including bimodal) covariate distributions. We assessed both anchored and unanchored estimators and examined the impact of adjustment model misspecification. We also applied these estimators to real-world data from the AKIKI and AKIKI-2 trials, comparing renal replacement therapy strategies in critically ill patients.

MAIC-1 demonstrated robust performance, remaining unbiased in the presence of moderate positivity violations and non-normal covariates, while MAIC-2 and PSW appeared more sensitive to positivity violations. All methods showed substantial bias when key confounders were omitted, emphasizing the importance of correct model specification. In real-world data, a consistent trend was found with MAIC-1 showing narrower confidence intervals with positivity violation.

Our findings support the cautious use of unanchored MAICs and highlight MAIC-1's resilience across moderate violations of assumptions. However, the method's limited flexibility underscores the need for careful use in real-world settings.

Keywords Indirect Treatment Comparisons, Matching-Adjusted Indirect Comparisons, Monte Carlo Simulations, Propensity Score Weighting, Transportability

1 Introduction

When estimating the relative effect of two treatments of interest, the gold standard is typically a randomized controlled trial (RCT). However, drawing evidence from an RCT is not always feasible due to various constraints, including circumstances that justify the use of a single-arm trial, proliferation of available treatment options, as well as the need to generate evidence rapidly in contexts of emerging therapies. Indirect Treatment Comparisons can be used to estimate relative effects of treatments that have been evaluated across different studies. They are used heavily in the context of Health Technology Assessment as substitutes when direct evidence is unavailable [14] [11] [26]. Since such comparisons rely on outcomes observed across distinct populations, they are inherently subject to confounding bias. One approach to mitigate this bias is to use a common treatment arm evaluated across two RCTs (Bucher Indirect Treatment Comparisons) [5]. This shared treatment arm is referred to as an anchor, and the resulting analyses are known as anchored comparisons. While this method cancels out the bias related to differences in the distribution of prognostic factors across trials, the estimated treatment effect remains biased in the presence of treatment-effect modifiers. Moreover, it requires by definition the presence of a common treatment arm. Consequently, an anchored comparison is impossible when the studies lack a common treatment arm—which will always be the case when a single-arm study is used in the comparison, an increasingly frequent situation [4] [1] [13].

Adjusted estimators are thus necessary to control for confounding and obtain an unbiased treatment effect. Such adjustment can be carried out using a balancing score (*e.g.* propensity score) or outcome regression models. Usually, these estimators would require having access to complete Individual Patient Data (IPD).

However, in the context of Health Technology Assessment, sponsors have access to IPD for the treatment that they are developing, but often they can only access published aggregated data for treatments developed by competitors. This explains the increasing use of Population-Adjusted Indirect Comparisons (PAIC), which are methodologies designed to estimate adjusted indirect treatment effects in pairwise comparisons when IPD is not available for all the studies informing the comparison. Two main classes of estimators are used in this context: Simulated Treatment Comparison and Matching-Adjusted Indirect Comparison (MAIC). Simulated Treatment Comparisons are outcome regression-based models estimating average conditional treatment effect estimates [7], while MAICs use a propensity-score-like reweighting approach to estimate marginal effects [24]. More recent methodological advancements such as Multilevel Network Meta-Regression allow estimating treatment effects in different target populations [17]. A large body of methodological literature has explored the properties and limitations of PAICs [18] [16] [20], and highlighted their growing practical application in that context ([25], [22] [8]). Thanks to their abilities to adjust for confounding, PAICs can be extended to situations where a common treatment arm is missing. While estimates yielded by the anchored form are deemed much more reliable [18], unanchored PAICs have been particularly popular due to their theoretical ability to derive treatment effect estimates even in situations where a common comparator arm is missing.

This work focuses exclusively on MAIC, as it remains the most widely used PAIC estimator in the literature and in applied settings [22], [25]. Briefly, MAIC estimates balancing weights using the method of moments and then reweights the Individual Patient Data (IPD) sample in a way that ensures that the weighted covariate moments (typically first, and sometimes second-order moment) match those of the Aggregated Data (AgD) trial. Relative treatment effect is subsequently estimated by comparing the weighted outcome from the IPD trial to the observed outcome in the AgD trial.

Nonetheless, despite MAICs appealing simplicity, the requirement to respect usual causal inference assumptions remains unchanged, that is exchangeability, consistency, and positivity [10]. Moreover, the lack of fully accessible individual data imposes important limitations. First, the target estimand is restricted by design to be the treatment effect in the AgD population. Second, the estimator itself may be constrained by the moments-matching approach (more details in the Methods section). Theoretical properties of estimators based on the method of moments may not hold when covariate distributions are bimodal or with heavy-tail distributions [28] [2]. Additionally, the logistic regression model used for weight estimation is fit on a subset of the data of interest (IPD trial), limiting statistical efficiency.

These constraints raise important concerns about the robustness of MAIC when its theoretical assumptions are violated. In such settings, a natural question arises: would access to full IPD from both trials yield a more reliable estimate as compared to partially AgD?

To date, however, no study has been conducted to compare MAICs to its full-IPD counterpart, *i.e.* Propensity Score Weighting (PSW) with an estimand targeting the AgD trial population, that is using average treatment effect in the controls weights [3].

This study addresses this gap by comparing the performance of MAIC and PSW across a range of scenarios using Monte Carlo simulations. Additionally, the relevance of these methods are assessed using empirical data from the AKIKI

and AKIKI-2 clinical trials, which compared strategies for initiating renal replacement therapy in critically ill patients with acute kidney injury.

2 Methods

2.1 Notation and Identification

The goal is to estimate the relative marginal effect of treatment A versus treatment B , noted Δ , as measured by an outcome Y .

These treatments have been evaluated in two separate clinical trials $T \in \{a, b\}$. Both a and b trials also evaluated a common treatment arm C . The treatment received is noted Z , with $Z_a \in \{A, C\}$, and $Z_b \in \{B, C\}$ in trials a and b respectively; as an example, an observation of Y under a treatment $Z = A$ in trial $T = a$ will be written $y_a(Z = A)$, simplified as $y_a(A)$ thereafter.

The population of interest is the source population of the b clinical trial. Therefore the target estimand for the indirect comparison is the treatment effect in the b population: Δ_b . It is known in the propensity score literature as the Average Treatment effect in the Control population (ATC) [29]. It entails specific implications for the positivity assumption by introducing an asymmetry in the required support for the baseline characteristics one wishes to balance across studies: under this weighting scheme, the positivity assumption holds as long as the covariate distribution in the b target trial is fully represented within the a index trial. Conversely the a trial covariate distribution may not need to be represented in the b trial. This contrasts with the Average Treatment Effect, which requires complete overlap in the covariate distribution between both trials.

In the unanchored case, an unbiased estimate of Δ_b is given by $\hat{\Delta}_b = g(\bar{y}_b(A)) - g(\bar{y}_b(B))$, with $\bar{y}_b(A)$ and $\bar{y}_b(B)$ being the sample mean of Y under respectively treatment A and B in the trial b , and g being an invertible link function that ensures additivity of outcomes. For instance, if Y is normally distributed, g would be the identity function ($g(x) = x$), and if Y is binary, g could be the identity function (estimating a difference of proportions), log function ($g(x) = \log(x)$) when estimating a log relative risk, or logit function ($g(x) = \log(\frac{x}{1-x})$) when estimating a log odds ratio. Because these Monte Carlo simulations' target estimand is the mean difference of continuous variables, we use the identity link and drop $g(\cdot)$ in what follows.

The estimation problem arises from the fact that the quantity $\bar{y}_b(A)$ is unobserved, that is treatment A has not been evaluated in trial b . Using the sample mean of Y under treatment A observed in a trial a , *i.e.* $\bar{y}_a(A)$, as a substitute for $\bar{y}_b(A)$ will yield a biased estimate of Δ_b if there exist prognostic factors (PF) in imbalance between a and b studies, which is most certainly the case when considering two separate clinical trials. Some of these PF may additionally influence the outcome differently depending on the treatment received, defining treatment-effect modifiers (TEM). Therefore, in practice, an unbiased estimation of Δ_b requires to estimate $\hat{\bar{y}}_b(A)$, that is to get an estimation of the average value of the outcome that would have been observed under treatment A in the b trial, *i.e.* in a sample with distributions of PF and TEM representative of the b source population. $\hat{\bar{y}}_b(A)$ is unbiased (and subsequently so is $\hat{\Delta}_b$) if the estimation process adequately takes into account the distribution of the PF and the TEM (that is all the confounders of the association between the treatment and the outcome) in the b trial population.

This requirement can be relaxed when using the common treatment arm C , which plays the role of an anchor. In such a case, the estimation of Δ_b can be obtained through a difference-in-difference between the outcomes observed under each treatment across the two trials a and b , that is $\hat{\Delta}_b = (\bar{y}_b(A) - \bar{y}_b(C)) - (\bar{y}_b(B) - \bar{y}_b(C))$. The unobserved quantity $(\bar{y}_b(A) - \bar{y}_b(C))$, that is the relative difference of the outcome Y under treatments A versus C in the b population, can be substituted by $(\hat{\bar{y}}_b(A) - \hat{\bar{y}}_b(C))$, that is its estimation would have treatment A had been evaluated in trial b . The quantity $(\hat{\bar{y}}_b(A) - \hat{\bar{y}}_b(C))$ is an unbiased estimate of $(\bar{y}_b(A) - \bar{y}_b(C))$ (and subsequently $\hat{\Delta}_b$ of Δ_b) if the estimation process adequately takes into account the distribution of the TEM—and not necessarily of the PF, in the b trial population.

2.2 Estimators

Three different treatment effect estimators will be used to estimate Δ_b . We define the unweighted estimator, the MAIC estimator, and the PSW estimator targeting the marginal effect in the trial b population. The goal of each estimator is to provide an estimate of Δ_b despite the lack of direct observations of $\bar{y}_b(A)$. Each estimator is assessed under both anchored and unanchored specifications.

2.2.1 Unweighted Estimator

The unweighted approach ignores cross-trial population differences and uses trial a 's observed outcome under a given treatment as a proxy for the outcome that would have been observed in trial b under that given treatment. This yields the unanchored case: $\hat{\Delta}_b = \bar{y}_a(A) - \bar{y}_b(B)$, and the anchored case: $\hat{\Delta}_b = (\bar{y}_a(A) - \bar{y}_a(C)) - (\bar{y}_b(B) - \bar{y}_b(C))$.

2.2.2 PSW Estimator

PSW reweights each subject i from trial a with a weight w_i so that the distribution of their baseline characteristics mirrors that of trial b . Each weight w_i is equal to the odds that individual i be included in trial b rather than in trial a , given their covariates. Individuals from trial b are assigned a weight of 1. If $e(x) = \Pr(T = b \mid X = x)$ is the probability of being in trial b , then for an individual with a vector of covariates X_i ,

$$w_i = \frac{\Pr(T_i = b \mid X_i)}{\Pr(T_i = a \mid X_i)} = \frac{e(X_i)}{1 - e(X_i)} \quad (1)$$

In practice, the estimate of $\frac{e(X_i)}{1 - e(X_i)}$ corresponds to the exponential of the linear predictor from a logistic regression model including covariates X_i , which models the probability of being included in trial b . The model is fitted by maximum likelihood using IPD from both trials.

Finally, we estimate the outcome under treatment z in the b trial population with:

$$\hat{y}_b(z) = \frac{\sum_i w_i y_i \mathbb{I}(Z_i = z)}{\sum_i w_i \mathbb{I}(Z_i = z)} \quad (2)$$

The treatment effect $\hat{\Delta}_b$ is then estimated by replacing the estimated quantities from the b trial with the estimated ones, as detailed in the Notations and Identification section.

2.2.3 MAIC Estimators

MAIC estimator targets the same estimand as PSW, that is the Average Treatment effect in the Control population. However, unlike PSW, it relies on moment-matching rather than full-distribution balancing to yield this treatment effect estimate. Each subject i from trial a is given a weight w_i such that the moments of the weighted distribution of baseline covariates X in trial a matches the reported baseline distribution moments in trial b . Formally, w_i is defined to be the odds that individual i be included in trial b rather than in trial a , given X_i . Keeping $e(x)$ definition we get:

$$\tilde{w}_i = \frac{\Pr(T_i = b \mid h(X_i))}{\Pr(T_i = a \mid h(X_i))} = \frac{e(X_i)}{1 - e(X_i)} \quad (3)$$

with h a function yielding the vector of moments for a vector of covariates X [23]. In practice, $\tilde{w}_i$ is obtained by solving method of moments equations that equate weighted covariate moments in trial a to the covariate moments reported for trial b . Let $h(X_b)$ be the set of relevant baseline summary measures available in trial b (e.g. means and eventually higher moments), MAIC estimates $\tilde{w}_i$ such that the weighted vector of baseline covariates X_a in trial a equates the vector of moments of baseline covariates X_b in trial b :

$$h(\tilde{w}_i X_{i,a}) = h(X_b) \quad (4)$$

In practice, this is done by finding the value of the vector α that minimizes the following sum: $\sum_i^n \exp(\alpha^T (X_i - h(X_b)))$. Two configurations were tested for the vector of moments $h(X_b)$: (i) first-order moments only and (ii) both first- and second-order moments. Including only the first moments corresponds, in practice, to matching covariate means, while including both the first and second moments corresponds to matching both means and variances. We consider these two settings to represent two different estimators, and we will refer to them as respectively MAIC-1 and MAIC-2 thereafter.

The estimated outcome under treatment $Z = z$ is then formed using Equation 2, and relative treatment effect Δ_b is estimated by substituting unobserved with estimated quantities.

2.2.4 Variance of the treatment effect

The computation of the variance of the treatment effect is identical for the three different estimators and is the sum of the variances of the outcome under each treatment used in the indirect comparison. Variance of the outcome in trial b under treatment z (weighted or unweighted) has been estimated with 2,000 non-parametric bootstrap iterations.

2.3 Data Generation Mechanisms (DGMs)

Following the previously introduced notation, Figure 1 depicts the causal Directed Acyclic Graph (DAG) of the true relationships between the variables used in our simulations: baseline covariates X_1 and X_2 , trial assignment T , treatment Z , and outcome Y .

From this DAG, we defined eight different DGMs that differ by the distributions of the covariates X_1 and X_2 and by the b trial assignment model. In every DGMs, X_1 is a PF and a TEM, and X_2 is a PF only—it does not influence the treatment effect. In every scenario, the outcome model will be:

$$Y = X_1 + 2X_2 + (1 + 2X_1)I(Z = A) + I(Z = B) + \epsilon$$

with $\epsilon \sim N(0, 1)$. The treatment Z is randomly allocated in a 1:1 ratio within each trial, with the two treatments arms being A and C in the a trial—*i.e.* $\mathbb{P}(Z = A|T = a) = \mathbb{P}(Z = C|T = a) = 0.5$, and B and C in the b trial—*i.e.* $\mathbb{P}(Z = B|T = b) = \mathbb{P}(Z = C|T = b) = 0.5$. Within each DGM, X_1 and X_2 are independent and identically distributed. The eight scenarios examined can be grouped into four pairs. Each pair is distinguished by X_1 and X_2 distributions and their effect on the trial assignment probability. Within each pair, one scenario presents positivity issues and the other not, as detailed below. Consistent with the explanation given in the Notation and Identification section, violation of the positivity assumption is defined as a lack of statistical support in trial a for the covariate distribution observed in trial b —which represents the target population. Below, we summarize the characteristics of the different situations explored by each scenario; additionally, the actual distribution of the covariate in trials a and b for each DGM is depicted in Figure 2.

DGM-1 and DGM-2 specifications

The first (DGM-1) and second (DGM-2) scenarios feature two covariates X_1 and X_2 with standard normal distributions. DGM-1 serves as a control case with no violation of positivity, meaning complete support of the target trial b in the source trial a . Conversely, DGM-2 introduces positivity violations due to a lack of statistical support, meaning part of X_1 and X_2 distributions in trial b are not represented within the trial a . In the following equations, the parameter δ controls the direction of the trial assignment mechanism. A value of $\delta = -1$ reflects scenarios with limited overlap across trials, mimicking positivity violations. $\delta = 1$ in DGM-1, and $\delta = -1$ in DGM-2.

$$\begin{cases} X_j & \sim N(0, 1), \quad j = \{1, 2\} \\ T & \sim \text{Bernoulli}(\pi_T) \text{ where } \pi_T = P(T = b | X_1, X_2) = g(\delta [\beta_0 + X_1 - 0.5X_1^2 + X_2 - 0.5X_2^2]) \end{cases} \quad (5)$$

For all the DGMs explored, $g(x) = (1 + \exp(-x))^{-1}$, and the constant β_0 is calibrated so that the marginal probability of being assigned to trial b is 0.5 in the overall population. Its numerical value is obtained using a root-finding algorithm.

DGM-3 and DGM-4 specifications To explore the impact of non-normally distributed covariate, scenarios DGM-3 and DGM-4 use right-truncated log-normal distributions. In these scenarios, $\delta = -1$ in DGM-3, and $\delta = 1$ in DGM-4, that is DGM-4 presents positivity issues.

$$\begin{cases} X_j & = \min(X'_j, 5) \text{ with } X'_j \sim \text{LogNormal}(0, 0.5), \quad j = \{1, 2\} \\ T & \sim \text{Bernoulli}(\pi_T) \text{ where } \pi_T = P(T = b | X_1, X_2) = g(\delta [2\beta_0 + 2X_1 + 2X_2]) \end{cases} \quad (6)$$

DGM-5, DGM-6, DGM-7 and DGM-8 specifications To examine the impact of bimodal covariate, scenarios five to eight generate two covariates with bimodal distributions. Scenarios 7 and 8 present more extreme bimodal distributions as compared to the fifth and sixth. DGM-5 and DGM-6 follow distributions laid out in Equation 7, and DGM-7 and DGM-8 follow the ones in Equation 8. $\delta = 1$ in DGM-5 and DGM-7, and $\delta = -1$ in DGM-6 and DGM-8, that is DGM-6 and DGM-8 present positivity issues.

$$\begin{cases} X_j & \sim \begin{cases} \mathcal{N}(0, 0.5) & \text{with probability } 0.5 \\ \mathcal{N}(3, 0.5) & \text{with probability } 0.5 \end{cases} \quad j = \{1, 2\} \\ T & \sim \text{Bernoulli}(\pi_T) \text{ where } \pi_T = P(T = b | X_1, X_2) = g(\delta [0.5\beta_0 + 0.5X_1 - 0.5X_1^2 + 0.5X_2 - 0.5X_2^2]) \end{cases} \quad (7)$$

$$\begin{cases} X_j & \sim \begin{cases} \mathcal{N}(0, 1) & \text{with probability } 0.5 \\ \mathcal{N}(3, 1) & \text{with probability } 0.5 \end{cases} & j = \{1, 2\} \\ T & \sim \text{Bernoulli}(\pi_T) \text{ where } \pi_T = P(T = b \mid X_1, X_2) = g\left(\delta \left[0.5\beta_0 + 0.5X_1 - 0.5X_1^2 + 0.5X_2 - 0.5X_2^2\right]\right) \end{cases} \quad (8)$$

2.3.1 Summary of scenarios

A summary of the different DGMs with regards to covariate distribution, positivity issues and theoretical treatment effect value is given in Table 1. The theoretical Δ_b value is the marginal treatment effect calculated using the real outcome model in a large population of 2×10^6 subjects.

	Covariate distribution	Positivity violation?	Theoretical treatment effect Δ_b
DGM-1	Normal distribution	No	$\Delta_b = 0.68$
DGM-2	Normal distribution	Yes	$\Delta_b = -0.55$
DGM-3	Lognormal distribution	No	$\Delta_b = 1.17$
DGM-4	Lognormal distribution	Yes	$\Delta_b = 3.53$
DGM-5	Bimodal distribution	No	$\Delta_b = 1.52$
DGM-6	Bimodal distribution	Yes	$\Delta_b = 2.27$
DGM-7	Bimodal distribution	No	$\Delta_b = 0.1$
DGM-8	Bimodal distribution	Yes	$\Delta_b = 3.52$

Table 1: Summary of the different data-generating mechanisms (DGMs) and corresponding theoretical treatment effect

2.4 Monte Carlo simulations settings

For each of the eight specified DGMs, 2,000 Monte Carlo iterations were performed. In each iteration, two trials of 250 patients per arm have been sampled from a super-population of 2×10^6 subjects, with a 1:1 randomization ratio (*i.e.* 500 patients per trial in the unanchored setting and 1,000 per trial in the anchored one).

Additionally, in each setting, three sets of variables were used to adjust for: $\{X_1\}$, $\{X_2\}$, or $\{X_1; X_2\}$. These different sets correspond, respectively, to adjusting for treatment effect modifiers, prognostic factors, or both. Consequently, and depending on the treatment effect estimator used, this results in scenarios where the trial-assignment model is either correctly specified or misspecified.

2.5 Performance metrics

The objective of the simulation study is to evaluate the bias and precision of each estimator of Δ_b . The estimator's performance in this regard will be evaluated using the following metrics:

2.5.1 Bias

$$\text{Bias} = \sum_{j=1}^{2000} \frac{\hat{\Delta}_{b,j} - \Delta_b}{2000} \quad (9)$$

The bias of an estimator is the difference between the expected value of $\hat{\Delta}_b$ and the true value of the parameter it is estimating Δ_b .

2.5.2 Root Mean Square Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{2000} \sum_{j=1}^{2000} (\hat{\Delta}_{b,j} - \Delta_b)^2} \quad (10)$$

The Root Mean Square Error (RMSE) is the square root of the average of the squared differences between the predicted and true values, and reflects both the bias and variability of the estimator.

2.5.3 Variability Ratio (VR)

$$VR = \frac{\sum_{j=1}^{2000} \frac{\sigma_{\Delta_b, j}}{2000}}{\sqrt{\sum_{j=1}^{2000} \frac{(\hat{\Delta}_{b, j} - \Delta_b)^2}{1999}}} \quad (11)$$

The variability ratio compares the average model-based estimation of the standard error σ_{Δ_b} to the empirical standard error, indicating potential under- or overestimation of uncertainty.

2.5.4 95% coverage rate ($CR_{95\%}$)

$$CR_{95\%} = \frac{1}{2000} \sum_{i=1}^{2000} \mathbb{I} \left(\Delta_b \in [\hat{\Delta}_{b, j} - 1.96\hat{\sigma}_{\Delta_b, j}, \hat{\Delta}_{b, j} + 1.96\hat{\sigma}_{\Delta_b, j}] \right) \quad (12)$$

The coverage rate is the proportion of confidence intervals that contain the true parameter value (and should be equal to 0.95 in the case of the 95% coverage rate), combining adequate estimation of the parameter and adequate estimation of uncertainty.

2.5.5 Post-weighting covariate distribution overlap

A graphical comparison of the unweighted and weighted covariate distributions between trials a and b was conducted to assess how the different estimators affected covariate balance, particularly with respect to the positivity assumption.

2.5.6 Software

The simulations were developed using the R programming language, version 4.5.1.

3 Results

3.1 Treatment effect estimations under different DGMs, assuming correct model specification

Results presented in this section use the correct trial assignment model—that is anchored estimators are adjusting on treatment-effect modifiers only (X_1), and unanchored estimators are adjusting on both prognostic factors and treatment-effect modifiers (X_1 and X_2).

Under DGM-1 (no overlap issue and covariate normally distributed), all adjusted estimators were unbiased (Figure 3). As expected, estimates yielded by unanchored unweighted estimators were significantly more biased than the ones from anchored unweighted ones. All adjusted estimators exhibited close to nominal coverage rates, indicating well-calibrated confidence intervals. Finally, in this situation, unanchored showed lower RMSE, that is, more precise estimates. VR appeared closer to 1 with anchored as compared to unanchored estimators, hinting at a better variance estimation, even though the VR variation was small across the different estimators.

Under DGM-2, the target population includes regions of the covariate space that are poorly supported in source trial a , leading to violation of the positivity assumption (Figure 3). In this situation, MAIC-1 and MAIC-2 produced unbiased estimates of the treatment effect, indicating robustness to moderate positivity assumption violations. RMSE was smaller for unanchored as compared to anchored MAICs. In contrast, PSW remained biased despite using the correct trial model, indicating larger sensitivity to violations of positivity assumptions. In such a case, RMSE was smaller with anchored as compared to unanchored. This suggests that anchoring on a common comparator arm helps mitigate the effects of limited common support, since it requires adjusting for TEM only (*i.e.* only X_1). Regarding variance and coverage, nominal coverage rate was achieved with both unanchored MAIC estimators, and slight undercoverage with the anchored versions. PSW, affected by residual bias, failed to achieve acceptable coverage, especially with the unanchored version.

Under DGM-3, which featured a lognormal distribution without positivity violations, estimators' performance was similar across all metrics (Figure 4). Notably, bias was constantly negligible, including for MAIC-1 and MAIC-2. Coverage was near nominal for anchored estimators, but unanchored PSW produced overly narrow confidence intervals, resulting in undercoverage.

Under DGM-4, which featured a lognormal distribution with positivity violation, estimator performance varied (Figure 4). MAIC-1 and MAIC-2 yielded unbiased point estimates, while PSW retained some residual bias. Unanchored MAIC-

1 demonstrated the highest precision. However, most estimators underestimated variability, leading to sub-nominal coverage rates for all estimators but unanchored MAIC-2 which overestimated variability, resulting in over-coverage.

Under DGM-5 and DGM-7 (with bimodal covariate distributions and no positivity violations), estimator performance was largely consistent with observations from DGM-1 (Figures 5 6). All estimators produced unbiased treatment effect estimates. Variability ratios remained close to unity across estimators, supporting well-calibrated standard errors and resulting in coverage rates near the nominal 95% level. Notably, the bimodal nature of covariates did not meaningfully degrade the performance of the MAIC estimators, especially MAIC-1 which yielded the most precise treatment effect estimations in terms of bias and coverage rate.

Under DGM-6 and DGM-8, which combines bimodal covariate distributions with a lack of covariate overlap, MAIC-1 and MAIC-2 showed the lowest bias and RMSE, outperforming PSW, which suffered from higher variability and bias (Figures 5 and 6). In terms of coverage, only MAIC-1 (both anchored and unanchored) approached nominal levels without bias, while all other estimators demonstrated under or overcoverage. This suggests that MAIC-1 happens to be robust in scenarios with non-normal covariate distributions and non-trivial positivity violation, whereas weighting strategies that aim at balance the full covariate distribution are more sensitive under limited covariate overlap.

Additionally it is worth noting that in these scenario with bimodal covariate, the method of moments adjusting for first and second-order moments (MAIC-2) failed to converge properly or yielded absurd estimates on some instances due to excessively low effective sample sizes after weighting: DGM-5: 1/2000 (unanchored), DGM-6: 73/2000 (anchored) and 294/2000 (unanchored), DGM-7: 1/2000 (unanchored), and DGM-8: 4/2000 (unanchored).

3.2 Comparisons between before and after weighting covariate distributions

An analysis of covariate distributions before and after weighting with anchored estimators revealed key differences across scenarios and estimators (Figure 7). Anchored results are presented in the main text because of the easier interpretation (adjusting for X_1 only). A corresponding figure showing weighting distributions for unanchored estimators is provided in Figure S1 S1. This figure displays the density distribution of the variable X_1 in trials a and b , stratified by estimator. A second Y-axis represents the individual propensity score within the a trial as a function of X_1 . For each DGM, these plots aggregate observations coming from a sample of 100 iterations.

In DGM-1, all three weighting estimators effectively reweight the covariate distributions, though their post-weighting distributions differ: MAIC-2 and PSW produce distributions that closely match the target distribution trial b , while MAIC-1 only aligns the covariate means, consistent with its theoretical design. A similar pattern is observed in DGM-5 and DGM-7: PSW and MAIC-2 again replicate the X_1 distribution of trial b post-weighting. Notably, MAIC-2 achieves this even with a bimodal distribution. As in DGM-1, the MAIC-1 post-weighting distribution of X_1 are not perfectly matching, but the means are equivalent.

In scenarios with positivity issue (*i.e.* DGM-2, DGM-4, DGM-6, and DGM-8) limited statistical support for the target population in trial a led to extreme propensity scores estimates under PSW, and to a lesser extent under MAIC-2 (*i.e.*, propensity scores close to 1, resulting in very large weights), reflecting their intrinsic tendency to tightly match the target distribution. In contrast, MAIC-1 produces fewer extreme scores, yielding more stable weights. This observation may explain the results regarding the relatively less biased treatment effect estimates yielded by MAIC-1 as compared to PSW and MAIC-2.

3.3 Sensitivity to incorrect adjustment model specification

Figure 8 shows the bias of different estimators depending on whether treatment effect estimation models were adequately specified—that is, whether the trial assignment model included all relevant confounders. The Y-axis indicates the covariates included in the weighting model: either X_1 , X_2 , or both X_1 and X_2 . In order to take into account all the relevant confounders, the anchored estimators should include X_1 (TEM), and the unanchored ones should include X_1 and X_2 (TEM and PF). Only MAIC-1 and PSW estimators are reported here in order to avoid overplotting, as the results for MAIC-2 were intermediate between these two estimators.

Bias due to model (mis-)specification was notably greater than that attributable to the choice of estimators. When key confounders were omitted, all estimators displayed substantial bias regardless of the weighting approach. Conversely, when all confounders were included, bias was markedly reduced across estimators. The only substantial difference in bias across different weighting methods with correct model specification was observed in the context of DGM-6 and DGM-8.

3.4 Empirical application

3.4.1 The AKIKI and AKIKI-2 randomized controlled trials

We conducted an empirical evaluation of the different estimators tested in the Monte Carlo simulations using data from the AKIKI and AKIKI-2 RCTs. Both trials investigated strategies for initiating renal replacement therapy (RRT) in critically ill patients with acute kidney injury, differing primarily in the timing of initiation. The AKIKI trial compared an "early" RRT initiation strategy, where RRT was started upon meeting predefined inclusion criteria, to a "delayed" strategy that deferred initiation until more severe conditions arose. AKIKI-2 extended this comparison by evaluating the same "delayed" strategy from AKIKI against a "more-delayed" approach further relaxing RRT initiation criteria. Inclusion criteria in the AKIKI-2 trial were identical to AKIKI's; however after inclusion patients first entered an observational phase until they eventually met the "delayed" strategy RRT initiation criteria, at which time they were randomized into either "delayed" or "more-delayed" RRT treatment arms. Figure 9 illustrates these three treatment arms evaluated across the two AKIKI trials.

Both trials found no statistically significant differences in survival at 60 days. The presence of a common comparator arm ("delayed" strategy), and comparable inclusion criteria made these trials well-suited for evaluating different indirect comparisons strategies.

To assess estimators, we alternatively considered AKIKI and AKIKI-2 as the target population. In alignment with the simulation settings, we evaluated both anchored and unanchored strategies, as well as estimators fitted with different sets of variables: prognostic factors (Age, Minimum Mean Arterial Pressure, SOFA score), treatment effect modifiers (Chronic Kidney Disease, SpO₂/FiO₂ ratio, 24h diuresis), or both. These variables sets were selected based on the clinical expertise of AKIKI trials investigators guidance and data analysis. Additionally, we considered two approaches for aligning the trials depending on the reference starting date considered in the AKIKI-2 trial: the date of start of follow-up (observational period), or the date of randomization. Using the date of inclusion (start of follow-up) is the most appropriate alignment since inclusion criteria were identical at that point in both AKIKI and AKIKI-2. However, if relying on published data, the date of randomization would likely be used instead. To ensure comparability of AKI severity at baseline across trials when aligning on AKIKI-2 inclusion date, we randomly assigned in a 1:1 ratio patients from AKIKI-2 observational data who were eligible for AKIKI but never reached AKIKI-2 randomization criteria. This emulated a hypothetical allocation to "delayed" vs. "more-delayed" RRT; for these patients who never reached the "delayed" (and thus "more-delayed") RRT initiation threshold, the observed outcomes correspond to those that would have been observed under randomization.

Crucially, the choice of alignment date has a major impact on the covariate distribution across trials. Aligning on the date of inclusion yields broadly comparable populations, whereas aligning on the date of randomization results in substantial differences in baseline characteristics, leading to positivity issues in some instances, thus reflecting the range of scenarios explored in our simulations.

3.4.2 Illustration of estimator behavior using AKIKI data

Figure 10 presents the results of the empirical evaluation of the different estimators using data from the AKIKI and AKIKI-2 clinical trials. The four panels differ by the target population and therefore the IPD used (panel A and C: target population is AKIKI and IPD is AKIKI-2, and panels B and D: target population is AKIKI-2 and IPD is AKIKI). These findings can be summarized as follows. (i) The variability across estimators within the same anchoring approach (anchored or unanchored) is considerably smaller than the variability observed between anchored and unanchored approaches. (ii) When unanchored comparisons are used and the adjustment set is expanded to include prognostic factors, results exhibit greater variability and become more sensitive to the alignment date.

Specifically, when AKIKI-2 is used as the target population (panels B and D), estimates are relatively consistent across estimators and alignment strategies, regardless of the confounding adjustment strategy. In contrast, when AKIKI is used as the target population and the date of randomization of AKIKI-2 is used as the follow-up starting date (Panel C), unanchored estimators exhibit markedly wider confidence intervals, particularly for MAIC-2 and PSW. This reflects a substantial reduction in effective sample size due to poorer covariate overlap, with MAIC-1 maintaining more stable estimates and narrower confidence intervals under these conditions. While we know from theory that this alignment date is incorrect, it may have been the one used in practical settings if basing the analysis on reported results of the RCT, which stresses out the sensitivity of these methods to the lack of statistical support.

4 Discussion

This study evaluated the performance of MAIC estimators in comparison to a PSW estimator that leverages full individual-level data, using Monte Carlo simulations across diverse scenarios. The goal was to examine how deviations from ideal conditions—such as violations of the positivity (overlap) assumption or non-normal covariate distributions—affect estimators’ performance. Several key findings emerged from our investigation. First, under ideal conditions with normally distributed covariates, sufficient overlap between trials, and correctly specified adjustment models, all population-adjusted estimators produced unbiased treatment effect estimates with coverage near the nominal 95% level. In this setting, the unanchored approach showed slightly better precision (lower RMSE) than anchored analysis, consistent with the intuition that using a common comparator involves an additional level of variability. Similarly, when introducing non-normal covariate distributions without overlap issues, all estimators maintained unbiasedness and nominal coverage, suggesting that the MAIC method of moments weighting remains relevant even with non-normal covariate distributions. This result offers some empirical reassurance regarding a previously noted concern that the theoretical properties of MAIC under non-standard covariate distributions were not well understood.

In contrast, when the DGM involved positivity violations, clear differences between estimators emerged, and the full-IPD PSW estimator exhibited persistent bias despite correct model specification. This indicates that even with all relevant covariates included, the IPD-based estimators struggled to extrapolate beyond the support of the data, consistently with known properties of PSW [30] [12]. The MAIC estimators, particularly the simpler variant adjusting only for the first moment (MAIC-1), by contrast, remained less biased and achieved acceptable coverage, suggesting that method of moments weighting can be more robust to moderate overlap issues. However, it must be noted that MAIC approaches encountered convergence issues in a small number of simulation iterations likely due to iterations with more extreme overlap issues.

We also investigated the impact of model misspecification on the estimators by deliberately omitting a confounder from the adjustment model. The results of this analysis demonstrated that bias was driven primarily by whether the model correctly included all treatment effect modifiers, rather than by the choice of weighting method, consistently with literature results [9] [6].

Our work builds upon and extends the existing literature on population-adjusted indirect comparisons. Previous applied studies had compared MAIC with estimators using only IPD, but mostly in the context of specific case studies without extensive simulation. For example, Park et al. used a case study in small-cell lung cancer to compare various population-adjustment estimators (MAIC, simulated treatment comparison, and doubly robust combination of the two) and estimators using full IPD (propensity weighting, outcome regression, and doubly robust)[15]. They reported that the resulting treatment effect estimates were fairly similar across methods, which was attributed to the fact that the two trial populations in their case study were not significantly different. Similarly, Wong et al. compared treatment outcomes for vedolizumab vs. adalimumab in ulcerative colitis using either a PSW or an MAIC approach, and found only minimal differences between the adjusted estimates from these two methods [27]. These empirical comparisons align with the portion of our findings showing that when populations have substantial overlap, MAIC and PSW yield concordant results.

However, prior to our work, there had not been a comprehensive simulation study systematically stress-testing MAIC against an IPD estimator across a range of less ideal scenarios. A recent study by Campbell et al. conducted two separate simulation exercises to compare weighting vs. outcome regression estimators, under settings with and without full-IPD availability, but they did not directly cross-compare aggregated data and full-IPD approaches[6].

The results confirm some of the theoretical expectations discussed in earlier methodological literature. For instance, methodological reviews by Phillippo et al., and Remiro-Azócar et al. have cautioned that unanchored indirect comparisons rely on strong assumptions (no unmeasured confounding) and may hardly be trustworthy in real-life applications [16] [20]. While not new, we provide additional evidence for these cautions. Our results emphasize that whether IPD are fully accessible is of a much lower importance as compared to having a common treatment arm to form the indirect comparison.

A strength of this work is the various data-generating mechanisms considered, namely comparing distribution shapes (normal vs. bimodal), while varying degree of population overlap. We also augmented the simulation study with a real-world example, with consistent results. To our knowledge, this is the first study to compare using Monte Carlo simulations a population-adjusted indirect comparison technique against its full-IPD counterpart. Additionally, by including both anchored and unanchored implementations, and comparing complete vs. biased models, we shed light on the relative importance of these different parameters in the eventual performance of the estimators.

Several limitations of our study should be acknowledged. First, our simulations focused on a continuous outcome and a linear model for the treatment effect. Most real-world applications involve non-collapsible effect measures (e.g. odds

ratios for binary outcomes or hazard ratios for time-to-event outcomes). The relative performance of MAIC and PS weighting in those contexts remains to be investigated. Second, the finding that MAIC outperformed the full-data PSW in scenarios with overlap violations (DGMs 2, 4, 6 and 8) must be interpreted in context. In practice, if an analyst has actually access to IPD from both trials, other estimands can be targeted which may be less subject to positivity issue depending on the context: Average Treatment Effect in the Treated, or Average Treatment Effect in the Overlap population [12]. Closely related to that aspect, some remedial measures could be taken, such as truncating portions of the sample distributions that lie outside the common support. By doing so, one can largely resolve positivity violations and may obtain more reliable estimates with IPD methods. However, this procedure also changes the target population: the estimand is no longer the treatment effect in the original target trial’s population, but rather in the overlapped sub-population that both trials share. Whether this modified estimand is of interest depends on the context. In our simulations, we did not implement trimming for the PSW estimator; we evaluated the method on the original target estimand even when overlap was poor, to observe the unmitigated impact of positivity violations. In a real analysis with full IPD, one would not usually proceed with a gross overlap violation without some corrective action. Therefore, situations where MAIC appeared to outperform PSW (in bias or coverage) may reflect settings where corrective measures could be applied in practice, potentially mitigating PSW’s shortcomings. In other words, our results do not imply that an aggregate-data approach is superior to an IPD approach in absolute terms, but rather that the method of moments actually yields satisfactory results in some situations with overlap issues. As a consequence, in situations where the estimand of interest is actually the one targeted by MAIC and when one assumes positivity issues, matching the first moment of the distribution may be of interest, regardless of the actual availability of individual data. Whether that estimand is actually the primary estimand of interest, particularly when compared with the Average Treatment Effect in the combined trials population, has been discussed elsewhere [19, 21].

We did not evaluate other approaches such as Simulated Treatment Comparison (STC) or meta-regression methods that incorporate both IPD and aggregate data, nor more complex extensions like multilevel network meta-regression. These methods could be examined in future work to see if they share similar properties or offer advantages in the scenarios we considered.

Finally, as with any simulation study, our scenarios are simplification of real-world situations. We assumed two covariates and a rather simple outcome model structure; actual studies may involve multiple effect modifiers or non-linear effects. Moreover, other non-normal covariate distributions could prove challenging for an MAIC estimator, especially a covariate distribution which moments may not be defined.

Despite these limitations, our findings have important implications for the conduct and interpretation of indirect treatment comparisons. They provide concrete evidence that when all assumptions are met, population-adjusted estimators like MAIC can perform just as well as conventional PSW that uses complete data, even in some cases maintaining accuracy in situations where the latter suffers (e.g. moderate overlap violations). This is an encouraging result when one must rely on aggregate data out of necessity since it suggests that, under reasonably good conditions, using MAIC does not inherently compromise the validity or precision of the estimated treatment effect. Consequently, the added value of access to full IPD lies in greater analytical flexibility. At the same time, our results emphasize that the primary determinants of validity are the upholding of causal inference assumptions—positivity, correct model specification and exchangeability—thus stressing out once again the inherent superiority of anchored as compared to unanchored comparisons.

Author contributions ASL, DH, JL conceived the study and developed the statistical methodology of the simulation study. ASL, DH, JL and SG conceived the empirical application. ASL implemented the simulations and empirical application. DH supervised the statistical analyses. ASL, DH, JL and SG interpreted the results. ASL drafted the first version of the manuscript. DH, JL and SG critically revised the manuscript. All authors read and approved the final version of the manuscript.

Financial disclosure None reported.

Conflict of interest ASL: Remunerated lecture for Pierre Fabre Laboratories on PAICs, unrelated to this work

List of Figures

1	Relationships among X_1 , X_2 , T , Z , and Y	13
2	Covariate distributions depending on DGM	14
3	Treatment effect estimation under DGM-1 and DGM-2	15
4	Treatment effect estimation under DGM-3 and DGM-4	16
5	Treatment effect estimation under DGM-5 and DGM-6 (asterisks indicate that estimators did not converge for some of the iterations: DGM-6, 73/2000 for anchored MAIC-2, and 294/2000 for unanchored MAIC-2)	17
6	Treatment effect estimation under DGM-7 and DGM-8 (asterisks indicate that estimators did not converge for some of the iterations: DGM-7, 1/2000 for unanchored MAIC-2; DGM-8 4/2000 for unanchored MAIC-2)	18
7	Anchored estimators: X_1 distribution before and after weighting + individual PS distribution depending on X_1 value	19
8	Bias of estimators depending on adequate model specification (the asterisk indicates that 3 iterations did not converge for unanchored MAIC-1)	20
9	AKIKI trials: treatment arms according to the RRT initiation time	21
10	Empirical data application on AKIKI RCTs: log(HR) estimate (95% CI) for the "more-delayed" strategy vs "early" strategy, using the "delayed" strategy as the anchor. The four panels differ on the target population (AKIKI vs AKIKI-2), and the alignment date according to the AKIKI-2 inclusion or randomization date. 95% Confidence intervals are not displayed on panel C for unanchored PSW and MAIC-2 because bounds failed outside of the plot limits.	22
S1	Unanchored estimators: X_1 and X_2 distributions before and after weighting + individual PS distribution depending on X_1 and X_2 values	26

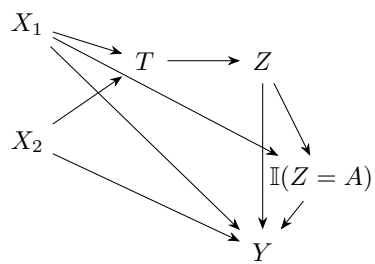


Figure 1: Relationships among X_1 , X_2 , T , Z , and Y .

Covariates (X1 and X2) distributions in a and b trials

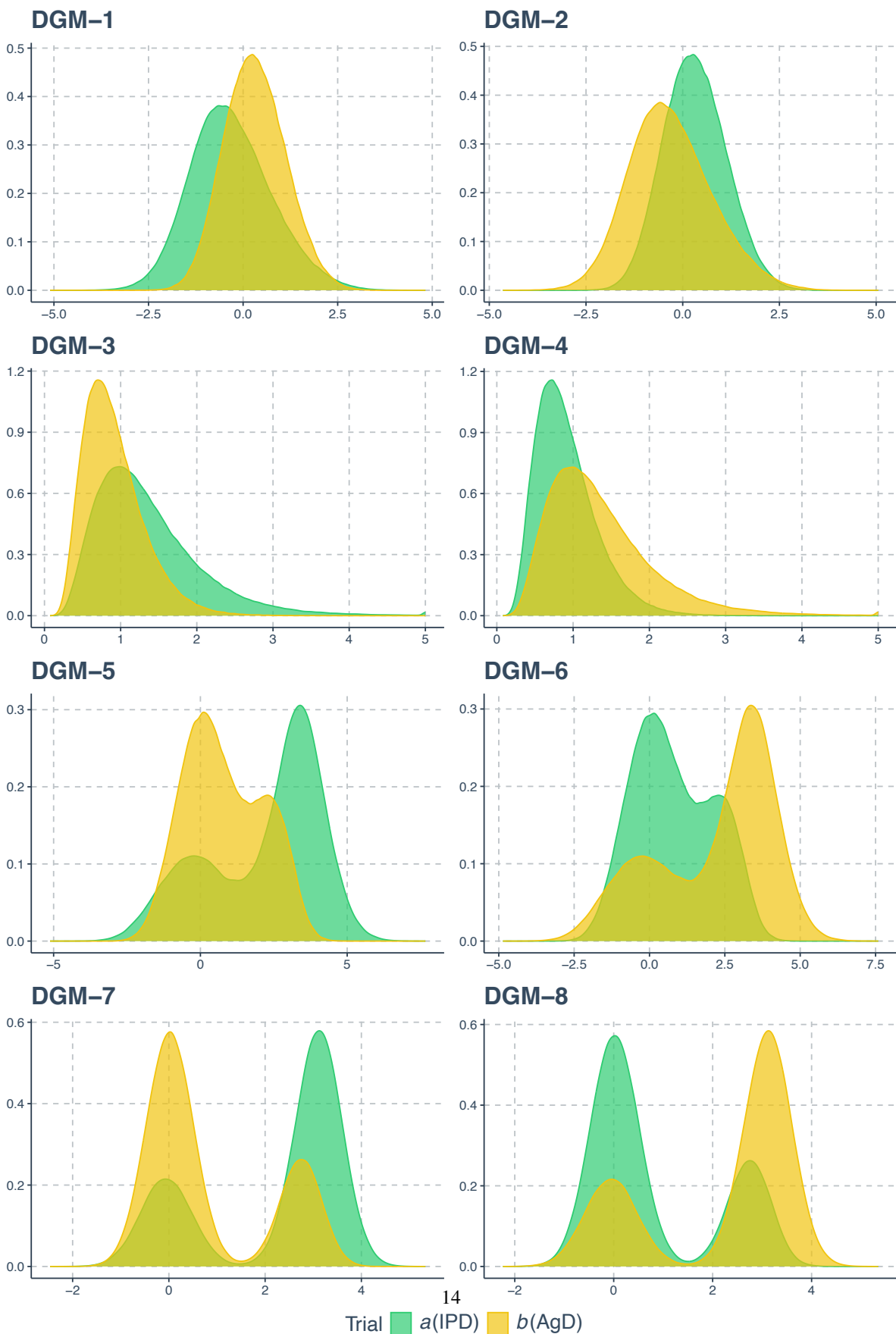


Figure 2: Covariate distributions depending on DGM

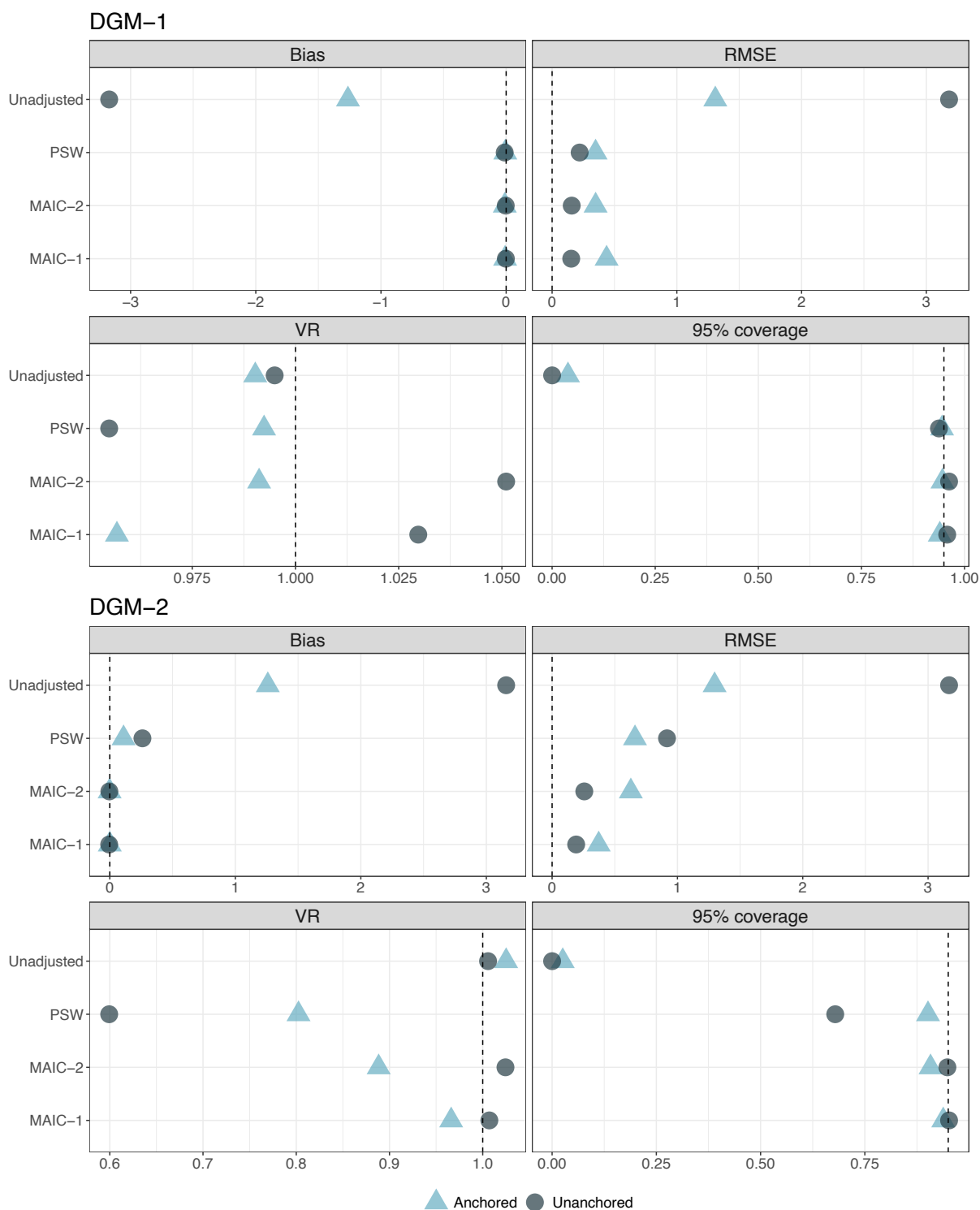


Figure 3: Treatment effect estimation under DGM-1 and DGM-2

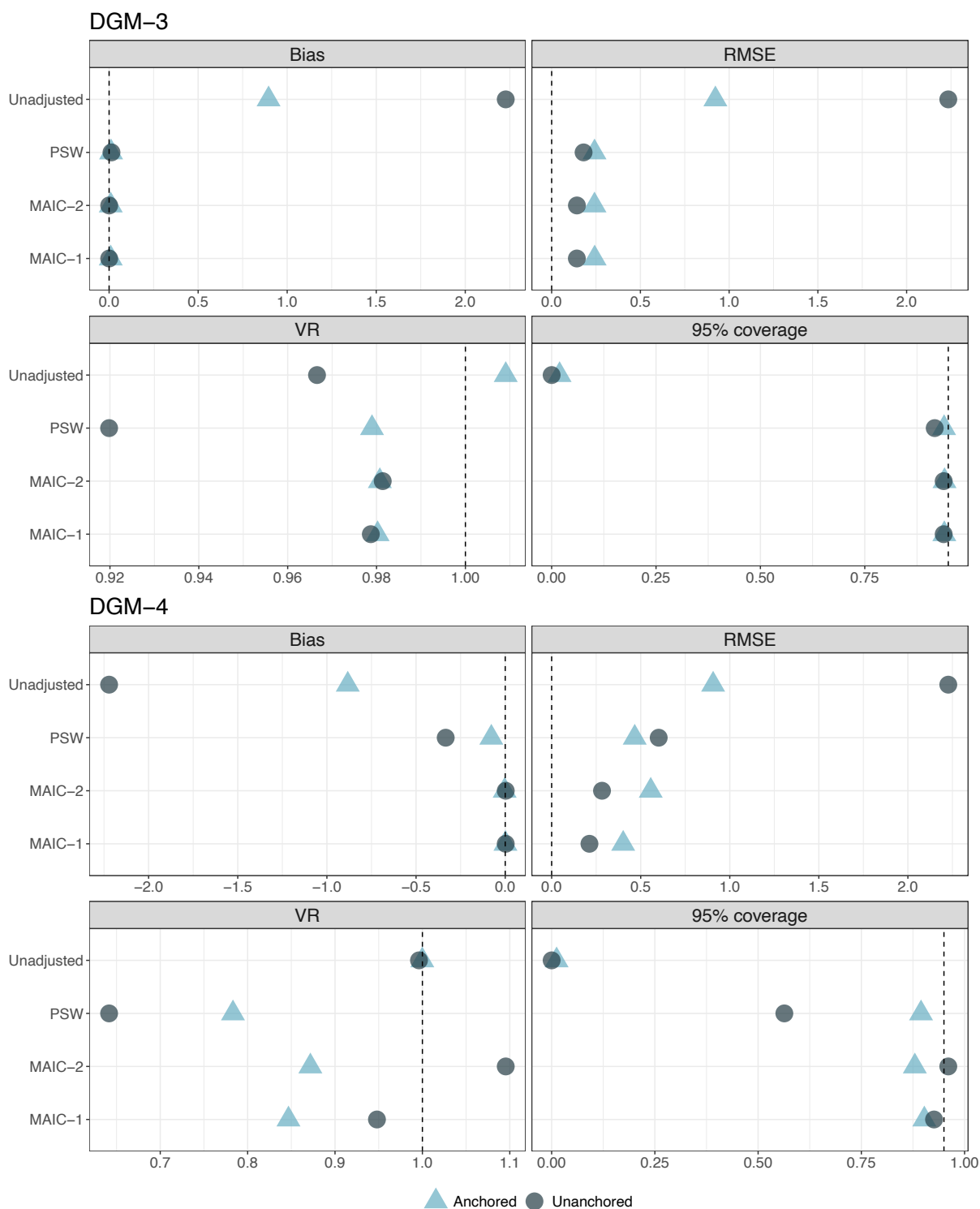


Figure 4: Treatment effect estimation under DGM-3 and DGM-4

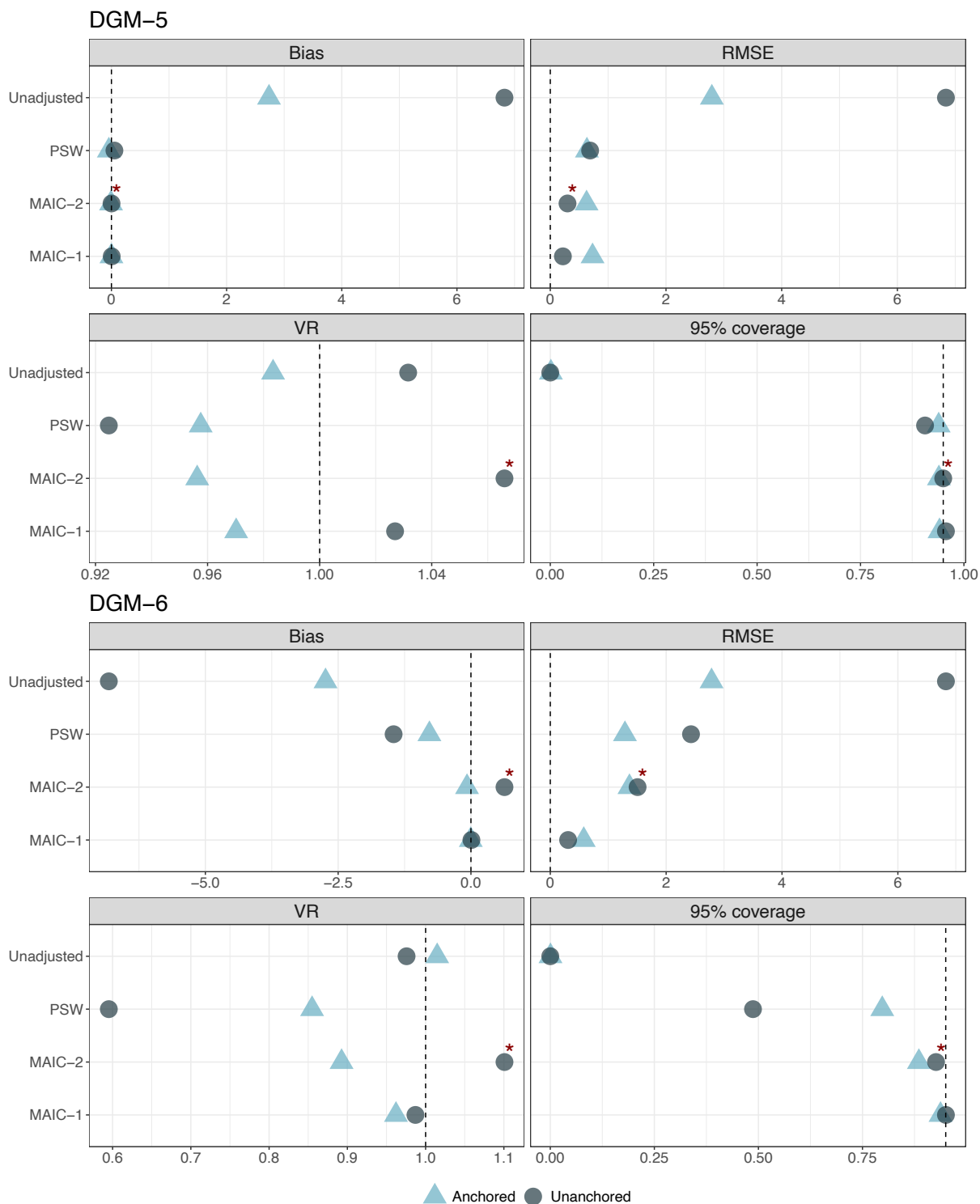


Figure 5: Treatment effect estimation under DGM-5 and DGM-6 (asterisks indicate that estimators did not converge for some of the iterations: DGM-6, 73/2000 for anchored MAIC-2, and 294/2000 for unanchored MAIC-2)

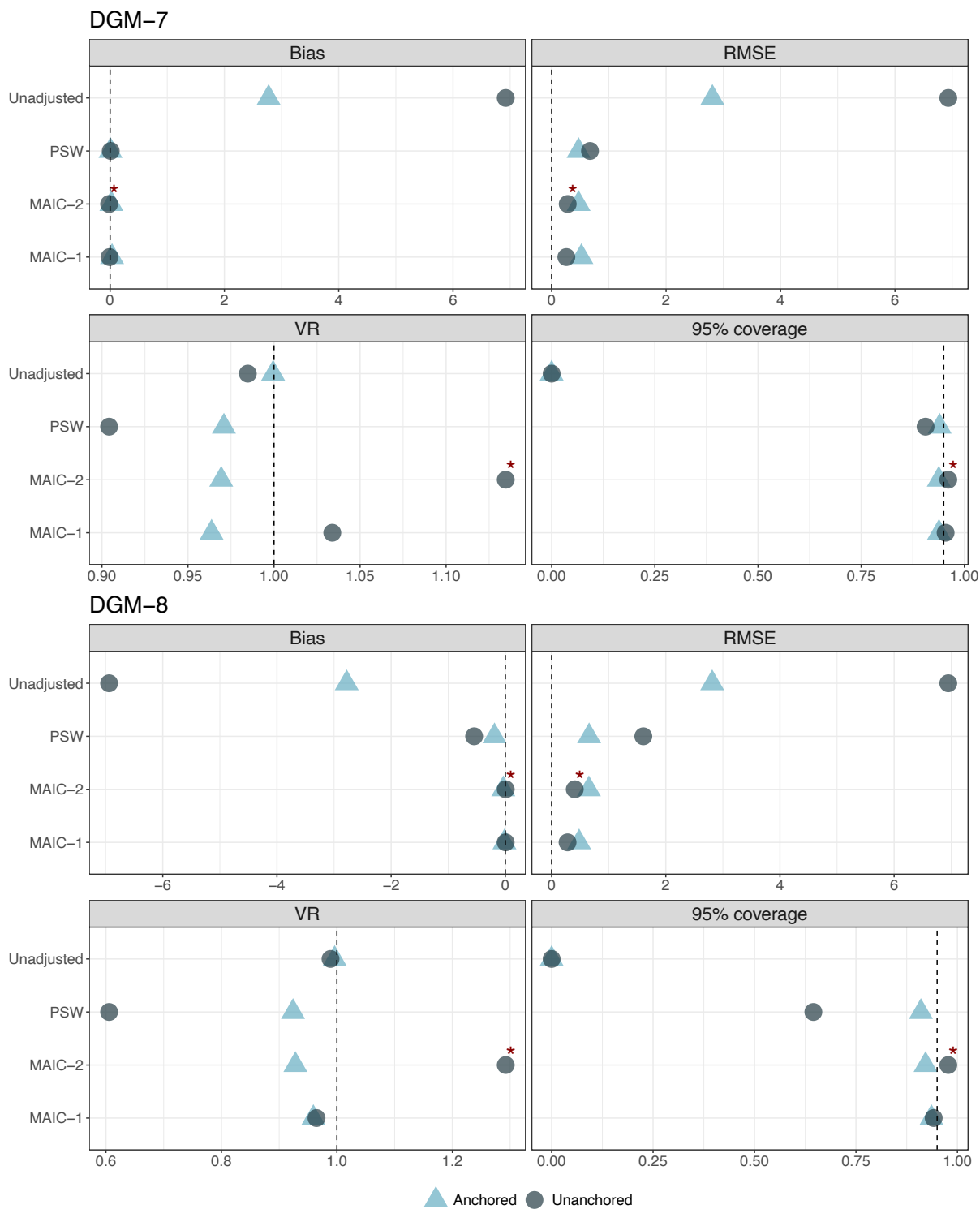


Figure 6: Treatment effect estimation under DGM-7 and DGM-8 (asterisks indicate that estimators did not converge for some of the iterations: DGM-7, 1/2000 for unanchored MAIC-2; DGM-8 4/2000 for unanchored MAIC-2)

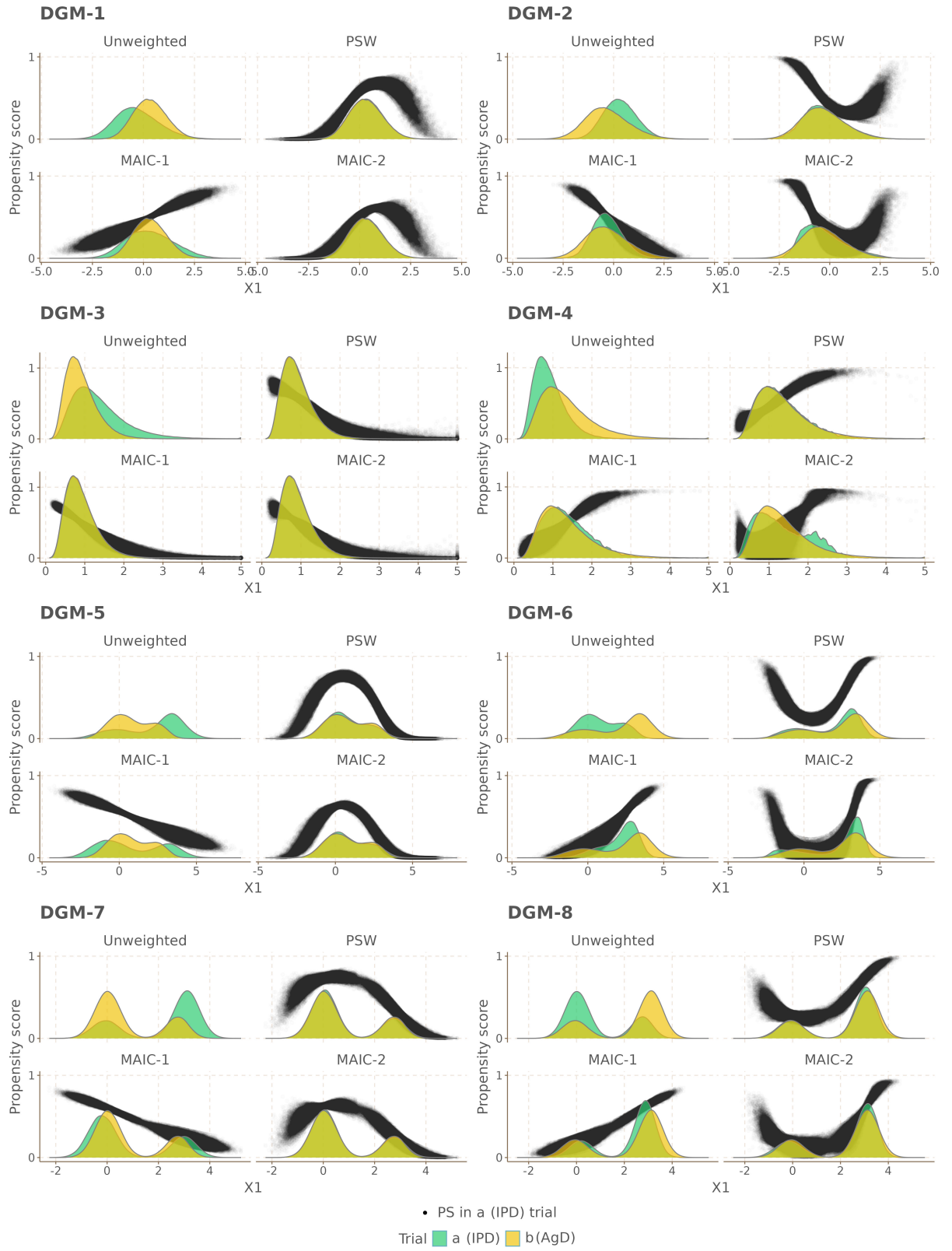


Figure 7: Anchored estimators: X_1 distribution before and after weighting + individual PS distribution depending on X_1 value

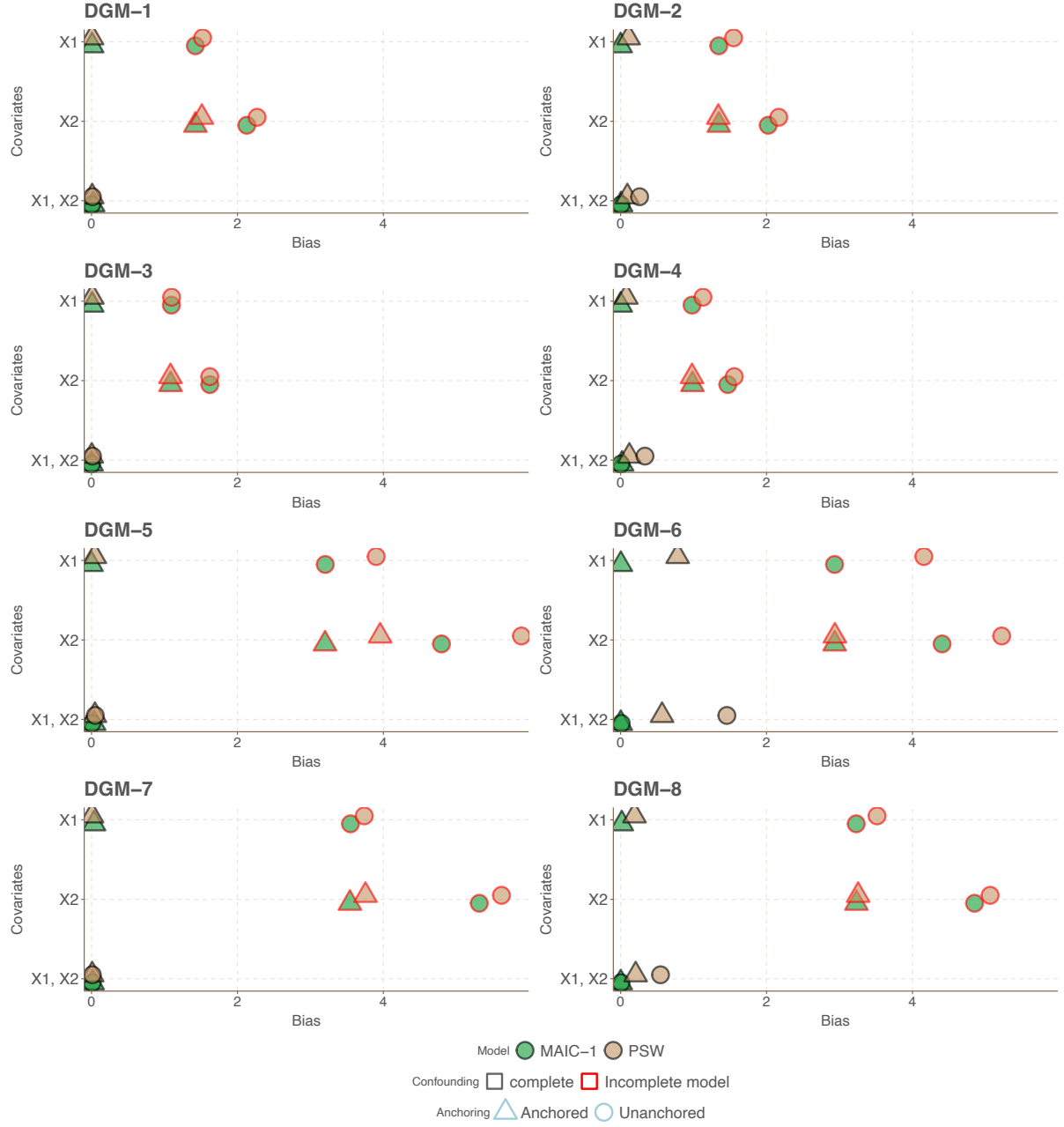


Figure 8: Bias of estimators depending on adequate model specification (the asterisk indicates that 3 iterations did not converge for unanchored MAIC-1)

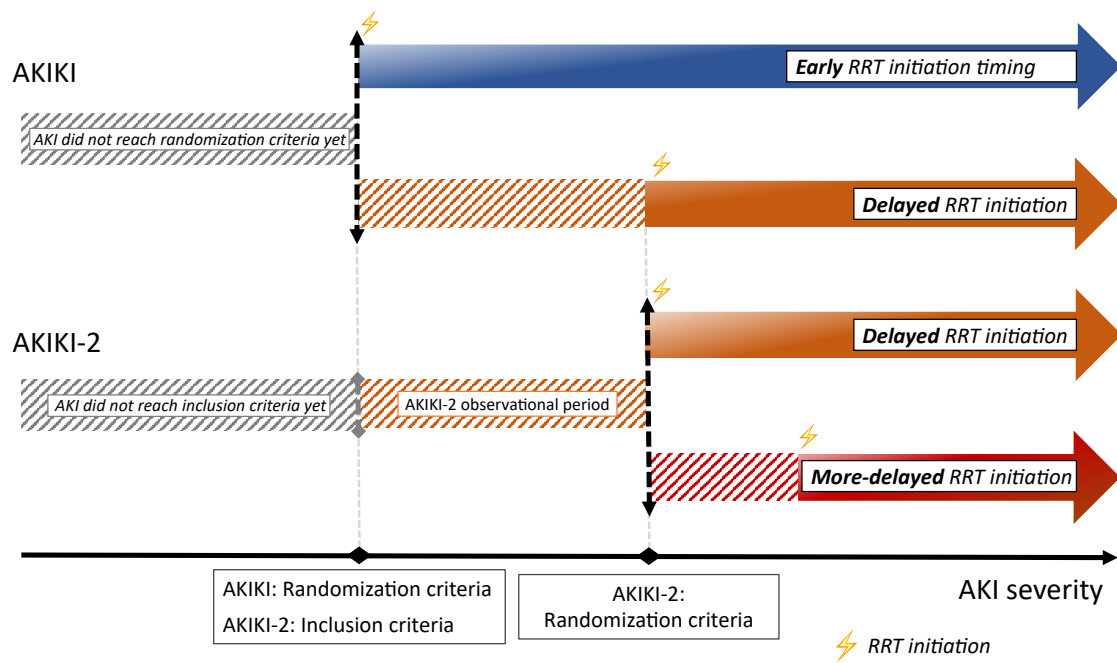


Figure 9: AKIKI trials: treatment arms according to the RRT initiation time

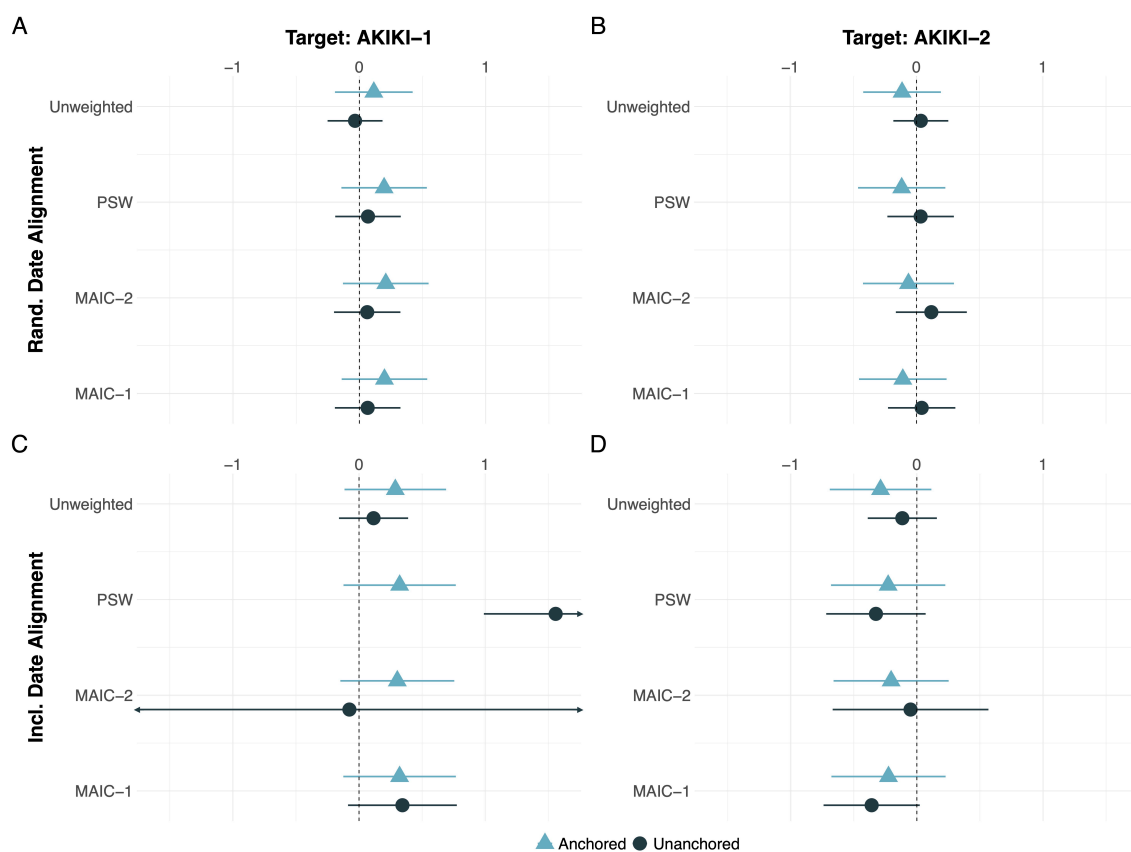


Figure 10: Empirical data application on AKIKI RCTs: log(HR) estimate (95% CI) for the "more-delayed" strategy vs "early" strategy, using the "delayed" strategy as the anchor. The four panels differ on the target population (AKIKI vs AKIKI-2), and the alignment date according to the AKIKI-2 inclusion or randomization date. 95% Confidence intervals are not displayed on panel C for unanchored PSW and MAIC-2 because bounds failed outside of the plot limits.

References

- [1] Sundeep Agrawal, Shaily Arora, Laleh Amiri-Kordestani, R. Angelo de Claro, Lola Fashoyin-Aje, Nicole Gormley, Tamy Kim, Steven Lemery, Gautam U. Mehta, Emma C. Scott, Harpreet Singh, Shenghui Tang, Marc R. Theoret, Richard Pazdur, Paul G. Kluetz, and Julia A. Beaver. Use of Single-Arm Trials for US Food and Drug Administration Drug Approval in Oncology, 2002-2021. *JAMA Oncology*, December 2022. ISSN 2374-2437. doi: 10.1001/jamaoncol.2022.5985.
- [2] William H. Asquith. L-moments and TL-moments of the generalized lambda distribution. *Computational Statistics & Data Analysis*, 51(9):4484–4496, May 2007. ISSN 0167-9473. doi: 10.1016/j.csda.2006.07.016.
- [3] Peter C. Austin. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, 46(3):399–424, May 2011. ISSN 1532-7906. doi: 10.1080/00273171.2011.568786.
- [4] Julia A. Beaver, Lynn J. Howie, Lorraine Pelosof, Tamy Kim, Jinzhong Liu, Kirsten B. Goldberg, Rajeshwari Sridhara, Gideon M. Blumenthal, Ann T. Farrell, Patricia Keegan, Richard Pazdur, and Paul G. Kluetz. A 25-Year Experience of US Food and Drug Administration Accelerated Approval of Malignant Hematology and Oncology Drugs and Biologics: A Review. *JAMA oncology*, 4(6):849–856, June 2018. ISSN 2374-2445. doi: 10.1001/jamaoncol.2017.5618.
- [5] Heiner C. Bucher, Gordon H. Guyatt, Lauren E. Griffith, and Stephen D. Walter. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of Clinical Epidemiology*, 50(6): 683–691, June 1997. ISSN 0895-4356. doi: 10.1016/S0895-4356(97)00049-8.
- [6] Harlan Campbell, Julie E. Park, Jeroen P. Jansen, and Shannon Cope. Estimating marginal treatment effects from observational studies and indirect treatment comparisons: When are standardization-based methods preferable to those based on propensity score weighting?, October 2023.
- [7] J. Jaime Caro and K. Jack Ishak. No Head-to-Head Trial? Simulate the Missing Arms:. *PharmacoEconomics*, 28 (10):957–967, October 2010. ISSN 1170-7690. doi: 10.2165/11537420-000000000-00000.
- [8] Cecília Menezes Farinasso, Vinícius Lins Ferreira, Flávia Cordeiro Medeiros, Aline Pereira da Rocha, Patrícia do Carmo Silva Parreira, Layssa Andrade Oliveira, Lays Pires Marra, Rosa Camila Lucchetta, and Haliton Alves de Oliveira. Matching-Adjusted Indirect Comparison Studies in Oncology: A Scoping Review Focused on Reporting Quality. *Value in Health Regional Issues*, 47:101088, May 2025. ISSN 2212-1099. doi: 10.1016/j.vhri.2025.101088.
- [9] Anthony James Hatswell, Nick Freemantle, and Gianluca Baio. The Effects of Model Misspecification in Unanchored Matching-Adjusted Indirect Comparison: Results of a Simulation Study. *Value in Health*, 23(6): 751–759, June 2020. ISSN 10983015. doi: 10.1016/j.jval.2020.02.008.
- [10] Miguel A Hernán and James M Robins. *Causal Inference: What If*. CRC Press, 2020.
- [11] Ataru Igarashi, Shiro Tanaka, Raf De Moor, Nan Li, Mariko Hirozane, David Bin-Chia Wu, Li Wen Hong, Dae Young Yu, Mahmoud Hashim, Brian Hutton, Krista Tantakoun, Christopher Olsen, Fatemeh Mirzayeh Fashami, Imtiaz A. Samjoo, and Chris Cameron. Indirect Treatment Comparisons in Healthcare Decision Making: A Targeted Review of Regulatory Approval, Reimbursement, and Pricing Recommendations Globally for Oncology Drugs in 2021-2023. *Advances in Therapy*, 42(1):52–69, January 2025. ISSN 1865-8652. doi: 10.1007/s12325-024-03013-6.
- [12] Fan Li, Laine E. Thomas, and Fan Li. Addressing Extreme Propensity Scores via the Overlap Weights. *American Journal of Epidemiology*, 188(1):250–257, January 2019. ISSN 0002-9262. doi: 10.1093/aje/kwy201.
- [13] Feinan Lu, Tao Wang, Ying Lu, and Jie Chen. Considerations for Single-Arm Trials to Support Accelerated Approval of Oncology Drugs, November 2024.
- [14] Bérengère Macabeo, Théophile Rotrou, Aurélie Millier, Clément François, and Philippe Laramée. The Acceptance of Indirect Treatment Comparison Methods in Oncology by Health Technology Assessment Agencies in England, France, Germany, Italy, and Spain. *PharmacoEconomics - Open*, 8(1):5–18, January 2024. ISSN 2509-4254. doi: 10.1007/s41669-023-00455-6.
- [15] Julie E. Park, Harlan Campbell, Kevin Towle, Yong Yuan, Jeroen P. Jansen, David Phillippo, and Shannon Cope. Unanchored Population-Adjusted Indirect Comparison Methods for Time-to-Event Outcomes Using Inverse Odds Weighting, Regression Adjustment, and Doubly Robust Methods With Either Individual Patient or Aggregate Data. *Value in Health*, 27(3):278–286, March 2024. ISSN 1098-3015. doi: 10.1016/j.jval.2023.11.011.
- [16] David M. Phillippo, Anthony E. Ades, Sofia Dias, Stephen Palmer, Keith R. Abrams, and Nicky J. Welton. Methods for Population-Adjusted Indirect Comparisons in Health Technology Appraisal. *Medical Decision Making*, 38(2):200–211, February 2018. ISSN 0272-989X, 1552-681X. doi: 10.1177/0272989X17725740.

- [17] David M. Phillippo, Sofia Dias, A. E. Ades, and Nicky J. Welton. Multilevel network meta-regression for general likelihoods: Synthesis of individual and aggregate data with applications to survival analysis, January 2024.
- [18] DM Phillippo, A. E. Ades, Sofia Dias, Stephen Palmer, Keith R. Abrams, and Nicky J. Welton. Methods for Population-Adjusted Indirect Comparisons in Submission to NICE. Technical Report 18, Decision Support Unit, National Institute for Health and Care Excellence, University of Sheffield, December 2016.
- [19] Antonio Remiro-Azócar. Target estimands for population-adjusted indirect comparisons. *Statistics in Medicine*, 41(28):5558–5569, December 2022. ISSN 0277-6715, 1097-0258. doi: 10.1002/sim.9413.
- [20] Antonio Remiro-Azócar, Anna Heath, and Gianluca Baio. Methods for population adjustment with limited access to individual patient data: A review and simulation study. *Research Synthesis Methods*, 12(6):750–775, 2021. ISSN 1759-2887. doi: 10.1002/jrsm.1511.
- [21] Estelle Russek-Cohen. Discussion of “target estimands for population-adjusted indirect comparisons” by Antonio Remiro-Azocar. *Statistics in Medicine*, 41(28):5573–5576, 2022. ISSN 1097-0258. doi: 10.1002/sim.9533.
- [22] Arnaud Serret-Larmande, Belkacem Zenati, Agnès Dechartres, Jérôme Lambert, and David Hajage. A methodological review of population-adjusted indirect comparisons reveals inconsistent reporting and suggests publication bias. *Journal of Clinical Epidemiology*, 163:1–10, November 2023. ISSN 0895-4356, 1878-5921. doi: 10.1016/j.jclinepi.2023.09.004.
- [23] James E. Signorovitch, Eric Q. Wu, Andrew P. Yu, Charles M. Gerrits, Evan Kantor, Yanjun Bao, Shiraz R. Gupta, and Parvez M. Mulani. Comparative effectiveness without head-to-head trials: A method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept. *PharmacoEconomics*, 28(10):935–945, 2010. ISSN 1179-2027. doi: 10.2165/11538370-000000000-00000.
- [24] Signorovitch J, Erder MH, Xie J, Sikirica V, Lu M, Hodgkins PS, and Wu EQ. Comparative effectiveness research using matching-adjusted indirect comparison: An application to treatment with guanfacine extended release or atomoxetine in children with attention-deficit/hyperactivity disorder and comorbid oppositional defiant disorder. *Pharmacoepidemiology and drug safety*, 21:130–7, 2012. ISSN 1099-1557 (Electronic).
- [25] Bang Truong, Lan-Anh T. Tran, Tuan Anh Le, Thi Thu Pham, and Tat-Thang Vo. Population adjusted-indirect comparisons in health technology assessment: A methodological systematic review. *Research Synthesis Methods*, 14(5):660–670, September 2023. ISSN 1759-2879, 1759-2887. doi: 10.1002/jrsm.1653.
- [26] Antoine Vanier, Judith Fernandez, Sophie Kelley, Lise Alter, Patrick Semenzato, Corinne Alberti, Sylvie Chevret, Dominique Costagliola, Michel Cucherat, Bruno Falissard, François Gueyffier, Jérôme Lambert, Etienne Lengliné, Clara Locher, Florian Naudet, Raphael Porcher, Rodolphe Thiébaut, Muriel Vray, Sarah Zohar, Pierre Cochat, and Dominique Le Guludec. Rapid access to innovative medicinal products while ensuring relevant health technology assessment. Position of the French National Authority for Health. *BMJ evidence-based medicine*, pages bmjebm–2022–112091, February 2023. ISSN 2515-4478. doi: 10.1136/bmjebm-2022-112091.
- [27] Emily C L Wong, Parambir S Dulai, John K Marshall, Vipul Jairath, Walter Reinisch, and Neeraj Narula. Matching-adjusted Indirect Comparisons vs Propensity Score Matching with Individual Patient-level Data to Estimate Treatment Efficacy. *Inflammatory Bowel Diseases*, page izad077, April 2023. ISSN 1078-0998. doi: 10.1093/ibd/izad077.
- [28] Quanfu Yu and Jun Xu. Harmonic transform-based density estimation method for uncertainty propagation and reliability analysis involving multi-modal distributions. *Mechanical Systems and Signal Processing*, 190:110113, May 2023. ISSN 0888-3270. doi: 10.1016/j.ymssp.2023.110113.
- [29] Yukang Zeng, Fan Li, and Guangyu Tong. Moving toward best practice when using propensity score weighting in survey observational studies, January 2025.
- [30] Yaqian Zhu, Rebecca A. Hubbard, Jessica Chubak, Jason Roy, and Nandita Mitra. Core Concepts in Pharmacoepidemiology: Violations of the Positivity Assumption in the Causal Analysis of Observational Data: Consequences and Statistical Approaches. *Pharmacoepidemiology and drug safety*, 30(11):1471–1485, November 2021. ISSN 1053-8569. doi: 10.1002/pds.5338.

A Appendix

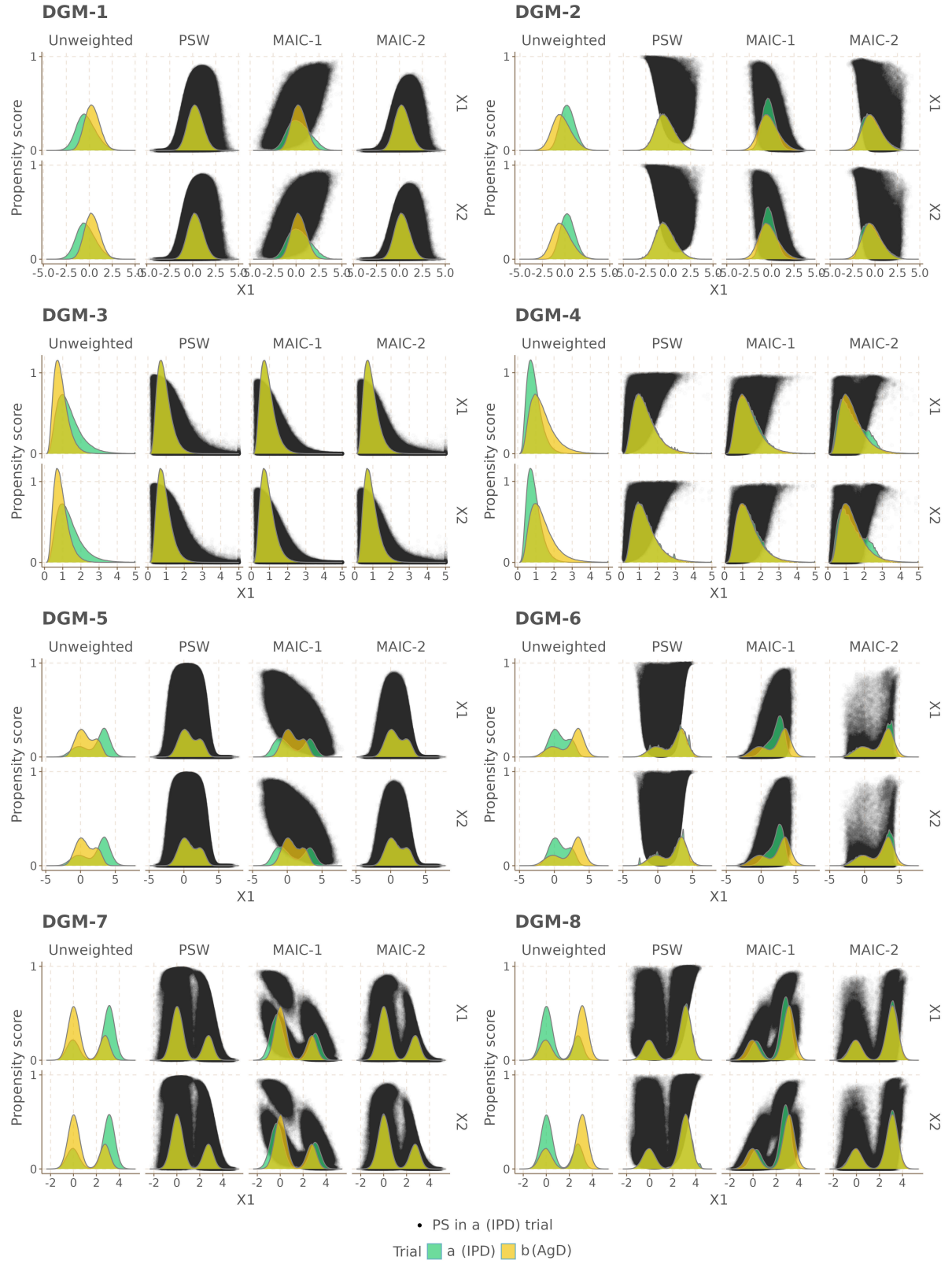


Figure S1: Unanchored estimators: X_1 and X_2 distributions before and after weighting + individual PS distribution depending on X_1 and X_2 values

Chapter 5

Discussion and perspectives

5.1 Findings summary

This work addressed the question of PAICs from two complementary perspectives: a meta-epidemiological one through a methodological review of published applications, and a methodological one, through the simulation and empirical evaluation of statistical methods. Taken together, these two angles provide both a description of current practice and add to the existing methodological literature evaluating the performance of PAICs.

The review was motivated by the contrast between the significant theoretical knowledge accumulated on PAIC methods and the limited practical knowledge regarding their implementation in published studies. The conduct of three other reviews on this subject, in parallel with ours, underscore the pressing need to better document the implementation of PAICs. Our initial objective in conducting this review was to provide a descriptive account of the field: documenting the dynamics of publications over time, the domains of publications, and to document the way in which PAIC methods were conducted and reported. However, the extensive publication and reporting bias was not anticipated but turned out to be the most noteworthy result of this work. The report of such a large publication bias calls into question the reliability of the body of evidence that partly underpins HTA and regulatory approvals. Conversely, one may regard it as stating the obvious since selective dissemination of favorable findings is a long-recognized phenomenon in medical research, and is particularly expected in the case of sponsored studies reporting unregistered post hoc analysis. However, another nuance to keep in mind is that there may be major differences between results presented in peer-reviewed articles and those contained in regulatory or HTA dossiers [91]. Indeed, these dossiers usually include a significantly larger body of results as compared to peer-reviewed publications, some of which may be less favorable

to the treatment or intervention being evaluated. It is unrealistic to expect that post-hoc analyses with negative or inconclusive findings, especially when unregistered and sponsor-initiated, would undergo the full peer-review process and ultimately reach publication. In contrast, such results may nonetheless be available within regulatory or HTA submissions, particularly when indirect comparisons are explicitly required by the authorities. Assessing the real magnitude and consequences of this publication bias would therefore require a systematic review of regulatory approval or HTA dossiers, comparing the results disclosed therein with those selectively published in the academic literature. Undertaking such a review would likely be highly resource-intensive, and may thus need to be restricted to a targeted sample of dossiers.

Additionally, we propose several suggestions that could help address the issue of selective reporting and publication bias. Health Authorities could encourage or require sponsors to prespecify (either at the statistical analysis planning stage or at the time of HTA submission) a transparent analytic framework for conducting indirect comparisons, including listing eligible studies and potential data sources, inclusion criteria, justification for covariate selection, adjustment methods, positivity checks, and sensitivity analyses. In line with this recommendation, independent third parties mandated and overseen by health authorities could be involved in the analysis when the indirect comparison is pivotal to the reimbursement decision. Technical solutions are being developed to mitigate privacy concerns when involving third-parties [92]. Additionally, encouraging de-identified data sharing could help with this objective; with the additional benefit of allowing more flexibility in the analysis if IPD is available for all studies in a given network, which allows using the most flexible analytical methods for transporting estimands across a wide range of populations of interest. Such calls have been coming from both the industry and regulatory bodies [93, 28].

Clinical trials data sharing in order to foster reanalysis is already an important and discussed topic in the literature [94, 95, 96, 97]. Hollis et al. 2016 stressed the importance of individual data sharing for maximizing the benefit of clinical trials to patients and society [98]. Some health technology developers require that their data be analyzed within the boundaries of a local system that doesn't allow for exporting data, and consequently doesn't allow for combining data coming from different sources. Interestingly, this specific context that could hinder the realization of IPD NMA because of the requirement to gather all IPD in one central location is not a constraint for conducting PAIC that by definition can be carried-out with one single IPD source data. This makes the initiative to develop multi-sponsor environments all the more relevant, which would manage both privacy constraints and transparency, reproducibility and trustworthiness of analyses results. The repeated calls to action spurred initiatives, with several high-profile industry players and academic

institutions having initiated the development of such centralized repositories [99, 100, 101]. Publicly funded institutions also embraced such requirement; one can mention the National, Heart, Lung, and Blood Institute which requires research funding application to provision a portion of the overall budget to fund sharing of the research data and developed their own data sharing platform alongside informatics solutions to query it [102, 103]. To date, no regulatory bodies mandates making anonymized IPD accessible. In the EU, the EMA *Policy 0070* mandates Health Technology Developers to provide IPD alongside market authorization applications, for the use of regulatory commission analysts, but there is currently no plans to make IPD more widely available [104]. Similar recommendations are in place in the US and in Canada [105, 106]. Finally, fostering greater involvement of academic consortia and publicly funded initiatives would increase confidence in results obtained using PAICs, by encouraging proactively evaluating key drug comparisons using PAIC and related methods, thereby complementing sponsor-submitted evidence.

This emphasis on sharing IPD echoes the work conducted in the second article featured in this thesis. The informal question that underlies this work was: “what is the cost of missing IPD”? At first glance, the findings might suggest that missing IPD is not a major concern, since MAICs appeared more robust than expected, comparatively to a weighting estimator using full IPD. However, this observation is not to be interpreted as an argument to justify relying on less-than-ideal estimators and reduce the granularity of information used in the analysis. We reiterate and further develop the arguments already emphasized in the discussion of this simulation study:

- **The ability to target different populations.** Consistent with a body of work discussing the matter, the most significant advantage in having access to full IPD is the ability to precisely define and estimate an effect in any relevant target population. This property is going to become increasingly important in the context of the incoming EU Joint Clinical Assessment (JCA)s, with the numerous PICOs resulting from the multitude of member states specific scope (*i.e.* under this new framework, each EU member state is entitled to define their own decision contexts when assessing new therapies regarding target population of interest, comparators, and outcomes) [107, 108]. Analysts note that this will further increase the requirements for ITC [109]. In this context, being able to estimate a treatment relative effect in various populations is paramount. MAIC and STC are, in contrast, lacking the possibility to transport the effect into any other population than the AgD one, thereby potentially limiting their use in the context of EU JCAs. ML-NMR extends population-adjusted methods to networks beyond pairwise comparisons, allowing transportability of the estimand to a chosen target population, under the additional assumption of shared effect-modifier

[14].

- **Greater flexibility in assumption checking.** Access to IPD provides unparalleled flexibility to rigorously interrogate the data and validate the underlying assumptions of the statistical models employed. More specifically, when having access to full IPD, analysts have greater flexibility to perform crucial assumption diagnostics, such as determining whether variables act as PF or TEM in the adjustment models, and to assess other requirements, including the positivity assumption and post-weighting covariate balance.
- **Access to full IPD also enables adjustment for PF or TEMs that are not reported in publications.** This may be a key advantage since published summaries may not report descriptive statistics for every clinically relevant variables, leaving PAIC analyses exposed to residual confounding if these unreported baseline characteristics are associated with the outcome and imbalanced between studies.
- **Sample size limitation.** Half available IPD comes with limitations in two forms. First, it leads to greater uncertainty in nuisance parameter estimation of the trial assignment or outcome models, and second it leads to a small ESS when using MAIC.

Hence, we want to stress again that the aforementioned results cannot be interpreted as an argument to justify slowing down initiatives fostering IPD sharing.

5.2 Perspectives and avenues for future research

This work explored the subject from two angles: a practical one, and a methodological one. Our work added to the existing literature important results that will represent significant contributions, we hope, to the rapidly-evolving field of PAIC. We list several avenues that can be explored to keep on building on this work.

Methodological review using HTA and regulatory approval dossiers.

Our methodological review focused on peer-reviewed articles. While the statistical properties of these methods make them useful in the context of academic epidemiological studies, they remain as of today mostly tools used in HTA and regulatory context. Consequently, a comprehensive review of these methods in the context of HTA would be of tremendous interest. Reviews do exist on this theme; one can mention Phillippo et al. (2019), that focused on submissions to NICE up to 2018 [20]. More recently Macabeo et al. (2024) focused on

submissions of ITC, including PAIC, in submissions to 5 European HTA bodies; however, the emphasis was mostly on the acceptance rate of such methods by these bodies [67]. To bridge our work and the ones we just mentioned, we think that a work focusing on comparing results published in the literature and the ones filed in HTA dossiers would provide another perspective on the publication bias that we reported in our study. Indeed, it is of utmost importance to understand whether such a skewed reporting bias may also be present in HTA dossiers, or whether it is merely a consequence of the fact that it is unrealistic to expect that unregistered analyses report unfavorable results for the study sponsor in peer-reviewed publications.

Updating conduct and reporting guidelines

In its two documents, *Practical Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons* and *Methodological Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons* the Member State Coordination Group on Health Technology Assessment (HTACG) issued recommendations on implementing and reporting PAIC [28, 70]. This EU-level guidance issues recommendations on the different types of existing ITC, their assumptions and their implementation recommendations. The aforementioned dossiers review may be used to further expand guidelines, with the inclusion of guidance on pre-registration of the analyses, recommendations on the causal model underlying the adjustment carried-out, the way information about confounders has been established (systematic review, expert knowledge), as well as an update to integrate the most recent methodological developments such as doubly-robust methods.

Moreover, similar to the STROBE statement for observational studies or CONSORT for clinical trials, the domain of PAIC could benefit from the creation of an international working group tasked with building consensus across academia, industry, and HTA bodies, with the goal of producing a guideline statement. Developing international standardized recommendations for the conduct and reporting of PAICs may improve the methodological rigor of such studies and provide regulatory bodies and HTA agencies clear benchmarks for evaluating submissions.

Expanding Monte Carlo simulations data generating mechanisms

In our simulation work, the DGM was constituted of two variables only (one PF and one TEM). The reason behind such a simple model was to facilitate results interpretation, namely regarding covariates overlap (or lack thereof), and complete versus incomplete model specification. However, this DGM can be considered an oversimplification of reality. It would be

especially interesting to conduct a simulation study which DGM would stick to the average number of variables and distributions found in general in real-life applications. This would allow for a finer exploration of the consequences of unmeasured covariates in adjustment models. Moreover, the inclusion of a greater number of covariates may change the outlook on the relative performance of MAIC and PSW because of the smaller sample size used to fit the trial assignment model when using MAIC. Closely related to the DGM is the outcome: we used continuous distributions in order to avoid statistical issues coming with using non-collapsible measure of association. However, this choice is not ideal; medical research is most of the times interested in binary or time-to-event outcomes; PAICs are no exception to the rule [21, 44]. While we also explored binary outcomes and difference of proportions in our simulations (results not published), the results could be expanded by comparing relative performance of weighting estimators with non-collapsible contrasts such as odds ratio and hazard ratio. Moreover, time-to-event analyses introduce right-censoring that bring additional challenges such as potential violations of proportional hazards after reweighting, competing-risk or trial-specific follow-up windows.

Finally, this full-IPD versus PAIC comparison could be further extended to investigate relative performance of regression-based outcome models, especially simulation-based STC and doubly robust methods combining trial-assignment and outcome modeling [110, 111, 18, 42].

Bibliography

- [1] Min-Hua Jen et al. “The Current Landscape of HTA Framework and Key Challenges”. In: *Statistics in Biopharmaceutical Research* 0.0 (), pp. 1–15. ISSN: null. DOI: 10.1080/19466315.2025.2495307.
- [2] *Pricing & Reimbursement of Drugs and HTA Policies in France*. Tech. rep. Haute Autorité de Santé, Mar. 2014.
- [3] An-Wen Chan et al. “SPIRIT 2025 Statement: Updated Guideline for Protocols of Randomised Trials”. In: *BMJ* 389 (Apr. 2025), e081477. ISSN: 1756-1833. DOI: 10.1136/bmj-2024-081477.
- [4] Sally Hopewell et al. “CONSORT 2025 Statement: Updated Guideline for Reporting Randomised Trials”. In: *BMJ* 389 (Apr. 2025), e081123. ISSN: 1756-1833. DOI: 10.1136/bmj-2024-081123.
- [5] *ICH E6 Good Clinical Practice - Scientific Guideline*. Tech. rep. International Council for Harmonisation, 2025.
- [6] Julia A. Beaver et al. “A 25-Year Experience of US Food and Drug Administration Accelerated Approval of Malignant Hematology and Oncology Drugs and Biologics: A Review”. In: *JAMA oncology* 4.6 (June 2018), pp. 849–856. ISSN: 2374-2445. DOI: 10.1001/jamaoncol.2017.5618.
- [7] Sundeep Agrawal et al. “Use of Single-Arm Trials for US Food and Drug Administration Drug Approval in Oncology, 2002-2021”. In: *JAMA Oncology* (Dec. 2022). ISSN: 2374-2437. DOI: 10.1001/jamaoncol.2022.5985.
- [8] *Reflection Paper on Establishing Efficacy Based on Single-Arm Trials Submitted as Pivotal Evidence in a Marketing Authorisation Application | Considerations on Evidence from Single-Arm Trials*. Tech. rep. European Medicines Agency, Sept. 2024.
- [9] Minyan Wang et al. “Single-Arm Clinical Trials: Design, Ethics, Principles”. In: *BMJ Supportive & Palliative Care* 15.1 (June 2024), e004984. ISSN: 2045-435X. DOI: 10.1136/spcare-2024-004984.
- [10] James E. Signorovitch et al. “Comparative Effectiveness without Head-to-Head Trials: A Method for Matching-Adjusted Indirect Comparisons Applied to Psoriasis Treatment with Adalimumab or Etanercept”. In: *PharmacoEconomics* 28.10 (2010), pp. 935–945. ISSN: 1179-2027. DOI: 10.2165/11538370-000000000-00000.

-
- [11] J. Jaime Caro and K. Jack Ishak. “No Head-to-Head Trial? Simulate the Missing Arms.” in: *PharmacoEconomics* 28.10 (Oct. 2010), pp. 957–967. ISSN: 1170-7690. DOI: 10.2165/11537420-000000000-00000.
- [12] K. Jack Ishak, Irina Proskorovsky, and Agnes Benedict. “Simulation and Matching-Based Approaches for Indirect Comparison of Treatments”. In: *PharmacoEconomics* 33.6 (June 2015), pp. 537–549. ISSN: 1170-7690, 1179-2027. DOI: 10.1007/s40273-015-0271-1.
- [13] David M. Phillippo et al. “Multilevel Network Meta-regression for Population-adjusted Treatment Comparisons”. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 183.3 (June 2020), pp. 1189–1210. ISSN: 0964-1998, 1467-985X. DOI: 10.1111/rssa.12579.
- [14] David M. Phillippo et al. “Validating the Assumptions of Population Adjustment: Application of Multilevel Network Meta-regression to a Network of Treatments for Plaque Psoriasis”. In: *Medical Decision Making* 43.1 (Jan. 2023), pp. 53–67. ISSN: 0272-989X. DOI: 10.1177/0272989X221117162.
- [15] DM Phillippo et al. *Methods for Population-Adjusted Indirect Comparisons in Submission to NICE*. Tech. rep. 18. University of Sheffield: Decision Support Unit, National Institute for Health and Care Excellence, Dec. 2016, p. 82.
- [16] Antonio Remiro Azócar. “Population-Adjusted Indirect Treatment Comparisons with Limited Access to Patient-Level Data”. PhD thesis. University College London, May 2021.
- [17] David Cheng, Eric Tchetgen Tchetgen, and James Signorovitch. “On the Double-Robustness and Semiparametric Efficiency of Matching-Adjusted Indirect Comparisons”. In: *Research Synthesis Methods* (Dec. 2022). ISSN: 1759-2887. DOI: 10.1002/jrsm.1616.
- [18] Julie E. Park et al. “Unanchored Population-Adjusted Indirect Comparison Methods for Time-to-Event Outcomes Using Inverse Odds Weighting, Regression Adjustment, and Doubly Robust Methods With Either Individual Patient or Aggregate Data”. In: *Value in Health* 27.3 (Mar. 2024), pp. 278–286. ISSN: 1098-3015. DOI: 10.1016/j.jval.2023.11.011.
- [19] Antonio Remiro-Azócar. “Two-Stage Matching-Adjusted Indirect Comparison”. In: *BMC Medical Research Methodology* 22.1 (Aug. 2022), p. 217. ISSN: 1471-2288. DOI: 10.1186/s12874-022-01692-9.
- [20] David M. Phillippo et al. “Population Adjustment Methods for Indirect Comparisons: A Review of National Institute for Health and Care Excellence Technology Appraisals”. In: *International Journal of Technology Assessment in Health Care* 35.03 (2019), pp. 221–228. ISSN: 0266-4623, 1471-6348. DOI: 10.1017/S0266462319000333.

- [21] Arnaud Serret-Larmande et al. “A Methodological Review of Population-Adjusted Indirect Comparisons Reveals Inconsistent Reporting and Suggests Publication Bias”. In: *Journal of Clinical Epidemiology* 163 (Nov. 2023), pp. 1–10. ISSN: 0895-4356, 1878-5921. DOI: 10.1016/j.jclinepi.2023.09.004.
- [22] Arnaud Serret-Larmande et al. *Assessing the Impact of Covariate Distribution and Positivity Violation on Weighting-Based Indirect Comparisons: A Simulation Study*. July 2025. DOI: 10.48550/arXiv.2507.12241. arXiv: 2507.12241 [stat].
- [23] David M. Phillippo et al. “Target Estimands for Efficient Decision Making: Response to Comments on ”Assessing the Performance of Population Adjustment Methods for Anchored Indirect Comparisons: A Simulation Study””. In: *Statistics in Medicine* 40.11 (May 2021), pp. 2759–2763. ISSN: 1097-0258. DOI: 10.1002/sim.8965.
- [24] Antonio Remiro-Azócar, Anna Heath, and Gianluca Baio. “Conflating Marginal and Conditional Treatment Effects: Comments on ”Assessing the Performance of Population Adjustment Methods for Anchored Indirect Comparisons: A Simulation Study””. In: *Statistics in Medicine* 40.11 (May 2021), pp. 2753–2758. ISSN: 1097-0258. DOI: 10.1002/sim.8857.
- [25] Antonio Remiro-Azócar. “Target Estimands for Population-Adjusted Indirect Comparisons”. In: *Statistics in Medicine* 41.28 (Dec. 2022), pp. 5558–5569. ISSN: 0277-6715, 1097-0258. DOI: 10.1002/sim.9413. arXiv: 2112.08023 [stat].
- [26] Guido W. Imbens and Joshua D. Angrist. “Identification and Estimation of Local Average Treatment Effects”. In: *Econometrica* 62.2 (Mar. 1994), p. 467. ISSN: 00129682. DOI: 10.2307/2951620. JSTOR: 2951620.
- [27] *Methods Guide for Health Technology Assessment*. Tech. rep. Canada’s Drug Agency, Mar. 2025.
- [28] *Methodological Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons*. Tech. rep. The Member State Coordination Group on Health Technology Assessment (HTACG), Mar. 2024.
- [29] Miguel A Hernán and James M Robins. “Estimating Causal Effects from Epidemiological Data”. In: *Journal of Epidemiology and Community Health* 60.7 (July 2006), pp. 578–586. ISSN: 0143-005X. DOI: 10.1136/jech.2004.029496.
- [30] Sander Greenland, James M. Robins, and Judea Pearl. “Confounding and Collapsibility in Causal Inference”. In: *Statistical Science* 14.1 (1999), pp. 29–46. ISSN: 0883-4237. JSTOR: 2676645.
- [31] Xavier de Luna and Per Johansson. “Testing for the Unconfoundedness Assumption Using an Instrumental Assumption”. In: *Journal of Causal Inference* 2.2 (Sept. 2014), pp. 187–199. ISSN: 2193-3685. DOI: 10.1515/jci-2013-0011.
- [32] David M. Eddy. “The Confidence Profile Method: A Bayesian Method for Assessing Health Technologies”. In: *Operations Research* 37.2 (Apr. 1989), pp. 210–228. ISSN: 0030-364X, 1526-5463. DOI: 10.1287/opre.37.2.210.

-
- [33] Julian P. T. Higgins and Anne Whitehead. “Borrowing Strength from External Trials in a Meta-Analysis”. In: *Statistics in Medicine* 15.24 (1996), pp. 2733–2749. ISSN: 1097-0258. DOI: 10.1002/(SICI)1097-0258(19961230)15:24<2733::AID-SIM562>3.0.CO;2-0.
- [34] Heiner C. Bucher et al. “The Results of Direct and Indirect Treatment Comparisons in Meta-Analysis of Randomized Controlled Trials”. In: *Journal of Clinical Epidemiology* 50.6 (June 1997), pp. 683–691. ISSN: 0895-4356. DOI: 10.1016/S0895-4356(97)00049-8.
- [35] *Wild-Type KRAS Is Required for Panitumumab Efficacy in Patients With Metastatic Colorectal Cancer* / *Journal of Clinical Oncology*.
- [36] Kendra K. Radtke et al. “Predicted Pharmacokinetic Interactions Between Hormonal Contraception and Single or Intermittently Dosed Rifampicin”. In: *The Journal of Clinical Pharmacology* 63.11 (2023), pp. 1283–1289. ISSN: 1552-4604. DOI: 10.1002/jcph.2303.
- [37] Giovanni Baranello et al. “Prognostic Factors and Treatment-Effect Modifiers in Spinal Muscular Atrophy”. In: *Clinical Pharmacology & Therapeutics* 110.6 (2021), pp. 1435–1454. ISSN: 1532-6535. DOI: 10.1002/cpt.2247.
- [38] Gary M. Clark. “Prognostic Factors versus Predictive Factors: Examples from a Clinical Trial of Erlotinib”. In: *Molecular Oncology* 1.4 (Apr. 2008), pp. 406–412. ISSN: 1574-7891. DOI: 10.1016/j.molonc.2007.12.001.
- [39] Harlan Campbell and Antonio Remiro-Azócar. *Doubly Robust Augmented Weighting Estimators for the Analysis of Externally Controlled Single-Arm Trials and Unanchored Indirect Treatment Comparisons*. Apr. 2025. DOI: 10.48550/arXiv.2505.00113. arXiv: 2505.00113 [stat].
- [40] Hans Fischer. *A History of the Central Limit Theorem : From Classical to Modern Probability Theory*. 2011. ISBN: 978-0-387-87856-0.
- [41] David M. Phillippo et al. “Equivalence of Entropy Balancing and the Method of Moments for Matching-adjusted Indirect Comparison”. In: *Research Synthesis Methods* 11.4 (July 2020), pp. 568–572. ISSN: 1759-2879, 1759-2887. DOI: 10.1002/jrsm.1416.
- [42] Landan Zhang, Sylwia Bujkiewicz, and Dan Jackson. “Four Alternative Methodologies for Simulated Treatment Comparison: How Could the Use of Simulation Be Re-Invigorated?” In: *Research Synthesis Methods* 15.2 (2024), pp. 227–241. ISSN: 1759-2887. DOI: 10.1002/jrsm.1681.
- [43] Sophie van Beekhuizen et al. “Indirect Treatment Comparisons in EUnetHTA Relative Effectiveness Assessments: Learnings and Recommendations for the Implementation of EU Joint Clinical Assessments”. In: *PharmacoEconomics - Open* (Apr. 2025). ISSN: 2509-4254. DOI: 10.1007/s41669-025-00575-1.
- [44] Bang Truong et al. “Population Adjusted-indirect Comparisons in Health Technology Assessment: A Methodological Systematic Review”. In: *Research Synthesis Methods* 14.5 (Sept. 2023), pp. 660–670. ISSN: 1759-2879, 1759-2887. DOI: 10.1002/jrsm.1653.

- [45] Anja Schiel. “Commentary on “Target Estimands for Population-Adjusted Indirect Comparisons””. In: *Statistics in Medicine* 41.28 (2022), pp. 5570–5572. ISSN: 1097-0258. DOI: 10.1002/sim.9517.
- [46] Estelle Russek-Cohen. “Discussion of “Target Estimands for Population-Adjusted Indirect Comparisons” by Antonio Remiro-Azocar”. In: *Statistics in Medicine* 41.28 (2022), pp. 5573–5576. ISSN: 1097-0258. DOI: 10.1002/sim.9533.
- [47] Antonio Remiro-Azócar et al. *Marginal and Conditional Summary Measures: Transportability and Compatibility across Studies*. Aug. 2025. DOI: 10.48550/arXiv.2507.21925. arXiv: 2507.21925 [stat].
- [48] Kelly Van Lancker, Frank Bretz, and Oliver Dukes. *The Use of Covariate Adjustment in Randomized Controlled Trials: An Overview*. June 2023. DOI: 10.48550/arXiv.2306.05823. arXiv: 2306.05823 [stat].
- [49] *Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products*. Tech. rep. Food and Drug Administration, Mon, 04/08/2024 - 11:18.
- [50] Daniela Golinelli et al. “Bias and Variance Trade-Offs When Combining Propensity Score Weighting and Regression: With an Application to HIV Status and Homeless Men”. In: *Health services & outcomes research methodology* 12.2-3 (June 2012), pp. 104–118. ISSN: 1387-3741. DOI: 10.1007/s10742-012-0090-1.
- [51] Bonnie E. Shook-Sa and Michael G. Hudgens. *Power and Sample Size for Marginal Structural Models*. Mar. 2020. DOI: 10.48550/arXiv.2003.05979. arXiv: 2003.05979 [stat].
- [52] David M. Phillippo et al. “Assessing the Performance of Population Adjustment Methods for Anchored Indirect Comparisons: A Simulation Study”. In: *Statistics in Medicine* 39.30 (Dec. 2020), pp. 4885–4911. ISSN: 0277-6715, 1097-0258. DOI: 10.1002/sim.8759.
- [53] Andrew J. Vickers and Gunnar Steineck. “Prognosis, Effect Modification, and Mediation”. In: *European urology* 74.3 (Sept. 2018), pp. 243–245. ISSN: 0302-2838. DOI: 10.1016/j.eururo.2018.05.027.
- [54] Antonio Remiro-Azócar. *Purely Prognostic Variables May Modify Marginal Treatment Effects for Non-Collapsible Effect Measures*. Oct. 2022. DOI: 10.48550/arXiv.2210.01757. arXiv: 2210.01757 [stat].
- [55] A. Laupacis, D. L. Sackett, and R. S. Roberts. “An Assessment of Clinically Useful Measures of the Consequences of Treatment”. In: *The New England Journal of Medicine* 318.26 (June 1988), pp. 1728–1733. ISSN: 0028-4793. DOI: 10.1056/NEJM198806303182605.
- [56] M. A. Hernan. “Causal Knowledge as a Prerequisite for Confounding Evaluation: An Application to Birth Defects Epidemiology”. In: *American Journal of Epidemiology* 155.2 (Jan. 2002), pp. 176–184. ISSN: 00029262. DOI: 10.1093/aje/155.2.176.

-
- [57] Sander Greenland, Judea Pearl, and James M. Robins. “Causal Diagrams for Epidemiologic Research”. In: *Epidemiology* 10.1 (Jan. 1999), p. 37. ISSN: 1044-3983.
- [58] Judea Pearl. “Causal Inference in Statistics: An Overview”. In: *Statistics Surveys* 3.none (Jan. 2009). ISSN: 1935-7516. DOI: 10.1214/09-SS057.
- [59] Karl D Ferguson et al. “Evidence Synthesis for Constructing Directed Acyclic Graphs (ESC-DAGs): A Novel and Systematic Method for Building Directed Acyclic Graphs”. In: *International Journal of Epidemiology* 49.1 (Feb. 2020), pp. 322–329. ISSN: 0300-5771. DOI: 10.1093/ije/dyz150.
- [60] Daniela Rodrigues et al. “Reflection on Modern Methods: Constructing Directed Acyclic Graphs (DAGs) with Domain Experts for Health Services Research”. In: *International Journal of Epidemiology* 51.4 (June 2022), pp. 1339–1348. ISSN: 0300-5771. DOI: 10.1093/ije/dyac135.
- [61] Wan Shakira Rodzlan Hasani et al. “Constructing Causal Pathways for Premature Cardiovascular Disease Mortality Using Directed Acyclic Graphs with Integrating Evidence Synthesis and Expert Knowledge”. In: *Scientific Reports* 14.1 (Nov. 2024), p. 28849. ISSN: 2045-2322. DOI: 10.1038/s41598-024-80091-0.
- [62] Brian C. Sauer et al. “A Review of Covariate Selection for Non-Experimental Comparative Effectiveness Research”. In: *Pharmacoepidemiology and Drug Safety* 22.11 (2013), pp. 1139–1145. ISSN: 1099-1557. DOI: 10.1002/pds.3506.
- [63] Center for Drug Evaluation and Research. *Developing Targeted Therapies in Low-Frequency Molecular Subsets of a Disease*. Apr. 2024.
- [64] Giorgio Scagliotti et al. “Treatment-by-Histology Interaction Analyses in Three Phase III Trials Show Superiority of Pemetrexed in Nonsquamous Non-Small Cell Lung Cancer”. In: *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer* 6.1 (Jan. 2011), pp. 64–70. ISSN: 1556-1380. DOI: 10.1097/JTO.0b013e3181f7c6d4.
- [65] Tien Hoang et al. “Does Histology Predict Survival of Advanced Non-Small Cell Lung Cancer Patients Treated with Platin-Based Chemotherapy? An Analysis of the Eastern Cooperative Oncology Group Study E1594”. In: *Lung cancer (Amsterdam, Netherlands)* 81.1 (July 2013), pp. 47–52. ISSN: 0169-5002. DOI: 10.1016/j.lungcan.2013.03.018.
- [66] Razvan Cristescu et al. “Tumor Mutational Burden Predicts the Efficacy of Pembrolizumab Monotherapy: A Pan-Tumor Retrospective Analysis of Participants with Advanced Solid Tumors”. In: *Journal for ImmunoTherapy of Cancer* 10.1 (Jan. 2022). ISSN: 2051-1426. DOI: 10.1136/jitc-2021-003091.
- [67] Bérengère Macabeo et al. “The Acceptance of Indirect Treatment Comparison Methods in Oncology by Health Technology Assessment Agencies in England, France, Germany, Italy, and Spain”. In: *PharmacoEconomics - Open* 8.1 (Jan. 2024), pp. 5–18. ISSN: 2509-4254. DOI: 10.1007/s41669-023-00455-6.

- [68] . IQWiG - *General Methods - Version 6.1*. Tech. rep. 6.1. Institute for Quality and Efficiency in Health Care, 2022.
- [69] Ioannis Katsoulis et al. “Perceptions of Indirect Treatment Comparisons as an Evidence Base in Oncology Decision-Making: Results of an International Survey of Health Technology Assessment and Payer Decision-Makers”. In: *Journal of Comparative Effectiveness Research* (Nov. 2024). DOI: 10.57264/cer-2024-0040.
- [70] *Practical Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons*. Tech. rep. The Member State Coordination Group on Health Technology Assessment (HTACG), Mar. 2024.
- [71] Tyler J. VanderWeele and Peng Ding. “Sensitivity Analysis in Observational Research: Introducing the E-Value”. In: *Annals of Internal Medicine* 167.4 (Aug. 2017), pp. 268–274. ISSN: 0003-4819, 1539-3704. DOI: 10.7326/M16-2607.
- [72] Samuel Aballéa et al. *Between Rigor and Relevance: Why the EU Hta Guidelines on Indirect Comparisons Miss the Mark*. June 2025. DOI: 10.20944/preprints202506.0161.v1.
- [73] “Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products”. In: (2023).
- [74] *Meta-Analyses of Randomized Controlled Clinical Trials to Evaluate the Safety of Human Drugs or Biological Products | FDA*. Tech. rep. U.S. Food & Drug Administration, Nov. 2018.
- [75] Shiro Tanaka et al. “A Targeted Review of Worldwide Indirect Treatment Comparison Guidelines and Best Practices”. In: *Value in Health* 27.9 (Sept. 2024), pp. 1179–1190. ISSN: 1098-3015. DOI: 10.1016/j.jval.2024.05.015.
- [76] *2020-2023 Value Assessment Framework*. Tech. rep. Institute for Clinical and Economic Review, Jan. 2020.
- [77] Dony Patel et al. “Use of External Comparators for Health Technology Assessment Submissions Based on Single-Arm Trials”. In: *Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research* 24.8 (Aug. 2021), pp. 1118–1125. ISSN: 1524-4733. DOI: 10.1016/j.jval.2021.01.015.
- [78] Daniel A. Ollendorf et al. “The Supercar Stays in the Garage: Factors Preventing Indirect Comparisons of Novel Medicines Targeting the Same Condition”. In: *Journal of Managed Care & Specialty Pharmacy* 26.3 (Mar. 2020), 10.18553/jmcp.2020.26.3.333. ISSN: 2376-0540. DOI: 10.18553/jmcp.2020.26.3.333.
- [79] Cecilia Menezes Farinasso et al. “Matching-Adjusted Indirect Comparison Studies in Oncology: A Scoping Review Focused on Reporting Quality”. In: *Value in Health Regional Issues* 47 (May 2025), p. 101088. ISSN: 2212-1099. DOI: 10.1016/j.vhri.2025.101088.

-
- [80] Owen Cassidy et al. “A Comparison of Relative-Efficacy Estimate(S) Derived From Both Matching-Adjusted Indirect Comparisons and Standard Anchored Indirect Treatment Comparisons: A Review of Matching-Adjusted Indirect Comparisons”. In: *Value in Health* 0.0 (July 2023). ISSN: 1098-3015, 1524-4733. DOI: 10.1016/j.jval.2023.07.001.
- [81] David M. Phillippo et al. *Multilevel Network Meta-Regression for General Likelihoods: Synthesis of Individual and Aggregate Data with Applications to Survival Analysis*. Jan. 2024. arXiv: 2401.12640 [stat].
- [82] Steve Kanters et al. “When Does the Use of Individual Patient Data in Network Meta-Analysis Make a Difference? A Simulation Study”. In: *BMC Medical Research Methodology* 21.1 (Jan. 2021), p. 21. ISSN: 1471-2288. DOI: 10.1186/s12874-020-01198-2.
- [83] Joy Leahy et al. “The Impact of Individual Patient Data in a Network Meta-Analysis: An Investigation into Parameter Estimation and Model Selection”. In: *Research Synthesis Methods* 9.3 (2018), pp. 441–469. ISSN: 1759-2887. DOI: 10.1002/jrsm.1305.
- [84] Emily C L Wong et al. “Matching-Adjusted Indirect Comparisons vs Propensity Score Matching with Individual Patient-level Data to Estimate Treatment Efficacy”. In: *Inflammatory Bowel Diseases* (Apr. 2023), izad077. ISSN: 1078-0998. DOI: 10.1093/ibd/izad077.
- [85] Ulrich Knief and Wolfgang Forstmeier. “Violating the Normality Assumption May Be the Lesser of Two Evils”. In: *Behavior Research Methods* 53.6 (2021), pp. 2576–2590. ISSN: 1554-351X. DOI: 10.3758/s13428-021-01587-5.
- [86] Jason W. Osborne and Elaine Waters. “Four assumptions of multiple regression that researchers should always test”. In: *Practical Assessment, Research, and Evaluation* 8.1 (Jan. 2002). ISSN: 1531-7714. DOI: 10.7275/r222-hv23.
- [87] G. E. P. BOX and G. S. WATSON. “Robustness to Non-Normality of Regression Tests †”. In: *Biometrika* 49.1-2 (Jan. 1962), pp. 93–106. ISSN: 0006-3444. DOI: 10.1093/biomet/49.1-2.93.
- [88] María J. Blanca et al. “Skewness and Kurtosis in Real Data Samples”. In: *Methodology* (Jan. 2013). ISSN: 1614-2241.
- [89] Roser Bono et al. “Non-Normal Distributions Commonly Used in Health, Education, and Social Sciences: A Systematic Review”. In: *Frontiers in Psychology* 8 (Sept. 2017). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2017.01602.
- [90] Theodore Micceri. “The Unicorn, the Normal Curve, and Other Improbable Creatures”. In: *Psychological Bulletin* 105.1 (1989), pp. 156–166. ISSN: 1939-1455. DOI: 10.1037/0033-2909.105.1.156.
- [91] Michael Köhler et al. “Information on New Drugs at Market Entry: Retrospective Analysis of Health Technology Assessment Reports versus Regulatory Reports, Journal Publications, and Registry Reports”. In: *BMJ* 350 (Feb. 2015), h796. ISSN: 1756-1833. DOI: 10.1136/bmj.h796.

- [92] Robert A. Smith, Paul P. Schneider, and Wael Mohammed. “Living HTA: Automating Health Economic Evaluation with R”. In: *Wellcome Open Research* 7 (2022), p. 194. ISSN: 2398-502X. DOI: 10.12688/wellcomeopenres.17933.2.
- [93] Ziren Jiang et al. “A Critical Assessment of Matching-Adjusted Indirect Comparisons in Relation to Target Populations”. In: *Research Synthesis Methods* 16.3 (May 2025), pp. 569–574. ISSN: 1759-2879, 1759-2887. DOI: 10.1017/rsm.2025.10.
- [94] Shanil Ebrahim et al. “Reanalyses of Randomized Clinical Trial Data”. In: *JAMA* 312.10 (Sept. 2014), pp. 1024–1032. ISSN: 0098-7484. DOI: 10.1001/jama.2014.9646.
- [95] David Henry and Tiffany Fitzpatrick. “Liberating the Data from Clinical Trials”. In: *BMJ (online)* 351 (Sept. 2015). DOI: 10.1136/bmj.h4601.
- [96] Harlan M. Krumholz and Eric D. Peterson. “Open Access to Clinical Trials Data”. In: *JAMA* 312.10 (Sept. 2014), pp. 1002–1003. ISSN: 0098-7484. DOI: 10.1001/jama.2014.9647.
- [97] Yusi Wei, Hande Y. Benson, and Muge Capan. *An Analytical Approach to Privacy and Performance Trade-Offs in Healthcare Data Sharing*. Aug. 2025. DOI: 10.48550/arXiv.2508.18513. arXiv: 2508.18513 [stat].
- [98] Sally Hollis et al. “Best Practice for Analysis of Shared Clinical Trial Data”. In: *BMC Medical Research Methodology* 16.1 (July 2016), pp. 15–22. ISSN: 1471-2288. DOI: 10.1186/s12874-016-0170-y.
- [99] Catrin Tudur Smith et al. “Sharing Individual Participant Data from Clinical Trials: An Opinion Survey Regarding the Establishment of a Central Repository”. In: *PLOS ONE* 9.5 (May 2014), e97886. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0097886.
- [100] Harlan M. Krumholz and Joanne Waldstreicher. “The Yale Open Data Access (YODA) Project — A Mechanism for Data Sharing”. In: *New England Journal of Medicine* 375.5 (Aug. 2016), pp. 403–405. ISSN: 0028-4793. DOI: 10.1056/NEJMp1607342.
- [101] National Academy of Medicine et al. “CASE STUDY: VIVLI”. In: *Sharing Health Data: The Why, the Will, and the Way Forward*. National Academies Press (US), 2022.
- [102] Jeran Stratford et al. “Best Practices for Clinical Trials Data Harmonization and Sharing on NHLBI bioData Catalyst (BDC) Learned from CONNECTS Network COVID-19 Studies”. In: *Journal of Clinical and Translational Science* 9 (Mar. 2025), pp. 1–22. DOI: 10.1017/cts.2025.52.
- [103] Stan Ahalt et al. “Building a Collaborative Cloud Platform to Accelerate Heart, Lung, Blood, and Sleep Research”. In: *Journal of the American Medical Informatics Association* 30.7 (2023), pp. 1293–1300.
- [104] *External Guidance on the Implementation of the European Medicines Agency Policy 0070 on the Publication of Clinical Data for Medicinal Products for Human Use*. Tech. rep. Domenico Scarlattiilaan 6 - 1083 HS Amsterdam - The Netherlands: European Medicine Agency, May 2025, p. 84.

-
- [105] Health Canada. *Public Release of Clinical Information: Guidance Document*. Guidance. Nov. 2021.
- [106] Center for Drug Evaluation and Research. *Study Data for Submission to CDER and CBER*. Tech. rep. FDA, Aug. 2024.
- [107] *Guidance on the Scoping Process*. Tech. rep. 1.0. Member State Coordination Group on Health Technology Assessment, Nov. 2024.
- [108] I. Buchholz et al. “PT28 Lost in PICO? A Simulation of the EU HTA Scoping Process”. In: *Value in Health* 26.12 (Dec. 2023), S392. ISSN: 1098-3015, 1524-4733. DOI: 10.1016/j.jval.2023.09.2054.
- [109] *Comparative Evidence for EU HTA*. Tech. rep. IQVIA, Apr. 2025.
- [110] Harlan Campbell et al. *Estimating Marginal Treatment Effects from Observational Studies and Indirect Treatment Comparisons: When Are Standardization-Based Methods Preferable to Those Based on Propensity Score Weighting?* Oct. 2023. DOI: 10.48550/arXiv.2301.09661. arXiv: 2301.09661 [stat].
- [111] Antonio Remiro-Azócar, Anna Heath, and Gianluca Baio. *Methodological Considerations for Novel Approaches to Covariate-Adjusted Indirect Treatment Comparisons*. May 2023. arXiv: 2305.08651 [stat].