



This is a repository copy of *Modeling the multi-state natural history of rare diseases with heterogeneous individual patient data: a simulation study*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/205230/>

Version: Published Version

---

**Article:**

Broomfield, J. [orcid.org/0000-0003-1846-2150](https://orcid.org/0000-0003-1846-2150), Abrams, K.R., Freeman, S. et al. (3 more authors) (2024) Modeling the multi-state natural history of rare diseases with heterogeneous individual patient data: a simulation study. *Statistics in Medicine*, 43 (1). pp. 184-200. ISSN 0277-6715

<https://doi.org/10.1002/sim.9949>

---

**Reuse**

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

**Takedown**


If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.



[eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk)  
<https://eprints.whiterose.ac.uk/>

## RESEARCH ARTICLE

# Modeling the multi-state natural history of rare diseases with heterogeneous individual patient data: A simulation study

Jonathan Broomfield<sup>1</sup>  | Keith R. Abrams<sup>2,3</sup> | Suzanne Freeman<sup>1</sup> |  
Nicholas Latimer<sup>4</sup> | Mark J. Rutherford<sup>1</sup> | Michael J. Crowther<sup>5</sup> | On behalf of Project  
HERCULES, the Cooperative International Neuromuscular Research Group investigators  
and Duchenne Regulatory Science Consortium members

<sup>1</sup>Biostatistics Research Group,  
Department of Population Health  
Sciences, University of Leicester,  
Leicester, UK

<sup>2</sup>Department of Statistics, University of  
Warwick, Coventry, UK

<sup>3</sup>Centre for Health Economics, University  
of York, York, UK

<sup>4</sup>School of Health and Related Research  
(ScHARR), University of Sheffield,  
Sheffield, UK

<sup>5</sup>Red Door Analytics, Stockholm, Sweden

## Correspondence

Jonathan Broomfield, Biostatistics  
Research Group, Department of  
Population Health Sciences, University of  
Leicester, Leicester, UK.  
Email: [jb781@le.ac.uk](mailto:jb781@le.ac.uk)

## Funding information

National Institute for Health and Care  
Research, Grant/Award Number:  
NIHR300984

Multi-state survival models are used to represent the natural history of a disease, forming the basis of a health technology assessment comparing a novel treatment to current practice. Constructing such models for rare diseases is problematic, since evidence sources are typically much sparser and more heterogeneous. This simulation study investigated different one-stage and two-stage approaches to meta-analyzing individual patient data (IPD) in a multi-state survival setting when the number and size of studies being meta-analyzed are small. The objective was to assess methods of different complexity to see when they are accurate, when they are inaccurate and when they struggle to converge due to the sparsity of data. Biologically plausible multi-state IPD were simulated from study- and transition-specific hazard functions. One-stage frailty and two-stage stratified models were estimated, and compared to a base case model that did not account for study heterogeneity. Convergence and the bias/coverage of population-level transition probabilities to, and lengths of stay in, each state were used to assess model performance. A real-world application to Duchenne Muscular Dystrophy, a neuromuscular rare disease, was conducted, and a software demonstration is provided. Models not accounting for study heterogeneity were consistently out-performed by two-stage models. Frailty models struggled to converge, particularly in scenarios of low heterogeneity, and predictions from models that did converge were also subject to bias. Stratified models may be better suited to meta-analyzing disparate sources of IPD in rare disease natural history/economic modeling, as they converge more consistently and produce less biased predictions of lengths of stay.

## KEYWORDS

multi-state model, natural history, rare diseases, simulation, survival analysis

## 1 | BACKGROUND

Natural history models describe the progression of a patient with a particular disease over their lifetime. This facilitates disease planning and management with patients,<sup>1</sup> and can also identify areas of rapid disease progression, against which new treatments may be targeted. These models are often used in a cost-effectiveness analysis (CEA) or health technology assessment (HTA),<sup>2</sup> by representing the current standard of care for patients and providing a baseline for such new treatments to be compared to in terms of their efficacy and cost. As such, the models need to be generalizable to future study populations, in case they differ from the population that the original natural history model is estimated from.

A popular method of natural history modeling, in particular for progressive diseases, is through multi-state models,<sup>3,4</sup> where different, clinically meaningful or financially distinct stages of the disease are represented as states that patients can transition through. Timescales of age or time since diagnosis are often used, since these are very important clinical predictors of how quickly patients progress through a disease.

To estimate such a model, a large natural history study will ideally have been conducted on a large cohort of patients in the population of interest, with health outcomes collected throughout follow-up to populate the different stages of the disease.<sup>2</sup> An example is the United States' Surveillance, Epidemiology and End Results (SEER) program.<sup>5</sup> The advantages of this approach are that a model estimated from such a study will serve as an accurate baseline against which to compare new treatments from a clinical trial in this population, and that the model is unlikely to be sparsely populated. Such cohorts are generally only available in more prevalent diseases like breast cancer.<sup>6,7</sup> When modeling the natural history of a rare disease, such a large study is unlikely to be feasible,<sup>8</sup> as recruitment is problematic due to the low prevalence. Running a smaller study would raise ethical issues since robust conclusions are unlikely to be reached.<sup>9</sup> As a result, constructing a natural history model of a rare disease often requires synthesis of different data sources.

Control arms of clinical trials can be synthesized, but this introduces a possibility of heterogeneity between data sources, for example if the clinical trials were conducted in different countries, where clinical practice may vary, or at different points in time, when standards of care may have improved. For rarer diseases, this possibility is increased as the sources of data lessen and so have to be pooled from increasingly different populations, such as by combining registry data with clinical trial data. Moreover, the data sources may relate to different stages of the disease, since follow-up is more likely to be limited (particularly from clinical trials) and there may be stages for which no data are available. In this situation, data could be elicited from professional opinion,<sup>2</sup> although this is of course much more prone to bias. Covariate information will also be much sparser in rare disease analysis, and is likely to be inconsistently recorded across multiple data sources.

This synthesis of evidence between disparate, small data sources can create problems for the generalizability of the natural history model. A model that does not account for the heterogeneity between data sources is likely to make inaccurate predictions. Moreover, it will not be adaptable to new study populations, which restricts the use of the model since it cannot then be used in a HTA of a treatment in a new population. Different methods of adjusting for this heterogeneity make different assumptions that affect the interpretation of predictions, such as through the incorporation of random effects that can result in conditional and marginal predictions.<sup>3,10</sup> It is not simple to determine which model/predictions should be used to provide a meaningful comparison to a new study in a HTA. It has been argued that conditional models should be the baseline from which both conditional and marginal predictions can be estimated.<sup>11</sup> Additionally, these methods are more computationally intensive, since the estimates of heterogeneity will be reliant upon just a few observations (the number of data sources), and so models may not converge and heterogeneity estimates may not be identifiable<sup>12</sup> in a rare disease context. This is an important measure to correctly estimate, since it is through this that the CEA/HTA can be conducted by providing a meaningful baseline comparison for the current standard of care in the disease.

This simulation study sought to identify the most appropriate methods for combining multiple disparate sources of multi-state data, when the number and size of the sources are small. The focus was on identifying methods that both correctly estimate the disease progression in the population(s) for which data are available, and estimate the heterogeneity between these populations, which allows the generalization of the natural history model to future studies in new populations. A real-world application of the methods was also conducted on a dataset from a variety of studies of Duchenne Muscular Dystrophy (DMD), a neuromuscular rare disease primarily affecting boys. This allowed a software implementation of the methods to be demonstrated, and highlights the different interpretations of study predictions from each model. A framework of models for multi-state meta-analysis of survival IPD is presented in Section 2, after which the motivating example of DMD is presented in Section 3. The details of the simulation study are discussed in Section 4 and the results are given in Section 5, followed by a discussion in Section 6.

## 2 | MULTI-STATE SURVIVAL IPD META-ANALYSIS METHODS

A general framework for meta-analyzing multi-state IPD survival data is presented, with different modeling options. The main distinction between the models is whether to adjust for study source in one or two stages, but assumptions around shared or independent transition parameters can also be varied. Weibull baseline hazards are considered, with proportional covariate effects, but the methods can easily generalise to alternative baselines, such as modeling the timescale using restricted cubic splines, or to more complex covariate relationships.

### 2.1 | Notation

The following notation is used. Patients are denoted by  $i$ , studies by  $j$  and transitions by  $k$ . The transitions in a multi-state model are labeled from 1 to  $K$ , with an associated transition-specific hazard function for a patient in a study  $h_{ijk}(t)$ .

### 2.2 | No adjustment

The simplest method in the framework is to assume no heterogeneity between studies. A common Weibull baseline hazard function, with scale and shape parameters  $\lambda_k$  and  $\gamma_k$  respectively, patient covariate information  $\mathbf{X}_{ijk}$  with associated effects  $\beta_k$  and observed event time  $t$ , is estimated for each transition with shared parameters across studies:

$$h_{ijk}(t) = \lambda_k \gamma_k t^{\gamma_k - 1} \exp(\mathbf{X}_{ijk} \beta_k)$$

This method does not adjust for study source, and so if a new study were being compared to a natural history model constructed with this method then there would be no adjustment possible to ensure comparability between the two populations. However, this restrictive assumption does increase the likelihood of model convergence, and making model predictions (and interpreting them) is more straightforward, since only fixed-effects parameters are estimated.

In this and subsequent models, as many covariates as desired can be included in the linear predictor as  $\mathbf{X}_{ijk} \beta_k$ . This formulation assumes transition-specific, proportional hazards for covariate effects that are common across studies (which also allows covariates that change over time to be included), but it is easy to relax this assumption to non-linear relationships or include interactions between covariates if desired.

### 2.3 | One-stage adjustment (frailty)

Many different one-stage methods exist to adjust for study source.<sup>3,13-17</sup> These methods fall under a wide variety of names, some of which are synonymous for identical methods. There are hierarchical methods, which introduce levels to the model to allow for effects at the study level and at the patient level.<sup>13,16</sup> There are random-effect methods, which allow parameters within the model to vary by study source but come from a common distribution.<sup>14</sup> Some of these are termed frailty models, since the random-effect parameter(s) can give a measure of whether studies contain more or less frail patients.<sup>3,12</sup> In some contexts, the data sources are not grouped by study, but rather by geographic location, or nested within individuals, and so the more general term of cluster is given to these groups.<sup>18,19</sup>

One of the simplest cases of one-stage adjustment for study source is to introduce random effects for the scale parameters  $\lambda_k$  in each (transition-specific) baseline hazard:

$$h_{ijk}(t) = \lambda_k \gamma_k t^{\gamma_k - 1} \exp(\alpha_{jk} + \mathbf{X}_{ijk} \beta_k),$$

$$\alpha_{jk} \sim N(0, \sigma_k^2)$$

This method assumes independent frailties for each transition, allowing the scale parameters to vary in each transition with different variances for each transition. This constricts baseline hazards to be proportional between studies. An overall mean scale  $\lambda_k$  is estimated conditional on zero frailty. Since no study will ever have precisely zero frailty, it may be more useful to estimate study-specific scale parameters through the empirical Bayes estimator of the random effects.<sup>20</sup> Then, if a new study is similarly frail to a study in the original analysis (ie, with similar relevant covariate distributions) then natural

history predictions from this original study can serve as a baseline, while benefiting from the increased strength/power by including the other original studies. This does rely on identifying an original study that is suitably similar to the new study. Another solution is to obtain marginal predictions by (numerically) integrating out the frailty. However, while this approach is available in software packages for single transitions in a frailty survival analysis, it has not been generalized to a multi-state setting due to the increased computational intensity that arises from transitions with delayed entry and the possibility of shared parameters across transitions. This study therefore focused on the estimates conditional on zero frailty.<sup>11</sup>

A random-effects model also complicates the likelihood estimation, since each study's contribution to the likelihood is now conditional on random effects, which need to be marginalised. Since  $\alpha_{jk}$  is assumed to follow a Normal distribution with mean 0, then the integral cannot be calculated in closed form, and so numerical integration must be used to estimate the integral and to then maximise the likelihood. An alternative to the Normal distribution is to assume Gamma-distributed random effects, which allow a closed form likelihood. Mean-variance adaptive Gauss-Hermite quadrature was used to evaluate this integral,<sup>16</sup> but this is a computationally intensive method, particularly when the number of studies is low. Other integration methods, such as non-adaptive Gauss-Hermite quadrature or Monte Carlo integration,<sup>21</sup> could be used, but in the context of a simulation study are hard to rigorously assess. It is worth noting that model convergence is often more of a computational limitation than a methodological one. In other words, a model that does not converge may perform better (ie, be less biased) than one that does, provided it can be made to converge. This could happen with novel software developments or altering techniques/software choices used for model fitting and estimation (such as optimising starting values), meaning that it is not worth disregarding a model purely because it does not converge.

The assumptions made by this model could be altered depending on beliefs about the nature of the data at hand. For instance, a shared frailty term  $\alpha_j \sim N(0, \sigma^2)$  could have been assumed, restricting the between-study variance of the scale parameters in each transition to be the same. This would require a stacked model with a likelihood dependent upon all transitions. Random effects could also have been considered on the shape parameters  $\gamma_k$ , or on the covariate effects to investigate if these vary between study—again, either independent to each transition or shared across all transitions.

## 2.4 | Two-stage models

### 2.4.1 | Proportional two-stage model

The one-stage frailty model discussed above can also be estimated in two stages. The first stage is to directly estimate a different scale parameter for each study on each transition:

$$h_{ijk}(t) = \lambda_{jk} \gamma_k t^{\gamma_k - 1} \exp(\mathbf{X}_{ijk} \boldsymbol{\beta}_k)$$

This gives  $J$  different scale parameters  $\lambda_{1k}, \dots, \lambda_{Jk}$  per transition, one for each study, as well as shared parameters  $\gamma_k, \boldsymbol{\beta}_k$  per transition. The second stage is then to obtain population-wide estimates of  $\lambda_k$  for each transition through a meta-analysis:

$$\hat{\lambda}_{jk} \sim N(\lambda_k, \sigma_k^2)$$

The marginal estimate of  $\lambda_k$  can now be used alongside the estimates of the fixed-effects parameters  $\gamma_k, \boldsymbol{\beta}_k$ . Uncertainty in these estimates should account for both within-study variability ( $\text{Var}(\lambda_{jk})$ ) and between-study variability ( $\sigma_k^2$ ). Fitting this model over two stages, rather than in one, reduces the computational demand and so increases the probability of convergence.

### 2.4.2 | Stratified two-stage model

The methods discussed so far have assumed proportional (or identical) baseline hazards between studies. An alternative to this is to consider stratified baseline hazards by studies. This method has often been adopted when interest lies in a treatment effect, since one can adjust for different study populations by baseline stratification but assume shared, or random, covariate effects.<sup>13,14</sup> A second stage to these methods is considered here where the baseline parameters are

also synthesized to obtain baseline estimates for the whole population. This method differs to the proportional two-stage method in that it stratifies both scale and shape parameters in the first stage:

$$h_{ijk}(t) = \lambda_{jk} \gamma_{jk} t^{\gamma_{jk}-1} \exp(X_{ijk} \beta_k)$$

Population-wide estimates are then obtained from the study-specific estimates through a multivariate meta-analysis:

$$\begin{pmatrix} \hat{\lambda}_{jk} \\ \hat{\gamma}_{jk} \end{pmatrix} \sim \text{MVN} \left( \begin{pmatrix} \lambda_k \\ \gamma_k \end{pmatrix}, \begin{pmatrix} \sigma_{\lambda,k}^2 & \rho_k \sigma_{\lambda,k} \sigma_{\gamma,k} \\ \rho_k \sigma_{\lambda,k} \sigma_{\gamma,k} & \sigma_{\gamma,k}^2 \end{pmatrix} \right)$$

The method presented above stratifies only Weibull baseline parameters; this could be extended to covariate effects or more complicated baselines if desired, although the same distribution would have to be applied to each study. Different distributions could be applied to different transitions if desired.

## 2.5 | Multi-state estimands

In a standard survival analysis, common estimands of interest include survival probabilities (the probability of not experiencing the event) and restricted mean survival time (the mean time spent without experiencing the event, calculated as the area under the survival probability function). The multi-state equivalent of these estimands are transition probabilities and lengths of stay (LOS).

Transition probabilities give the probability of being in a state at a later time  $\tau$  in follow-up, given a state occupancy at an earlier time. A special case of these probabilities is calculated from time 0 when patients all begin in state 1; these are referred to as state occupancy probabilities. If the process  $G(t)$  represents the state a patient is in at time  $t$ , then these probabilities are defined as:

$$p_k(\tau) = P(G(\tau) = k | G(0) = 1) \quad (1)$$

LOS represent the mean times spent in each state up to this later follow-up time, and can be obtained from calculating the areas under the curve of transition probabilities:

$$\text{LOS}_k(\tau) = \int_0^\tau p_k(u) du \quad (2)$$

Transition probabilities (and thus LOS) will depend on the estimated hazard functions between states. In some simple cases, closed-form expressions of transition probabilities can be written with analytic solutions available. However, it is often more straightforward, and in some cases essential, to use a numerical approach instead.<sup>22</sup> In this study, a simulation approach (separate to the simulation of data to evaluate the different methods) was adopted to estimate transition probabilities and LOS. Once hazard functions  $h_{ijk}(t)$  were estimated from a given model, the disease progression of many patients were simulated according to these estimates and transition probabilities were calculated as the proportion of patients in each state at given points in follow-up. LOS could then be calculated from the areas under these transition probabilities.

When predicting these measures from a random-effects model, the choice of conditional or marginal predictions must be made. Conditional predictions may be easier to compute, and useful for study-specific inference, but frailty predictions might be more interpretable for the general population.

## 3 | APPLICATION TO DMD DATA

### 3.1 | Introduction to the data

The four methods in the dataset were applied to a real-world dataset of patients with DMD, made available from the C-Path Duchenne Regulatory Science Consortium (D-RSC).<sup>23</sup> An overview of the dataset is provided in Table 1, with



TABLE 1 Overview of the DMD dataset.

Study	Region	Study type	<i>n</i>	Age range	Study period
Pooled	International	—	1005	4–34	2004–2018
1	USA	Natural history	38	4–31	10 years follow-up
2	USA	Clinical test data	22	4–14	2009
3	International	Natural history	427	4–34	2006–2009
4	International	Trial placebo arm	113	7–15	2013–2015
5	USA	Natural history	81	5–18	2010–2018
6	International	Trial placebo arm	57	5–16	2008–2009
7	International	Trial placebo arm	114	7–15	2013–2014
8	International	Clinical trial	63	4–12	2004–2007

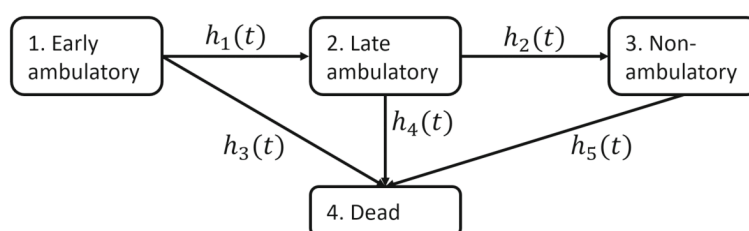


FIGURE 1 Assumed multi-state structure of the DMD dataset.

further information in Supplementary Material 2. Example code is provided in Supplementary Material 3, applied to a simulated dataset (also available).

There were eight studies from a range of locations; three of the studies were conducted in the USA, while the remaining five contained international populations with patients from North America, South America, Europe, Asia and Oceania. The studies were a combination of natural history cohorts and clinical trials, with large variation in the total follow-up; follow-up from the natural history studies typically lasted longer than from the clinical trials. As a result, there is likely to be a fair amount of heterogeneity between the data sources.

The dataset contained information on test scores such as the 6-min walk distance<sup>24</sup> and forced vital capacity, from which states were determined corresponding to earlier (ambulatory) and later (non-ambulatory and ventilated) stages of DMD. No mortality data were present, and so reconstructed mortality IPD were instead used from a systematic literature review of 12 international studies on mortality in patients with DMD.<sup>25</sup> Figure 1 shows the multi-state structure that was assumed to demonstrate and contrast the methods.

The five transitions in the model have been labeled sensibly, but any ordering of transition labels can be used (so long as they are specified in the correct order in the transition matrix).

### 3.2 | Comparison of models

The DMD dataset informed transitions 1 and 2, and the reconstructed mortality dataset informed transitions 3, 4, and 5. Patients in the mortality dataset were mapped to intermediate states 1–3 using the age distributions of patients in these states in the DMD dataset. While the amount of follow-up between studies varied, the models are able to account for this by estimating each transition hazard from follow-up contributions of any patients observed to make (or be at risk of making) the transition. Covariate information (beyond that which was used to map patients onto each state) was not available consistently across studies to be included in the model specification. In the Supplementary Materials, the example code is applied to a simulated dataset where two covariates are available, and so these covariates are included in the model demonstration and comparison to highlight differences when covariates are available in the real world.

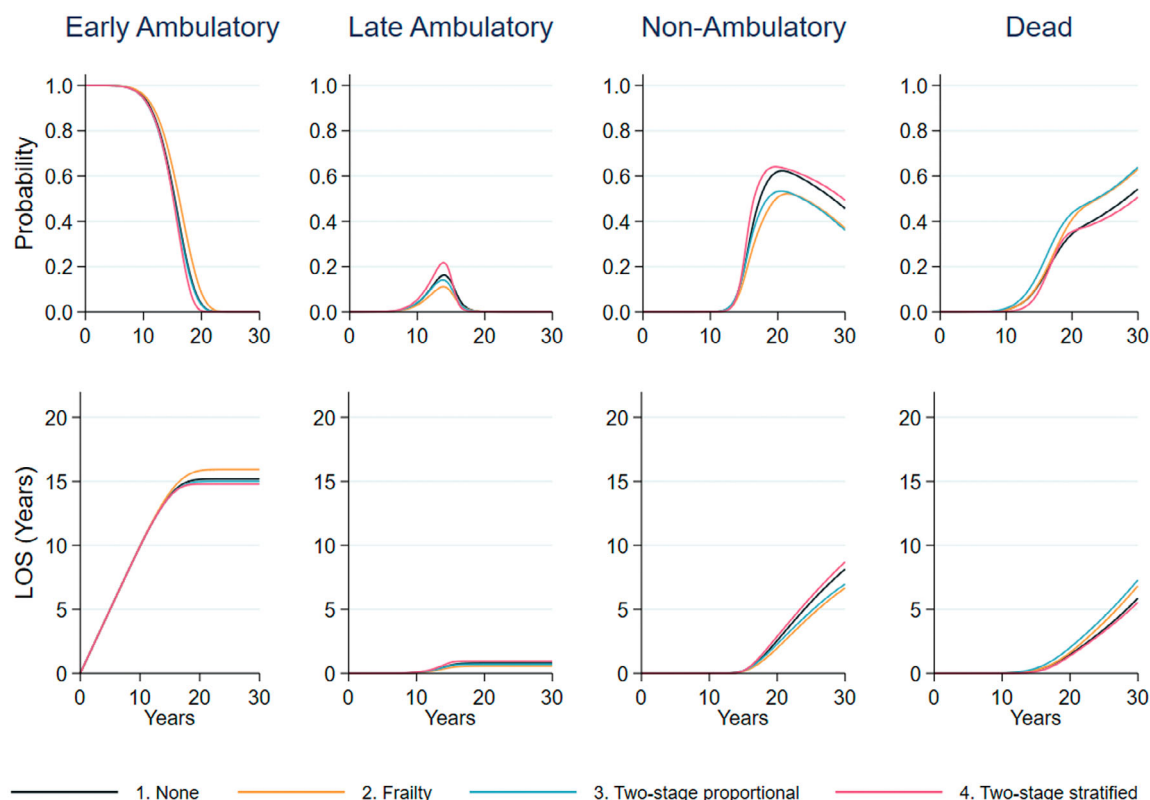


FIGURE 2 Model predictions of transition probabilities and lengths of stay in the DMD states.

TABLE 2 Lengths of stay in the DMD states at age 30.

Method	Early ambulatory	Late ambulatory	Non-ambulatory	Dead
None	15.18 (14.71, 15.68)	0.80 (0.48, 1.33)	8.20 (6.57, 10.24)	5.81 (4.64, 7.27)
Frailty	15.92 (15.54, 16.30)	0.57 (0.36, 0.92)	6.67 (5.33, 8.34)	6.84 (5.62, 8.33)
Two-stage proportional	14.97 (14.52, 15.44)	0.72 (0.45, 1.15)	6.88 (5.58, 8.48)	7.43 (6.19, 8.92)
Two-stage stratified	14.79 (13.85, 15.80)	0.95 (0.56, 1.61)	8.70 (5.64, 13.40)	5.56 (2.85, 10.87)

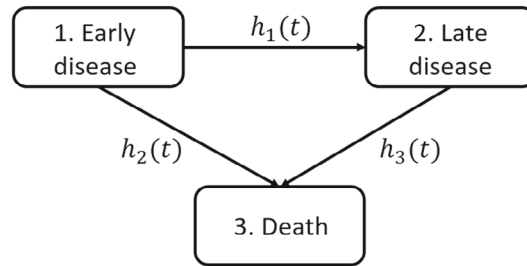
The estimates of transition probabilities and LOS are presented across the four health states in Figure 2. Table 2 presents LOS at 30 years of age across the health states with confidence intervals to quantify the uncertainty in model predictions.

There is some discrepancy between all four models in both transition probabilities and LOS. For instance, after 30 years, the no-adjustment model predicted 8.2 years spent in the non-ambulatory state, the frailty model predicted 6.7 years, the two-stage proportional model predicted 6.9 years and the two-stage stratified predicted 8.7 years. A range of 2 years between the four models is significant in terms of a possible error in a reference population to which a new treatment is being compared, and may lead to inappropriate decision-making from a HTA that relied on an incorrect model. Some of the difference in predictions between models could be because predictions from the frailty model are conditional on zero frailty, rather than marginal, although they do align with the two-stage proportional model's predictions, suggesting this conditional estimate may be close to what the marginal frailty predictions would be.

## 4 | SIMULATION STUDY METHODS

The simulation study followed the “ADEMP” structure proposed by Morris et al,<sup>26</sup> this section details the aims, data generation, estimands, methods, and performance measures of the study.





**FIGURE 3** Illness-death model, with transitions  $h_k(t)$  between states.

## 4.1 | Aims

The primary aim was to determine which method(s) should be used to model the natural history of a rare disease and quantify progression through its different stages. This was assessed by both model accuracy (comparing predicted estimates of transition probabilities and LOS to known truths) and convergence, since methods that correctly account for the nature of the available data may be too complex to estimate reliably, potentially requiring a more restrictive but robust method to be considered instead. Another factor of interest was how accurately methods capture the heterogeneity between study sources, which in rare diseases is likely to be far greater since data sources are typically more heterogeneous and more infrequent. This measure is hard to compare across different methods since different assumptions cause the interpretation of heterogeneity parameters to vary.

## 4.2 | Data generation

Biologically plausible data were simulated based on two exemplar disease trajectories: one with relatively slow progression through intermediate states, such as Duchenne Muscular Dystrophy (DMD),<sup>27</sup> where patients live with the disease and its effects for a number of years (median survival of 28.1 years<sup>25</sup>); and one with more rapid progression, such as infantile-onset spinal muscular atrophy (SMA),<sup>28</sup> where median survival from diagnosis can be just 8 months. This is important to assess the performance of the models across a realistic range of disease types, since rare diseases are typically heterogeneous. Individual patient data (IPD) on disease progression in DMD were available from the Critical Path (C-Path) Institute,<sup>23</sup> from which biologically plausible parameters were derived. Plausible parameters for disease progression in SMA were derived from a published natural history study.<sup>28</sup>

The underlying data structure was an illness-death model, with forward transitions allowed only (patients can move from state 1 directly to state 3 (a terminal state) or via state 2 (an intermediate state)). Figure 3 shows the structure of this model. In the context of the example disease trajectories (DMD and SMA), state 1 corresponds to an early disease state, where mortality rates are lower and quality of life is higher, and state 2 to a late disease state, where the reverse is true. IPD were assumed to be available from each simulated data source.

Studies were simulated independent of one another to allow the data-generating mechanism (DGM) to assume either the same or different (transition-specific) baseline parameters and covariate effects in each study. The baseline hazard function for each transition was assumed to follow a Weibull distribution, and two covariate effects were included—one continuous, one binary. Once the data had been generated for each study, they were pooled into one dataset for analysis.

Firstly, it was assumed that each study had the same baseline parameters and covariate effects (DGM 1, no heterogeneity). This meant that for transition  $k \in \{1, 2, 3\}$ , the hazard function for patient  $i$  in study  $j$ , with covariate values  $X_{1,ijk} \in \mathbb{R}$ ,  $X_{2,ijk} \in \{0, 1\}$  respectively, is:

$$h_{ijk}(t) = \lambda_k \gamma_k t^{\gamma_k - 1} \exp(X_{1,ij} \beta_{1,k} + X_{2,ij} \beta_{2,k})$$

The next two DGMs introduced heterogeneity between studies by allowing the baseline parameters to vary randomly. This was done by taking random draws  $\alpha_{\lambda,ijk}$  and  $\alpha_{\gamma,ijk}$  from  $N(0, \sigma_{\lambda,k}^2)$  and  $N(0, \sigma_{\gamma,k}^2)$  distributions respectively for each study on the log-scale (allowing for the baseline hazard parameters in one study to vary differently across each transition). The baseline parameters were then modified in the following two ways:

**TABLE 3** Parameter values considered in the simulation study.

DGM (Heterogeneity)	$\sigma_{\lambda,k}^2, \sigma_{\gamma,k}^2$	$(\lambda_1, \lambda_2, \lambda_3), (\gamma_1, \gamma_2, \gamma_3)$	$j$	$n_j$	$X_1, X_2$	$\beta_{1,k}, \beta_{2,k}$
1 (None)	—, —	(1.5e – 5, 1.5e – 8, 0.0054)	3	1000	N(0, 10 <sup>2</sup> ),	0.01, 0.5
2A (Proportional)	0.35 <sup>2</sup> , —	(4.2, 6.6, 1.7)	5	100	Bin(1, 0.5)	
2B (Proportional)	1 <sup>2</sup> , —	(0.2, 0.2, 0.2), (1.2, 1.2, 1.2)	10	1000 × 1,		
3A (Stratified)	0.05 <sup>2</sup> , 0.05 <sup>2</sup>			100 × (j – 1)		
3B (Stratified)	0.1 <sup>2</sup> , 0.1 <sup>2</sup>					

$$\begin{aligned}
 \lambda_{jk} &= \lambda_k \exp(\alpha_{\lambda,jk}) \\
 \rightarrow h_{ijk}(t) &= \lambda_{jk} \gamma_k t^{\gamma_k-1} \exp(X_{1,ij} \beta_{1,k} + X_{2,ij} \beta_{2,k}) \\
 \lambda_{jk} &= \lambda_k \exp(\alpha_{\lambda,jk}), \gamma_{jk} = \gamma_k \exp(\alpha_{\gamma,jk}) \\
 \rightarrow h_{ijk}(t) &= \lambda_{jk} \gamma_{jk} t^{\gamma_{jk}-1} \exp(X_{1,ij} \beta_{1,k} + X_{2,ij} \beta_{2,k})
 \end{aligned}$$

Thus, the true baseline parameters for the whole population remained  $\lambda_k$  and  $\gamma_k$  but there was heterogeneity between study populations. In DGM 2, the baseline hazards were proportional between trials since they differ only by the multiplicative term  $\exp(\alpha_{\lambda,jk})$ . In DGM 3, the hazards were stratified between trials since each trial has its own baseline parameters  $\lambda_{jk}$  and  $\gamma_{jk}$  on transition  $k$ . However, the nature of the data generation meant that the true hazard functions (and thus true estimands, which were predictions based on these hazard functions) were known—both in terms of the overall population and in each study.

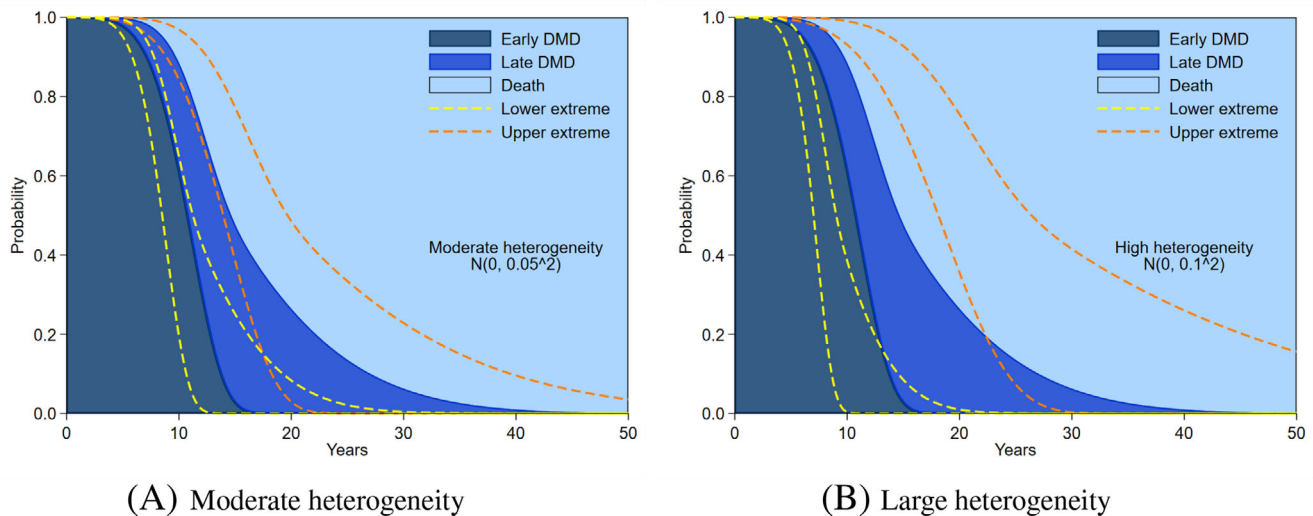
Table 3 lists the different parameter values that were varied factorially in the simulation study.

There were  $5 \times 2 \times 3 \times 3 \times 1 \times 1 = 90$  different scenarios considered in total. For the five different variances across the three DGMs given in Table 3, the following scenarios were considered. Three, five and 10 studies were simulated for each permutation of parameter values; once with 1000 patients in each study, once with 100 patients in each study and once with 1000 patients in one study and 100 patients in the other studies. True parameter transition values for  $(\lambda_1, \lambda_2, \lambda_3)$  and  $(\gamma_1, \gamma_2, \gamma_3)$  were set to (0.000015, 0.000000015, 0.0054) and (4.2, 6.6, 1.7), then (0.2, 0.2, 0.2) and (1.2, 1.2, 1.2) respectively. The first set of these values correspond to progression rates for a disease such as DMD, where every patient will have died by the age of 50, and have been informed by collapsing the DMD data on disease progression into two stages: early DMD, where the patient can still walk, and late DMD, where the patient can no longer walk and will likely require ventilation support and assistance with hand-to-mouth functions. The second set of values correspond to progression in a disease like SMA, where 50% of patients will have died or moved to permanent ventilatory support within 1 year of birth, and almost all patients will have died within 10 years. Covariates  $X_1$  and  $X_2$  were simulated from N(0, 10<sup>2</sup>) and Bin(1, 0.5) distributions for each patient in each study, representing a mean-centred continuous covariate and a binary covariate. These were assumed to have effects of  $\beta_{1,k} = 0.01$  and  $\beta_{2,k} = 0.5$  on all of the log-hazard functions. In DGM 1 there are no random-effect parameters; in DGM 2, moderate and large values of 0.35<sup>2</sup> and 1<sup>2</sup> were assigned to  $\sigma_{\lambda,k}^2$  (proportional heterogeneity); in DGM 3, moderate and large values of (0.05<sup>2</sup>, 0.05<sup>2</sup>) and (0.1<sup>2</sup>, 0.1<sup>2</sup>) were assigned to  $(\sigma_{\lambda,k}^2, \sigma_{\gamma,k}^2)$  (stratified heterogeneity). Figure 4 shows the transition probabilities that would be observed from the 2.5<sup>th</sup> and 97.5<sup>th</sup> centiles of the Normal distributions with moderate and high variances in DGM 3, motivating the choices of the variance parameters in DGM 3. Variance parameters for DGM 2 were determined in a similar manner.

The main mechanism of the simulation study was to vary both the number of simulated studies and the sample sizes in these studies, to investigate the methods in a rare disease context. Administrative censoring was introduced at 50 years to ensure no extrapolation beyond observed follow-up in the calculation of transition probabilities and LOS was required, allowing the analyses to be focused on the performance of the meta-analysis methods. The *survsim* command was used to simulate multi-state data from the three hazard functions.<sup>29</sup>

### 4.3 | Estimands

The estimands were population-level transition probabilities to and LOS in each state at 10 and 20 years of follow-up for the first set of baseline parameter values (DMD context), and at 2 and 4 years for the second set (SMA context).



**FIGURE 4** Plausible range of transition probabilities under moderate ( $\sigma_{\lambda,k}^2 = \sigma_{\gamma,k}^2 = 0.05$ ) and high ( $\sigma_{\lambda,k}^2 = \sigma_{\gamma,k}^2 = 0.1$ ) variance in shape and scale parameters (DGM 3). (a) Moderate heterogeneity; (b) Large heterogeneity.

These relatively long follow-up times were chosen as the context of the study is natural history modeling, meaning longer follow-up is required to map out the progression of the disease across a patient's lifetime after diagnosis. It was assumed that all patients began in state 1 at origin, equating the transition probabilities to state occupancy probabilities. These estimands were calculated for patients with covariate values of 0 for the continuous variable and 1 for the binary variable. This study used the `predictms` command<sup>30</sup> to numerically estimate transition probabilities and LOS. The true values of the estimands were calculated numerically from the known population-level parameters ( $\lambda_k, \gamma_k, \beta_{1,k}, \beta_{2,k}$ ) for each transition.

A key objective of an economic evaluation is often to estimate the mean quality adjusted life years (QALYs) gained that are associated with a new treatment compared to the standard treatment. Therefore, estimates of mean survival are required (in each state), motivating the use of LOS in each state as an estimand. However, it is possible to accurately estimate the mean LOS for each health state, but to over-/under-predict earlier in a state and under-/over-predict later, which is why using transition probabilities at different points in follow-up as an estimand is also important, since these would reflect this inaccuracy in predictions. The transition probabilities are measured from origin, using a continuous timescale—in an economic context, they have the same interpretation as the proportion of patients in each state after 2 and 4, or 10 and 20 years in a Markov trace table.

#### 4.4 | Methods

Four different methods were used to estimate transition-specific hazards; one approach that did not account for studies, as well as a one-stage approach and two two-stage approaches for adjusting for study source. The first model is termed the no adjustment model, the second is the frailty model, the third is the two-stage proportional model and the fourth is the two-stage stratified model. All models estimated shared transition-specific covariate effects between studies for the continuous and binary covariate, since this is how the data were simulated.

In the frailty model, five-point quadrature was used first to numerically evaluate the likelihood integral. If this did not converge then 11-point and 21-point quadrature were used, consistent with the simulation study used to develop the method.<sup>16</sup> This model (along with the proportional two-stage method) is the true method for analyzing data from DGM 2, where heterogeneity was simulated proportionally between studies. The simulation study therefore provides inference on the identifiability of random effects in rare disease settings, where the low number of studies (which equate to the number of data points from which to identify the variance parameter) is problematic. The no adjustment model corresponds to the true method for analysing data from DGM 1, while the stratified two-stage method is the true method for DGM 3.

The one-stage method included in the simulation study assumed proportional baseline hazards between studies. An alternative to this is to consider stratified baseline hazards by studies. This method has often been adopted when interest lies in a treatment effect, since one can adjust for different study populations by baseline stratification but assume

shared, or random, covariate effects.<sup>13,14</sup> However, this study considered a second stage to these methods where the baseline parameters are also synthesized to obtain baseline estimates for the whole population. Two different methods were considered, the first of which stratified only scale parameters (thus making baselines proportional between studies):

Transition probabilities and LOS were calculated using `predictms`<sup>30</sup> from the estimates  $(\hat{\lambda}_k, \hat{\gamma}_k, \hat{\beta}_{1,k}, \hat{\beta}_{2,k})$ .

## 4.5 | Performance measures

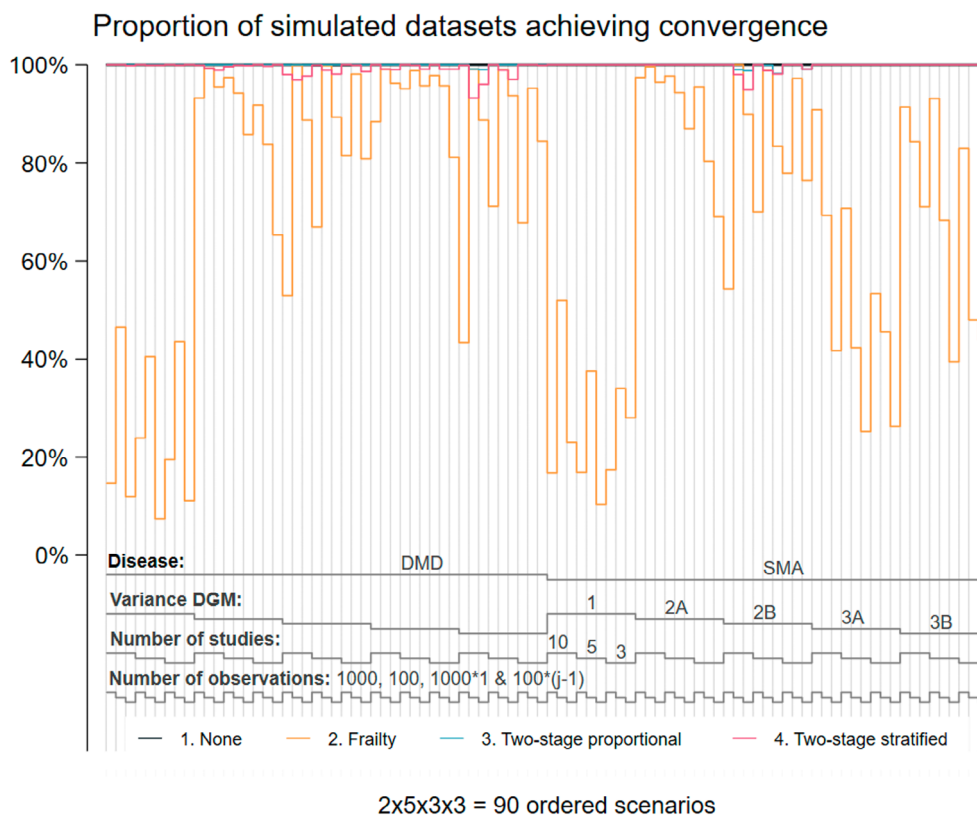
The performance of the estimated transition probabilities and LOS were measured by bias, coverage and empirical standard error, with uncertainty in these quantified by Monte Carlo standard errors (MCSEs).<sup>26</sup> The probability of convergence was also used to assess the performance of each method. A maximum of 30 iterations of log-likelihood convergence were allowed for each method before convergence was deemed to have failed.

## 5 | RESULTS

### 5.1 | Model convergence

The proportion of simulated datasets which converged across all simulation scenarios is presented in Figure 5.

Convergence was very poor for the frailty model, particularly when there was no underlying heterogeneity between studies (meaning the frailty model is trying to overfit the data). This is observed in Figure 5 as the yellow line is lower in scenarios with no study heterogeneity (DGM 1). However, even when the frailty model was the correct model to fit (under proportional heterogeneity in DGM 2), model convergence was still not perfect; for the scenario of a DMD-like disease with 10 studies of 1000 observations each, model convergence was 53% when under moderate scale-parameter heterogeneity and 43% under large scale-parameter heterogeneity. In general, this poor convergence was due to the small number



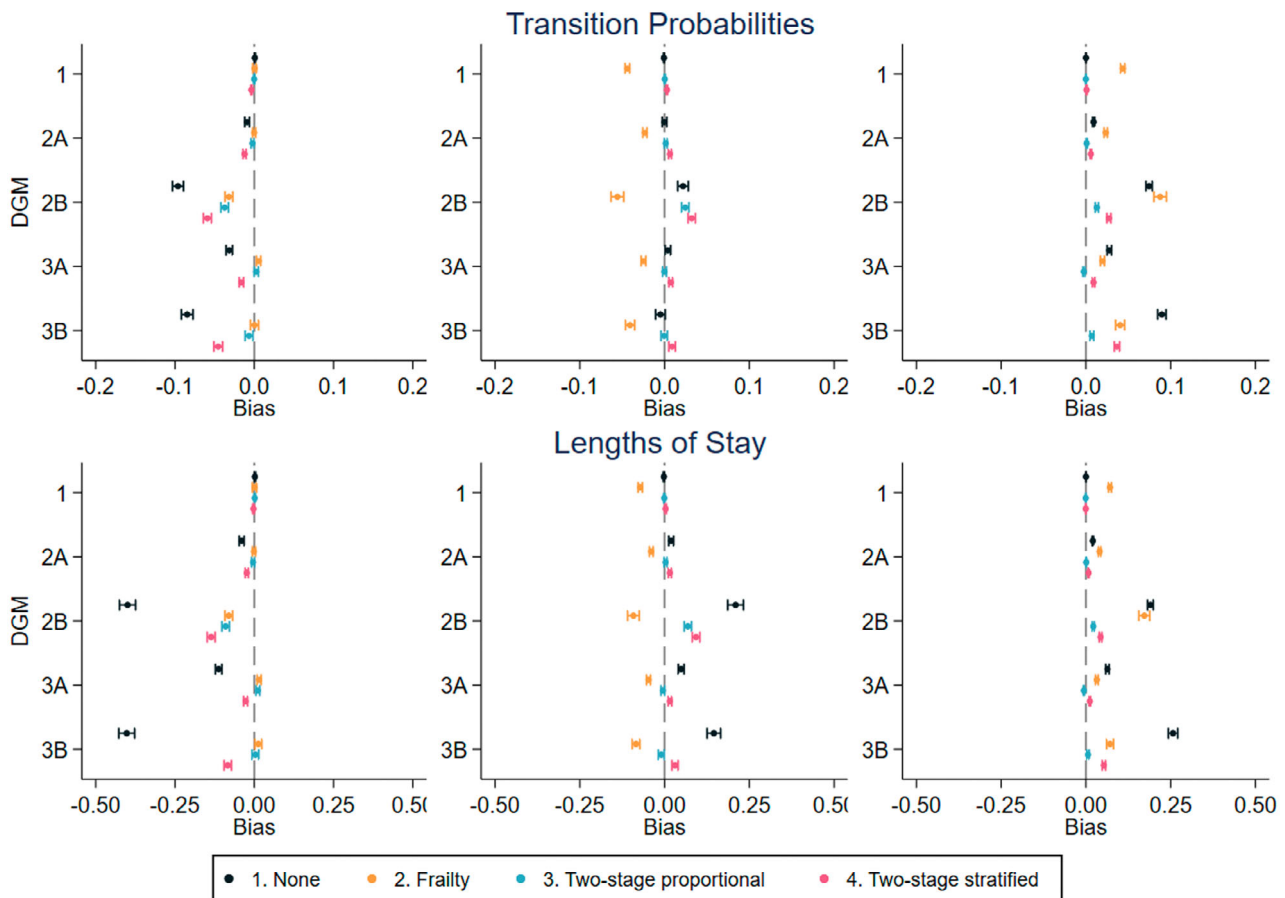
**FIGURE 5** Convergence of models across all scenarios. The key along the bottom of the figure indicates the simulation scenario corresponding to the observed convergence for each model.

of studies that were being meta-analyzed (3, 5, or 10) since this is the effective sample size from which the frailty term is being estimated. Convergence for other models was either perfect or almost perfect; the two-stage proportional model converged at least 98% of the time in all scenarios, and the two-stage stratified model converged at least 93% of the time.

It is possible that in some of the situations where the frailty model did not converge, model convergence could have been achieved (eg, through specifying a greater number of integration points, a lower tolerance criterion for log-likelihood convergence than Stata's default of  $1e-7$ , or a different integration technique, such as non-adaptive Gauss-Hermite quadrature). However, including too many options in a simulation study is impractical, particularly when considering such a range of scenarios. These results highlight the possible difficulties that may occur when attempting to fit a one-stage frailty model to heterogeneous IPD studies of rare diseases. It is likely, though not guaranteed, that model performance would be similar to model performance in scenarios where the model does converge, since it is the same model being fitted with the same assumptions, only tweaking methods for estimating the model parameters. To assess this, 10 simulated datasets across three different scenarios were investigated where the frailty model failed to converge in the simulation study. The model could be made to converge by increasing the number of integration points in the numerical integration of the likelihood function. The bias observed from the 30 converged frailty models across the three scenarios was very close to the bias observed from the subset of models in the same scenarios that converged in the simulation study.

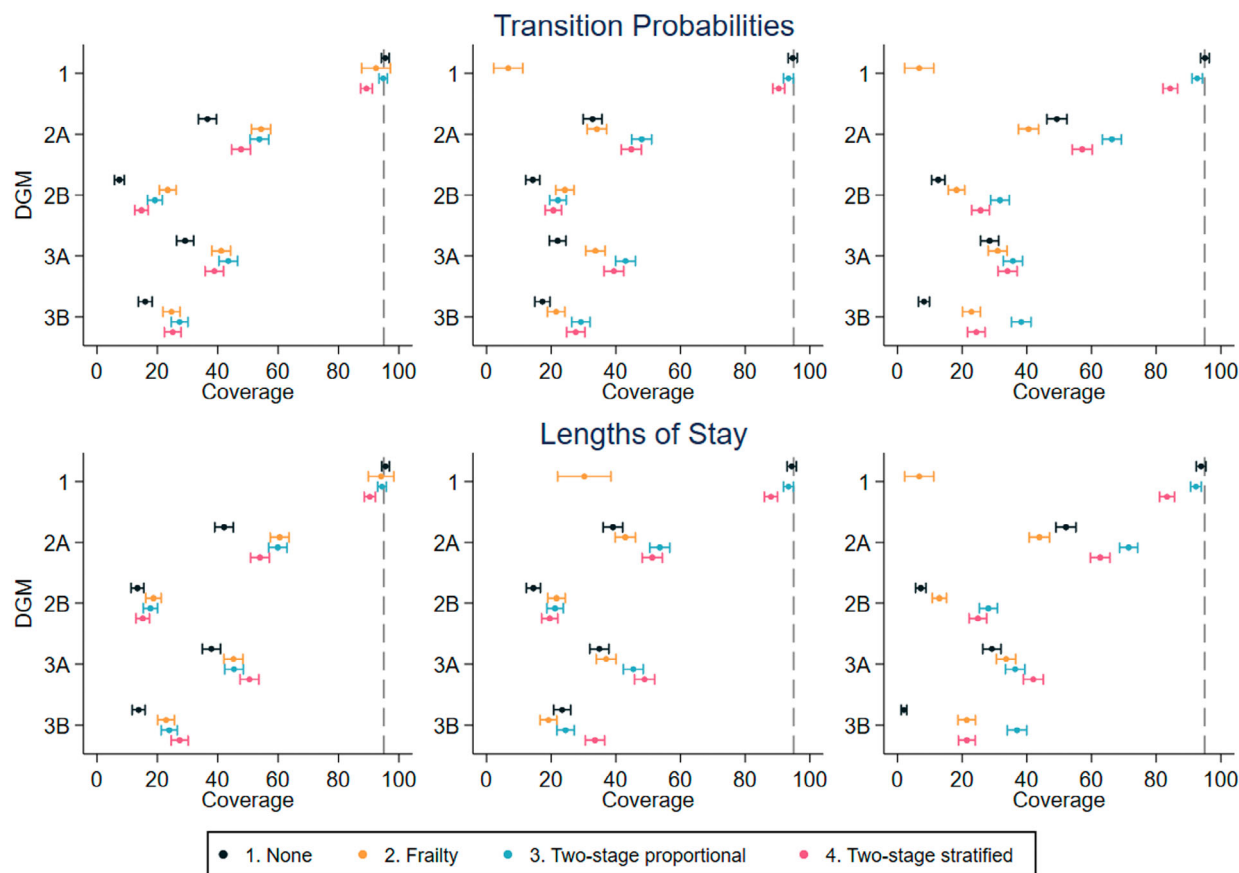
## 5.2 | Bias and coverage

Figures 6 and 7 show respectively the bias and coverage (with confidence intervals calculated using MCSEs) of transition probabilities and LOS at 10 years of follow-up for a disease similar to DMD. The results are shown across methods for scenarios in which there was one study of 1000 patients and nine of 100 patients. Results are presented only for the subset



**FIGURE 6** Bias of transition probabilities and LOS at 10 years for a disease similar to DMD. Scenarios with one study of 1000 patients and nine studies of 100.





**FIGURE 7** Coverage of transition probabilities and LOS at 10 years for a disease similar to DMD. Scenarios with one study of 1000 patients and nine studies of 100.

of models that converged. Bias and coverage are both presented to investigate whether coverage is poor even in situations with minimal bias. Convergence is also shown to highlight the limitations of interpreting results where convergence is low.

Supplementary Figures 1 and 2 show respectively the bias of transition probabilities and LOS at both follow-up points for diseases similar to DMD (10 and 20 years) and SMA (2 and 4 years). The results are shown across methods for all scenarios (again, only for the subset of models that converged). The model with no adjustment for study source was generally the most biased in predicting transition probabilities and LOS for the DGMs shown in Figure 6. In scenarios with high heterogeneity (DGMs 2B and 3B) LOS in state 1 were negatively biased by up to 0.4 years (5 months), which also caused positive bias in states 2 and 3. However, the frailty model performed more poorly than the other models across all scenarios for predictions in states 2 and 3, particularly at later points in follow-up (Supplementary Figure 2) and even when it was the correct model to fit. In some scenarios, a mean bias of  $-2.0$  years spent in state 2 and  $2.0$  years in state 3 was observed from this model, which suggests that even frailty models that converge may not give reliable predictions. This could be due to the fact that the frailty estimates are conditional on zero frailty, rather than marginal over the population. It highlights the lack of identifiability of proportional heterogeneity when the number of studies is small. Some bias was observed from the two-stage models but generally this was of a much lower magnitude. The bias of the two-stage models is lower in states 2 and 3 than the bias of the no adjustment and frailty models. However, there were some scenarios where predictions of LOS in state 1 were more biased for the two-stage models than for the frailty model. In general, the direction of bias was consistent across methods; if predictions in one state were negatively biased for one model, they were negatively biased for all models, and vice versa.

Coverage was good across all scenarios under no heterogeneity, except for transition probabilities to and LOS in states 2 and 3 for the frailty model (Figure 7). Since there was a small amount of bias observed in these scenarios, it is likely that this is what is causing the poor coverage. This suggests that the methods can satisfactorily quantify uncertainty in model predictions under scenarios where there is very minimal or no study heterogeneity. However, once heterogeneity is present, coverage is very poor, and gets worse as this heterogeneity increases.



The calculation of uncertainty in the model predictions only takes into account fixed-effect variability, rather than any estimated random-effect variability. When no random effects are being estimated (in the first model), this method cannot be improved and used to quantify the uncertainty. However, for the other methods, the estimated variance between studies needs to be taken into account in order to obtain suitably wide confidence intervals. This is not currently possible using standard software packages, and so could not be implemented feasibly into the simulation study. A demonstration of how to obtain confidence intervals that account for the random effects as well as the fixed effects for model predictions from the one-stage frailty model is given in Supplementary Material 6. Another factor affecting the poor coverage is that the percentage biases are quite high for some model predictions (particularly in states 2 and 3), due to the true values of these predictions being close to zero. The full simulation results are available in Supplementary Material 5, including empirical standard errors of the estimands from each simulated scenario.

## 6 | DISCUSSION

This paper shows the importance of adjusting for study source when modeling the natural history of rare diseases, despite the difficulties that can arise in parameter estimation and interpretation. It is likely that data sources collected when attempting to estimate the natural history of a rare disease will be heterogeneous in a number of ways, and a failure to account for this in the modelling stage can lead to biases in estimating the time spent in disease/health states. This has a direct impact on the HTA of new or existing treatments, since times spent in states will be used to calculate QALYs based on the utilities of these states. While biological plausibility should always be considered in model selection, if there is disagreement between model predictions then the model that is more flexible in between-study variability (ie, the two-stage stratified model in this study) should be used since it is likely to be capturing more of the heterogeneity between studies than the alternatives.

The lower absolute bias of population-level transition probabilities and LOS observed for the two-stage stratified model in all scenarios considered, particularly for later states, suggests that this approach is the most adept across a range of between-study heterogeneity, although simulation under alternative, non-Weibull data generating mechanisms (such as flexible parametric models) may be required to assess this further. The largest absolute bias when estimating LOS observed for this method was  $-0.14$  years spent in state 1, which was a percentage bias of just  $-1.6\%$ . In contrast, the model with no adjustment for study source was biased by  $-0.40$  years in the equivalent scenario, corresponding to a percentage bias of  $-4.7\%$ . Higher percentage biases were observed for this model, and other models, in other states when the true values were closer to zero. Transition probabilities were more robust to a lack of adjustment for study heterogeneity, but LOS are used more frequently in a HTA. Confidence intervals were poorly estimated due to a lack of adjustment for random-effect heterogeneity.

Model convergence is also an important limitation to consider in the methods discussed. While convergence is not a criteria on which to base model selection, models such as the frailty model may be limited in their usefulness in the setting of rare disease natural history modeling due to their poor convergence. The advantage of models that do not account for study source is that their simplicity makes them much more likely to converge. The fixed-effects models in the first stage of the two-stage models, which stratify either scale or shape and scale Weibull parameters by study, also had high convergence, but the second stage of these models (the meta-analysis of these stratified parameters) was slightly more prone to non-convergence. However, there are options to alleviate this on a case-by-case basis, for example by altering the estimation technique used (the default in *mvmeta* is restricted maximum likelihood, but unrestricted maximum likelihood and multivariate methods of moments can also be used<sup>31</sup>) or the starting values of the variance-covariance matrix, which in the simulation study was set to an identity matrix. Similarly, while convergence of the one-stage frailty model was poor in the simulation study, this can be improved on an individual basis. However, it is possible that the model will never converge, particularly when underlying heterogeneity is low and when the number of studies is low. The non-convergence of the frailty model in a large proportion of the simulated scenarios is a key finding of this paper, particularly even when it was the true underlying model used to simulate the data. As a result of the high levels of observed non-convergence, very limited conclusions should be drawn on the other estimands calculated from the frailty model.

A further motivation to use methods that account for study source is that study-specific model predictions may be able to be made. A failure to do so creates the risk of overfitting, since future study populations will not be represented by the natural history model. For example, in the frailty model, empirical Bayes means of random-effect terms can be estimated in Stata, from which study-specific scale parameters can be obtained. In the two-stage approaches, in order to still include

information from the other studies, study-level covariates could be included to observe how baseline parameters may vary across the studies' population(s). Once these study-specific estimates have been obtained, they can be used as a reference for a new study population that may be more similar in nature to one of the current studies, rather than using the pooled predictions that may not be generalizable. However, deciding which study is the most similar in nature to the new study is not obvious, and relies on consistent study-level covariate information being available across all the studies. It would also be preferable to incorporate such covariate information into the model to obtain a more representative baseline comparison for the new study. In the absence of such information, though, using study-specific predictions may still be preferable to population-wide predictions, which may average over a wide range of geographic locations or time periods and so not be representative of the new study population.

The methods have all been considered in a frequentist approach, but it would be possible to translate them to a Bayesian framework. With the development of Bayes software programs and languages such as Jags and Stan, Bayesian IPD meta-analysis is increasingly feasible (implemented via Markov Chain Monte Carlo algorithms), and may have some particular use in a rare disease setting, where sparser transitions could be fed by prior clinical beliefs. This is a worthwhile and planned area of future research.

The simulation study considered 90 different variations in data-generation factorially, allowing for a wide range of biologically plausible scenarios to be considered. Scenarios not considered in the simulation study were more complex baseline hazards and covariate relationships, varying the follow-up time between studies, extensions beyond the illness-death model considered in the study (although it is likely that these results will generalise to more complex disease structures), and pooling studies with IPD together with studies that only have aggregate data available. It is worth noting that the models all correctly accounted for covariate effects in the simulation study, since these were simulated under proportional hazards. Similar model performance might be expected in scenarios where covariate effects are not proportional, since the models have the flexibility to account for this if desired. Future work could investigate the performance of models under misspecification of more complicated baseline and covariate relationships and differing lengths of study follow-up.

This paper has compared different approaches to modeling the natural history of rare diseases with multiple sources of IPD, and demonstrated that a lack of accounting for the (probable) heterogeneity between these sources will lead to bias in population-level model predictions that economic decisions rely upon. Two-stage approaches that stratify baseline parameters by study source were shown to be the most robust, with the least bias, even when other models were true to the underlying data generation, while also being a more reliable choice in terms of model convergence. One-stage approaches struggled particularly under low heterogeneity and with small numbers of studies. A software implementation of the methods considered was demonstrated on a real-world dataset of patients with DMD, and highlighted the disparity between model predictions. This demonstrated the importance of accounting for heterogeneity, and difficulties in implementation under current software, when quantifying the uncertainty of model predictions such as LOS. Further analysis is required to investigate how reliable the estimates of heterogeneity are, and how to take these into account to provide suitable estimates of uncertainty around model predictions.

## ACKNOWLEDGEMENTS

The authors would like to thank C-Path Duchenne Regulatory Science Consortium members and study team members from participating CINRG sites, details of whom can be found in Supplementary Material 4. This study was supported by the National Institute for Health and Care Research (NIHR) Applied Research Collaboration East Midlands (ARC EM) and Leicester NIHR Biomedical Research Centre (BRC). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. This research used the SPECTRE High Performance Computing Facility at the University of Leicester.

## FUNDING INFORMATION

This study was funded by an NIHR Doctoral Research Fellowship NIHR300984.

## DATA AVAILABILITY STATEMENT

Further details of the C-Path D-RSC datasets used are in Supplementary Material 2. D-RSC data access applications can be made via <https://c-path.org/programs/d-rsc/overview/database-access/>. Access to the mortality data set is available on request from the corresponding author.

## ORCID

Jonathan Broomfield  <https://orcid.org/0000-0003-1846-2150>

## REFERENCES

1. Kirchhoff K, Hammes B, Kehl K, Briggs L, Brown R. Effect of a disease-specific planning intervention on surrogate understanding of patient goals for future medical treatment. *J Am Geriatr Soc*. 2010;58(7):1233-1240.
2. Dias S, Welton N, Sutton A, Ades A. NICE DSU technical support document 5: evidence synthesis in the baseline natural history model. *Natl Inst Health Care Excell*. 2011. <https://www.ncbi.nlm.nih.gov/books/NBK310368/>
3. Yen A, Chen T, Duffy S, Chen C. Incorporating frailty in a multi-state model: application to disease natural history modelling of adenoma-carcinoma in the large bowel. *Stat Methods Med Res*. 2010;19(5):529-546.
4. Laird A, Hubbard R, Inoue L. Multi-state models for natural history of disease. UW Biostatistics Working Paper Series. 2013 Working Paper 399.
5. Warren J, Klabunde C, Schrag D, Bach P, Riley G. Overview of the SEER-medicare data: content, research applications, and generalizability to the United States elderly population. *Med Care*. 2002;40(8):IV3-IV18.
6. Zahl P, Gotzche P, Maehlen J. Natural history of breast cancers detected in the Swedish mammography screening programme: a cohort study. *Lancet Oncol*. 2011;12(12):1118-1124.
7. Wang Z, Wang H, Ding X, Chen X, Shen K. A large-cohort retrospective study of metastatic patterns and prognostic outcomes between inflammatory and non-inflammatory breast cancer. *Ther Adv Med Oncol*. 2020;12:1758835920932674.
8. Zechmeister-Koss I, Schnell-Inderst P, Zauner G. Appropriate evidence sources for populating decision analytic models within health technology assessment (HTA): a systematic review of HTA manuals and health economic guidelines. *Med Decis Making*. 2014;34(3):288-299.
9. Behera M, Kumar A, Soares H, Sokol L, Djulbegovic B. Evidence-based medicine for rare diseases: implications for data interpretation and clinical trial design. *Cancer Control*. 2007;14(2):160-166.
10. Balan T, Putter H. A tutorial on frailty models. *Stat Methods Med Res*. 2020;29(11):3424-3454.
11. Lee Y, Nelder J. Conditional and marginal models: another view. *Stat Sci*. 2004;19(2):219-238.
12. Putter H, Houwelingen H. Frailties in multi-state models: are they identifiable? Do we need them? *Stat Methods Med Res*. 2015;24(6):675-692.
13. Tudur Smith C, Williamson P, Marson A. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. *Stat Med*. 2005;24(9):1307-1319.
14. Crowther M, Riley R, Staessen J, Wang J, Gueyffier F, Lambert P. Individual patient data meta-analysis of survival data using Poisson regression models. *BMC Med Res Methodol*. 2012;12:34.
15. Bijwaard G. Multistate event history analysis with frailty. *Demogr Res*. 2014;30:1591-1620.
16. Crowther M, Look M, Riley R. Multilevel mixed effects parametric survival models using adaptive gauss-Hermite quadrature with application to recurrent events and individual participant data meta-analysis. *Stat Med*. 2014;33:3844-3858.
17. Gasperoni F, Ieva F, Paganoni A, Jackson C, Sharples L. Evaluating the effect of healthcare providers on the clinical path of heart failure patients through a semi-Markov, multi-state model. *BMC Health Serv Res*. 2020;20(1):533.
18. Bakoyannis G. Nonparametric analysis of nonhomogeneous multistate processes with clustered observations. *Biometrics*. 2020;77(2):1-14.
19. Zhang H, Kelvin E, Carpio A, Hauser W. A multistate joint model for interval-censored event-history data subject to within-unit clustering and informative missingness, with application to neurocysticercosis research. *Stat Med*. 2020;39:3195-3206.
20. Yau K. Multilevel models for survival analysis with random effects. *Biom J*. 2001;57:96-102.
21. Crowther M. Merlin — a unified modeling framework for data analysis and methods development in Stata. *Stat J*. 2020;20(4):763-784.
22. Olariu E, Cadwell K, Hancock E, Trueman D, Chevrou-Severac H. Current recommendations on the estimation of transition probabilities in Markov cohort models for use in health care decision-making: a targeted literature review. *Clin Outcomes Res*. 2007;9:537-546.
23. C-Path. Duchenne Regulatory Science Consortium. <https://c-path.org/programs/d-rsc/> 2017.
24. Enright P. The six-minute walk test. *Resp Care*. 2003;48:783-785.
25. Broomfield J, Hill M, Crowther M, Abrams K. Life expectancy in Duchenne muscular Dystrophy: reproduced individual patient data meta-analysis. *Neurology*. 2021;97(23):e2304-e2314.
26. Morris T, White I, Crowther M. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074-2102.
27. Sussman M. Duchenne muscular dystrophy. *J Am Acad Orthop Sur*. 2002;10(2):138-151.
28. Kolb S, Coffey C, Yankey J, et al. Natural history of infantile-onset spinal muscular atrophy. *Ann Neurol*. 2010;82(6):883-891.
29. Crowther M. Simulating time-to-event data from parametric distributions, custom distributions, competing-risk models and general multi-state models. *Stat J*. 2021;22(1):3-24.
30. Crowther M, Lambert P. Parametric multistate survival models: flexible modelling allowing transition-specific distributions with application to estimating clinically useful measures of effect differences. *Stat Med*. 2017;36(29):4719-4742.
31. Jackson D, Riley R, White I. Multivariate meta-analysis: potential and promise. *Stat Med*. 2011;30(20):2481-2498.

**SUPPORTING INFORMATION**

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

**How to cite this article:** Broomfield J, Abrams KR, Freeman S, et al. Modeling the multi-state natural history of rare diseases with heterogeneous individual patient data: A simulation study. *Statistics in Medicine*. 2023;1-17. doi: 10.1002/sim.9949