Research Synthesis Methods (2025), **00**: 1–15
doi:10.1017/rsm.2025.10059

Research Synthesis Methods

RESEARCH ARTICLE

The hazards of using hazard ratios from proportional hazard models in indirect treatment comparisons

Ziren Jiang[1], Jialing Liu[1], Weili He[2], Joseph Cappelleri[3], Satrajit Roychoudhury[3], Yong Chen[4,5] and Haitao Chu[1,3]

[1]Division of Biostatistics and Health Data Science, University of Minnesota Twin Cities, Minneapolis, USA
[2]Medical Affairs and Health Technology Assessment Statistics, Data and Statistical Sciences, AbbVie Inc, USA
[3]Data Sciences and Analytics, Pfizer Inc, USA
[4]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, USA
[5]The Center for Health AI and Synthesis of Evidence (CHASE), University of Pennsylvania, Philadelphia, PA, USA

Corresponding author: Haitao Chu; Email: chux0051@umn.edu

Received: 31 March 2025; **Revised:** 17 October 2025; **Accepted:** 22 October 2025

Keywords: indirect treatment comparison; matching-adjusted indirect comparison; population-adjusted indirect comparison; statistical transitivity; time-to-event outcome

Abstract

Indirect treatment comparison (ITC) is widely used to estimate the comparative effectiveness of treatments when head-to-head trials are unavailable. For the typical scenario of anchored ITC where one trial compares drug A to drug C (AC trial) and another compares drug B to drug C (BC trial), the comparative effectiveness of drugs A versus B is calculated by subtracting (or dividing) the relative treatment effect of A versus C in the AC trial by that of B versus C in the BC trial, assuming the covariate distributions in both trials are balanced. This operation is valid only if the chosen effect measure is transitive, that is, in a three-arm randomized trial of drugs A, B, and C, the direct treatment effect of A versus B equals the indirect treatment effect of A versus B through their comparisons to C. For survival outcomes, many ITCs use the hazard ratio (HR) as the effect measure. In this article, we demonstrate that HR is generally not transitive and should be used with caution. As more reliable alternatives, we recommend effect measures with better transitivity properties: the restricted mean survival time (RMST) difference, the landmark survival probability difference (or ratio) at a prespecified time point, and the average hazard with survival weights (AH-SW) difference.

Highlights

What is already known?

- ITC, including population-adjusted indirect comparison (PAIC) and network meta-analysis, are widely used to estimate the comparative effectiveness of interventions when direct head-to-head trials are unavailable.
- HR are commonly used as effect measures in ITCs involving time-to-event outcomes.

What is new?

- We introduce the concept of *statistical transitivity* as a key criterion for evaluating effect measures in ITC and highlight its importance in interpreting comparative effectiveness.
- Through theoretical proofs and practical examples, we demonstrate the risks of using HRs in ITC, especially when the proportional hazard assumption does not hold across all treatment arms.

Potential impact for RSM readers

- We caution against the indiscriminate use of HRs in ITC for time-to-event outcomes and recommend alternative measures with better transitivity properties, such as RMST, landmark survival probability, or the AH-SW differences or ratios.
- We present a decision-making flowchart for survival outcome ITCs and emphasize explicit evaluation of both clinical and statistical transitivity to assess validity.

1. Introduction

In medical product development, there are often multiple approved treatments for the same disease condition. Comparing their relative effectiveness within specific populations is crucial, particularly for securing reimbursement from public or private sectors, where a drug must be evaluated against the standard of care (SoC) through a health technology assessment (HTA).[1] While randomized controlled trials (RCTs) are the gold standard for treatment comparisons, it is often impractical to conduct head-to-head trials for all available drugs. As an alternative, indirect treatment comparison (ITC) methods leverage existing trial data to compare treatments.

In a typical anchored indirect comparison of drug A versus drug B, a common comparator C such as a placebo, is present in both trials (referred to as the AC and BC trials). Given the potential imbalance of effect modifiers, ITC requires balancing the covariate distributions across the two trial populations. However, due to data availability constraints, individual participant data (IPD) are often available for only one trial, while the other provides only aggregate-level data (AgD). To address this issue, researchers have developed the population-adjusted indirect comparison (PAIC) methods, which account for covariate distribution differences between the AC and BC trials. Among various PAIC approaches,[2,3] matching-adjusted indirect comparison (MAIC) has gained popularity in the HTA submissions.[4,5] In an anchored indirect comparison of drug A versus drug B, MAIC estimates a set of balancing weights for the AC trial (where IPD is available) to ensure the weighted population matches the BC trial population based on summary statistics from AgD. Comparative effectiveness is then calculated by subtracting or dividing the weighted relative effect of drug A versus drug C by the relative effect of drug B versus drug C.

An essential requirement for the validity of indirect comparison between drug A and B through a common comparator C is the *transitivity* of relative effects. This assumes that the true relative effect of drug A versus drug B equals the difference (or ratio) between the true relative effects of drug A versus drug C and drug B versus drug C. This assumption (sometimes described as *exchangeability*, *consistency*, or *similarity* across studies) has been well known in other ITC scenarios such as network meta-analysis.[6–11] Salanti[10] highlights two key requirements for the validity of transitivity in ITC. First, the drug C must be comparable in both the AC trial and BC trials. For instance, if drug C represents the current SoC at the time of the trials, differences in SoC due to advancements over time could violate transitivity if the trials are not conducted simultaneously. Second, all effect modifiers must be balanced between the AC trial and BC trial populations. Salanti et al.[9] provide a cautionary example showing that violations of transitivity, particularly differences in study populations, can fundamentally compromise network meta-analysis (NMA) results. Song et al.[6] systematically compared results from direct head-to-head meta-analyses with those from adjusted indirect comparisons and underscored that the validity of such comparisons relies on the similarity of the trial populations and interventions. Ades et al.[8] discussed the subtle differences between the network meta-analysis and pairwise meta-analysis and argued that their validity depends on the exchangeability (encompassing concepts of homogeneity, similarity, and consistency) of included trials. Cipriani et al.[12] underscores that the reliability of network meta-analysis depends critically on the assumption that studies are comparable in all important respects and provides guidance for evaluating, documenting, and, where necessary, adjusting for differences to support valid indirect comparisons. Donegan et al.[13] emphasize that systematically evaluating

https://doi.org/10.1017/rsm.2025.10059 Published online by Cambridge University Press

and transparently reporting population similarity are essential for ensuring the reliability of indirect comparisons in evidence synthesis.

Existing literature discussed the transitivity assumption primarily from a scientific or clinical perspective, particularly focused on the similarity and consistency of the interventions, study populations, or effect measures. In this article, we further examine the transitivity assumption from a statistical perspective. Specifically, for a relative effect to be transitive, the *effect measure* itself must be a statistically transitive measurement. In other words, the validity of the ITC depends not only on the specific scientific background, but also on the selection of statistical measurement. In Section 2, we give a formal definition of statistical transitivity in the context of ITC and offer insights into its implications.

This article focuses on anchored ITC for time-to-event outcomes. The hazard ratio (HR)[14] is one of the most popular measures for comparing the efficacy of two drugs in survival analysis. In the comparison of drug A versus drug C, the Cox model treats the baseline hazard (i.e., the hazard for drug C) as a nuisance parameter and only estimates the HR between the two drugs. Our literature review indicates that most existing studies use the HR as the metric for ITC when dealing with time-to-event outcomes. For example, Aouni et al.[15] compared different matching strategies and penalty factors for MAIC, using the ratio of HR as the indirect comparison measure. Remiro-Azocar et al.[16] and Weber et al.[17] conducted comprehensive simulation studies to compare various indirect comparison methods, while both of which used the log HR as the effect measure. Leahy and Walsh[18] proposed using MAIC in Bayesian network meta-analysis for covariate adjustment with an HR model. Park et al.[19] introduced a doubly robust approach for indirect comparison with time-to-event outcomes, also using HR as the comparison metric. Beyond MAIC, HR has also been widely adopted as the primary comparison metric in network meta-analysis.[20–24]

While we acknowledge and appreciate these contributions to ITC literature, this cautionary note highlights that the HR is not a transitive measure and should be used with caution for ITC. Alternatively, we propose using the restricted mean survival time (RMST) difference, the landmark survival probability difference (or ratio) or the average hazard with survival weights (AH-SW) difference as more appropriate measures.

2. The statistical transitivity of an effect measurement

In an anchored ITC, transitivity refers to the property that the comparative effectiveness of drugs A versus B can be inferred through a common comparator C, that is, $\mu_{AB} = \mu_{AC} - \mu_{BC}$ (or $\mu_{AB} = \mu_{AC}/\mu_{BC}$ for a ratio effect measure) where μ_{AB} represents the comparative effectiveness measurement (e.g., mean difference, risk difference, or risk ratio) between drug A and drug B.[10] However, the validity of transitivity in an anchored indirect comparison relies on several factors: 1) the common comparator drug C must be consistent (i.e., similar) in both the AC and BC trials, 2) all effect modifiers must be balanced (after adjustment) in the populations of the AC and BC trials, and 3) the effect measurement itself must be statistically transitive (the focus of this article).

The first requirement can be addressed by carefully examining the design and timing of both trials. For a detailed discussion, readers are referred to Salanti.[10] Various ITC approaches[3,25–28] have been proposed to address the second requirement, typically assuming that all relevant effect modifiers are included in the adjustment. For a comprehensive review of MAIC approaches, readers can refer to Jiang et al.[29]

This article primarily focuses on the third requirement: the statistical transitivity of an effect measurement. To illustrate this concept, consider a randomized three-arm trial that includes drugs A, B, and C. First, we define two key concepts in this context: the direct comparison and the indirect comparison. The direct comparison of drug A versus drug B, denoted as μ^D_{AB}, is calculated directly from the outcome data of drugs A and B using metrics such as risk difference, HR, or other measures. In contrast, the indirect comparison, denoted as μ^I_{AB}, is derived by contrasting the effects of drug A versus drug C and drug B versus drug C (i.e., $\mu^I_{AB} = \mu^D_{AC} - \mu^D_{BC}$ or $\mu^I_{AB} = \mu^D_{AC}/\mu^D_{BC}$). Notably, a three-arm trial allows for both direct and indirect comparisons of drug A versus drug B. The indirect comparison

https://doi.org/10.1017/rsm.2025.10059 Published online by Cambridge University Press

in a three-arm trial mirrors the ITC by hypothetically creating the AC and BC trials, ensuring that drug C is identical in both trials.

The primary rationale for ITC is to estimate the comparative effectiveness of two treatments when direct comparison results are unavailable. Therefore, we ideally expect the direct comparison of drug A versus drug B to match the indirect comparison of drug A versus drug B with a common comparator drug C. This leads to the following definition of the statistical transitivity of a measurement.

Definition: (Statistical transitivity of an effect measurement).

A comparative effect measurement, denoted as $\mu\left(\{Y_i^A\}_{i=1}^{n_A}, \{Y_i^B\}_{i=1}^{n_B}\right)$, where $\{Y_i^A\}_{i=1}^{n_A}$ represents the data from drug A and $\{Y_i^B\}_{i=1}^{n_B}$ represents the data from drug B, is considered transitive in an ITC if, in the three-arm trial scenario, the direct comparison of drug A versus drug B is identical to the indirect comparison of drug A versus drug B. For a ratio effect measure, this means that $\mu\left(\{Y_i^A\}_{i=1}^{n_A}, \{Y_i^B\}_{i=1}^{n_B}\right) = \frac{\mu(\{Y_i^A\}_{i=1}^{n_A}, \{Y_i^C\}_{i=1}^{n_C})}{\mu(\{Y_i^B\}_{i=1}^{n_B}, \{Y_i^C\}_{i=1}^{n_C})}$ for arbitrary data of drugs, A, B, and C.

Note that the indirect comparison of drug A versus drug B can also be expressed as the difference (instead of the ratio) of two treatment effects, in which case, $\mu\left(\{Y_i^A\}_{i=1}^{n_A}, \{Y_i^B\}_{i=1}^{n_B}\right) = \mu\left(\{Y_i^A\}_{i=1}^{n_A}, \{Y_i^C\}_{i=1}^{n_C}\right) - \mu\left(\{Y_i^B\}_{i=1}^{n_B}, \{Y_i^C\}_{i=1}^{n_C}\right)$, depending on the chosen effect measure. It is evident that many effect measurements used in ITC are transitive under this definition. For example, the risk/mean difference $\mu\left(\{Y_i^A\}_{i=1}^{n_A}, \{Y_i^C\}_{i=1}^{n_C}\right) = \frac{1}{n_A}\sum_{i=1}^{n_A} Y_i^A - \frac{1}{n_B}\sum_{i=1}^{n_B} Y_i^B$ is a transitive measurement. In Section 3, we will show that the HR is not a transitive measurement for time-to-event outcomes and thus should be used with caution in ITC.

We emphasize that the transitivity of measurements is crucial in ITCs. Transitivity refers to the property that allows us to integrate over the common comparator C when comparing drugs A and B indirectly, provided all other assumptions are met. If a measurement is transitive, then the information from comparator C can be completely accounted for, assuming that 1) drug C is similar in both the AC and BC trials, and 2) there is no difference in the distribution of effect modifiers in the populations of the AC and BC trials.

However, if a measurement is not statistically transitive, then the choice of comparator C can affect the result of indirect comparison of A versus B, even in the ideal scenario of a randomized three-arm trial. Therefore, we recommend that indirect comparisons be performed only using statistically transitive measurements whenever possible.

3. HR should be used with caution in indirect treatment comparisons

3.1. HR derived from the Cox Proportional Hazard (PH) model is not a transitive measurement: an illustrative example

In this section, we focus on the transitivity of the HR within the context of ITC. Consider a three-arm trial with arms A, B, and C. Denote the hazard functions for drugs A, B, and C as $h_A(t)$, $h_B(t)$, and $h_C(t)$, respectively. The "model free" HR $\lambda_{AB}(t)$ at each specific time point t can be defined as $\lambda_{AB}(t) = \frac{h_A(t)}{h_B(t)}$. It should be noted that the HR function, $\lambda(t)$, is a transitive measurement if each hazard function is estimated independently (e.g., using the Kaplan–Meier estimator):

$$\widehat{\lambda}_{AB}(t) = \frac{\widehat{h}_A(t)}{\widehat{h}_B(t)} = \frac{\widehat{h}_A(t)}{\widehat{h}_C(t)} \Big/ \frac{\widehat{h}_B(t)}{\widehat{h}_C(t)} = \frac{\widehat{\lambda}_{AC}(t)}{\widehat{\lambda}_{BC}(t)}$$

for any t. However, the HR function $\lambda_{AB}(t)$ is hard to interpret and thus rarely reported in real clinical trials. Instead, it is common to assume $\lambda(t)$ does not vary with time t (i.e., the Cox PH assumption) and

https://doi.org/10.1017/rsm.2025.10059 Published online by Cambridge University Press

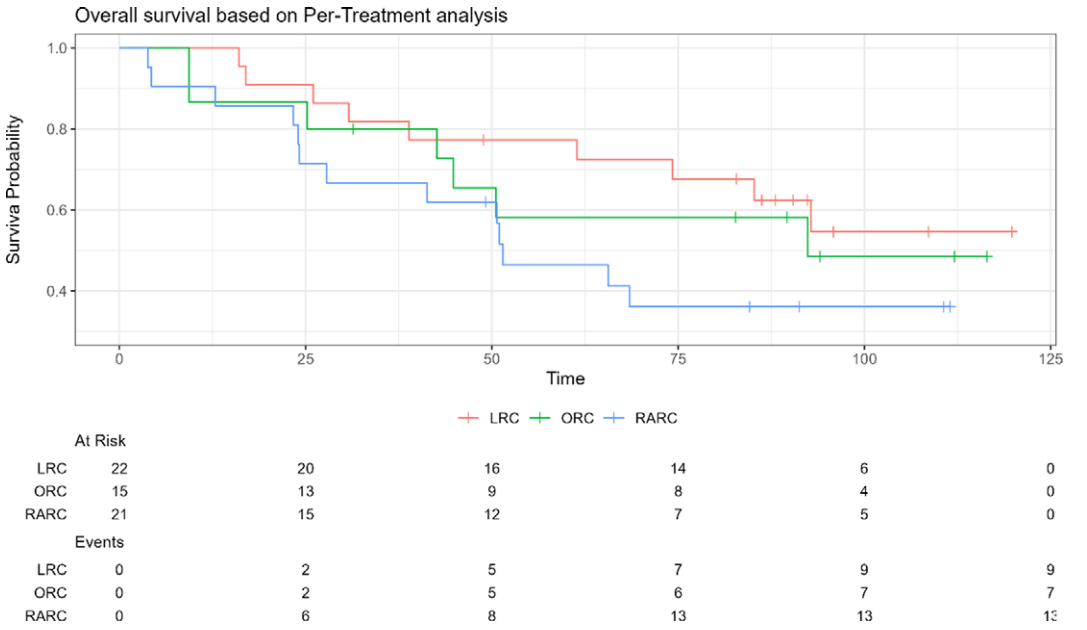

Figure 1. *Kaplan–Meier survival curves for the illustrative example of radical cystectomy. ORC, LRC, and RARC refer to the open radical cystectomy, laparoscopic radical cystectomy, and robotic-assisted radical cystectomy.*

defines this common ratio as HR λ. In the remainder of this article, "hazard ratio" refers to this constant λ rather than the time-varying $\lambda(t)$.

We begin by demonstrating that the HR λ is not a transitive measurement through an illustrative example involving bladder cancer. Open radical cystectomy (ORC) is widely regarded as the gold standard treatment for patients with muscle-invasive bladder cancer. Recent advancements in medical technology have introduced minimally invasive techniques, such as laparoscopic radical cystectomy (LRC) and robotic-assisted radical cystectomy (RARC).[30] In this example, we use data from a randomized three-arm trial comparing ORC, LRC, and RARC in terms of the overall survival outcomes.[31] The IPD for the time-to-event outcome is reconstructed from the published Kaplan–Meier survival curves using the R package IPDfromKM.[32] Figure 1 displays the reconstructed Kaplan–Meier survival curves. All the analyses are performed using R version 4.3.2.[33]

Denote RARC as drug A, ORC as drug B, and LRC as drug C. We calculate the pairwise HR for A versus B, A versus C, and B versus C using the *coxph* function in the R package *survival*.[34] The estimated pairwise log HRs are $\widehat{\mu}^D_{AC} = 0.701$, $\widehat{\mu}^D_{BC} = 0.247$, and $\widehat{\mu}^D_{AB} = 0.394$, which are consistent with the published results. It is evident that, in this three-arm trial, the direct comparison (log HR) of drug A versus B, $\widehat{\mu}^D_{AB} = 0.394$, does not equal the indirect comparison of A versus B through drug C, because $\widehat{\mu}^I_{AB} = \widehat{\mu}^D_{AC} - \widehat{\mu}^D_{BC} = 0.701 - 0.247 = 0.454$. This discrepancy indicates that neither the log of the HR nor the HR itself is a transitive measurement for ITC. Based on this illustrative example, we have the following observations:

Property 1 (HR from Cox PH model is not a statistically transitive measurement). HR does not satisfy the definition of a transitive measurement for ITC as outlined in Section 2 and, therefore, is not a transitive measurement.

Property 1 shows that in a randomized three-arm trial, the indirect HR does not always match the direct HR, indicating that the HR λ under the Cox proportional hazard model is not inherently transitive.

https://doi.org/10.1017/rsm.2025.10059 Published online by Cambridge University Press

3.2. The expectation of a HR does not maintain transitivity

One might suspect that this discrepancy is simply due to sampling variability, implying that the two HRs could be identical with an infinitely large sample size. However, as we will show in the next theorem, even the expected values of the HRs are non-transitive if the proportional hazards assumption does not hold among drugs A, B, and C. Therefore, the indirect comparison of HR is not an unbiased estimator if the proportional hazards assumption is violated, which also depends on the common comparator C.

Theorem 2 (Transitivity of HR under expectation). *Let $h_A(t), h_B(t)$, and $h_C(t)$ be the hazard function for treatments A, B, and C, respectively, where we assume that they have the same follow-up time $t \in [0, T]$. Denote the sample size for each treatment group as n_A, n_B,and n_C. Additionally, denote the expectation of the estimated HR of treatment A versus treatment B under the Cox proportional hazards model as* HR_{AB} *(with* HR_{AC} and HR_{BC} *defined similarly). Then, the expectation of the HR is transitive in the following manner:*

$$\mathrm{HR}_{AB} = \mathrm{HR}_{AC}/\mathrm{HR}_{BC}$$

for any n_A, n_B,and n_C, if and only if *the proportional hazard assumption is satisfied among the three arms, that is,*

$$h_A(t) = \alpha h_B(t) = \beta h_C(t)$$

for some scalars $\alpha, \beta > 0$.

Proof: See the Appendix for the proof.

In other words, the expected value of HR is transitive only when the proportional hazards assumption holds for all the three included groups. The proof of Theorem 2 further clarifies why the direct and indirect HRs may differ. Because the HR represents an average of the pointwise HRs over time, the influence of drug C is not fully neutralized in the indirect comparison. Consequently, the direct and indirect comparisons yield different values. Essentially, the detailed information embedded in the baseline hazard functions of the A-C and B-C comparisons cannot be cancelled out, rendering the HR a non-transitive measurement in ITC.

3.3. A simulated paradoxical example for using the HR in ITC

In this subsection, we present a simulated data example to illustrate how a paradox can arise between the indirect and direct HRs when the proportional hazards assumption is violated. Specifically, in a three-arm trial, the direct HR comparing drug A with drug B favors drug B, while the indirect HR, derived through drug C as a common comparator, suggests an advantage for drug A.

To demonstrate this, consider three drugs A, B, and C with corresponding hazard functions $h_1(t), h_2(t)$, and $h_3(t)$ where $t \in [0, 12]$. For drug A and drug C, we assume proportional hazards with $h_1(t) = 0.2$ and $h_3(t) = 0.8$ for all $t \in [0, 12]$. For drug B, we assume the hazard function changes with $h_2(t) = 0.1$ for $t \in [0, 3]$ and $h_2(t) = 0.35$ for $t \in (3, 12]$. Using the hazard functions, we simulate $n = 2,000$ subjects for each group (we select a large sample size to ensure that all the estimated relative effects are not impacted by random variability) and estimate the HRs $\widehat{\mathrm{HR}}_{AB}$, $\widehat{\mathrm{HR}}_{AC}$, and $\widehat{\mathrm{HR}}_{BC}$. The estimated direct HRs are $\mu^D_{AB} = \widehat{\mathrm{HR}}_{AB} = 0.896$ with the corresponding 95% confidence interval$(0.840, 0.956)$, suggesting that drug A is statistically significant worse than drug B. However, for the indirect HR, we have $\mu^I_{AB} = \frac{\widehat{\mathrm{HR}}_{AC}}{\widehat{\mathrm{HR}}_{BC}} = 0.238/0.184 = 1.296$ with a 95% confidence interval of $(1.117, 1.440)$, suggesting that drug A is significantly better than drug B. The corresponding R code for simulating the data is included in the Supplementary Material.

This example highlights the issue of non-transitivity when using HRs in ITC. When direct and indirect comparisons can yield conflicting conclusions, it raises concerns about the reliability of ITC

https://doi.org/10.1017/rsm.2025.10059 Published online by Cambridge University Press

results. In this case, while the hazards for drugs A and C remain constant and follow a proportional pattern, the hazard for drug B changes over time. As demonstrated in Section 3.2, a key driver of this paradoxical phenomenon is the violation of the proportional hazards assumption. However, in real-world applications of ITC, individual patient data (IPD) for the BC trial are often unavailable, making it challenging to assess the validity of this assumption. As illustrated in this example, even if the proportional hazards assumption holds in the AC trial, it does not necessarily prevent the occurrence of such paradoxical results. This highlights the need for caution when interpreting ITC findings.

3.4. Different follow-up times in the ITC

In the previous subsections, we focused on the transitivity issue under the assumption of a common follow-up time for all three drugs in a hypothetical three-arm trial. We demonstrated through Property 1 and Theorem 2 that the HR is not a transitive measure and should be used with caution in ITC. However, there is another important issue related to the HRs that can affect the validity of ITC, which arises from differences in follow-up time between trials.

The HR depends on follow-up time: As demonstrated by previous studies (e.g., Hernan[35]), the HR is highly sensitive to the duration of follow-up. The estimated HR is essentially a weighted average of the time-dependent HR over the entire follow-up period. This means that the value of HR can change depending on the length of the follow-up time considered. This matters for ITC because the AC and BC trials often have different follow-up periods. Consequently, the C arm in each trial is observed over different time windows, producing HR estimates that are not directly comparable, even if drug C is identical across trials and populations are well balanced.

The discrepancy in follow-up times further disrupts the transitivity of the HR in ITC. If follow-up time is different in the AC trial compared to the BC trial, then the HR derived from the AC trial (comparing drug A with drug C) and the HR from the BC trial (comparing drug B with drug C) may not be directly comparable. Even if drug C is the same in both trials, the HR estimates in both trials will be influenced by the respective follow-up times. Thus, if HRs are used in ITC (despite the concerns in previous sections), they should be estimated over a common follow-up period across the AC and BC trials. This can be done by digitizing the KM survival curve in the BC trial and re-estimating the HR during a comparable follow-up time period.

4. Alternative measurements for ITC with time-to-event outcomes

As demonstrated in Section 3, the HR is not a statistically transitive measurement for ITC. This limitation highlights the need for alternative approaches to compare treatments in a time-to-event context. In this section, we recommend three alternative measurements, RMST and landmark survival probability difference (or ratio), and the average hazard with survival weights difference (or ratio) which are fully transitive and easy to compute. We illustrate how these measurements can be used in the scenario of matching-adjusted treatment comparison (MAIC), though they are generally applicable for other ITC approaches.

4.1. Notation and setup for anchored MAIC

Let us consider an anchored MAIC, focusing on the comparison between drugs A and B with a common comparator drug C. We assume that researchers only have the individual-level data for AC trial: $\left\{\boldsymbol{X}_i^t, T_i^t, \delta_i^t\right\}_{i=1}^{n_t}$, where $t = A, C$ indicate the treatment (control) allocation; ,$i = 1, \ldots, n_t$ the index of trial participants; $\boldsymbol{X}_i^t$ is the vector of all effect modifiers that need to be adjusted; T_i^t is the right-censored event time; δ_i^t is the censoring indicator with $\delta_i^t = 0$ if T_i^t is censored and $\delta_i^t = 1$ if T_i^t corresponds to an event. For the BC trial, researchers only have the published summary-level data $\left\{\widehat{\boldsymbol{X}}^t, \widehat{\mu}^{BC(BC)}\right\}$ where $\widehat{\boldsymbol{X}}^t, t = B, C$ denote the sample mean of the covariates in group $t = B, C$

https://doi.org/10.1017/rsm.2025.10059 Published online by Cambridge University Press

and $\widehat{\mu}^{BC(BC)}$ be the estimated treatment effect of drug B versus drug C in the population of BC trial. For time-to-event outcomes, the summary-level data usually includes published Kaplan–Meier survival curves with sample size information (number of individuals at risk). If the BC trial does not report the RMST (or landmark survival probability) difference, but does report Kaplan–Meier survival curves, one can easily digitalize the KM curves and calculate the corresponding measurement based on the reconstructed IPD. For further details on how to digitize a Kaplan–Meier curve and calculate the corresponding measurements, one can refer to Liu et al.,[32] which provides a comprehensive guide, along with an R package for digitizing KM curves.

MAIC aims to estimate a set of balancing weights $\boldsymbol{w} = \left\{w_1^A, \ldots, w_{n_A}^A, w_1^C, \ldots, w_{n_C}^C\right\}$, with $\sum_{\boldsymbol{i}} w_i^A = 1$, and $\sum_{\boldsymbol{i}} w_i^C = 1$, for each participant in the AC trial such that the weighted population of treatment and control groups in the AC trial aligns with that in the BC trial in terms of the reported sample mean $\widehat{\boldsymbol{X}}^B$ and $\widehat{\boldsymbol{X}}^C$, that is, $\sum_{i=1}^{n_A} w_i^A \boldsymbol{X}_i^A = \widehat{\boldsymbol{X}}^B$ and $\sum_{i=1}^{n_C} w_i^C \boldsymbol{X}_i^C = \widehat{\boldsymbol{X}}^C$. There is also an alternative matching strategy which, instead of matching the treated groups and control group separately, matches the entire AC trial population with the entire BC trial population. The weights can be estimated through various approaches, including the original MAIC with the method of moments, the MAIC with the largest effective sample size,[28] and two-stage MAIC method,[27] see Jiang et al.[29] for a comprehensive review of the approaches. Under the assumption that all effect modifiers are included and that the correlations between covariates are similar between the AC trial and BC trial, the weighted IPD can be used to generate the comparative effectiveness of drug A versus drug C in the population of the BC trial.

4.2. Matching-adjusted indirect comparison with restricted mean survival time difference

The RMST[36] is a model-free measurement that reflects the average survival time up to a prespecified fixed follow-up time. Compared to the HR, RMST is more flexible as it does not require the proportional hazards assumption. RMST is also easier to interpret clinically, since it directly relates to the average survival time within a fixed time interval.[36–38] As RMST represents the area under the survival curve up to a specified time, the RMST difference of drug B versus drug C (and its corresponding standard error) can be easily calculated using the reconstructed IPD from the reported Kaplan–Meier survival curve. For estimating the RMST difference of drug A versus drug C using the weighted IPD with weights $\boldsymbol{w} = \left\{w_1^A, \ldots, w_{n_A}^A, w_1^C, \ldots, w_{n_C}^C\right\}$ estimated from one of the MAIC methods, the RMST can be calculated as the area under the estimated weighted Kaplan–Meier survival curve.[39–41]

Suppose the event in the drug A's arm occurs at D distinct times $t_1^A < t_2^A < \cdots < t_D^A$, then the weighted Kaplan–Meier estimator of the survival function can be expressed as

$$\widehat{S}^A(t) = \begin{cases} 1, & \text{if } t < t_1^A, \\ \prod_{t_j^A \le t} \left(1 - \frac{\tilde{\theta}_j^A}{\theta_j^A}\right), & \text{otherwise,} \end{cases}$$

where $\tilde{\theta}_j^A = \sum_{i:T_i^A = t_j^A} w_i^A \delta_i^{AC}$ and $\theta_j^A = \sum_{i:T_i^A > t_j^A} w_i^A$ be the weighted number of events and the weighted number of individuals at risk for drug A at time t_j^A. Then, the weighted RMST of drug A with threshold time τ can be calculated as the area under $\widehat{S}^A(t)$ with $t \le \tau$, that is, $\widehat{\mu}^A(\tau) = \int_{t=0}^{\tau} \widehat{S}^A(t)dt$. The corresponding variance can be estimated either through nonparametric bootstrap method or using the formula provided in Conner et al.,[41] namely,

$$\widehat{V}(\widehat{\mu}) = \sum_{j:t_j^A \le \tau} \left[\sum_{i=j}^{\tau} \widehat{S}^A(t_i)(t_{i+1} - t_i)\right]^2 \frac{\tilde{\theta}_j^A}{M_j^A\left(\theta_j^A - \tilde{\theta}_j^A\right)},$$

https://doi.org/10.1017/rsm.2025.10059 Published online by Cambridge University Press

where $M_j^A = \frac{\sum_{i:T_i^A \geq t_j^A} w_i^A}{\sum_{i:T_i^A \geq t_j^A} \left(w_i^A\right)^2}$. Similarly, the weighted RMST of drug C $\widehat{\mu}^C(\tau)$ can be estimated using $\widehat{S}^C(t)$ and the RMST difference of drug A versus drug C in the population of the BC trial is defined as

$$\widehat{\mu}^{AC(BC)}(\tau) = \widehat{\mu}^A(\tau) - \widehat{\mu}^C(\tau),$$

and the indirect comparison of the RMST of drug A versus drug B in the population of the BC trial can be derived as the difference between $\widehat{\mu}^{AC(BC)}(\tau)$ and $\widehat{\mu}^{BC(BC)}(\tau)$.

4.3. MAIC with landmark survival probability difference

Landmark survival probability refers to the survival probabilities $S(t)$ for groups of patients at specific time points, or landmarks t, during a follow-up period. For each prespecified time point, the landmark survival probability serves as a transitive measurement of the survival outcome. Based on the estimated survival function $\widehat{S}(t)$ introduced in the previous section, the landmark survival probability can be calculated for any treatment arm with any time point $t < t_{\max}$, where $t_{\max}$ be the maximum time point with the observed event. The corresponding variance of $\widehat{S}(t)$ can be estimated using the variance formula in Xie and Liu.[40]

4.4. Matching-adjusted indirect comparison with the average hazard with survival weights difference (or ratio)

Uno and Horiguchi[42] recently proposed an alternative measurement that summarizes the group-specific hazard information, which they termed as the AH-SW. The AH-SW ratio for a given threshold time τ is defined as

$$\eta_k(\tau) = \frac{F_k(\tau)}{R_k(\tau)},$$

where $k \in \{A, B, C\}$, $F_k(\tau)$ is the cumulative density function at time τ, and $R_k(\tau)$ is the RMST with threshold time τ. AH-SW represents a person-time incidence rate that is independent of random censoring. Uno and Horiguchi[42] proposed to estimate $\eta_k(\tau)$ by $\widehat{\eta}_k(\tau) = \frac{\widehat{F}_k(\tau)}{\widehat{R}_k(\tau)}$ where $\widehat{F}_k(\tau)$ and $\widehat{R}_k(\tau)$ are estimated nonparametrically through the Kaplan–Meier estimator. By estimating $\eta_k(\tau)$ through the nonparametric Kaplan–Meier estimator, $\widehat{\eta}_k(\tau)$ is clearly a transitive measurement. Since we introduced the variance estimation for $\widehat{F}_k(\tau)$ and $\widehat{R}_k(\tau)$ under MAIC weights, the variance of $\widehat{\eta}_k(\tau)$ can either be calculated analytically or through bootstrap.

4.5. Revisit the ORC example

Here, we revisit the illustrative ORC example and illustrate how to compute the weighted RMST/landmark survival probability and the corresponding variance. As the data are from a three-arm randomized trial, we randomly simulate the weights for each subject following a log-normal distribution. Then, we estimate the weighted Kaplan–Meier survival curve using the simulated weights. Figure 2 displays the estimated KM curve. The weighted RMST and its standard error are estimated using the "akm.rmst" function provided by Conner et al.[41] The estimated RMST after adjustment with $\tau = 70$ are 57.17 for LRC group (with standard error 8.50), 60.75 for the ORC group (with standard error 8.47), and 50.44 for the RARC group (with standard error 9.99). The landmark survival probability at $t = 70$ is 0.689 for LRC (with standard error 0.203), 0.752 for ORC (with standard error 0.169), and 0.269 for the RARC group (with standard error 0.166). It is clear to see that both RMST and landmark survival probability are transitive, since

$$\mu^{AB} = \mu^A - \mu^B = \left(\mu^A - \mu^C\right) - \left(\mu^B - \mu^C\right) = \mu^{AC} - \mu^{BC},$$

where μ is either the RMST or the landmark survival probability.

https://doi.org/10.1017/rsm.2025.10059 Published online by Cambridge University Press

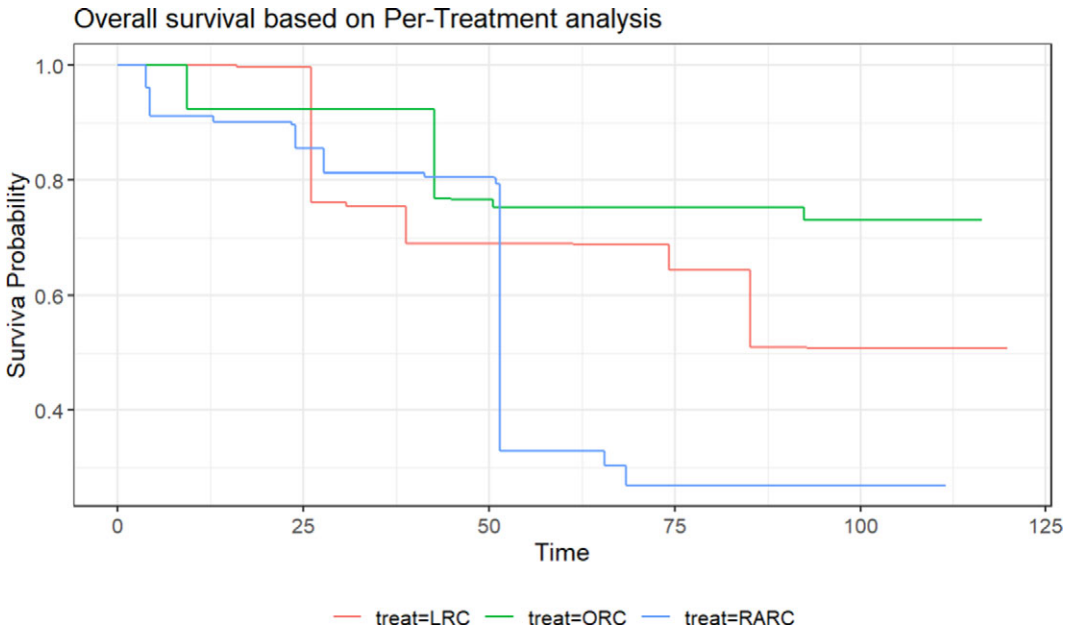

Figure 2. *Weighted Kaplan–Meier survival curve with randomly generated weights.*

It should be noted that RMST and landmark survival probability are transitive only if they are estimated nonparametrically (i.e., without using the Cox PH model) which is the common approach for the estimation. In the Supplementary Code (see Supplementary Material), we illustrate that the RMST can be non-transitive if it is estimated by the Cox proportional hazard model.

5. Discussion

In this article, we highlight a critical issue with the use of HR in ITCs: the HR is not a statistically transitive measurement. This discrepancy primarily arises due to the differences in how the baseline hazard functions are estimated in the direct and indirect comparisons. Specifically, the Cox proportional hazards model, which is commonly used to estimate HRs, treats the baseline hazard function as a nuisance parameter. Because of this, the baseline hazard functions are implicitly different in each pairwise comparison (A vs. B, A vs. C, and B vs. C). As a result, the indirect comparison of drugs A and B can be affected by the choice of the common comparator C. Theorem 2 demonstrated that if the proportional hazards assumption holds for all three drugs (A, B, and C) in the three-arm trial, the expected value of the estimated HR would be transitive. This means that the HRs calculated through direct and indirect comparisons would align. However, in practice, ensuring that the proportional hazards assumption holds across both the AC and BC trials is difficult. This assumption cannot be assessed using aggregate data alone, especially when individual patient data (IPD) is not available for the BC trial.

Even if both the AC and BC trials have tested the assumption (i.e., failed to reject the PH assumption), the HR should still be used with caution because the power of the PH test is likely very low.[43] Indeed, Stensrud and Hernán[44] argued that the proportional hazard assumption seldom holds in medical studies. In addition to its non-transitivity, the HR also suffers from *non-collapsibility*; its value can change when conditioning on additional covariates, even in the absence of confounding,[45] and *lack causal interpretability*, as it represents the average of instantaneous relative risk among individuals who remain at risk, a quantity influenced by the evolving composition of the risk set over time. Given these limitations, we recommend that the HR be used with caution in ITC.

https://doi.org/10.1017/rsm.2025.10059 Published online by Cambridge University Press

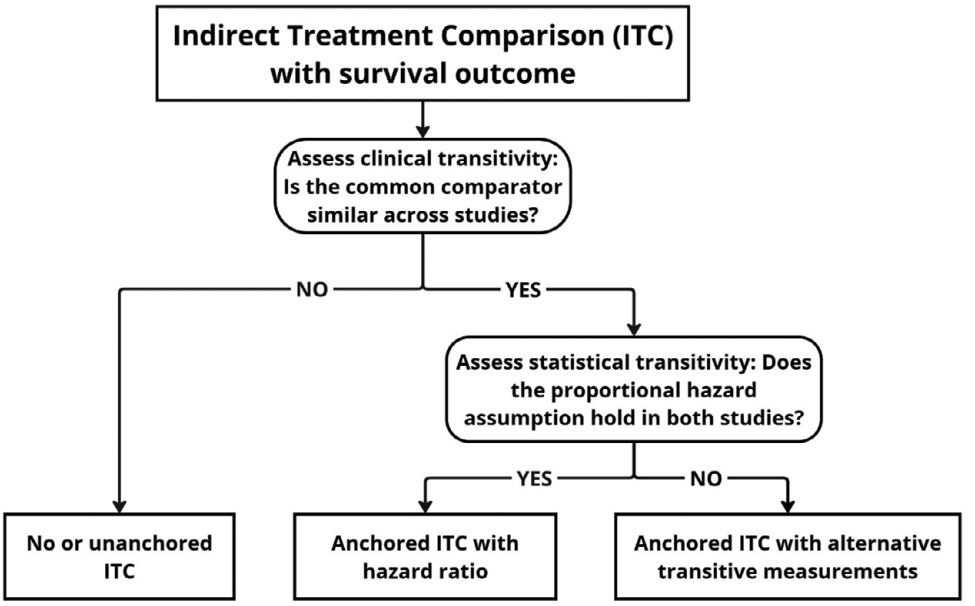

Figure 3. *Decision-making flowchart for performing indirect treatment comparison with survival outcomes.*

While RMST and landmark survival probability difference (or ratio) offer alternatives to the HR, they are not without limitations. As discussed in the literature, RMST can only summarize survival information up to a specific time point, typically the smallest follow-up time across the groups. This limitation means that when comparing drug A and drug B using RMST in a scenario where the AC trial follows up for 4 years and the BC trial follows up for 3 years, only data for the first 3 years can be included in the comparison. This results in the exclusion of potentially valuable information from the AC trial, especially if the treatment effects extend beyond the shorter follow-up period.

Another limitation of RMST is that it is a nonparametric measure, and thus, it cannot inform about survival beyond the observed study period. This makes RMST unsuitable for long-term analyses, especially when policymakers or healthcare require evidence for extended timeframes. For this reason, we suggest that when performing ITC, multiple measures—RMST, landmark survival probability, and HR—should be considered together. These measures are relatively easy to compute when individual patient data can be reconstructed from Kaplan–Meier curves.

Beyond this manuscript's focus on statistical transitivity, clinical transitivity is also frequently overlooked in ITC practice. Truong et al.[46] found that only 4.9% of population-adjusted ITCs explicitly evaluated whether the common comparator was consistent across trials. Accordingly, we propose a decision-making flowchart (Figure 3) for ITC with survival outcomes: First, assess clinical transitivity by verifying the comparability of the common comparator across studies; if satisfied, assess statistical transitivity before using the HR. When the proportional hazards assumption fails in either trial, alternative metrics with better transitivity properties—such as RMST differences (or ratios), landmark survival probability differences (or ratios), or other measures (e.g., AH-SW difference or ratio) — should be considered. Given the limitations of any single metric, we recommend sensitivity analyses using alternative measures regardless of the primary choice, because they can provide complementary insights.

This article primarily focuses on anchored indirect comparison where there exists a common comparator group in two separate trials. For unanchored indirect comparisons, survival outcomes can be directly compared between drugs A and B after adjusting for covariates. In this case, HRs can be calculated by weighting the IPD from trial A and the reconstructed IPD from trial B. However, as

https://doi.org/10.1017/rsm.2025.10059 Published online by Cambridge University Press

pointed out by Phillippo et al.,[1] conducting an unanchored indirect comparison requires including all prognostic variables and effect modifiers that influence the outcome, which can be very challenging in practice. Therefore, we also strongly recommend performing an anchored indirect comparison whenever possible, as it offers a more robust and reliable approach.

Author contributions. Conceptualization: Z.J., H.C.; Investigation: Z.J., J.L., W.H., J.C., S.R., Y.C., H.C.; Methodology: Z.J., J.L., H.C.; Resources: W.H., J.C., S.R., Y.C., H.C.; Supervision: W.H., J.C., S.R., Y.C., H.C.; Validation: J.L., W.H., J.C., S.R., Y.C., H.C.; Visualization: Z.J., J.L.; Writing—original draft: Z.J., J.L.; Writing—review and editing: Z.J., J.L., W.H., J.C., S.R., Y.C., H.C.

Competing interest statement. W.H. is employed by Abbvie; J.C., S.R., and H.C. are employed by Pfizer. They may own stocks of their company. However, all of the contents in this manuscript are strictly educational, instructive, and methodological, not involving any real medicinal intervention. The remaining authors declare no competing interests.

Data availability statement. The data used in the illustrative example were digitalized from the cited paper using the R IPDfromKM[32] package.

Funding statement. The authors declare that no specific funding has been received for this article.

Supplementary material. To view supplementary material for this article, please visit http://doi.org/10.1017/rsm.2025.10059.

References

[1] Phillippo D, Ades T, Dias S, Palmer S, Abrams KR, Welton N. NICE DSU technical support document 18: methods for population-adjusted indirect comparisons in submissions to NICE. Published 2016. https://sheffield.ac.uk/media/34216/download.

[2] Caro JJ, Ishak KJ. No head-to-head trial? Simulate the missing arms. *PharmacoEconomics*. 2010;28(10): 957–967. https://doi.org/10.2165/11537420-000000000-00000.

[3] Signorovitch JE, Wu EQ, Yu AP, et al. Comparative effectiveness without head-to-head trials: a method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept. *PharmacoEconomics*. 2010;28(10): 935–945. https://doi.org/10.2165/11538370-000000000-00000.

[4] Phillippo DM, Dias S, Elsada 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): 221–228. https://doi.org/10.1017/S0266462319000333.

[5] Jiang Z, Liu J, Alemayehu D, et al. A critical assessment of matching-adjusted indirect comparisons in relation to target populations. *Res Synth Methods*. 2025: 1–6. https://doi.org/10.1017/rsm.2025.10.

[6] Song F. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *BMJ*. 2003;326(7387): 472–472. https://doi.org/10.1136/bmj.326.7387.472.

[7] Glenny A, Altman D, Song F, et al. Indirect comparisons of competing interventions. *Health Technol Assess*. 2005;9(26). https://doi.org/10.3310/hta9260.

[8] Ades AE, Welton NJ, Dias S, Phillippo DM, Caldwell DM. Twenty years of network meta-analysis: continuing controversies and recent developments. *Res Synth Methods*. 2024;15(5): 702–727. https://doi.org/10.1002/jrsm.1700.

[9] Salanti G, Marinho V, Higgins JPT. A case study of multiple-treatments meta-analysis demonstrates that covariates should be considered. *J Clin Epidemiol*. 2009;62(8): 857–864. https://doi.org/10.1016/j.jclinepi.2008.10.001.

[10] Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. *Res Synth Methods*. 2012;3(2): 80–97. https://doi.org/10.1002/jrsm.1037.

[11] Jansen JP, Naci H. Is network meta-analysis as valid as standard pairwise meta-analysis? It all depends on the distribution of effect modifiers. *BMC Med*. 2013;11(1). https://doi.org/10.1186/1741-7015-11-159.

[12] Cipriani A, Higgins JPT, Geddes JR, Salanti G. Conceptual and technical challenges in network meta-analysis. *Ann Intern Med*. 2013;159(2): 130–137. https://doi.org/10.7326/0003-4819-159-2-201307160-00008.

[13] Donegan S, Williamson P, Gamble C, Tudur-Smith C. Indirect comparisons: a review of reporting and methodological quality. *PLoS ONE*. 2010;5(11):e11054. https://doi.org/10.1371/journal.pone.0011054.

[14] Cox DR. Regression models and life-tables. *J Royal Stat Soc Ser B*. 1972;34(2): 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.

[15] Aouni J, Gaudel-Dedieu N, Sebastien B. Matching-adjusted indirect comparisons: application to time-to-event data. *Stat Med*. 2021;40(3): 566–577. https://doi.org/10.1002/sim.8789.

[16] Remiro-Azócar A, Heath A, Baio G. Methods for population adjustment with limited access to individual patient data: a review and simulation study. *Res Synth Methods*. 2021;12(6): 750–775.

[17] Weber D, Jensen K, Kieser M. Comparison of methods for estimating therapy effects by indirect comparisons: a simulation study. *Med Decis Mak*. 2020;40(5): 644–654. https://doi.org/10.1177/0272989X20929309.
[18] Leahy J, Walsh C. Assessing the impact of a matching-adjusted indirect comparison in a Bayesian network meta-analysis. *Res Synth Methods*. 2019;10(4): 546–568. https://doi.org/10.1002/jrsm.1372.
[19] Park JE, Campbell H, Towle K, 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. *Value Health*. 2024;27(3): 278–286. https://doi.org/10.1016/j.jval.2023.11.011.
[20] Nazarzadeh M, Bidel Z, Canoy D, et al. Blood pressure lowering and risk of new-onset type 2 diabetes: an individual participant data meta-analysis. *Lancet*. 2021;398(10313): 1803–1810. https://doi.org/10.1016/S0140-6736(21)01920-6.
[21] Efthimiou O, Taipale H, Radua J, et al. Efficacy and effectiveness of antipsychotics in schizophrenia: network meta-analyses combining evidence from randomised controlled trials and real-world data. *Lancet Psychiatry*. 2024;11(2): 102–111. https://doi.org/10.1016/S2215-0366(23)00366-8.
[22] Mastrantoni L, Chiaravalli M, Spring A, et al. Comparison of first-line chemotherapy regimens in unresectable locally advanced or metastatic pancreatic cancer: a systematic review and Bayesian network meta-analysis. *Lancet Oncol*. 2024;25(12): 1655–1665. https://doi.org/10.1016/S1470-2045(24)00511-4.
[23] Wankhede D, Yuan T, Kloor M, Halama N, Brenner H, Hoffmeister M. Clinical significance of combined tumour-infiltrating lymphocytes and microsatellite instability status in colorectal cancer: a systematic review and network meta-analysis. *Lancet Gastroenterol Hepatol*. 2024;9(7): 609–619. https://doi.org/10.1016/S2468-1253(24)00091-8.
[24] Zhao Y, He Y, Wang W, et al. Efficacy and safety of immune checkpoint inhibitors for individuals with advanced EGFR-mutated non-small-cell lung cancer who progressed on EGFR tyrosine-kinase inhibitors: a systematic review, meta-analysis, and network meta-analysis. *Lancet Oncol*. 2024;25(10): 1347–1356. https://doi.org/10.1016/S1470-2045(24)00379-6.
[25] Phillippo DM, Dias S, Ades AE, et al. Multilevel network meta-regression for population-adjusted treatment comparisons. *J Royal Stat Soc Ser A*. 2020;183(3): 1189–1210.
[26] Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. Methods for population-adjusted indirect comparisons in health technology appraisal. *Med Decis Mak*. 2018;38(2): 200–211. https://doi.org/10.1177/0272989X17725740.
[27] Remiro-Azócar A. Two-stage matching-adjusted indirect comparison. *BMC Med Res Methodol*. 2022;22(1): 1–16.
[28] Jackson D, Rhodes K, Ouwens M. Alternative weighting schemes when performing matching-adjusted indirect comparisons. *Res Synth Methods*. 2021;12(3): 333–346.
[29] Jiang Z, Cappelleri JC, Gamalo M, Chen Y, Thomas N, Chu H. A comprehensive review and shiny application on the matching-adjusted indirect comparison. *Res Synth Methods*. 2024: jrsm.1709. https://doi.org/10.1002/jrsm.1709.
[30] Bochner BH, Dalbagni G, Sjoberg DD, et al. Comparing open radical cystectomy and robot-assisted laparoscopic radical cystectomy: a randomized clinical trial. *Eur Urol*. 2015;67(6): 1042–1050. https://doi.org/10.1016/j.eururo.2014.11.043.
[31] Khan MS, Omar K, Ahmed K, et al. Long-term oncological outcomes from an early phase randomised controlled three-arm trial of open, robotic, and laparoscopic radical cystectomy (CORAL). *Eur Urol*. 2020;77(1): 110–118. https://doi.org/10.1016/j.eururo.2019.10.027.
[32] Liu N, Zhou Y, Lee JJ. IPDfromKM: reconstruct individual patient data from published Kaplan-Meier survival curves. *BMC Med Res Methodol*. 2021;21(1): 111.
[33] R Core Team. R: a language and environment for statistical computing. Published 2023. Available at https://www.R-project.org/. Accessed December 1, 2025.
[34] Therneau TM. A package for survival analysis in R. Published 2024. Available at https://CRAN.R-project.org/package=survival. Accessed December 1, 2025.
[35] Hernán MA. The hazards of hazard ratios. *Epidemiology*. 2010;21(1): 13–15. https://doi.org/10.1097/EDE.0b013e3181c1ea43.
[36] Royston P, Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Methodol*. 2013;13(1): 152. https://doi.org/10.1186/1471-2288-13-152.
[37] Tian L, Jin H, Uno H, et al. On the empirical choice of the time window for restricted mean survival time. *Biometrics*. 2020;76(4): 1157–1166. https://doi.org/10.1111/biom.13237.
[38] Jiang Z, Lu C, Liu J, et al. Nonconcurrent controls in platform trials: can we borrow their concurrent observation data? *Stat Biopharm Res*. 2023: 1–10. https://doi.org/10.1080/19466315.2023.2267502.
[39] Winnett A, Sasieni P. Adjusted Nelson–Aalen estimates with retrospective matching. *J Am Stat Assoc*. 2002;97(457): 245–256. https://doi.org/10.1198/016214502753479383.
[40] Xie J, Liu C. Adjusted Kaplan–Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Stat Med*. 2005;24(20): 3089–3110. https://doi.org/10.1002/sim.2174.
[41] Conner SC, Sullivan LM, Benjamin EJ, LaValley MP, Galea S, Trinquart L. Adjusted restricted mean survival times in observational studies. *Stat Med*. 2019;38(20): 3832–3860. https://doi.org/10.1002/sim.8206.
[42] Uno H, Horiguchi M. Ratio and difference of average hazard with survival weight: new measures to quantify survival benefit of new therapy. *Stat Med*. 2023;42(7): 936–952. https://doi.org/10.1002/sim.9651.
[43] Austin PC. Statistical power to detect violation of the proportional hazards assumption when using the Cox regression model. *J Stat Comput Simul*. 2018;88(3): 533–552. https://doi.org/10.1080/00949655.2017.1397151.

https://doi.org/10.1017/rsm.2025.10059 Published online by Cambridge University Press

[44] Stensrud MJ, Hernán MA. Why test for proportional hazards? *JAMA*. 2020;323(14): 1401. https://doi.org/10.1001/jama.2020.1267.
[45] Daniel R, Zhang J, Farewell D. Making apples from oranges: comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biom J*. 2021;63(3): 528–557. https://doi.org/10.1002/bimj.201900297.
[46] Truong B, Tran LT, Le TA, Pham TT, Vo T. Population adjusted-indirect comparisons in health technology assessment: a methodological systematic review. *Res Synth Methods*. 2023;14(5): 660–670. https://doi.org/10.1002/jrsm.1653.
[47] Schemper M, Wakounig S, Heinze G. The estimation of average hazard ratios by weighted Cox regression. *Stat Med*. 2009;28(19): 2473–2489. https://doi.org/10.1002/sim.3623.

Appendix

Theorem 2 (Transitivity of hazard ratio under expectation). *Let $h_A(t), h_B(t)$, and $h_C(t)$ be the hazard function for drugs A, B, and C, respectively, where $t \in [0, T]$ denotes time. Let the sample size for treatment groups be denoted as n_A, n_B,and n_C. The expectation of the estimated hazard ratio (HR) of drug A versus drug B under the Cox model is denoted as HR_{AB}, with HR_{AC} and HR_{BC} defined similarly. Then, the expectation of the HR follows a transitive property such that*

$$\mathrm{HR}_{AB} = \mathrm{HR}_{AC}/\mathrm{HR}_{BC}$$

for any values of n_A, n_B,and n_C, if and only if the hazard functions for three groups are proportional. Specifically, this means that $h_A(t) = \alpha h_B(t) = \beta h_C(t)$ where α and $\beta > 0$ are constants.

Proof. According to Schemper et al.,[47] under the nonproportional hazards, the Cox model estimates the average HR across all event times, with an expectation given by

$$\mathrm{HR}_{AB} = \frac{\int_0^T \frac{h_A(t)}{h_A(t)+h_B(t)} f_{AB}(t)dt}{\int_0^T \frac{h_B(t)}{h_A(t)+h_B(t)} f_{AB}(t)dt},$$

where $h_A(t)$ and $h_B(t)$ are the hazard functions for drugs A and B, and $f_{AB}(t)$ is the event density for the combined population of drugs A and B. Similarly, the corresponding relationships for other HRs can be expressed in a similar form. □

$$\mathrm{HR}_{AC} = \frac{\int_0^T \frac{h_A(t)}{h_A(t)+h_C(t)} f_{AC}(t)dt}{\int_0^T \frac{h_C(t)}{h_A(t)+h_C(t)} f_{AC}(t)dt},$$

and

$$\mathrm{HR}_{BC} = \frac{\int_0^T \frac{h_B(t)}{h_B(t)+h_C(t)} f_{BC}(t)dt}{\int_0^T \frac{h_C(t)}{h_B(t)+h_C(t)} f_{BC}(t)dt}.$$

It is important to note that the event density for the combined population of drugs A and B, $f_{AB}(t)$, depends on the relative sample sizes of treatment groups A and B. If n_A is changed while n_B remains fixed, the density $f_{AB}(t)$ will be different unless the hazard functions for both groups are equal $h_A(t) = h_B(t)$.

Now, assuming the proportional hazard assumption holds for the three treatment groups, $h_A(t) = \alpha h_B(t) = \beta h_C(t)$ for some constants $\alpha, \beta > 0$, we can derive the following:

$$\mathrm{HR}_{AB} = \frac{\int_0^T \frac{h_A(t)}{h_A(t)+h_B(t)} f_{AB}(t)dt}{\int_0^T \frac{h_B(t)}{h_A(t)+h_B(t)} f_{AB}(t)dt} = \frac{\int_0^T \frac{\alpha h_B(t)}{\alpha h_B(t)+h_B(t)} f_{AB}(t)dt}{\int_0^T \frac{h_B(t)}{\alpha h_B(t)+h_B(t)} f_{AB}(t)dt} = \alpha.$$

Similarly, $\mathrm{HR}_{AC} = \beta$ and $\mathrm{HR}_{BC} = \beta/\alpha$. Thus, we have the relation $\mathrm{HR}_{AB} = \mathrm{HR}_{AC}/\mathrm{HR}_{BC}$, which is independent of the sample sizes n_A, n_B, and n_C.

https://doi.org/10.1017/rsm.2025.10059 Published online by Cambridge University Press

To demonstrate the opposite direction of the assertion, assume $h_A(t)$ and $h_B(t)$ are not proportional, meaning that $h_A(t)/h_B(t)$ varies with time t. We will show that the relationship $\mathrm{HR}_{AB} = \mathrm{HR}_{AC}/\mathrm{HR}_{BC}$ cannot hold for all values of n_A, n_B,and n_C.

Assume that $\mathrm{HR}_{AB} = \mathrm{HR}_{AC}/\mathrm{HR}_{BC}$ holds for the sample sizes with the ratio $n_A : n_B : n_C = a : b : c$ with the corresponding densities f_{AB}, f_{AC}, and f_{BC}. Now, consider the scenario where the sample size ratio is changed to $n_A : n_B : n_C = a' : b : c$, resulting in new density functions f'_{AB} and f'_{AC}. Since the $h_A(t)$ and $h_B(t)$ are not proportional, the term $\frac{h_A(t)}{h_A(t)+h_B(t)}$ will change over time t. Therefore, for some value of n'_a, the following relationship will hold:

$$\mathrm{HR}_{AB} = \frac{\int_0^T \frac{h_A(t)}{h_A(t)+h_B(t)} f_{AB}(t)dt}{\int_0^T \frac{h_B(t)}{h_A(t)+h_B(t)} f_{AB}(t)dt} \neq \frac{\int_0^T \frac{h_A(t)}{h_A(t)+h_B(t)} f'_{AB}(t)dt}{\int_0^T \frac{h_B(t)}{h_A(t)+h_B(t)} f'_{AB}(t)dt} = \mathrm{HR}'_{AB}.$$

Now consider the new hazard ratio HR'_{AC} with the updated density function f'_{AC}. If $h_A(t)$ and $h_C(t)$ are proportional, HR'_{AC} will be the same as HR_{AC} as it is unaffected by the sample sizes. In this case, we have

$$\mathrm{HR}'_{AC}/\mathrm{HR}_{BC} = \mathrm{HR}_{AC}/\mathrm{HR}_{BC} = \mathrm{HR}_{AB} \neq \mathrm{HR}'_{AB}.$$

This shows that the HR is not transitive.

If $h_A(t)$ and $h_C(t)$ are not proportional, HR'_{AC} can differ from HR_{AC}. For the HR to be transitive, we would need $\mathrm{HR}'_{AC}/\mathrm{HR}_{BC} = \mathrm{HR}'_{AB}$. However, this cannot hold for all ratios of $a' \neq a$. To understand why, note that the density function $f_{AB}(t)$ is the mixed density of the combined populations of treatments A and B. A change in the sample size for treatment A will result in a similar change in both $f_{AB}(t)$ and $f_{AC}(t)$. However, since the ratios $h_A(t)/h_B(t)$ and $h_A(t)/h_C(t)$ are different over some intervals, the change in sample size n_A will certainly have different magnitudes of influence on HR'_{AC} and HR'_{AB}. As a result, the equation $\mathrm{HR}'_{AC}/\mathrm{HR}_{BC} = \mathrm{HR}'_{AB}$ cannot always hold.

Therefore, the only scenario in which the HR is transitive under expectation is when all treatment groups have proportional hazards.

Cite this article: Jiang Z, Liu J, He W, Cappelleri J, Roychoudhury S, Chen Y, Chu H. The hazards of using hazard ratios from proportional hazard models in indirect treatment comparisons. *Research Synthesis Methods*. 2025;00: 1–15. https://doi.org/10.1017/rsm.2025.10059