Journal of the Royal Statistical Society Series A: Statistics in Society, 2025, **00**, 1–20
https://doi.org/10.1093/jrsssa/qnaf169

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series A
Statistics in Society

Original Article

Multilevel network meta-regression for general likelihoods: synthesis of individual and aggregate data with applications to survival analysis

David M. Phillippo[1], **Sofia Dias**[2], **A. E. Ades**[1] and **Nicky J. Welton**[1]

[1]Bristol Medical School, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol BS8 2PS, UK
[2]Centre for Reviews and Dissemination, University of York, York YO10 5DD, UK

Address for correspondence: David M. Phillippo, Bristol Medical School, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol BS8 2PS, UK. Email: david.phillippo@bristol.ac.uk

Abstract

Network meta-analysis combines aggregate data (AgD) from multiple randomized controlled trials, assuming that any effect modifiers are balanced across populations. Individual participant data (IPD) meta-regression is the 'gold standard' method to relax this assumption, however IPD are frequently only available in a subset of studies. Multilevel network meta-regression (ML-NMR) extends IPD meta-regression to incorporate AgD studies whilst avoiding aggregation bias. However, implementation of this method so far has required the aggregate-level likelihood to have a known closed form, which has prevented application to time-to-event outcomes. We extend ML-NMR to individual-level likelihoods of any form, by integrating the individual-level likelihood function over the AgD covariate distributions to obtain the respective marginal likelihood contributions. We illustrate with two examples of time-to-event outcomes: modelling progression-free survival in newly diagnosed multiple myeloma using flexible baseline hazards with cubic M-splines, and a simulated comparison showing the performance of ML-NMR with little loss of precision from a full IPD analysis. Extending ML-NMR to general likelihoods, including for survival outcomes, greatly increases the applicability of the method. R and Stan code is provided, and the methods are implemented in the *multinma* R package.

Keywords: effect modification, indirect comparison, individual participant data, network meta-analysis, population adjustment

1 Introduction

Healthcare decision-making requires reliable estimates of the relative effectiveness of all relevant treatments in a given population. Standard indirect comparison and network meta-analysis methods are commonly used to synthesize evidence from multiple trials, each of which may compare only a subset of the treatments of interest, under the assumption that there is no imbalance in effect-modifying variables between the trials (Bucher et al., 1997; Dias et al., 2011b; Higgins & Whitehead, 1996; Lu & Ades, 2004). However, when effect modification is present these methods may be biased. The 'gold standard' approach to adjust for effect modifiers and relax this assumption is network meta-regression with individual participant data (IPD) available for all studies (Berlin et al., 2002; Dias et al., 2011a; Lambert et al., 2002; Riley et al., 2010). However, this level of data availability is rare—particularly in contexts such as health technology assessment where multiple treatments are of interest. Population adjustment methods have, therefore, been proposed that use IPD from the subset of studies where it is available, and published aggregate data (AgD)

Received: August 6, 2024. **Revised:** September 23, 2025. **Accepted:** September 24, 2025

from the rest (Phillippo et al., 2016, 2018). A substantial majority of applications of population adjustment analyses to date involve survival or time-to-event data; one recent review found that 72% of population adjustment analyses in technology appraisal submissions to the National Institute for Health and Care Excellence in England involved survival outcomes (Phillippo et al., 2019). However, the set of population adjustment methods that are currently applicable to survival outcomes are faced with significant limitations, and methods that address these limitations have not yet been extended to handle survival data.

Matching-adjusted indirect comparison (MAIC) is a widely used population adjustment method that re-weights individuals in one IPD study to match the covariate distribution in an AgD study (Ishak et al., 2015; Phillippo et al., 2016; Signorovitch et al., 2010). Since IPD are only available from one of the studies weights are typically estimated using the method of moments, which has been shown to be equivalent to an entropy-balancing approach (Phillippo et al., 2020c), although alternatives have been proposed (Jackson et al., 2020). While MAIC is currently the most widely used population adjustment method, including for survival outcomes (Phillippo et al., 2019), it is limited to a pairwise indirect comparison scenario with one IPD study and one AgD study and cannot readily be extended to incorporate larger networks of studies and treatments (Phillippo et al., 2016). Moreover, population-adjusted estimates can only be produced for the AgD study population; while this may be of interest for commercial reasons, this is not typically representative of the target population for a treatment decision (Phillippo et al., 2016).

Simulated treatment comparison (STC) is an alternative approach based on regression adjustment, where a regression model fitted in the IPD study is used to predict outcomes on each treatment from the IPD study in the AgD study population (Caro & Ishak, 2010; Ishak et al., 2015; Phillippo et al., 2016). The typical approach is to simply 'plug-in' the mean covariate values to produce predictions. However, when the model is nonlinear in the covariates this results in aggregation bias. Moreover, when the outcome measure is noncollapsible, such as hazard ratios or odds ratios, this results in bias due to combining incompatible conditional and marginal effect measures (Phillippo et al., 2021; Remiro-Azócar et al., 2021). Simulation can be used to avoid these biases (Caro & Ishak, 2010), however as originally proposed this incurs additional sampling variation by simulating a limited number of participants in the aggregate trial. A more sophisticated form of STC based on G-computation via simulation from the joint covariate distribution in the AgD study has been developed that addresses these issues, with variance estimation handled by bootstrapping or embedding in a Bayesian analysis (Remiro-Azócar et al., 2022). Similar simulation-based approaches have recently been published (Ren et al., 2024; Zhang et al., 2024). However, like MAIC, all of these approaches are only applicable to pairwise indirect comparisons and cannot produce estimates for target populations other than that represented by the AgD study.

Multilevel network meta-regression (ML-NMR) is a population adjustment method that extends IPD network meta-regression to incorporate evidence from both IPD and AgD sources (Phillippo, 2019; Phillippo et al., 2020a). Aggregation bias is avoided by integrating the individual-level model over the joint covariate distribution in the AgD studies, in contrast to previous meta-regression approaches (Donegan et al., 2013; Saramago et al., 2012; Sutton et al., 2008) that combine IPD and AgD by simply 'plugging in' mean covariate values from the AgD studies. Unlike MAIC and STC, ML-NMR can coherently synthesize evidence from networks of any size, and crucially for decision-making can produce population-adjusted estimates of relative or absolute effects in any target population of interest. Moreover, in larger networks, key assumptions regarding unobserved effect modifiers and effect modifier interactions can be assessed using ML-NMR, whereas these are untestable assumptions under all approaches when performing pairwise indirect comparisons (Phillippo et al., 2022). ML-NMR is an extension of the standard network meta-analysis (NMA) framework (Dias et al., 2011b; Higgins & Whitehead, 1996; Lu & Ades, 2004), reducing to IPD network meta-regression if IPD are available from all studies, and to AgD NMA when no covariates are included in the model. Phillippo et al. (2020a) construct the aggregate-level model for ML-NMR in two steps: (i) deriving the aggregate likelihood from the individual likelihood, using standard results on the sums of random variables and (ii) integrating the individual-level model over the covariate distribution in the aggregate population to form the aggregate-level model, using a general numerical approach based on quasi-Monte Carlo integration. However, derivation of the aggregate likelihood is not straightforward in general and may even be intractable, since analytic results for the sums of random variables are only available for some special cases (e.g. Normal, Poisson, or

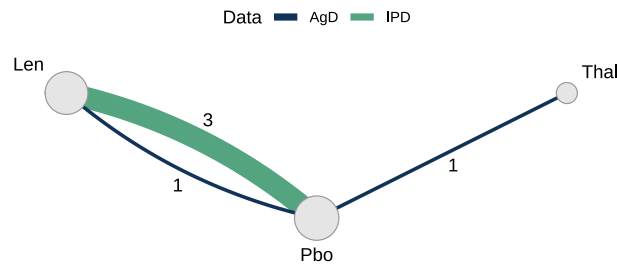

Figure 1. Network of five studies comparing lenalidomide or thalidomide to placebo for treatment of newly diagnosed multiple myeloma. IPD were available from three studies, and AgD from two studies. Edge widths and numbers indicate the number of studies making each comparison, and the size of each node corresponds to the number of individuals randomized to each treatment.

Bernoulli distributions, Phillippo et al., (2020a), or ordered categorical distributions Phillippo et al., 2022). Most notably, this is the case for the analysis of survival outcomes where the aggregate likelihood cannot be derived analytically. As it stands, therefore, ML-NMR cannot be applied to survival outcomes which is a major practical limitation of the method.

In this paper, we aim to address this limitation by extending the ML-NMR framework to individual-level likelihoods of any general form. We begin by describing a motivating example comparing maintenance treatments for newly diagnosed multiple myeloma. We then set out the ML-NMR framework in a general form based on the likelihood contributions from different sources of data. We directly integrate the individual-level likelihood function over the joint covariate distribution to obtain the likelihood contributions for the AgD studies. This approach does not require the form of the aggregate-level likelihood to be analytically tractable, or even known. We then use this approach to describe ML-NMR models for censored time-to-event outcomes with general survival and hazard functions. Finally, we demonstrate these ideas in practice: first with the newly diagnosed multiple myeloma example, and then with a simulated comparison showing performance against full IPD network meta-regression.

2 Example: newly diagnosed multiple myeloma

As a motivating example, we compare progression-free survival on lenalidomide vs. thalidomide maintenance treatment after autologous stem cell transplant (ASCT) for patients with newly diagnosed multiple myeloma (Leahy & Walsh, 2019). These treatments were not compared head-to-head in a single randomized controlled trial, but instead were both compared separately to placebo in five studies, forming the evidence network shown in Figure 1. IPD are available from three trials of lenalidomide vs. placebo, with only published AgD available from the thalidomide vs. placebo trial and one further lenalidomide trial.

Summaries of four clinically relevant baseline characteristics are given in Table 1: age, international staging system (ISS) stage (stage III vs. stage I–II), response post-ASCT (complete response or very good partial response vs. other), and sex (male or female). These covariates were considered to be potential effect modifiers in a previous analysis (Leahy & Walsh, 2019), and are not well-balanced across study populations which may lead to biased estimates of treatment effects if these are not accounted for.

This network was previously analysed by Leahy and Walsh (2019), who applied multiple MAIC analyses before combining in a NMA. However, there are several disadvantages with this approach: in particular, only the IPD studies are adjusted and the constancy of relative effects assumption is still required to combine the AgD studies, and estimates can only be produced for a weighted-average of the AgD study populations. An analysis using ML-NMR can address these issues, coherently synthesizing the available evidence whilst adjusting for effect modifiers, and producing estimates relevant to specific target populations of interest.

Table 1. Baseline characteristics of studies included in the ML-NMR analysis of progression-free survival after ASCT for newly diagnosed multiple myeloma

Study/treatment	Sample size	Age (years)	ISS stage III (%)	Response CR/VGPR (%)	Male (%)
Attal2012*					
Placebo	307	54.22 (5.24)	15.96	54.07	57.98
Lenalidomide	307	54.35 (6.06)	23.78	54.72	55.37
McCarthy2012*					
Placebo	229	57.39 (5.56)	18.34	71.18	55.46
Lenalidomide	231	57.93 (6.33)	27.27	62.34	52.38
Palumbo2014*					
Placebo	125	54.44 (8.98)	12.00	38.40	63.20
Lenalidomide	126	53.90 (9.69)	10.32	42.06	46.03
Jackson2019					
Placebo	864	64.63 (9.40)	19.21	83.10	62.15
Lenalidomide	1,137	65.17 (8.94)	24.80	82.59	61.65
Morgan2012					
Placebo	410	63.92 (9.01)	36.34	71.71	61.95
Thalidomide	408	65.59 (8.38)	31.86	74.51	61.52

Note. Statistics are mean and standard deviation for the continuous covariate age, and percent for the categorical covariates. *Individual participant data available.

3 ML-NMR for general likelihoods

Consider a network of J randomized controlled trials, each investigating a subset $\mathscr{K}_j$ of K treatments. If IPD are available from each of the J studies, then we can estimate a standard IPD network meta-regression model, which may be written as

$$y_{ijk} \sim \pi_{\text{Ind}}(\theta_{ijk}), \tag{1a}$$

$$g(\theta_{ijk}) = \eta_{jk}(\boldsymbol{x}_{ijk}) = \mu_j + \boldsymbol{x}_{ijk}^{\mathrm{T}}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \gamma_k, \tag{1b}$$

with IPD outcomes y_{ijk} for individuals $i = 1, \ldots, N_{jk}$ in study $j = 1, \ldots, J$ receiving treatment $k \in \mathscr{K}_j$ given the likelihood distribution $\pi_{\text{Ind}}(\theta_{ijk})$. The link function $g(\cdot)$ links the likelihood parameter θ_{ijk} to the linear predictor $\eta_{jk}(\boldsymbol{x}_{ijk})$, with covariates $\boldsymbol{x}_{ijk}$. The parameters μ_j are study-specific intercepts, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_{2,k}$ are regression coefficients for prognostic and effect modifying covariates, respectively, and γ_k are individual-level treatment effects. We set $\boldsymbol{\beta}_{2,1} = \gamma_1 = 0$ for the reference treatment 1.

By specifying an individual-level model (1), with a likelihood, link function, and linear predictor, we are also specifying an individual-level likelihood function, conditional on the covariate values for each individual. Letting $\boldsymbol{\xi}$ denote the set of all model parameters $\{\mu_j, \boldsymbol{\beta}_1, \boldsymbol{\beta}_{2,k}, \gamma_k : \forall\, j, k\}$, we denote the individual conditional likelihood function by $L^{\text{Con}}_{ijk\,|\,\boldsymbol{x}}(\boldsymbol{\xi}; y_{ijk}, \boldsymbol{x}_{ijk})$. The form of this individual conditional likelihood function follows from the chosen individual-level model.

To extend the IPD network meta-regression model (1) into a ML-NMR model that incorporates evidence from AgD studies, we integrate the individual conditional likelihood function over the joint covariate distribution in an AgD study to obtain an individual *marginal* likelihood function, describing the likelihood where individual outcomes are known but individual covariates are not (only summary covariate distributions). Integrating the individual conditional likelihood function over the joint covariate distribution $f_{jk}(\cdot)$ on treatment k in study j, we obtain the individual marginal likelihood function

$$L^{\text{Mar}}_{ijk}(\boldsymbol{\xi}; y_{ijk}) = \int_{\mathfrak{X}} L^{\text{Con}}_{ijk\,|\,\boldsymbol{x}}(\boldsymbol{\xi}; y_{ijk}, \boldsymbol{x}) f_{jk}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}, \tag{2}$$

which no longer depends on $\boldsymbol{x}$. In other words, for an individual on treatment k in study j with outcome y_{ijk}, if we do not know their individual covariates $\boldsymbol{x}_{ijk}$ but only the distribution $f_{jk}(\cdot)$, their likelihood contribution is given by (2). This integration may be performed using quasi-Monte Carlo integration, as described previously (Phillippo et al., 2020a). With a set of $\tilde{N}$ integration points $\tilde{\boldsymbol{x}}_{jk}$ drawn from $f_{jk}(\cdot)$, the individual marginal likelihood function (2) is evaluated as

$$L^{\text{Mar}}_{ijk}(\boldsymbol{\xi}; y_{ijk}) \approx \tilde{N}^{-1} \sum_{\tilde{\boldsymbol{x}}} L^{\text{Con}}_{ijk\,|\,\boldsymbol{x}}(\boldsymbol{\xi}; y_{ijk}, \boldsymbol{x}). \tag{3}$$

In practice, it is likely that only marginal covariate summaries are available from the AgD studies instead of the full joint distribution $f_{jk}(\cdot)$, but we can reconstruct the joint distribution given assumed forms for the marginal covariate distributions and the correlation matrix, for example assuming that these are the same as those observed in the IPD studies (Phillippo et al., 2020a). Simulation studies with binary outcomes have found that the results of ML-NMR analyses are not sensitive to the assumptions used in reconstructing the joint distribution (Phillippo et al., 2020b); we expect this result to hold for other outcomes including time-to-event.

If we have summary outcomes $y_{\bullet jk}$ on a given treatment k in study j, we can attempt to derive a corresponding *aggregate* marginal likelihood function as the product of the individual marginal likelihood functions (2), up to a normalizing constant,

$$L^{\text{Mar}}_{\bullet jk}(\boldsymbol{\xi}; y_{\bullet jk}) \propto \prod_{i=1}^{N_{jk}} L^{\text{Mar}}_{ijk}(\boldsymbol{\xi}; y_{ijk}), \tag{4}$$

where the subscript $\bullet$ denotes quantities that have been aggregated over individuals. If the result can be rearranged in terms of $y_{\bullet jk}$, we can then use $L^{\text{Mar}}_{\bullet jk}(\boldsymbol{\xi}; y_{\bullet jk})$ to evaluate the aggregate marginal likelihood function. For example, we demonstrate with binary outcomes in Appendix A online supplementary material, where the aggregate marginal likelihood is shown to be equal to the Binomial likelihood used previously for ML-NMR by Phillippo et al. (2020a). Similar results can be obtained for categorical outcomes with a Multinomial likelihood (Phillippo, 2019) and count outcomes with a Poisson likelihood, again obtaining the same results as using standard relations on the sums of random variables. However, this may not be possible, in general.

By working directly with the likelihood contributions from each level of the model, we avoid having to explicitly derive the form of the aggregate likelihood. The full ML-NMR model for general likelihoods may be written using (2) and (4) as

Individual:

$$L^{\text{Con}}_{ijk\,|\,\boldsymbol{x}}(\boldsymbol{\xi}; y_{ijk}, \boldsymbol{x}_{ijk}) = \pi_{\text{Ind}}(y_{ijk} \mid \theta_{ijk}) \tag{5a}$$

$$g(\theta_{ijk}) = \eta_{jk}(\boldsymbol{x}_{ijk}) = \mu_j + \boldsymbol{x}^T_{ijk}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \gamma_k. \tag{5b}$$

Aggregate:

$$L^{\text{Mar}}_{ijk}(\boldsymbol{\xi}; y_{ijk}) = \int_{\mathfrak{X}} L^{\text{Con}}_{ijk\,|\,\boldsymbol{x}}(\boldsymbol{\xi}; y_{ijk}, \boldsymbol{x}) f_{jk}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \tag{5c}$$

$$L^{\text{Mar}}_{\bullet jk}(\boldsymbol{\xi}; y_{\bullet jk}) \propto \prod_{i=1}^{N_{jk}} L^{\text{Mar}}_{ijk}(\boldsymbol{\xi}; y_{ijk}), \tag{5d}$$

where in a Bayesian analysis, prior distributions are placed over each of the parameters $\mu_j, \boldsymbol{\beta}_1, \boldsymbol{\beta}_{2,k}$, and γ_k. For the analyses in this paper, we will use non- or weakly informative prior distributions

which do not unduly influence the posterior distribution, as is often desired in decision-making applications. However, the analyst may—and indeed should—select prior distributions that are appropriate for the situation at hand and assess sensitivity to reasonable alternatives as appropriate. Computationally, we fit these models in Stan by directly coding the log likelihood contributions with a `target +=` statement (Stan Development Team, 2023).

3.1 Application to survival analysis

We now apply this general framework to derive ML-NMR models for survival or time-to-event outcomes. Consider that every study provides a pair $y_{ijk} = \{t_{ijk}, c_{ijk}\}$ of outcome times t_{ijk} and censoring indicators c_{ijk} for each individual i in study j receiving treatment k, where $c_{ijk} = 1$ if an individual experiences the event or $c_{ijk} = 0$ if they are censored. For the AgD studies, these data could be obtained by digitizing published Kaplan–Meier curves and reconstructing the event and censoring times using an algorithm such as that described by Guyot et al. (2012). Individual covariate information $\boldsymbol{x}_{ijk}$ is available for every individual in the IPD studies, but for the AgD studies only the joint distribution of the covariates at baseline $f_{jk}(\cdot)$ is available (likely reconstructed from reported marginal summaries Phillippo et al., 2020a).

The individual conditional likelihood contributions for each time t_{ijk} in the IPD are

$$L^{\text{Con}}_{ijk\,|\,\boldsymbol{x}}(\boldsymbol{\xi};\, t_{ijk},\, c_{ijk},\, \boldsymbol{x}_{ijk}) = S_{jk}(t_{ijk}\,|\,\boldsymbol{x}_{ijk}) h_{jk}(t_{ijk}\,|\,\boldsymbol{x}_{ijk})^{c_{ijk}}, \tag{6}$$

where $S_{jk}(t\,|\,\boldsymbol{x})$ and $h_{jk}(t\,|\,\boldsymbol{x})$ are the survival and hazard functions conditional on covariates $\boldsymbol{x}$, which may take any form. For illustration, a Weibull proportional hazards model has survival and hazard functions

$$\begin{aligned} S_{jk}(t\,|\,\boldsymbol{x}) &= \exp\left(-t^{\nu_j} \exp\left(\eta_{jk}(\boldsymbol{x})\right)\right), \\ h_{jk}(t\,|\,\boldsymbol{x}) &= \nu_j t^{\nu_j - 1} \exp\left(\eta_{jk}(\boldsymbol{x})\right), \end{aligned}$$

where ν_j is a study-specific shape parameter. In practice, the choice of model may be based on model fit statistics (see Section 3.2) and plausibility of extrapolations. Notice that we stratify the baseline hazard by study to respect randomization, e.g. with study-specific shape parameters ν_j, akin to the stratification of the study-specific intercepts μ_j in the linear predictor. Appendix B online supplementary material details survival and hazard functions for all survival models currently implemented in the *multinma* R package (Phillippo, 2024), including a full range of parametric proportional hazards and accelerated failure time models, and flexible baseline hazards models based on M-splines or piecewise exponentials.

Using equation (2), the individual marginal likelihood contributions for each event/censoring time in the AgD studies are

$$\begin{aligned} L^{\text{Mar}}_{ijk}(\boldsymbol{\xi};\, t_{ijk},\, c_{ijk}) &= \int_{\mathfrak{x}} L^{\text{Con}}_{ijk\,|\,\boldsymbol{x}}(\boldsymbol{\xi};\, t_{ijk},\, c_{ijk},\, \boldsymbol{x}) f_{jk}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \\ &= \int_{\mathfrak{x}} S_{jk}(t_{ijk}\,|\,\boldsymbol{x}) h_{jk}(t_{ijk}\,|\,\boldsymbol{x})^{c_{ijk}} f_{jk}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{7}$$

We evaluate this integral using quasi-Monte Carlo integration following equation (3) as

$$L^{\text{Mar}}_{ijk}(\boldsymbol{\xi};\, t_{ijk},\, c_{ijk}) \approx \tilde{N}^{-1} \sum_{\tilde{\boldsymbol{x}}} S_{jk}(t_{ijk}\,|\,\tilde{\boldsymbol{x}}) h_{jk}(t_{ijk}\,|\,\tilde{\boldsymbol{x}})^{c_{ijk}}. \tag{8}$$

3.2 Model comparison

Model comparison for Bayesian network meta-analyses is typically performed using the Deviance Information Criterion (DIC) (Dias et al., 2011b; Spiegelhalter et al., 2002). However, the general ML-NMR model equation (5) may not have a closed-form aggregate-level likelihood, which means that the usual p_D complexity penalty cannot be evaluated. Instead, the DIC may be calculated using the p_V penalty proposed by Gelman et al. (2013), or the Watanabe-Akaike Information

Criterion (WAIC) or Leave-One-Out Information Criterion (LOOIC) (Vehtari et al., 2016) can be used, all of which are calculated directly from the log likelihood contributions. We choose to use LOOIC here, as it (and its approximation WAIC) evaluates predictive performance over the entire posterior distribution rather than only at a point estimate and works well when the posterior is not approximately Normal, unlike DIC (Vehtari et al., 2016).

3.3 Assessing integration error

ML-NMR models are typically implemented using Quasi-Monte Carlo integration via Sobol' sequences to evaluate the integral for the aggregate-level model (Phillippo et al., 2020a). Phillippo et al. (2020a) previously suggested assessing the accuracy of the numerical integration by plotting the empirical integration error over the entire posterior distribution for increasing values of $\tilde{N}$. Whilst this approach may be suitable when the aggregate-level model is of the form (5d) and can be simplified into a single integral per AgD study arm, it becomes untenable in practice when the aggregate-level model is of the form (5c), and there is one integral for every individual in each AgD study. In this case, there may be hundreds or even thousands of such individuals and corresponding integration error plots, and the computational burden of saving and plotting the cumulative integration points quickly becomes unfeasible.

Instead, we propose the algorithm in Appendix C online supplementary material to ensure that $\tilde{N}$ is sufficient using the $\widehat{R}$ convergence statistic (Vehtari et al., 2020). Based on the usual practice of fitting $C > 1$ chains in parallel (usually $C = 4$), we use $\tilde{N}$ integration points for one half of the chains and $\tilde{N}/2$ for the other half. We then check convergence with $\widehat{R}$ within each half set and between all chains together, to determine convergence of both MCMC and numerical integration.

Values of $\tilde{N}$ that are powers of 2 are recommended as these are expected to be particularly efficient (Owen, 2013). The sufficient value of $\tilde{N}$ will vary depending on the model. In our experience, a value of $\tilde{N} = 64$ strikes a conservative balance between sufficient accuracy and increased runtime, and should be sufficient for many models to only require a single run. The *multinma* R package (Phillippo, 2024) implements the above algorithm (with $\tilde{N} = 64$ by default) and provides user-friendly warnings when the number of integration points is detected to be insufficient.

3.4 Checking model assumptions

The key assumption underlying all anchored population adjustment approaches is *conditional constancy of relative effects*, which requires that there are no unobserved effect modifiers in imbalance between the included study populations and between these and the target population (Phillippo et al., 2016). With ML-NMR, we can assess this assumption using standard techniques from the network meta-analysis literature, where residual heterogeneity or inconsistency may indicate a violation of this assumption (Phillippo et al., 2020a, 2022). Residual heterogeneity can be assessed using a random effects model (Dias et al., 2011b), replacing γ_k in equation (5) by a study-specific random effect $\delta_{jk} \sim \mathrm{N}(\gamma_k, \tau^2)$, where τ is the between-studies standard deviation. For studies with more than two arms, a multivariate Normal random effects distribution is required to account for the correlation between relative effects (Dias et al., 2011b; Phillippo et al., 2020a). Residual inconsistency can be assessed using unrelated mean effects or node-splitting models (Dias, 2011c). For example, an unrelated mean effects model replaces γ_k in equation (5) by $\gamma_{t_{j1}k}$, where t_{j1} is the treatment in arm 1 of study j and we set $\gamma_{kk} = 0$ for all k. We note that, as is the case for standard NMA, these approaches to detect residual heterogeneity and inconsistency may have low power. Phillippo et al. (2022) demonstrate the practical application of these techniques to ML-NMR models using the *multinma* R package.

In practice, we often find that there are insufficient data to estimate independent effect modifier interaction terms $\boldsymbol{\beta}_{2,k}$ for each treatment. Where this is the case, we typically rely on the *shared effect modifier assumption* for a set of treatments $\mathscr{T}$, and define the effect modifier interaction terms to be equal for all treatments within this set, $\boldsymbol{\beta}_{2,k} = \boldsymbol{\beta}_{2,\mathscr{T}}\ \forall\ k \in \mathscr{T}$ (Phillippo et al., 2016, 2020a). This assumption is only likely to be reasonable when treatments belong to the same class, sharing a mode of action (Phillippo et al., 2016). Regulatory bodies as decision-makers typically require strong biological or clinical rationale to justify this assumption (HTA Coordination Group , 2024), and with such evidence this has been accepted by decision-makers (e.g. TA1013, National Institute for Health and Care Excellence, 2024). Phillippo et al. (2022) demonstrate how the

shared effect modifier assumption may be relaxed and assessed one covariate at a time, which is less data-intensive than fitting a model with independent interactions for all covariates at once. When the shared effect modifier assumption or other identifying assumptions are not plausible, and there are insufficient data to identify the model, then ML-NMR is restricted to producing estimates in the aggregate study population(s)—the same as MAIC and STC.

When fitting time-to-event models, the suitability of the proportional hazards assumption (or the analogous accelerated failure time assumption) should be assessed. We assess this assumption by letting the baseline hazard vary between the arms of each study. For parametric models like the Weibull model, this means allowing independent shape parameters v_{jk} to vary by treatment arm as well as by study. For a flexible M-spline hazard model, this means allowing independent spline coefficient vectors $\boldsymbol{\alpha}_{jk}$ by arm as well as by study.

3.5 Producing population-average estimates for a target population

For decision-making, we must produce estimates of quantities of interest, such as population-average treatment effects or survival probabilities, in a target population relevant to the decision. The decision target population need not be represented by one of the studies in the network; indeed, it is likely best represented by a registry or cohort study conducted in the population of interest (Phillippo et al., 2016).

Population-average conditional treatment effects $d_{ab(P)}$ between each pair of treatments a and b in a population P are produced by integrating contrasts of the linear predictor over the joint covariate distribution $f_{(P)}(\boldsymbol{x})$, which due to linearity reduces to plugging-in mean covariate values $\bar{\boldsymbol{x}}_{(P)}$,

$$\begin{aligned} d_{ab(P)} &= \int_{\mathfrak{X}} (\eta_{(P)b}(\boldsymbol{x}) - \eta_{(P)a}(\boldsymbol{x})) f_{(P)}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \\ &= \gamma_b - \gamma_a + \bar{\boldsymbol{x}}_{(P)}^{\mathrm{T}} (\boldsymbol{\beta}_{2,b} - \boldsymbol{\beta}_{2,a}). \end{aligned} \tag{9}$$

The primary marginal quantity of interest is the population-average marginal survival function, also called the standardized survival function, from which we can also produce a range of other marginal estimates. The population-average marginal survival probability $\bar{S}_{(P)k}(t)$ on treatment k in population P at time t is found by integrating the individual-level survival function $S_{(P)k}(t \mid \boldsymbol{x})$ over the joint covariate distribution $f_{(P)}(\boldsymbol{x})$ at each time t,

$$\bar{S}_{(P)k}(t) = \int_{\mathfrak{X}} S_{(P)k}(t \mid \boldsymbol{x}) f_{(P)}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}. \tag{10}$$

This integral can be calculated using the same quasi-Monte Carlo numerical integration approach described earlier, using a set of integration points drawn from the joint distribution $f_{(P)}(\boldsymbol{x})$, analogously to (3). In the likely scenario that only marginal covariate summaries are available, again we can reconstruct the joint covariate distribution from assumed forms for the marginal distributions and correlation matrix (Phillippo et al., 2020a). We also require information on the distribution of the baseline hazard in the target population P, that is distributions for the linear predictor intercept parameter $\mu_{(P)}$ and any additional parameters of the survival function such as the Weibull shape parameter $v_{(P)}$ or M-spline coefficients $\boldsymbol{\alpha}_{(P)}$. Estimates of these parameters may not be available directly for an external target population. If instead we have (reconstructed) Kaplan–Meier data available for outcomes on a reference treatment in the target population (along with the summary covariate distribution), then these data may be included in the model as a single-arm study at the synthesis stage through equation (7); this will allow the parameters of the baseline hazard in this population to be estimated, but will not contribute information to any other model parameters. Otherwise, estimates may be borrowed from a study in the network where the properties of the baseline hazard are deemed to be representative of the target population.

From this marginal survival function, we can then produce a range of other marginal estimates. The corresponding population-average marginal hazard function is a weighted average of the individual-level hazard functions,

$$\bar{h}_{(P)k}(t) = \frac{\int_{\mathfrak{X}} S_{(P)k}(t \mid \boldsymbol{x}) h_{(P)k}(t \mid \boldsymbol{x}) f_{(P)k}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}}{\bar{S}_{(P)k}(t)}, \tag{11}$$

weighted by the probability of surviving to time t. Again, this integral can be calculated using quasi-Monte Carlo numerical integration. The corresponding population-average marginal cumulative hazard function is

$$\bar{H}_{(P)k}(t) = -\log\left(\bar{S}_{(P)k}(t)\right). \tag{12}$$

Quantiles of the population-average marginal survival times are found by solving

$$\bar{S}_{(P)k}(t^{(\alpha)}_{(P)k}) = 1 - \alpha, \tag{13}$$

to find $t^{(\alpha)}_{(P)k}$ for the α% quantile, which can be achieved using numerical root finding.

Means or restricted means of the population-average marginal survival times are found by integrating the marginal survival function up to a restricted time horizon t^*,

$$\text{RMST}_{(P)k}(t^*) = \int_0^{t^*} \bar{S}_{(P)k}(t)\,\mathrm{d}t, \tag{14}$$

with $t^* = \infty$ for population-average mean marginal survival time, which is typically evaluated using quadrature; we use the implementation in the *flexsurv* R package (Jackson, 2016).

Contrasts of the above quantities are population-average marginal treatment effects $\Delta_{ab(P)}(t)$. For example, the ratio of population-average marginal hazard functions (11) for two treatments a and b forms a population-average marginal hazard ratio,

$$\Delta^{\text{HR}}_{ab(P)}(t) = \frac{\bar{h}_{(P)b}(t)}{\bar{h}_{(P)a}(t)}. \tag{15}$$

In a similar fashion, we can also create population-average median survival time ratios or differences, or differences in population-average (restricted) mean survival times.

All the quantities (10)–(15) are *marginal*, being derived from the population-average marginal survival function $\bar{S}_{(P)k}(t)$. These depend on the distributions of the baseline hazard and of all covariates (not just those that are effect-modifying). Furthermore, the population-average marginal hazard ratios $\Delta^{\text{HR}}_{ab(P)}(t)$ also vary over time; if hazards are proportional conditional on covariates (prognostic or effect modifying) this means that, mathematically, proportional hazards *cannot* hold at the marginal level. In contrast, $d_{ab(P)}$ are population-average *conditional* treatment effects which depend only on the distribution of *effect-modifying* covariates. $d_{ab(P)}$ are constant over time and do not depend on the distribution of baseline hazard or the distribution of purely prognostic covariates. The population-average conditional treatment effects can be interpreted as the average of the individual-level treatment effects in the target population P, the average effect of moving each individual in the population from treatment a to b. The population-average marginal treatment effects can be interpreted in terms of the effects of treatment on the overall marginal survival curve in the population.

4 Application to newly diagnosed multiple myeloma example

We now apply these methods to the network of five studies comparing lenalidomide to placebo or thalidomide to placebo as maintenance treatment for newly diagnosed multiple myeloma, shown in Figure 1 (Leahy & Walsh, 2019). The outcome of interest is progression-free survival after autologous stem cell transplant (ASCT). IPD as individual event/censoring times and covariates are available from three studies; AgD as event/censoring times from digitized Kaplan–Meier curves and overall covariate summaries are available from two studies.

Since we did not have access to original IPD from the three IPD studies, for illustration we instead constructed synthetic data that resemble the original IPD using published Kaplan–Meier curves and regression coefficients. This process is detailed in Appendix D online supplementary material.

4.1 Newly diagnosed multiple myeloma: methods

Instead of making parametric assumptions about the form of the baseline hazard, we propose a novel approach using M-splines to flexibly model the baseline hazard over time. This approach builds on previous applications of M-splines for flexible baseline hazard models in other contexts (Brilleman et al., 2020; Jackson, 2023) and is described in detail in Appendix B.1.4 online supplementary material.

The survival and hazard functions for the M-spline model are given by

$$S_{jk}(t \mid \boldsymbol{x}) = \exp\left(-\boldsymbol{\alpha}_j^T \boldsymbol{I}_\kappa(t, \boldsymbol{\zeta}_j) \exp\left(\eta_{jk}(\boldsymbol{x})\right)\right), \tag{16a}$$

$$h_{jk}(t \mid \boldsymbol{x}) = \boldsymbol{\alpha}_j^T \boldsymbol{M}_\kappa(t, \boldsymbol{\zeta}_j) \exp\left(\eta_{jk}(\boldsymbol{x})\right), \tag{16b}$$

where $\boldsymbol{\alpha}_j$ is a study-specific vector of spline coefficients, $\boldsymbol{M}_\kappa(t, \boldsymbol{\zeta}_j)$ is the M-spline basis of order κ with a study-specific knot sequence $\boldsymbol{\zeta}_j$ evaluated at time t, and $\boldsymbol{I}_\kappa(t, \boldsymbol{\zeta}_j)$ is the corresponding integrated M-spline basis (an I-spline basis; see Appendix B.1.4 online supplementary material). The basis polynomials have degree $\kappa - 1$, so a basis of order $\kappa = 4$ corresponds to a cubic M-spline basis; a piecewise exponential baseline hazards model is a special case with degree zero ($\kappa = 1$).

To avoid overfitting, we propose a novel weighted random walk prior distribution on the inverse-softmax transformed spline coefficients,

$$\boldsymbol{\alpha}_j = \text{softmax}(\boldsymbol{\alpha}_j^*), \tag{17a}$$

$$\alpha_{j,l}^* = c_{j,l} + \sum_{m=1}^{l} u_{j,m} \quad \forall\, l = 1, \ldots, L + \kappa - 1, \tag{17b}$$

$$u_{j,l} \sim \text{N}(0, \sigma_j^2 w_{j,l}) \quad \forall\, l = 1, \ldots, L + \kappa - 1, \tag{17c}$$

where L is the number of internal knots, and the softmax (or multinomial logit) transform is $\text{softmax}(\boldsymbol{\alpha}_j^*) = [1,\ \exp(\boldsymbol{\alpha}_j^*)^T]^T / (1 + \sum_{l=1}^{L+\kappa-1} \exp(\alpha_{j,l}^*))$. The random walk is centred around a prior mean vector $\boldsymbol{c}_j$ that corresponds to a constant baseline hazard (see Appendix B.1.4 online supplementary material), borrowing an idea of Jackson (2023) who derived $\boldsymbol{c}_j$ to use instead for the prior mean of a random effect on $\boldsymbol{\alpha}_j$. The weights $w_{j,l}$ are derived from the distance between each pair of knots (see Appendix B.1.4 online supplementary material), following a similar approach to the Bayesian P-splines proposed by Li and Cao (2022) except that we additionally normalize the weights to sum to 1. The weights serve to make the prior invariant to the number and location of the knots, even if they are unevenly spaced, and to the timescale, greatly simplifying the specification of a hyperprior distribution for the random walk standard deviation σ_j. The random walk standard deviation σ_j controls the amount of smoothing and shrinkage of the spline coefficients; as σ_j approaches zero the baseline hazard becomes smoother (less 'wiggly') and approaches a constant baseline hazard. We allow σ_j to be estimated from the data, giving this a weakly informative hyperprior distribution $\sigma_j \sim \text{half-N}(0, 1^2)$.

We adjust for four clinically relevant covariates considered to be potential effect modifiers by Leahy and Walsh (2019): age, ISS stage (stage III vs. stage I–II), response post-ASCT (complete response or very good partial response vs. other), and sex (male or female). The distributions of these covariates in each study at baseline are given in Table 1. Due to the lack of data on thalidomide (only a single AgD study), we make the shared effect modifier assumption between the two active treatments in order to identify the effect modifying treatment–covariate interactions (Phillippo et al., 2016, 2020a). Since thalidomide and lenalidomide are in the same class of treatments, this assumption may be reasonable.

We fit a cubic M-spline model with seven internal knots placed at evenly spaced quantiles of the uncensored survival times in each study, plus boundary knots at time 0 and the last event/censoring time in each study. The number of knots is set to be larger than we might expect to need, since any potential for overfitting is avoided by shrinkage through the random walk prior. To ensure that seven knots are sufficient, we also fit a model with ten internal knots for comparison. We

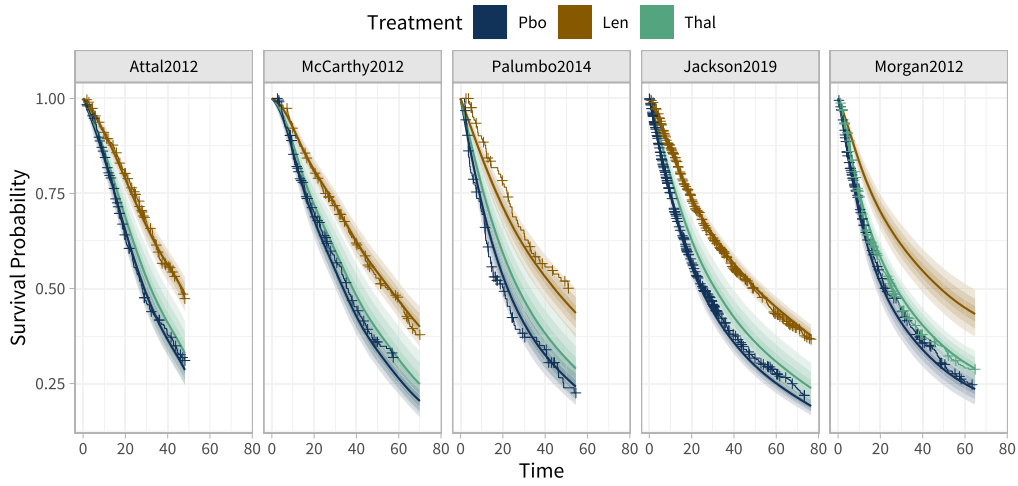

Figure 2. Estimated survival curves on each treatment in each study population, under a cubic M-spline model. Shaded bands indicate the 50%, 80%, and 95% Credible Intervals for the survival curves (thick lines), overlaid on the unadjusted Kaplan–Meier curves from the treatments in each study (thin lines).

assess the proportional hazards assumption by fitting models with spline coefficients $\boldsymbol{\alpha}_{jk}$ stratified by treatment arm as well as by study. We give noninformative $N(0, 100^2)$ prior distributions to every parameter in the linear predictor. We also fit unadjusted NMA models with the same M-spline baseline hazard for comparison.

Analyses were carried out in R version 4.3.1 (R Core Team, 2023) and Stan version 2.26.23 (Carpenter et al., 2017). Analysis code and data are available from https://github.com/dmphillippo/ML-NMR-general-likelihoods-paper. Two sets of analysis codes are provided: one that fits the models via the user-friendly *multinma* R package (Phillippo, 2024), making these techniques accessible to a broad audience; and another that fits the models by calling Stan directly, which is likely to be useful for those who wish to further modify or extend the code. The data are also available in the *multinma* R package along with a vignette that walks through the analysis (Phillippo, 2024). Using *multinma*, the ML-NMR models take around 1.25 hr each to fit on a modern laptop (Intel Core Ultra 7 165H 5 GHz, 32 GB RAM); the unadjusted NMA models take around 4 min each.

4.2 Newly diagnosed multiple myeloma: results

The estimated population-average survival curves in each study population are shown in Figure 2, overlaid with the observed (unadjusted) Kaplan–Meier curves. These show a good visual fit to the observed data, with the possible exception of the lenalidomide arm of Palumbo 2014 where the unadjusted Kaplan–Meier estimate lies consistently above the population-adjusted estimate. This is likely due to the slight baseline imbalance in Palumbo 2014 between arms, with the lenalidomide arm having 17% fewer males than the placebo arm. The unadjusted Kaplan–Meier curves do not account for this difference, whereas the population-adjusted survival estimates from the ML-NMR model do. The population-average median survival times corresponding to these population-average survival curves are given in Table E.1 online supplementary material. The posterior means for the median survival estimates vary across populations between 20.75 and 33.30 months on placebo, 26.55 and 38.44 months on thalidomide, and 44.95 and 55.92 months on lenalidomide.

To assess whether seven knots are sufficient, we also fit a model with ten knots. Comparing the model fit in Table E.2 online supplementary material, we find that there is no substantial difference between the models. The LOOIC is slightly worse for the model with ten knots, but not substantially, due to a slight increase in the effective number of parameters p_{LOO}; however, the random walk prior distribution is behaving as expected and controlling the overall complexity through

Table 2. Estimated population-average conditional log hazard ratios and 95% Credible Intervals in each study population from the cubic M-spline model

Study	Lenalidomide vs. placebo	Thalidomide vs. placebo	Thalidomide vs. lenalidomide
Attal2012	−0.59	−0.13	0.47
	(−0.74, −0.45)	(−0.38, 0.13)	(0.23, 0.69)
McCarthy2012	−0.62	−0.16	0.47
	(−0.74, −0.51)	(−0.38, 0.07)	(0.23, 0.69)
Palumbo2014	−0.64	−0.18	0.47
	(−0.80, −0.48)	(−0.44, 0.09)	(0.23, 0.69)
Jackson2019	−0.69	−0.22	0.47
	(−0.81, −0.57)	(−0.43, −0.01)	(0.23, 0.69)
Morgan2012	−0.68	−0.22	0.47
	(−0.81, −0.56)	(−0.42, −0.01)	(0.23, 0.69)

shrinkage. This is also apparent when looking at the individual-level baseline hazard functions (Figure E.1 online supplementary material) and the corresponding population-average marginal hazard functions (Figure E.2 online supplementary material) which are very similar between models. We also check the LOOIC within each study separately (Table E.3 online supplementary material) to ensure that no studies individually are better fit with a higher number of knots, which could be missed when looking overall. We conclude that seven internal knots are sufficient, both overall and within each study in the network.

To assess the proportional hazards assumption, we modify the M-spline model to stratify the spline coefficients $\boldsymbol{\alpha}_{jk}$ on the baseline hazard by treatment arm as well as by study. Comparing the overall model fit between the models with and without the proportional hazards assumption (Table E.4 online supplementary material), we see that the LOOIC is lower for the proportional hazards model. Again, we also check the LOOIC within each study separately (Table E.5 online supplementary material), to ensure that the proportional hazards assumption is reasonable within each study in the network. We conclude that the proportional hazards assumption is reasonable here. For comparison, we also fitted unadjusted models with no covariates (i.e. a standard network meta-analysis) both with and without the proportional hazards assumption. Whilst there was little difference in the overall model fit (Table E.6 online supplementary material), the nonproportional hazards model did have a substantially lower LOOIC in the Jackson 2019 study (Table E.7 online supplementary material). Including the covariates in the ML-NMR analysis, even though they are fixed and not time-varying, is sufficient to remove this proportional hazards violation, and the ML-NMR model is a much better overall fit than the unadjusted NMA.

The estimated population-average conditional log hazard ratios from the ML-NMR model (with seven internal knots and proportional hazards) are given in Table 2. Both lenalidomide and thalidomide are consistently estimated to be more effective than placebo in each of the study populations, however the 95% credible intervals for the thalidomide vs. placebo comparison cross zero in both AgD study populations (Jackson 2019 and Morgan 2012), where both relative effects vs. placebo are estimated with slightly more uncertainty. The thalidomide vs. lenalidomide relative effect estimates are constant across all populations (0.47, 95% CrI 0.24–0.71), due to the shared effect modifier assumption.

5 Simulated example

To illustrate the performance of this approach, let us consider an artificial example of simulated survival outcomes in a population-adjusted indirect comparison of two treatments B and C via a common comparator A. Since the data are simulated, we can compare the results and performance of ML-NMR to a full IPD NMA and to the known true values. We simulate outcomes from a Weibull model including three prognostic and effect-modifying covariates (two continuous and one binary); full details are given in Appendix F online supplementary material.

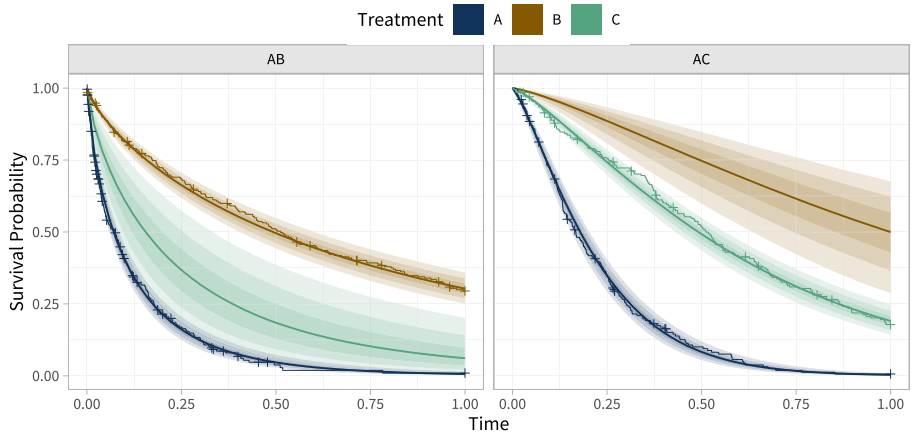

Figure 3. ML-NMR estimated survival curves on each treatment in each study population, under a Weibull model. Shaded bands indicate the 50%, 80%, and 95% Credible Intervals for the survival curves (thick lines), overlaid on the unadjusted Kaplan–Meier curves from the treatments in each study (thin lines).

5.1 Simulated example: methods

We fit Exponential, Weibull, and Gompertz proportional hazards models (Appendix B.1 online supplementary material) in the general ML-NMR framework adjusting for all three covariates. We place noninformative N(0, 100^2) prior distributions on every parameter in the linear predictor, and a weakly informative half-N(0, 10^2) prior distribution on the Weibull and Gompertz shape parameters. For comparison, we also fit the corresponding IPD NMA models with IPD from both studies. We also carry out MAIC and STC analyses, commonly used for population-adjustment in this two-study scenario. For the STC, we use the simulation approach of Remiro-Azócar et al. (2022). Lastly, we perform a standard (nonpopulation adjusted) indirect comparison, formed from the log hazard ratios estimated in each study separately using a Weibull model adjusted only for prognostic factors, reflecting 'best case' common practice (i.e. correct form of parametric model, fully adjusted for prognostic factors).

Analyses were carried out in R version 4.3.1 (R Core Team, 2023) and Stan version 2.26.23 (Carpenter et al., 2017). Full analysis code and data are provided at https://github.com/dmphillippo/ML-NMR-general-likelihoods-paper, again in two formats: one that fits the models via *multinma* R package (Phillippo, 2024), and another that fits the models by calling Stan directly. Using *multinma*, the ML-NMR models take around 90 s each to fit on a modern laptop (Intel Core Ultra 7 165 H 5 GHz, 32 GB RAM); the IPD NMA models take around 4 s each.

5.2 Simulated example: results

Inspecting the LOOIC model comparison statistics in Table F.1 online supplementary material, we see that the Weibull model has the lowest LOOIC for both ML-NMR and IPD NMA, and the standard error of the difference suggests that the Weibull model is a substantially better fit than either the Exponential or Gompertz models in both the ML-NMR and IPD NMA scenarios. Comparing individual LOOIC contributions between the ML-NMR and IPD NMA models reveals that individual observations are fitted similarly well under each model (Figure F.2 online supplementary material).

The estimated population-average survival curves on each treatment in each study population under the Weibull model fitted using ML-NMR are shown in Figure 3, overlaid on the unadjusted Kaplan–Meier curves. Visually, the estimated survival curves are a good fit to the observed data. Table 3 presents the estimated population-average conditional log hazard ratios (HRs) for each pairwise comparison in each population, along with the true values from the simulation. The ML-NMR estimates agree well with both the IPD NMA and the true values, and the *B* vs. *A* and *C* vs. *A* estimates within the *AB* and *AC* study populations, respectively, are unchanged in point estimate or standard error. Standard errors for comparisons not observed in the data are

Table 3. Table of estimated population-average conditional log hazard ratios and 95% Credible Intervals from the ML-NMR model and the full IPD NMA, alongside the true log hazard ratios, in the *AB* and *AC* study populations

		Comparison		
Study	**Method**	***B* vs. *A***	***C* vs. *A***	***C* vs. *B***
AB	Truth	−1.62	−0.92	0.70
	ML-NMR	−1.53	−0.62	0.90
		(−1.74, −1.30)	(−1.19, −0.06)	(0.28, 1.52)
	IPD NMA	−1.54	−0.67	0.87
		(−1.76, −1.32)	(−1.12, −0.23)	(0.36, 1.37)
AC	Truth	−2.07	−1.37	0.70
	ML-NMR	−2.20	−1.29	0.90
		(−2.76, −1.63)	(−1.54, −1.05)	(0.28, 1.52)
	IPD NMA	−2.17	−1.31	0.87
		(−2.63, −1.70)	(−1.54, −1.08)	(0.36, 1.37)

Note. ML-NMR = multilevel network meta-regression; IPD NMA = individual participant data network meta-analysis.

slightly increased (by 2%–6%) using ML-NMR compared to full IPD NMA, which is expected due to the reduced information available.

Due to noncollapsibility, the standard indirect comparison, MAIC, and STC analyses cannot be compared to the population-average conditional log hazard ratios in Table 3, as these methods can only estimate marginal quantities. Instead, we compare the estimated restricted mean survival times up until the end of follow up ($t^* = 1$) on each treatment in each study population, which have the same interpretation as a marginal quantity under each of the five models. The MAIC had an approximate effective sample size of 5.4 (matching both means and variances for continuous covariates; 6.8 matching means only), and bootstrapping to calculate standard errors and credible intervals was highly unstable (45% of iterations failed). The results of the MAIC were therefore considered unusable and are not presented here. Restricted mean survival time estimates for the remaining four methods are displayed in Table 4. The results from the ML-NMR and IPD NMA agree closely, with nearly identical posterior means and credible intervals; the estimates of treatment *B* in the *AC* population and treatment *C* in the *AB* population are slightly more uncertain from the ML-NMR model due to the reduced information available. The STC produces similar estimates to ML-NMR in the *AC* population, with a similar level of uncertainty. However, STC cannot produce estimates for all treatments in the *AB* population, or any other target population of interest to decision makers. Lastly, the standard indirect comparison produces estimates that are clearly biased: differences in effect modifiers between the populations are not accounted for, and so the difference in restricted mean survival time between treatments *B* and *C* is underestimated in both populations.

Examining the parameters from the ML-NMR and IPD NMA models in Table F.2 online supplementary material, we see that these agree closely with each other and recover the true parameter values well.

6 Discussion

In this paper, we extended the ML-NMR framework to handle general likelihoods where the aggregate-level likelihood may not have a closed form. This greatly expands the range of models which can be fitted, including time-to-event outcomes which are common in technology appraisals (Phillippo et al., 2019). As in Phillippo et al. (2020a), we began with a fully specified individual-level model. However, instead of explicitly deriving the form of the aggregate likelihood via standard results on the sums of random variables, we proceeded by directly integrating the individual conditional likelihood function over the covariate distribution to obtain the individual marginal

Table 4. Table of estimated restricted mean survival times and 95% Credible Intervals on each treatment from the ML-NMR model, full IPD NMA, STC, and standard indirect comparison, in the *AB* and *AC* study populations

		Treatment		
Study	**Method**	***A***	***B***	***C***
AB	ML-NMR	0.13	0.54	0.27
		(0.11, 0.16)	(0.49, 0.58)	(0.15, 0.43)
	IPD NMA	0.13	0.54	0.28
		(0.11, 0.16)	(0.49, 0.58)	(0.17, 0.42)
	STC	0.13	0.54	–
		(0.11, 0.16)	(0.49, 0.58)	
	Standard IC	0.13	0.54	0.47
		(0.11, 0.16)	(0.49, 0.58)	(0.39, 0.56)
AC	ML-NMR	0.22	0.75	0.53
		(0.20, 0.25)	(0.63, 0.84)	(0.49, 0.58)
	IPD NMA	0.22	0.74	0.53
		(0.20, 0.25)	(0.63, 0.82)	(0.49, 0.58)
	STC	0.22	0.78	0.54
		(0.20, 0.25)	(0.67, 0.90)	(0.49, 0.58)
	Standard IC	0.23	0.59	0.54
		(0.20, 0.25)	(0.52, 0.66)	(0.49, 0.58)

Note. ML-NMR=multilevel network meta-regression; IPD NMA= individual participant data network meta-analysis; STC= simulated treatment comparison; IC= indirect comparison.

likelihood function. This is then used in one of two ways, depending on the data available, with different levels of generality.

Firstly, when the aggregate data consist of individual outcomes but only summary covariate information (such as survival data reconstructed from Kaplan–Meier curves), the aggregate part of the model is fitted directly using the individual marginal likelihood contributions. In this case, the method is fully general: individual conditional likelihood functions of any form can be integrated numerically to evaluate the individual marginal likelihood function.

Secondly, when the aggregate data consist of summary outcomes and summary covariate information, the individual marginal likelihood contributions are multiplied together to obtain the aggregate marginal likelihood contributions for the summary outcomes. Evaluation of the aggregate marginal likelihood contributions requires that these can be expressed in terms of the summary outcomes (as demonstrated in Appendix A online supplementary material), which is only straightforward for discrete outcomes. However, the aggregate-level likelihood has a known closed form for many continuous individual-level likelihoods common in practice (Phillippo et al., 2020a).

Data at different levels of aggregation are encountered across a wide range of research areas, not just in the healthcare decision-making context considered in this paper. The general approach that we propose of directly integrating an individual-level likelihood to obtain an aggregate-level likelihood has, to our knowledge, not been considered before and is broadly applicable wherever data at different levels of aggregation are encountered, allowing coherent modelling across the levels of aggregation in a manner that avoids aggregation biases. Moreover, whilst we have used a Bayesian framework here, these ideas might also be applied to frequentist likelihoods or partial likelihoods, for example to estimate a frequentist multilevel Cox model; this is an interesting area for further research.

We found close agreement between the results of ML-NMR and full IPD NMA in our simulated example, which both successfully recovered the true values. Furthermore, the lack of IPD in the *AC* study did not greatly reduce precision for ML-NMR compared to IPD NMA; the standard errors of population-average log hazard ratios were the same for comparisons observed within each study

population, and only slightly increased for the unobserved comparisons. The conclusions of the model selection process were identical, in both cases correctly selecting the Weibull model. Nevertheless, this scenario is only a single instance. A full simulation study could further validate the performance of ML-NMR for survival analysis, and investigate the impact of invalid assumptions. However, we expect the conclusions of previous simulation studies on binary outcomes to apply broadly to ML-NMR models of general forms, including survival analysis (Phillippo et al., 2020b).

The additional IPD available to IPD NMA does offer further possibilities for analysis. For example, we required the shared effect modifier assumption in the simulated example to identify the ML-NMR model. In the interests of a fair comparison between ML-NMR and IPD NMA, both methods made use of this assumption in this analysis which was known to hold due to the simulated setup. However, IPD NMA could relax this assumption and estimate separate effect modifier interaction coefficients $\boldsymbol{\beta}_{2,B}$ and $\boldsymbol{\beta}_{2,C}$. In this scenario, since we know that $\boldsymbol{\beta}_{2,B} = \boldsymbol{\beta}_{2,C}$, the standard errors for IPD NMA would have been inflated by the unnecessarily more flexible model. The shared effect modifier assumption was also used in the newly diagnosed multiple myeloma example, again due to insufficient data to estimate separate treatment-covariate interactions for thalidomide. In this case, the assumption may be reasonable, since lenalidomide and thalidomide both belong to the same class of treatments. However, when treatments are not in the same class this assumption is likely to be much less plausible (Phillippo et al., 2016). Even when this assumption does not hold, we still expect population-average estimates in the AgD study population to be unbiased (Phillippo et al., 2020b). In larger treatment networks, it can be possible to assess and relax the shared effect modifier assumption in ML-NMR (Phillippo et al., 2022). When all studies across the network report relative effect estimates within subgroups, network meta-interpolation has recently been proposed to combine these in a manner that relaxes the shared effect modifier assumption (Harari et al., 2023). Ongoing work aims to utilize subgroup results and regression estimates, where available from trial reports, to support the estimation of ML-NMR models and reduce reliance on the shared effect modifier assumption in practical applications.

When working with a noncollapsible treatment effect measure, such as hazard ratios or survival time ratios for time-to-event outcomes (or odds ratios for binary outcomes), population-average conditional treatment effects $d_{ab(P)}$ and population-average marginal treatment effects $\Delta_{ab(P)}(t)$ are not equal and have different interpretations (Daniel et al., 2021; Kahan et al., 2014). Most notably, the population-average marginal treatment effects $\Delta_{ab(P)}(t)$ vary over time and depend on the distribution of all prognostic factors, effect modifiers, and baseline hazard in population P. The population-average conditional effects $d_{ab(P)}$ are constant over time and do not depend on the distribution of prognostic factors or baseline hazard in population P. Moreover, different population adjustment methods target different estimands. MAIC, and STC based on simulation or G-computation, can only produce marginal estimates. STC based on plugging in mean covariate values is biased for both estimands, and targets neither a conditional or marginal estimand correctly. Network meta-interpolation suffers similar biases to plug-in means STC, targeting neither a conditional or marginal estimand correctly, and furthermore cannot typically produce absolute estimates (e.g. survival curves or any derivative quantities) which are often required in a decision-making setting. At present, ML-NMR is the only population-adjustment method that can produce both conditional and marginal estimates, as well as absolute estimates, depending on the requirements for decision-making.

Leahy and Walsh (2019) analysed the newly diagnosed multiple myeloma example using multiple MAIC analyses followed by Bayesian NMA. The inherent limitations of such types of analyses have been described previously (Phillippo et al., 2016). In particular, when there are multiple AgD studies, a choice must first be made over which AgD study population to match to. Then, combining the network of MAIC-adjusted studies and AgD studies in a NMA requires an assumption of constancy of relative effects (i.e. that there are no effect modifiers in imbalance between these different populations), which is precisely the assumption that a population-adjusted analysis seeks to relax. Finally, the resulting estimates are only applicable in a population defined as some weighted average of the included AgD study populations, which may not represent the decision target population. The ML-NMR analysis addresses each of these issues: it coherently combines evidence from the IPD and AgD studies, accounting for differences between the populations of each study including the AgD studies, and can produce estimates in any target population for decision-making.

In both examples that we considered, event/censoring times were available from each individual in the aggregate studies, e.g. reconstructed from Kaplan–Meier plots (Guyot et al., 2012). If these are not available but instead only conditional log hazard ratios are reported (or log survival time ratios for accelerated failure time models), these may be synthesized directly using a Normal likelihood. For example, for the conditional log HR of treatment b vs. treatment a in study j the likelihood would be $\mathrm{N}(\eta_{jb}(\boldsymbol{x}_j^*) - \eta_{ja}(\boldsymbol{x}_j^*), s_{jab}^2)$, where s_{jab} is the standard error of the log HR and $\boldsymbol{x}_j^*$ is the vector of covariates at the reference levels used in study j. Studies with three or more arms would require the correlations between log HRs to be accounted for in the likelihood (Dias et al., 2011b). The limitation of this approach is the reported log hazard ratios must be adjusted in the same manner as the rest of the ML-NMR model. In theory, it should be possible to instead synthesize reported marginal summary outcomes such as marginal median survival times or marginal (restricted) mean survival times by application of Equations (13) and (14). This remains an area for further research.

We have only considered adjusting for covariates measured at baseline: time-varying covariates were not considered since it is likely that, in the aggregate studies, summary covariate information is available only available at baseline and not throughout follow-up. The inclusion of time-varying covariates in a survival model is often an attempt to correct for observed nonproportionality (i.e. failure of the proportional hazards or accelerated failure time assumption). However, such problems may be symptomatic of other issues such as omitted covariates, an incorrect functional form for a covariate, or using an inappropriate model form (e.g. a proportional hazards model when an accelerated failure time model would be more appropriate) (Therneau & Grambsch, 2000). Notably, the solutions for these issues can be dealt with within the ML-NMR framework we have described, without requiring further information on time-varying covariates. Indeed, in the newly diagnosed multiple myeloma example, we found evidence for nonproportional hazards in one study when fitting an unadjusted NMA, but adjusting for baseline covariates in the ML-NMR analysis was sufficient to remove this.

Stratifying the baseline hazard by study is imperative for respecting randomization within studies, in the same way that we must stratify the intercepts by study in the linear predictor. In this paper, we considered further stratifying the baseline hazard by treatment arm as a way to detect nonproportionality. If nonproportionality is still present after covariate adjustment, however, the model with baseline hazards stratified by study and treatment arm is of limited use for prediction of absolute effects, since survival curves (and all the ensuing summaries) can only be produced for treatments already observed in a population. Instead, the models considered here can be extended to incorporate a regression model on the shape of the baseline hazard. This opens up a further rich and flexible class of models, where departures from nonproportionality can be modelled and absolute predictions can once again be made for any treatment in any population. Such models are already implemented in the *multinma* R package (Phillippo, 2024).

For the newly diagnosed multiple myeloma example we used M-splines to flexibly model the baseline hazard, which is the first time that such a model has been applied to network meta-analysis of survival outcomes. We proposed a novel random walk prior distribution for the inverse-softmax transformed spline coefficients, which controls the level of smoothing and avoids overfitting through shrinkage. This may be applied to M-spline models in any context and has several advantages over previous approaches. Brilleman et al. (2020) used a Dirichlet prior directly on the spline coefficients, but this does not induce any smoothing or shrinkage and requires careful selection of the number and position of the knots. Jackson (2023) used a random effect on the inverse-softmax transformed spline coefficients, centred around a constant baseline hazard, aiming to induce shrinkage and avoid overfitting; however, we found that in practice this did not achieve sufficient shrinkage, with the model complexity and 'wiggliness' continuing to increase as the number of knots increased, leading to overfitting. Our random walk prior distribution does induce sufficient shrinkage to avoid overfitting, as demonstrated in the example, allowing the analyst to simply choose a 'large enough' number of knots and have the model shrink to an appropriate complexity based on the data. Li and Cao (2022) proposed Bayesian P-splines using a weighted (zero mean) random walk to allow for unevenly-spaced knots and make the prior invariant to knot positioning; we further normalized the knots to also make the prior invariant to the number of knots and timescale. This greatly simplifies specification of a hyperprior for the random walk standard deviation, since this no longer depends on the number of knots or the timescale, and ensures that unevenly-spaced knots do not affect smoothing or shrinkage behaviour.

Extending the ML-NMR framework to general likelihoods greatly increases the applicability of this approach, including to the very common scenario of population adjustment for survival outcomes. The Stan code that we have developed and provided in the supplementary materials is modular, and all that is required to fit a range of alternative models in the ML-NMR framework is to specify the form of the survival and hazard functions for the individual-level model. Once these have been specified, the numerical integration step to obtain the individual marginal likelihood remains the same, and is automatically implemented in the Stan code. Whilst not described here, it is also straightforward to account for left censoring, interval censoring, and left truncation (delayed entry) in this framework in the standard manner by considering the appropriate contributions from the survival function (e.g. as summarized by Brilleman et al., 2020), and all of these are implemented in the *multinma* R package (Phillippo, 2024). The *multinma* R package provides a user-friendly interface to implementing ML-NMR, AgD NMA, and IPD NMA models for a wide range of data types, supporting the uptake of these methods by analysts in practical applications.

Conflicts of interest: None declared.

Funding

This work was supported by the UK Medical Research Council, under grant numbers MR/P015298/1, MR/R025223/1, and MR/W016648/1.

Data availability

Analysis code and data are available from https://github.com/dmphillippo/ML-NMR-general-likelihoods-paper.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series A*.

References

Berlin J. A., Santanna J., Schmid C. H., Szczech L. A., & Feldman H. I. (2002). Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: Ecological bias rears its ugly head. *Statistics in Medicine*, *21*(3), 371–387. https://doi.org/10.1002/sim.1023

Brilleman S. L., Elci E. M., Novik J. B., & Wolfe R. (2020). 'Bayesian survival analysis using the rstanarm R package', arxiv, arxiv:2002.09633, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2002.09633

Bucher H. C., Guyatt G. H., Griffith L. E., & Walter S. D. (1997). The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of Clinical Epidemiology*, *50*(6), 683–691. https://doi.org/10.1016/S0895-4356(97)00049-8

Caro J. J., & Ishak K. J. (2010). No head-to-head trial? Simulate the missing arms. *PharmacoEconomics*, *28*(10), 957–967. https://doi.org/10.2165/11537420-000000000-00000

Carpenter B., Gelman A., Hoffman M. D., Lee D., Goodrich B., Betancourt M., Brubaker M., Guo J., Li P., & Riddell A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Daniel R., Zhang J., & Farewell D. (2021). Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biometrical Journal*, *63*(3), 528–557. ISSN 1521-4036. https://doi.org/10.1002/bimj.v63.3

Dias S., Sutton A. J., Welton N. J., & Ades A. E. (2011a). *NICE DSU technical support document 3: Heterogeneity: Subgroups, meta-regression, bias and bias-adjustment* (Technical Report). National Institute for Health and Care Excellence. http://www.nicedsu.org.uk

Dias S., Welton N. J., Sutton A. J., & Ades A. E. (2011b). *NICE DSU technical support document 2: A generalised linear modelling framework for pair-wise and network meta-analysis of randomised controlled trials* (Technical Report). National Institute for Health and Care Excellence. http://www.nicedsu.org.uk

Dias S., Welton N. J., Sutton A. J., Caldwell D. M., Lu G., & Ades A. E. (2011c). *NICE DSU technical support document 4: Inconsistency in networks of evidence based on randomised controlled trials* (Technical Report). National Institute for Health and Care Excellence. http://www.nicedsu.org.uk.

Donegan S., Williamson P., D'Alessandro U., Garner P., & Smith C. T. (2013). Combining individual patient data and aggregate data in mixed treatment comparison meta-analysis: Individual patient data may be beneficial if only for a subset of trials. *Statistics in Medicine*, *32*(6), 914–930. https://doi.org/10.1002/sim.5584

Gelman A., Carlin J. B., Stern H. S., Dunson D. B., Vehtari A., & Rubin D. B. (2013). *Bayesian data analysis*. 3rd ed.). Chapman & Hall/CRC Texts in Statistical Science. (CRC Press. ISBN 9781439898208.

Guyot P., Ades A. E., Ouwens M. J. N. M., & Welton N. J. (2012). Enhanced secondary analysis of survival data: Reconstructing the data from published Kaplan-Meier survival curves. *BMC Medical Research Methodology*, *12*(1), article number 9. https://doi.org/10.1186/1471-2288-12-9

Harari O., Soltanifar M., Cappelleri J. C., Verhoek A., Ouwens M., Daly C., & Heeg B. (2023). Network meta-interpolation: Effect modification adjustment in network meta-analysis using subgroup analyses. *Research Synthesis Methods*, *14*(2), 211–233. ISSN 1759-2887. https://doi.org/10.1002/jrsm.1608

Higgins J. P. T., & Whitehead A. (1996). Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine*, *15*(24), 2733–2749. https://doi.org/10.1002/(ISSN)1097-0258

HTA Coordination Group (2024). *Practical guideline for quantitative evidence synthesis: Direct and indirect comparisons* (Technical Report). European Commission. https://health.ec.europa.eu/publications/practical-guideline-quantitative-evidence-synthesis-direct-and-indirect-comparisons_en.

Ishak K. J., Proskorovsky I., & Benedict A. (2015). Simulation and matching-based approaches for indirect comparison of treatments. *PharmacoEconomics*, *33*(6), 537–549. https://doi.org/10.1007/s40273-015-0271-1

Jackson C. (2023). survextrap: A package for flexible and transparent survival extrapolation. *BMC Medical Research Methodology*, *23*(1), article number 282. ISSN 1471-2288. https://doi.org/10.1186/s12874-023-02094-1

Jackson C. H. (2016). flexsurv: A platform for parametric survival modeling in R. *Journal of Statistical Software*, *70*(8), 1–33. https://doi.org/10.18637/jss.v070.i08

Jackson D., Rhodes K., & Ouwens M. (2020). Alternative weighting schemes when performing matching-adjusted indirect comparisons. *Research Synthesis Methods*, *12*(3), 333–346. https://doi.org/10.1002/jrsm.1466

Kahan B. C., Jairath V., Doré C. J., & Morris T. P. (2014). The risks and rewards of covariate adjustment in randomized trials: An assessment of 12 outcomes from 8 studies. *Trials*, *15*(1), article number 139. https://doi.org/10.1186/1745-6215-15-139

Lambert P. C., Sutton A. J., Abrams K. R., & Jones D. R. (2002). A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *Journal of Clinical Epidemiology*, *55*(1), 86–94. https://doi.org/10.1016/S0895-4356(01)00414-0

Leahy J., & Walsh C. (2019). Assessing the impact of a matching-adjusted indirect comparison in a Bayesian network meta-analysis. *Research Synthesis Methods*, *10*(4), 546–568. https://doi.org/10.1002/jrsm.1372

Li Z., & Cao J. (2022). 'General p-splines for non-uniform b-splines', arxiv, arxiv:2201.06808, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2201.06808

Lu G. B., & Ades A. E. (2004). Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine*, *23*(20), 3105–3124. https://doi.org/10.1002/sim.1875

National Institute for Health and Care Excellence (2024). TA1013: Quizartinib for induction, consolidation and maintenance treatment of newly diagnosed FLT3-ITD-positive acute myeloid leukaemia. Committee papers, National Institute for Health and Care Excellence, https://www.nice.org.uk/guidance/ta1013/.

Owen A. B. (2013). *Monte Carlo theory, methods and examples*. https://artowen.su.domains/mc/.

Phillippo D. M. (2019). *Calibration of Treatment Effects in Network Meta-Analysis using Individual Patient Data* [PhD thesis]. University of Bristol. https://research-information.bris.ac.uk/.

Phillippo D. M. (2024). *multinma: Network Meta-Analysis of Individual and Aggregate Data in Stan*. https://cran.r-project.org/package=multinma. R package.

Phillippo D. M., Ades A. E., Dias S., Palmer S., Abrams K. R., & Welton N. J. (2016). *NICE DSU technical support document 18: Methods for population-adjusted indirect comparisons in submission to NICE* (Technical Report). National Institute for Health and Care Excellence. http://www.nicedsu.org.uk.

Phillippo D. M., Ades A. E., Dias S., Palmer S., Abrams K. R., & Welton N. J. (2018). Methods for population-adjusted indirect comparisons in health technology appraisal. *Medical Decision Making*, *38*(2), 200–211. https://doi.org/10.1177/0272989X17725740

Phillippo D. M., Dias S., Ades A. E., Belger M., Brnabic A., Saure D., Schymura Y., & Welton N. J. (2022). Validating the assumptions of population adjustment: Application of multilevel network meta-regression to a network of treatments for plaque psoriasis. *Medical Decision Making*, *43*(1), 53–67. https://doi.org/10.1177/0272989X221117162

Phillippo D. M., Dias S., Ades A. E., Belger M., Brnabic A., Schacht A., Saure D., Kadziola Z., & Welton N. J. (2020a). Multilevel network meta-regression for population-adjusted treatment comparisons. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *183*(3), 1189–1210. https://doi.org/10.1111/rssa.12579

Phillippo D. M., Dias S., Ades A. E., & Welton N. J. (2020b). Assessing the performance of population adjustment methods for anchored indirect comparisons: A simulation study. *Statistics in Medicine*, *39*(30), 4885–4911. https://doi.org/10.1002/sim.8759

Phillippo D. M., Dias S., Ades A. E., & Welton N. J. (2020c). Equivalence of entropy balancing and the method of moments for matching-adjusted indirect comparison. *Research Synthesis Methods*, *11*(4), 568–572. https://doi.org/10.1002/jrsm.1416

Phillippo D. M., Dias S., Ades A. E., & Welton N. J. (2021). Target estimands for efficient decision making: Response to comments on "assessing the performance of population adjustment methods for anchored indirect comparisons: A simulation study". *Statistics in Medicine*, *40*(11), 2759–2763. https://doi.org/10.1002/sim.8965

Phillippo D. M., Dias S., Elsada A., Ades A. E., & Welton N. J. (2019). Population adjustment methods for indirect comparisons: A review of national institute for health and care excellence technology appraisals. *International Journal of Technology Assessment in Health Care*, *35*(03), 221–228. https://doi.org/10.1017/S0266462319000333

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Remiro-Azócar A., Heath A., & Baio G. (2021). Conflating marginal and conditional treatment effects: Comments on "assessing the performance of population adjustment methods for anchored indirect comparisons: A simulation study". *Statistics in Medicine*, *40*(11), 2753–2758. https://doi.org/10.1002/sim.8857

Remiro-Azócar A., Heath A., & Baio G. (2022). Parametric g-computation for compatible indirect treatment comparisons with limited individual patient data. *Research Synthesis Methods*, *13*(6), 716–744. https://doi.org/10.1002/jrsm.1565

Ren S., Ren S., Welton N. J., & Strong M. (2024). Advancing unanchored simulated treatment comparisons: A novel implementation and simulation study. *Research Synthesis Methods*, *15*(4), 657–670. ISSN 1759-2887. https://doi.org/10.1002/jrsm.1718

Riley R. D., Lambert P. C., & Abo-Zaid G. (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. *British Medical Journal*, *340*(feb05 1), c221. https://doi.org/10.1136/bmj.c221

Saramago P., Sutton A. J., Cooper N. J., & Manca A. (2012). Mixed treatment comparisons using aggregate and individual participant level data. *Statistics in Medicine*, *31*(28), 3516–3536. https://doi.org/10.1002/sim.5442

Signorovitch J. E., Wu E. Q., Yu A. P., Gerrits C. M., Kantor E., Bao Y. J., Gupta S. R., & Mulani P. M. (2010). 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. https://doi.org/10.2165/11538370-000000000-00000

Spiegelhalter D. J., Best N. G., Carlin B. P., & van der Linde A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Stan Development Team (2023). *Stan Language Reference Manual*, https://mc-stan.org/users/documentation/

Sutton A. J., Kendrick D., & Coupland C. A. C. (2008). Meta-analysis of individual- and aggregate-level data. *Statistics in Medicine*, *27*(5), 651–669. https://doi.org/10.1002/sim.2916

Therneau T. M., & Grambsch P. M. (2000). *Modeling survival data: Extending the Cox model*. Statistics for biology and health. Springer-Verlag. ISBN 0387987843.

Vehtari A., Gelman A., & Gabry J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari A., Gelman A., Simpson D., Carpenter B., & Bürkner P.-C. (2020). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*, *16*(2), 667–718. https://doi.org/10.1214/20-ba1221

Zhang L., Bujkiewicz S., & Jackson D. (2024). Four alternative methodologies for simulated treatment comparison: How could the use of simulation be re-invigorated? *Research Synthesis Methods*, *15*(2), 227–241. ISSN 1759-2887. https://doi.org/10.1002/jrsm.1681