Extrapolation of efficacy and other data to support the development of new medicines for children: a systematic review of methods

Journal:	*Statistical Methods in Medical Research*
Manuscript ID	SMM-15-0132.R2
Manuscript Type:	Review Article
Keywords:	Adult, Bayesian, Bridging study, Commensurability, Extrapolation, Paediatric, Prior distribution, Random-effects meta-analysis
Abstract:	Objective: When developing new medicines for children, the potential to extrapolate from adult data to reduce the experimental burden in children is well recognised. However, significant assumptions about the similarity of adults and children are needed for extrapolations to be biologically plausible. We reviewed the literature to identify statistical methods that could be used to optimise extrapolations in paediatric drug development programmes. Methods: Web of Science was used to identify papers proposing methods relevant for using data from a `source population' to support inferences for a `target population'. Four key areas of methods development were targeted: paediatric clinical trials; trials extrapolating efficacy across ethnic groups or geographic regions; the use of historical data in contemporary clinical trials; and using short-term endpoints to support inferences about long-term outcomes. Results: Searches identified 626 papers of which 52 met our inclusion criteria. From these we identified 102 methods comprising 58 Bayesian and 44 frequentist approaches. Most Bayesian methods (n=54) sought to use existing data in the source population to create an informative prior distribution for a future clinical trial. Of these, 46 allowed the source data to be down-weighted to account for potential differences between populations. Bayesian and frequentist versions of methods were found for assessing whether key parameters of source and target populations are commensurate (n=34). Fourteen frequentist methods synthesised data from different populations using a joint model or a weighted test statistic. Conclusions: Several methods were identified as potentially applicable to paediatric drug development. Methods which can accommodate a heterogeneous target population and which allow data from a source population to be down-weighted are preferred. Methods assessing the

	commensurability of parameters may be used to determine whether it is appropriate to pool data across age groups to estimate treatment effects.

Extrapolation of efficacy and other data to support the development of new medicines for children: a systematic review of methods

Ian Wadsworth[†*], Lisa V. Hampson[†], Thomas Jaki[†]

[†] Department of Mathematics & Statistics, Fylde College, Lancaster University, Lancaster, LA1 4YF, UK

[*] Corresponding author: E-mail: i.wadsworth@lancaster.ac.uk.

Abstract

Objective: When developing new medicines for children, the potential to extrapolate from adult data to reduce the experimental burden in children is well recognised. However, significant assumptions about the similarity of adults and children are needed for extrapolations to be biologically plausible. We reviewed the literature to identify statistical methods that could be used to optimise extrapolations in paediatric drug development programmes.

Methods: Web of Science was used to identify papers proposing methods relevant for using data from a 'source population' to support inferences for a 'target population'. Four key areas of methods development were targeted: paediatric clinical trials; trials extrapolating efficacy across ethnic groups or geographic regions; the use of historical data in contemporary clinical trials; and using short-term endpoints to support inferences about long-term outcomes.

Results: Searches identified 626 papers of which 52 met our inclusion criteria. From these we identified 102 methods comprising 58 Bayesian and 44 frequentist approaches. Most Bayesian methods (n=54) sought to use existing data in the source population to create an informative prior distribution for a future clinical trial. Of these, 46 allowed the source data to be down-weighted to account for potential differences between populations. Bayesian and frequentist versions of methods were found for assessing whether key parameters of source and target populations are commensurate (n=34). Fourteen frequentist methods synthesised data from different populations using a joint model or a weighted test statistic.

Conclusions: Several methods were identified as potentially applicable to paediatric drug development. Methods which can accommodate a heterogeneous target population and which allow data from a source population to be down-weighted are preferred. Methods assessing the commensurability of parameters may be used to determine whether it is appropriate to pool data across age groups to estimate treatment effects.

Keywords: Adult; Bayesian; Bridging study; Commensurability; Extrapolation; Paediatric; Prior distribution; Random-effects meta-analysis

1 Introduction

Extrapolation has been defined as extending data and conclusions available from studies conducted in a 'source population' to make or support inferences for a 'target population' [1]. Extrapolating from existing data, also commonly referred to as bridging or borrowing strength, is common in drug development. Examples include incorporating historical data into the analysis of contemporary clinical trials [2–4] and, more controversially, using information on a drug's short-term effect to draw conclusions about its long-term effect [5]. Alternatively, one may seek to test the efficacy of a medicine in a new geographic region when data are available confirming it is beneficial for patients from another locality. In such cases, it may suffice to conduct a smaller 'bridging' study in the new region that will collect efficacy and safety data to support the extrapolation of data from other localities to this site [6].

For extrapolations to be appropriate, source and target populations should be similar in terms of the key parameter(s) of interest. Extrapolations are 'complete', in the sense that existing data obviate the need to collect data from the target population, when there is strong prior opinion that differences between populations are small. Such opinion may be informed by pre-clinical work or experiences of developing related drugs or treating related patient groups. When there is greater uncertainty about the biological plausibility of similarities, 'partial' extrapolations may be more acceptable. A partial strategy would stipulate that existing data in the source population be complemented by supportive data in the target population generated by a reduced drug development programme. This reduced programme would be targeted to fill in gaps in existing knowledge or to verify similarities about which there is most uncertainty. To illustrate how an extrapolation strategy might be selected, suppose that data from the standard of care arm of several historical trials are available to inform the design and analysis of a new study. If investigators are confident that the standard of care has changed little over time and response rates have been stable, the historical data may be used as the control arm of the new (single-arm) trial. Otherwise, the historical data may be used to augment data from the new study, which would be designed as a randomised controlled trial (RCT) but would allocate fewer patients to control. Making full use of existing data can have important implications for the efficiency and feasibility of drug development in difficult to study populations such as rare diseases or groups where there are ethical and practical barriers to trial recruitment.

The use of extrapolation to facilitate the development of safe and effective medicines for children has received much attention [7–10]. Adult data are often available at the time development of a new medicine begins in children. Moreover, trials in children can be more challenging to conduct due to practical constraints on available sample sizes and pharmacokinetic sampling [11]. There is also a common perception that recruitment into paediatric trials will be challenging, although this has been contradicted by recent research finding that parents and practitioners are willing to enter children into trials

[12]. Dunne et al. [7] discuss the paediatric study decision tree [8, 10] shown in Figure 1, which is an algorithmic approach to determining which additional data are needed in children to support paediatric licensing decisions. The level of extrapolation is determined by whether adults and children can be assumed to be similar in terms of key characteristics, such as disease progression and the pharmacokinetic-pharmacodynamic (PK-PD) relationship of the drug. While this framework clearly identifies scenarios in which different extrapolation strategies are appropriate, it neither accommodates uncertainty about extrapolation assumptions nor allows for differences between age groups of children. To capture the heterogeneity of growth, development and pharmacokinetics in the population, the ICH E11 guideline [10] suggests one possible age grouping: preterm newborn infants, term newborn infants (0 to 27 days), infants and toddlers (28 days to 23 months), children (2 to 11 years) and adolescents (12 to 16/18 years, dependent on region). Batchelor and Marriott [13] state that there may be age related changes in drug pharmacokinetics caused by anatomical and physiological differences between younger and older children and adults. However, Stephenson [14] notes that adults' and children's responses to many drugs have much in common. The European Medicines Agency (EMA) [1] has proposed a general framework for extrapolation allowing for the incorporation of uncertainty about assumptions. This framework stipulates that an extrapolation concept, containing explicit hypotheses on expected differences between populations, should inform the development of an extrapolation plan. This plan will detail which additional data will be generated in the target population, and these data should, in turn, be used to verify the extrapolation concept.

This paper describes the findings of a systematic review conducted to identify statistical methods that can be used to optimise extrapolations in paediatric drug development. We sought methods relevant for using data from a source population to support inferences for a target population. To provide focus for the literature search, we restricted our attention to publications developing methods in the context of four applications in which extrapolations are common, namely, paediatric clinical trials; trials extrapolating efficacy across ethnic groups or geographic regions; the use of historical data in contemporary clinical trials; the use of short-term endpoints to support inferences about long-term outcomes. The rest of the paper proceeds as follows. Section 2 outlines the strategy used to identify relevant papers and methods which are briefly summarised in Section 3. In Section 4, we give a detailed account of the methods found, grouped according to four common approaches. We conclude in Section 5 with a discussion of the suitability of these methods for making extrapolations in paediatric drug development.

2 Methods

Articles were identified by searching the Science Citation Index Expanded (SCI- EXPANDED) database of the Web of Science. Searches were restricted to English language papers listed on Web of Science prior to 31st January 2014 in the following cat-

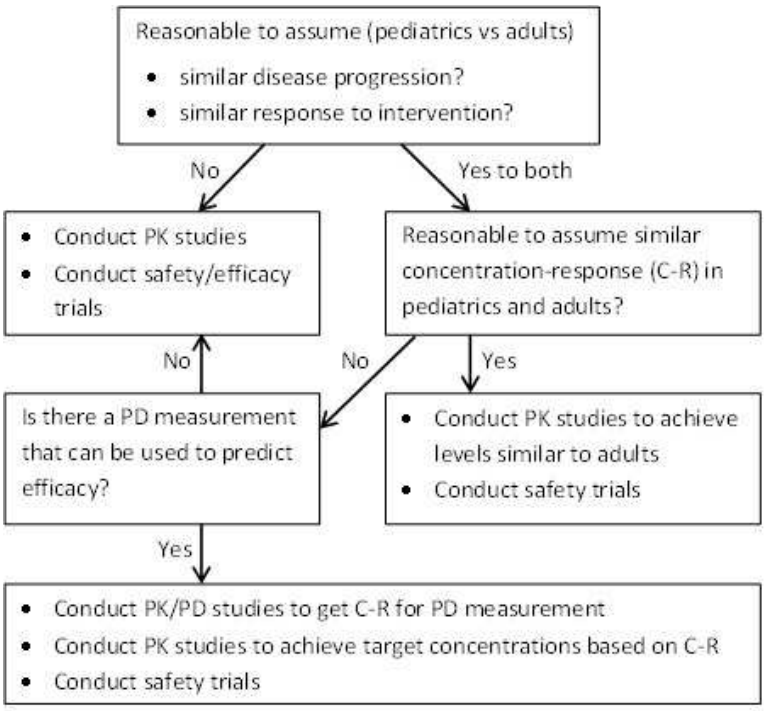


Figure 1: Paediatric Study Decision Tree: Image reproduced from [8]

egories: biology; mathematical and computational biology; mathematics (applied, interdisciplinary applications); medical informatics; research and experimental medicine; pediatrics; and statistics and probability. Preliminary searches were also made of other databases (JSTOR, PubMed) but no additional relevant articles were found. Separate searches of the SCI-EXPANDED database were made to identify potentially relevant papers proposing statistical methods for: a) incorporating historical data into contemporary clinical trials; b) using data on short-term endpoints to support inferences on long-term outcomes; c) paediatric clinical trials; and d) bridging clinical trials. Since there was considerable overlap between the search terms needed to identify papers on the last two topics, these were combined so that a total of three separate searches were made. Search terms can be found in the web based materials accompanying this manuscript (Supplementary Appendix A). We searched for papers containing these search terms either in the title, abstract or keywords.

Articles identified using this search strategy were then screened, first by title and then by abstract. At each stage the following types of manuscripts were excluded: a) conference proceedings; b) reports of clinical trials; c) reports of meta-analyses or evidence synthesis analyses; and d) papers unrelated to medical statistics (returned because one search term, 'bridge', occurs in many contexts). A full text review of the remaining articles was then performed. At this stage manuscripts were excluded if they did not consider statistical methods; if they used source population data only to inform

the design of a future trial; or if they considered trials using a historical control arm without consideration of possible differences between populations. From each paper we extracted details of all statistical methods relevant for extrapolating data from a source population to support inferences for a target population. Methods for establishing whether data from source and target populations are consistent were regarded as relevant, assuming that if commensurability is established it would be appropriate to analyse data pooled across populations. A data extraction form (Supplementary Appendix B) was completed for each statistical method and the number of methods extracted from each paper was recorded. When identical methods were found in more than one paper, we recorded the method as it appeared in the earliest publication. Papers presenting only duplicate methods were excluded from the review. Data were extracted by one author (IW) seeking guidance from others (LVH, TJ) where necessary.

3 Results

Searches identified 52 papers satisfying the stated inclusion/exclusion criteria as summarised in Figure 2, from which we extracted 102 methods. A single method was extracted from each of 34 papers. Of the remaining papers, eight presented two methods each, while 10 presented three or more methods each.

Methods can be categorised into four main areas: i) paediatric drug development (5 of 102 methods); ii) use of historical data in contemporary clinical trials (48 of 102); iii) bridging trials extrapolating efficacy data between ethnic groups or geographic regions (43 of 102); and iv) the use of short-term data to support inferences on long-term outcomes (6 of 102). This is displayed in Figure 3. All five methods in category i) considered extrapolating information from an adult source population to support inferences about children. Of the 48 methods in category ii), 25 sought to extrapolate from a historical control group to support conclusions about control response rates in a contemporary patient group. Of the 43 methods in category iii), 14 took as the target population an unstudied patient group in a new geographic region and sought to borrow strength from existing data on patients in another geographic region for whom the treatment had already been shown to be efficacious. One further method in this category evaluated the consistency of data in two ethnic groups of patients. The remaining 28 methods in category iii) were proposed to assess the consistency of treatment effects across regions of a multi-regional clinical trial (MRCT).

Of the 102 methods, 100 expected data from the source and target populations to make inferences about key parameters in the latter group, and as such are appropriate for making partial extrapolations. An example of a method that did not expect data from the target population, Nedelman et al. [15] suggest that a necessary condition for using adult efficacy data to support conclusions about the efficacy of oxcarbazepine

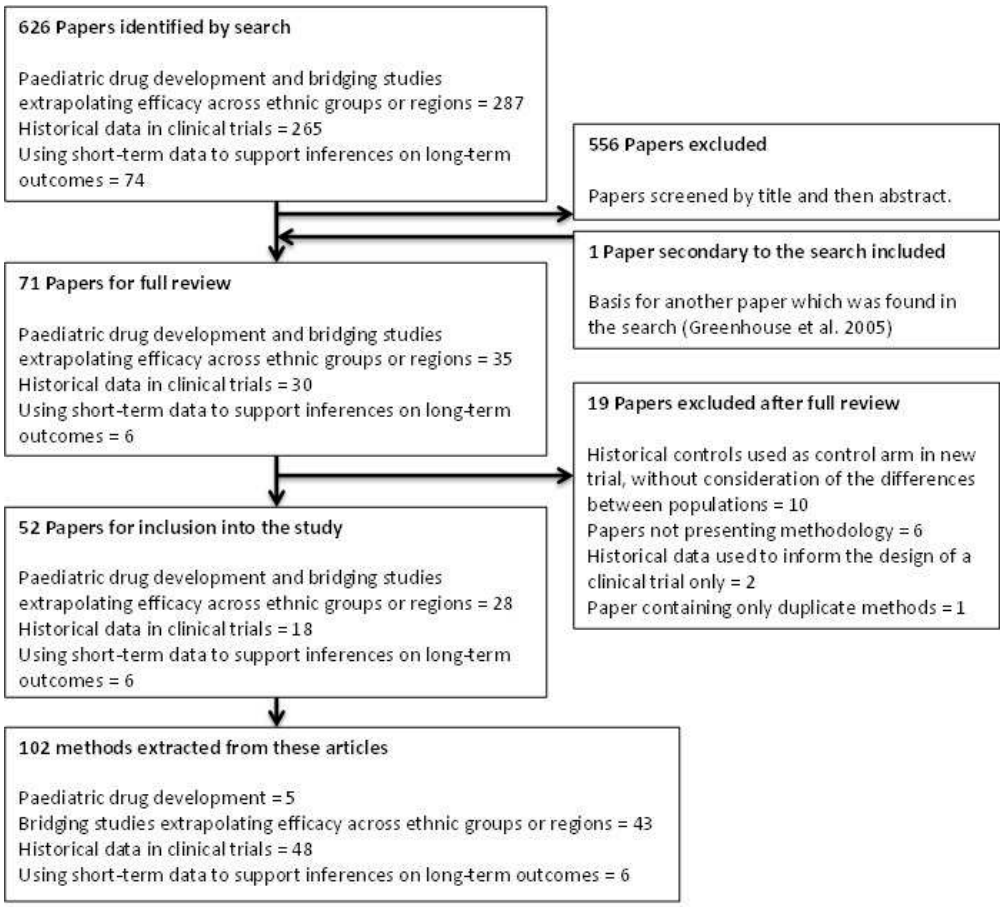


Figure 2: Flow diagram of systematic review results

as a monotherapy for children with epilepsy, is that PK-PD relationships should be similar in adults and children receiving oxcarbazepine as an add-on therapy.

None of the methods found considered extrapolating safety data across populations. Instead all methods expected either efficacy or PD data (100 of 102) or PK data (2 of 102). In the context of paediatric drug development, this may be due to the fact that the paediatric study decision tree stipulates that safety data must be collected in children regardless of one's confidence in extrapolation assumptions. Most methods (100 of 102) sought to make comparisons between treatments while two methods were proposed in the context of dose-finding trials.

4 Thematic analysis of methods for extrapolation

Methods were first classified according to the type of statistics used, that is, Bayesian or frequentist statistics. Categories were then refined to form three broad groups of ap-

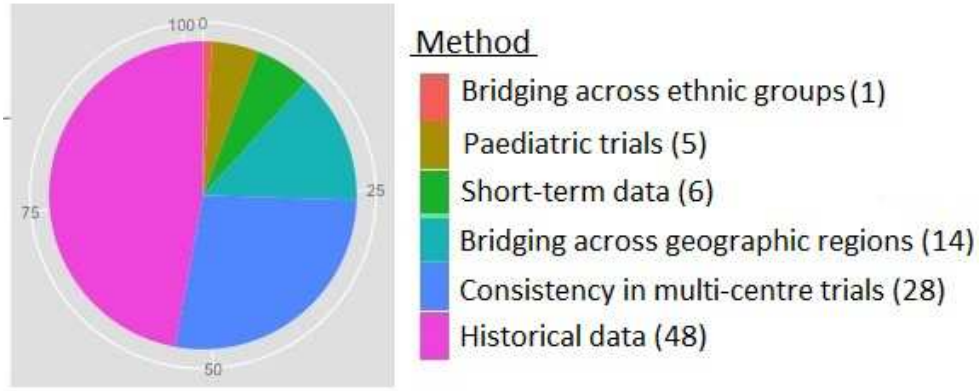


Figure 3: Plot showing distribution of methods across four main areas

proaches, namely, Bayesian methods using existing data to create an informative prior distribution for a parameter of a target population; Bayesian and frequentist methods assessing the commensurability of parameters of source and target populations; frequentist methods synthesizing data across populations using a joint model or weighted test statistic. Further details of the extrapolation methods are given below.

In all descriptions of methods, we will index parameters and data from the source (target) population by a subscript S (T). Therefore, x_S (x_T) will denote data from a source (target) population which depends on an unknown parameter θ_S (θ_T). When θ_S and θ_T are assumed equal, we will refer to their common value as θ. When several datasets are available from a source population, we will let H denote the total number of datasets available and n_{hS} denote the size of dataset h, $h = 1, \ldots, H$. Throughout, $\pi(\cdot)$ will be used to denote a general prior or posterior probability density function (pdf).

4.1 Bayesian methods

Searches identified 58 Bayesian methods from 25 papers [2–4, 17–22, 24–39]. Of these, 54 methods [2–4, 17–22, 24–35] sought to create an informative prior for θ_T while four [36–39] assessed the consistency of treatment effects or PK responses between the source and target populations.

4.1.1 Using existing data in a source population to create a prior for θ_T

All methods in this category sought to augment data from a future trial in the target population (x_T) with existing data from one or more studies in the source population (x_S). For example, θ_T and θ_S could be response rates on the standard of care available to patients in a new and historical trial, respectively. In this setting, differences between θ_T and θ_S may arise due to differences between trial protocols, advances in medical care

or demographic shifts in the patient population over time. More generally, the source data will be useful for learning about θ_T only if the clinical effects of treatments in the source and target populations patients are similar. Of the 54 methods which used x_S to create an informative prior for θ_T, most proposed discounting these data to account for potential differences. Thirty-one methods [2–4, 17–22, 24, 25] considered differences between θ_T and θ_S, and formulated priors for θ_T which when updated with emerging data from the new trial adaptively weight x_S according to the commensurability of x_S and x_T. Fifteen methods adopted a fixed non-adaptive approach to down-weight x_S. Eight methods did not down-weight x_S at all, so that the final posterior distribution for θ_T would attribute equal weight to the source and target population data.

Adaptive down-weighting of data from the source population

Most approaches in this category were proposed for incorporating data from a historical trial into a contemporary study. One approach which has received much attention is the power prior and 10 variations on this were found [2, 3, 17, 18]. Power priors are formed by raising the likelihood of the historical data to a power $a_0 \in [0, 1]$. More formally, assuming parameters are consistent across populations, let $L(\theta \mid x_S)$ denote the likelihood of the source data and let $\pi_0(\theta)$ represent the prior for θ held before these data became available. Then the hierarchical power prior for θ after observing x_S [2] is

$$\pi^{PP}(\theta, a_0|x_S) \propto L(\theta|x_S)^{a_0}\pi_0(\theta)\pi(a_0). \quad (1)$$

The prior for a_0 captures prior uncertainty about the commensurability of parameters of the historical and contemporary data. Ibrahim and Chen [2] suggest placing a beta, truncated gamma or normal prior on a_0. Once data from the new trial become available, they are used to update (1) using Bayes theorem to derive a posterior distribution for θ and a_0 given x_S and x_T. Both datasets are used to learn about a_0 and thus determine the contribution of the historical data to the marginal posterior distribution for θ. If x_S and x_T are commensurate, in the sense that they are consistent with the hypothesis that $\theta_T = \theta_S$, greater posterior weight will be placed on powers close to 1, in which case observations from both datasets are regarded as equally informative for θ_T and pooled. Conflicting datasets will result in information from x_S being discarded as greater posterior weight is placed on powers close to 0. Ibrahim and Chen [2] extend π^{PP} in (1) to incorporate data from multiple historical studies. Versions accommodating data following generalized linear fixed and mixed effect models, proportional hazards models and cure rate models are also derived.

It has been noted that the hierarchical power prior in (1) violates the likelihood principle since it omits the normalising constant for a_0 [16, 17]. Modifying (1) to incorporate the normalising constant $C(a_0) = \left\{\int L(\theta|x_S)^{a_0}\pi_0(\theta)d\theta\right\}^{-1}$, we obtain

$$\pi^{MPP}(\theta, a_0|x_S) = C(a_0)L(\theta|x_S)^{a_0}\pi_0(\theta)\pi(a_0), \quad (2)$$

which Hobbs et al. [3] refer to as the modified power prior (MPP). Chen et al. [18] extend the MPP to accommodate several historical datasets, as well as binary and normally distributed data. Hobbs et al. [3] modify the MPP in (2) by extending the Bayesian model for x_S and x_T to incorporate a parameter τ measuring the correlation between parameters of the historical and contemporary data, and stipulating that $\theta_T \mid \theta_S, \tau \sim N(\theta_S, 1/\tau)$ and $a_0 \mid \tau \sim \text{Beta}(g(\tau), 1)$. Here g is a positive function specified by the analyst which is small for τ close to 0 and large when τ is large. Thus given the historical and contemporary data are commensurate (inconsistent), the prior distribution for a_0 is concentrated about powers close to 1 (0). From this model one can derive the location commensurate power prior (LCPP) as

$$\pi^{LCPP}(\theta_T, a_0, \tau | x_S) \propto \pi(a_0|\tau)\,\pi(\tau) \int \frac{[L(\theta_S|x_S)]^{a_0}}{\int [L(\theta_S|x_S)]^{a_0} d\theta_S} \times \sqrt{\tau}\phi((\theta_T - \theta_S)\sqrt{\tau})\, d\theta_S,$$

where ϕ is the pdf of a standard normal variable and $\pi(\tau)$ is a vague prior on τ. Once the new study has been completed, conflicting historical and contemporary data consistent with small τ will lead to an adaptive down-weighting of x_S in the marginal posterior for θ_T.

A similar Bayesian model for x_S and x_T is assumed to derive the commensurate prior (CP) for θ_T [3]. Again modelling conditional prior opinion on θ_T as $\theta_T \mid \theta_S, \tau \sim N(\theta_S, 1/\tau)$, the CP for θ_T given x_S and θ_S is

$$\pi^{CP}(\theta_T, \tau | x_S, \theta_S) \propto L(\theta_S|x_S) \times \sqrt{\tau}\phi((\theta_T - \theta_S)\sqrt{\tau})\,\pi_0(\theta_T)\,\pi(\tau). \quad (3)$$

Once data from the new trial become available, the posterior density for (θ_T, τ) given x_T and x_S is proportional to (3) multiplied by $L(\theta_T \mid x_T)$. If the historical and contemporary data are consistent with $\tau \approx 0$, the historical data are discarded and the marginal posterior distribution for θ_T tends towards the distribution that would result from updating the initial prior for θ_T with x_T. On the other hand, if data are consistent with $\tau \approx \infty$, the marginal posterior for θ_T converges to the posterior that would result from pooling x_T and x_S to update $\pi_0(\theta_T)$ assuming $\theta_T = \theta_S$. Hobbs et al. [19, 20] suggest defining $\pi(\tau)$ in (3) as a conditionally conjugate prior distribution or using a 'spike and slab' prior. Alternatively, an empirical Bayesian approach can be adopted, replacing τ by its marginal maximum likelihood estimate (MLE) [19]. Hobbs et al. [20] use the commensurate prior [3] to incorporate historical control data into a new adaptive RCT. The randomisation ratio between the novel treatment and control is updated group sequentially on the basis of the current effective sample size of the historical data: more patients are randomised to the novel treatment when there is weak evidence of heterogeneity between the historical and contemporary control data. The CP-approach has been extended to accommodate a variety of data types, including responses following general linear mixed effect models, and generalised linear models with fixed or mixed effects. Hobbs et al. illustrate this approach with applications to

binary, survival and count data.

Hobbs et al. [3] adapt the CP in (3) for the case of normally distributed data to propose a location commensurate prior (LCP), assuming historical patient responses have mean μ_S and variance σ_S^2, and data from the new trial have mean μ_T and variance σ_T^2. If no information is available for μ_S before the historical trial, so that $\pi_0(\mu_S) \propto 1$, the posterior distribution for μ_S after observing a historical dataset of size n_S with sample mean $\bar{x}_S$ would be $\mathrm{N}(\bar{x}_S, \hat{\sigma}_S^2 n_S^{-1})$, replacing σ_S^2 by its MLE. Before the new trial data become available, we model $\mu_T \mid \mu_S, \tau \sim N(\mu_S, 1/\tau)$. Placing a non-informative prior on σ_T^2 and a vague prior on τ, we obtain the LCP:

$$\pi^{LCP}(\mu_T, \sigma_T^2, \tau | x_S) \propto \sqrt{\tau}\phi\left(\frac{\mu_T - \bar{x}_S}{\sqrt{(\tau^{-1} + \hat{\sigma}_S^2 n_S^{-1})}}\right)\sigma_T^{-2}\pi(\tau).$$

Updating the LCP with x_T, the weight attributed to the historical data by the posterior distribution for (μ_T, σ_T^2) will depend on the consistency of x_S and x_T with the claim that $\mu_S = \mu_T$. Hobbs et al. [3] extend the LCP to derive the location scale commensurate prior (LSCP): the weighting of the historical data depends upon the consistency of x_S and x_T with the claim that $\mu_T = \mu_S$ and $\sigma_T^2 = \sigma_S^2$.

Meta-analytic predictive (MAP) priors are an approach to combining data across several heterogeneous source populations to formulate an informative prior for θ_T. The use of historical control data potentially allows for the randomisation of fewer contemporary patients to control in a future RCT. Two methods developed this approach [4, 21], synthesising data from the control arms of several historical trials in a Bayesian random-effects meta-analytical model to derive the posterior predictive distribution for the parameter of interest in the control group of a new study. Models are formulated assuming parameters of the historical and contemporary datasets are exchangeable. Suppose there are H historical trials generating estimates $x_{S1}, \ldots, x_{SH}$ of $\theta_{S1}, \ldots, \theta_{SH}$. If patient responses are normally distributed, θ_{Sh} is the expected response on control in historical trial h, or it may be the log-odds of response on control if outcomes are binary. Neuenschwander et al. [4] assume parameter estimates are normally distributed with known standard errors $s_{S1}, \ldots, s_{SH}$. A Bayesian random-effects meta-analytic model is:

$$\begin{aligned} X_{Sh} \mid \theta_{Sh} &\sim N(\theta_{Sh}, s_{Sh}^2), \quad \text{for } h = 1, \ldots, H, \\ \theta_{S1}, \ldots, \theta_{SH}, \theta_T | \theta^\star, \nu^2 &\sim N(\theta^\star, \nu^2), \\ \theta^\star &\sim \pi(\theta^\star), \\ \nu^2 &\sim \pi(\nu^2). \end{aligned} \tag{4}$$

In the special case that ν is known, the posterior distribution of $\theta^\star$ given the historical data is

$$\theta^\star | x_{S1}, \ldots, x_{SH}, \nu \sim N\left(\frac{\sum w_h x_{Sh}}{\sum w_h}, \frac{1}{\sum w_h}\right),$$

where $w_h = (s_{Sh}^2 + \nu)^{-1}$. Before the new trial begins, the prior distribution of θ_T is its posterior predictive distribution given the historical data. If ν is known, this distribution is

$$\theta_T | x_{S1}, \ldots, x_{SH}, \nu \sim N\left(\frac{\sum w_h x_{Sh}}{\sum w_h}, \frac{1}{\sum w_h} + \nu^2\right).$$

Neuenschwander et al. [4] recommend using priors for ν to check the sensitivity of conclusions in a fully Bayesian meta-analysis. Gsteiger et al. [21] extend this method to derive the MAP prior for the log mean count on control in a new trial when count data are overdispersed and follow a negative binomial model. Chen et al. [18] propose a similar method for normally distributed and binary data which synthesises historical and contemporary data within a Bayesian random-effects meta-analytic model. Hobbs et al. [19] state that when $H = 1$, there is a one-to-one relationship between the commensurability parameter τ in (3) and the between-study variance ν in model (4).

Cuffe [22] considers a new RCT extrapolating from a single historical study to support inferences for the expected response on control. Responses from n_S (historical) and n_T (contemporary) control patients are summarised by the sample means x_S and x_T, respectively. These statistics are assumed to follow a Bayesian random-effects model

$$\begin{gathered} X_S \mid \theta_S \sim N(\theta_S, \sigma^2/n_S) \quad \text{and} \quad X_T \mid \theta_T \sim N(\theta_T, \sigma^2/n_T), \\ \theta_S, \theta_T \mid \theta^\star \sim N(\theta^\star, \sigma^2/n_b), \\ \theta^\star \sim N(0, \sigma_1^2), \end{gathered} \tag{5}$$

where σ^2 is assumed known and σ_1^2 is chosen to be large. It follows that the posterior marginal expectation of θ_T is

$$\lim_{\sigma_1 \to \infty} \mathbb{E}(\theta_T | x_T, x_S) = \frac{n_b n_S}{2n_S n_T + n_b n_T + n_b n_S} x_S + \frac{2n_S n_T + n_b n_T}{2n_S n_T + n_b n_T + n_b n_S} x_T. \tag{6}$$

Model (5) indexes the between-trial variance, and thus the degree of information borrowed from x_S to estimate θ_T, by the parameter n_b. Since this will often be unknown, Cuffe adopts an empirical Bayesian approach, evaluating the posterior expectation of θ_T at

$$\hat{n}_b = (n_m/d_m) \max\{d_m - |x_T - x_S|, 0_+\}, \tag{7}$$

so that the historical data contribute to our estimation of θ_T only if the discrepancy between these and the new data is less than a pre-specified maximum tolerable difference (d_m). The maximum influence of the historical data, attained when $x_S = x_T$, is pre-specified as n_m. The condition 0_+ in (7) ensures n_b is strictly positive. On conclusion of the contemporary RCT, data on the experimental treatment are summarised by the statistic x_a. A classical frequentist analysis is then conducted to test for a treatment effect, comparing x_a with an estimate of $\mathbb{E}(\theta_T \mid x_T, x_S)$ derived substituting $\hat{n}_b$ into (6). Cuffe finds that incorporating historical control data into the analysis of a

contemporary RCT may actually reduce the power to detect a clinically relevant effect if the critical value of the frequentist test must be adjusted to ensure adequate type I error rate control under all possible values of $(\theta_T - x_S)$. Viele et al. [23] describe the results of a simulation study comparing methods for incorporating control data from a single historical trial into the analysis of a contemporary RCT. The authors find that, in general, incorporating historical control data does have benefits for increasing power and reducing the type I error rate when $|\theta_T - x_S|$ is close to 0, although how far this 'sweet spot' extends before losses in power or increases in type I error rate are incurred depends on the method used for extrapolation.

Mixture priors are another approach for using existing data to create an informative prior distribution for θ_T. Two methods [24, 25] use mixture priors to augment data from a future clinical trial in a new geographic region with data, x_S, from an area that has previously been studied. These methods set the prior for the treatment effect in the new region as

$$\pi(\theta_T \mid x_S) = \omega\, \pi_1(\theta_T) + (1 - \omega)\, \pi_2(\theta_T),$$

where $\pi_1(\cdot)$ is an informative prior derived from x_S, and $\pi_2(\cdot)$ is a non-informative distribution used to dilute the information for θ_T obtained from x_S so that $\pi(\theta_T \mid x_S)$ has heavy tails. Hsiao et al. [24] recommend that the mixing proportion ω be fixed by the regulatory authority of the new region. This weight may be specified in view of differences between the new and previously studied regions in terms of intrinsic and extrinsic ethnic factors. The corresponding posterior distribution for θ_T will also be a mixture distribution, with components that are the posterior distributions if $\pi_1(\theta_T)$ or $\pi_2(\theta_T)$ were the priors, and weights that are a function of the data, such that more weight is given to the posterior that would result from updating the prior component most commensurate with x_T. Hobbs et al. [3] also consider mixture priors, proposing a prior for the mean and variance of patient responses in a new trial which is a mixture of m LSCPs with fixed pairs of commensurability parameters $(\tau_1, \gamma_1), \ldots, (\tau_m, \gamma_m)$ and fixed weighting proportions $\omega_1, \ldots, \omega_m$. This method allows for the consideration of different plausible relationships between the location and scale parameters of the historical and contemporary data.

Non-adaptive down-weighting of data from the source population

Fifteen methods [2, 3, 25–33] used existing data from a source population to formulate an informative prior for θ_T, down-weighting these data in a non-adaptive, pre-specified manner. The power prior can be considered in this category if a_0 in (1) is taken to be a fixed constant and Hobbs et al. [3] refer to this approach as the conditional power prior (CPP). Six methods [2, 3, 25–27] propose power priors with fixed a_0. Ibrahim and Chen [2] propose a variation on this approach for the case that historical data are from a single trial and patient responses follow an arbitrary regression model. Neither

paper discusses how to choose a_0 [2, 3]. De Santis [26] defines a geometric prior, raising the likelihood of data from a single historical trial to a power $a_0 = r/n_S$, where r is a constant specified by the analyst. The author also modifies this approach to weight different historical datasets by different fractions when they differ in their relevance to the new trial. De Santis [26] illustrates how the geometric prior can be used to inform early stopping decisions in a new Bayesian clinical trial. Rietbergen et al. [27] consider the CPP incorporating data from several historical studies, assigning data from each study a weight elicited from expert opinion. Gandhi et al. [25] consider the CPP for the purposes of incorporating existing binary data from a geographic region in which a drug has been shown to be effective into the analysis of a bridging trial conducted in a new region. The authors recommend performing sensitivity analyses to explore the impact on inferences of different choices of weights. Hobbs et al. [3] also provide a variation on the commensurate prior described in the previous subsection which treats τ as fixed.

Schoenfeld et al. [28] augment data from a clinical trial in children with data from a completed adult trial, assuming parameters of adult and paediatric data are samples from a normal population distribution with mean $\theta^\star$ and known variance ν^2. The choice of ν^2 reflects opinion on between population differences. This method is equivalent to the conditional power prior when data are available from one adult study: if data from more than one adult trial are available, these should be summarised by a single estimate derived from a meta-analysis of adult studies. Schoenfeld et al. [28] also consider an approach for determining the sample size needed to ensure the Bayesian paediatric trial incorporating adult data has high Bayesian power. Augmenting paediatric data with adult data means that fewer children may be required.

Chen et al. [29] derive a Bayesian empirical prior distribution for a treatment effect θ_T in a specific local region of a MRCT which borrows strength from data from other trial sites. The prior $\theta_T \sim N(\hat{\mu}, \sigma^2)$ is specified by defining $\hat{\mu}$ as the global treatment effect estimate found by averaging across effect estimates obtained from each trial region. Meanwhile, σ^2 is taken to be a linear function of the variance of the region-specific effect estimates, where smaller values of the coefficient of the interregional variance allow for more borrowing of strength across regions. Chen et al. [29] recommend that this coefficient be specified ahead of time and chosen to reflect the consensus opinion of the local regulatory authority and the trial sponsor.

Six other methods in this category of approach shift the location and/or inflate the standard error of an estimate of θ_S to create an informative prior for θ_T while discounting the source population data [30–33]. For example, French et al. [32] formulate a normal prior distribution for θ_T with mean equal to the MLE of θ_S obtained from x_S, and standard deviation equal to four times the standard error of the MLE; the authors propose using this prior for the Bayesian interim monitoring of a trial which will terminate with a conventional frequentist analysis. Whitehead et al. [33] consider Bayesian sample size calculations, using historical placebo data to create an informa-

tive prior distribution for the expected response on placebo in the new trial. This prior is normally distributed, with the mean taken to be the mean response from the historical placebo group and precision chosen to reflect how many patients the prior should represent.

No down-weighting of data from the source population

Eight methods [25, 31, 32, 34, 35] used data from a source population to create an informative prior distribution for θ_T without any down-weighting. Thus, once available, data from the target and source populations are pooled to derive a posterior distribution for θ_T.

4.1.2 Assessing consistency between source and target populations

Four Bayesian methods were proposed to assess the consistency of parameters in source and target populations [36–39]. Pei and Hughes [36] seek to assess whether candidate doses for adults and children result in similar percentages of patients experiencing low levels of a drug; inferences are made testing whether the proportion of children recording PK levels below a quantile estimated from adult data is non-inferior or equivalent to a design value. Tsou et al. [37] use Bayesian most plausible prediction [40] to assess the consistency of treatment effect estimates generated by a new clinical trial comparing an experimental treatment (E) with control (C) in a new geographic region, and reference studies which have demonstrated the advantage of E versus C in an original geographic region, under the assumption of normally distributed treatment effect estimates. The difference between treatment group sample means for the bridging trial, $\hat{\theta}_T$, is said to be consistent with the results of the H reference studies, denoted by $\hat{\boldsymbol{\theta}}_S = (\hat{\theta}_{S1}, \ldots, \hat{\theta}_{SH})$, if and only if

$$p(\hat{\theta}_T \mid \hat{\boldsymbol{\theta}}_S) \geq \rho_B \min\{p(\hat{\theta}_{Sh} \mid \hat{\boldsymbol{\theta}}_{\mathcal{S}\setminus h}); h = 1, \ldots, H\},$$

where $p(\hat{\theta}_T \mid \hat{\boldsymbol{\theta}}_S)$ is the posterior predictive probability of $\hat{\theta}_T$ given the results of all reference studies, $\hat{\boldsymbol{\theta}}_{\mathcal{S}\setminus h}$ is the vector of reference effect estimates excluding $\hat{\theta}_{Sh}$, and $\rho_B > 0$ is a pre-specified constant which reflects the prior confidence of the regulatory authority in the commensurability of data from the new and original geographic regions. Posterior predictive probabilities are derived assuming a non-informative prior distribution for the common treatment effect θ before any data are observed. The posterior predictive probability $p(\hat{\theta}_T \mid \hat{\boldsymbol{\theta}}_S)$ therefore provides a measure of the plausibility of $\hat{\theta}_T$ given the previous trial results. Chow et al. [38] also use posterior predictive probabilities to assess the consistency of data from a bridging trial and reference studies. Gould et al. [39] propose an approach whereby the results of a bridging study are judged to be consistent with those of the reference studies if they fall within contours or regions of the posterior predictive distribution derived from the reference confirmatory trials. The sample size of the bridging trial may be chosen to find an acceptable

balance between the producer risk, that is, the probability of incorrectly rejecting a conclusion of consistency, and the consumer risk, which is the probability of incorrectly concluding consistency.

4.2 Frequentist methods

Forty-four frequentist methods were identified [15, 36–38, 41–45, 47–56, 66–76] of which 11 methods [41–45, 47–52] synthesised data from source and target populations in a joint model, three methods [53–55] combined data across populations through a weighted test statistic, and 30 methods [15, 36–38, 49–52, 56, 66–76] proposed criteria to assess the consistency of estimates of key parameters in different populations.

4.2.1 Joint model incorporating data from source and target populations

Five methods [41–45] proposed using short-term data to support inferences about a long-term endpoint assuming simple models to relate observations on different outcomes. In this setting, θ_T and θ_S could represent long- and short-term treatment effects, or characterise the distribution of the two endpoints. Several authors extrapolate from short-term data to inform early stopping decisions for sequential trials. Hampson and Jennison [41] seek to increase the efficiency of group sequential tests (GSTs) monitoring a long-term outcome by incorporating data on a correlated short-term endpoint so as to increase the Fisher information available for θ_T at each interim analysis. MLEs of θ_T are found maximising the joint likelihood of x_S and x_T assuming pairs of responses on the same patient follow a bivariate normal distribution. No assumption is made about the form of the relationship between the short- and long-term responses other than that they are correlated. The authors derive optimal designs and show that incorporating data on a highly correlated short-term endpoint can reduce the expected sample size of a trial by around 5% of the fixed sample size when the time to availability of the short-term endpoint is at least half that of the long-term endpoint. A similar problem is considered by Galbraith & Marschner [42], who incorporate into GSTs repeated measurements of a continuous endpoint taken at an arbitrary number of follow-up times. The vector of repeated measurements for each individual is assumed to follow a multivariate normal distribution, with correlations between the measurements being exploited to improve estimation and inference associated with the long-term measurement. Marschner and Becker [43] increase the interim information available for a long-term response probability by incorporating data on a short-term binary endpoint, deriving the MLE of the long-term response rate from the joint likelihood of the combined dataset. The values of the short- and long-term endpoints may be associated, however, a patient's short-term response does not necessarily determine their long-term response.

Stallard [44] uses observations on short- and long-term endpoints to support early stopping and treatment selection decisions in a seamless Phase II/III clinical trial.

Responses on the same patient are assumed to follow a bivariate normal distribution, fitted using the double regression method of Engel and Walstra [46]. Wüst and Kieser [45] also consider bivariate normal outcomes and derive a more precise estimator of the variance of the long-term outcome incorporating short- and long-term data. Using this improved estimator to inform blinded sample size adjustments at an interim analysis reduces the variability of the final trial sample size when compared to using long-term data alone.

Six methods [47–52] synthesize data from source and target populations using a frequentist random-effects model. Thall and Simon [47] combine historical and contemporary control data via a univariate random-effects meta-analysis while Arends et al. [48] model short-term and long-term outcomes from trials using a multivariate random effects model. Chen et al. [49] and Ko [50] use a random effects model to accommodate heterogeneity between regions and test for an overall treatment effect. Liu et al. [51] use a random effects model to test for similarity or non-inferiority between treatment effects in different regions. Ko [52] models survival data from different regions using a proportional hazards model with frailties to allow patients in different regions to have varying underlying hazards of experiencing an event.

4.2.2 Combining data across populations in a weighted test-statistic

Three methods [53–55] propose making final inferences about the efficacy of a new treatment in a new geographic region on the basis of a test-statistic combining information from the source and target populations. Suppose Z_T and Z_S are standardised test statistics comparing mean responses on a new treatment and placebo in a new and original region, respectively. For reasonable sample sizes, Z_T and Z_S follow at least approximately standard normal distributions. Lan et al. [53] propose a weighted Z statistic for testing efficacy across regions, $H_0 : \theta = 0$, defined as,

$$Z_w = \sqrt{\omega}\, Z_S + \sqrt{1-\omega}\, Z_T,$$

with $0 \leq \omega \leq 1$. Chow et al. suggest that $|Z_w| > z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of a standard normal distribution, implies the results of the bridging study are consistent with those of the reference study which demonstrated efficacy of the new treatment relative to placebo in the original geographic region. The weight ω should be pre-specified by the regulatory agency, although Lan et al. [53] suggest that this weight may be based on evidence of efficacy established in the original region.

4.2.3 Assessing the consistency of data from source and target populations

Thirty methods were proposed to assess the consistency of data from different populations. Chen et al. [56] survey nine methods in their systematic review for testing the commensurability of a treatment effect across regions of a MRCT, of which we extracted eight. These methods comprised 'Global methods' assessing consistency based on a

test-statistic combining data across all trial regions; 'Multivariate quantitative' methods assessing consistency by considering all pairwise differences between region-specific effect estimates; and 'Multivariate qualitative methods' assessing whether patients from all trial regions can benefit from a new treatment. All eight methods assumed patient responses to be normally distributed. Let Δ_j be the difference in mean response on treatments E and C in trial region j, for $j = 1, \ldots, s$. Then, $\Delta = \sum_{j=1}^{s} n_j \Delta_j / n$ is the overall treatment effect for the trial, where n_j is the number of patients per treatment in the jth region and n is the total number of patients per treatment.

One Global method is Cochran's Q statistic [57] for testing the null hypothesis $H_0 : \Delta_1 = \Delta_2 = \ldots = \Delta_s = \Delta$, against the alternative that at least one Δ_j is different. Treatment effects are judged to be consistent if we fail to reject H_0, that is, if

$$Q = \sum_{j=1}^{s} \frac{(\hat{\Delta}_j - \hat{\Delta})^2}{2/n_j} < \chi^2_{s-1;1-\alpha},$$

where $\chi^2_{s-1;1-\alpha}$ is the $(1-\alpha)$ quantile of a central chi-square distribution with $(s-1)$ degrees of freedom. The test of H_0 based on the Q statistic is well known to have low power [58] in certain situations; for example, in the current context, when the total information available for estimating Δ is low or there are large imbalances between the contributions of different centres to this total information. Higgins' I^2 statistic [59], defined as $I^2 = 100(1-(s-1)/Q)$, measures the degree of inconsistency between $\Delta_1, \ldots, \Delta_s$. However, interpretation of I^2 can be problematic since it increases as a non-linear function of the between-centre heterogeneity [60]. This statistic also depends on the within-centre precision [61] and the number of centres, s, such that under H_0, $E(I^2) = -200/(s-3)$ if $s > 3$ [60]. An alternative measure of consistency not found by this review but pointed out by a reviewer is $H^2 = Q/(s-1)$ [62], which is independent of the number of centres.

Global test statistics can also be used to test for a qualitative interaction between the treatment effect and trial regions. The Gail-Simon test [63] of $H_0 : \{\Delta_j \geq 0, \text{ for all } j = 1, \ldots, s\} \cup \{\Delta_j < 0, \text{ for all } j = 1, \ldots, s\}$ rejects the null hypothesis if $\min(Q^+, Q^-)$ exceeds a critical value c, where

$$Q^- = \sum_{j=1}^{s} \frac{\hat{\Delta}_j^2}{2/n_j} I(\hat{\Delta}_j > 0), \quad Q^+ = \sum_{j=1}^{s} \frac{\hat{\Delta}_j^2}{2/n_j} I(\hat{\Delta}_j < 0).$$

Chen et al. [56] also review Multivariate quantitative methods which test $H_0 : \Delta_1 = \ldots = \Delta_s = \Delta$ and declare treatment effects as consistent if there are no significant pairwise differences between effect estimates, that is, if

$$|\hat{\Delta}_i - \hat{\Delta}_j| < z_{\alpha/2} \sqrt{2(n_j + n_i)/(n_i n_j)} \quad \text{for } i, j = 1, \ldots, s, \; i \neq j.$$

A variation on this approach has been proposed for testing $H_0 : |\Delta_1 - \Delta| > m$ or ... or $|\Delta_s - \Delta| > m$ [64], where rejecting H_0 implies that all regional effects lie within an equivalence margin m of Δ.

Multivariate qualitative methods reviewed by Chen et al. [56] include testing $H_0 : \Delta_1 \leq \delta\Delta$ or ... or $\Delta_s \leq \delta\Delta$, to determine whether all regional effects are non-inferior to the global treatment effect, proposed by Liu et al. [51]. One further method is based on confidence interval coverage which declares the treatment effect to be consistent across regions if $\hat{\Delta}_j > \pi\hat{\Delta} - z_{\alpha/2}\sqrt{2/n_j}$ for $j = 1, \ldots, s$. The Pharmaceuticals and Medical Devices Agency (PMDA) suggest declaring consistency if a positive trend is observed, that is, if $\hat{\Delta}_j > 0$ for all $j = 1, \ldots, s$, or if $\hat{\Delta}_j > \delta\hat{\Delta}$ for all $j = 1, \ldots, s$ [65]. The PMDA recommend setting $\delta \geq 0.5$ although Chen et al. comment that this may be too conservative when several trial regions are included. This literature review found 15 further methods [37, 49, 50, 52, 66–73] proposing consistency criteria similar to the PMDA method. For example, let $\hat{\Delta}$, $\hat{\Delta}_{\mathcal{S}\setminus j^\star}$ and $\hat{\Delta}_{j^\star}$ denote the treatment effect estimates derived from pooling data across all trial regions, all regions excluding region $j^\star$, and region $j^\star$ alone, respectively. Ko et al. [69] consider several alternative criteria for determining whether a new treatment should be deemed efficacious in region $j^\star$ when there is strong statistical evidence to reject $H_0 : \Delta = 0$. For example, investigators may pre-specify one of the following criteria for their study: 1) $\hat{\Delta}_{j^\star} \geq \rho\hat{\Delta}_{\mathcal{S}\setminus j^\star}$; 2) $\hat{\Delta}_{j^\star} \geq \rho\hat{\Delta}$; 3) $\rho \leq \hat{\Delta}_{j^\star}/\hat{\Delta}_{\mathcal{S}\setminus j^\star} \leq 1/\rho$; or 4) $\rho \leq \hat{\Delta}_{j^\star}/\hat{\Delta} \leq 1/\rho$. Here $\rho \in (0, 1)$ may be pre-specified by the regulatory agency of region $j^\star$. Alternatively, Chen et al. [74] derive standardised weighted least squares residuals from $\hat{\Delta}_1, \ldots, \hat{\Delta}_s$ and use these to create Q-Q plots for assessing consistency between regional treatment effects. Pei and Hughes [36] propose a frequentist version of their method described in Section 4.1.2 which seeks to assess whether candidate doses for adults and children result in similar percentages of patients experiencing low levels of a drug.

Hsiao et al. [75] propose two-stage designs for bridging trials. The trial begins recruiting patients from the original region. If efficacy in this region is confirmed at the interim analysis, the trial proceeds to recruit patients from the new region in Stage 2. Otherwise the trial terminates early for lack of benefit. On conclusion of the trial, data accumulated from both regions are pooled and analysed to test a one-sided null hypothesis of no treatment effect. If the result of Stage 1 is similar to the pooled result of Stage 2, the result from the new region is declared consistent with that from the original region and we conclude that the new treatment is effective in both localities.

Cai et al. [76] propose evaluating the similarity of data from clinical trials performed in different ethnic populations using a 'distribution adjusted mean'. This method assumes that there is a covariate Y prognostic for the primary endpoint which differs in distribution between the two ethnic groups. If Y is continuous, its domain can be partitioned into intervals and the relative frequency of each interval in the target pop-

ulation is recorded. These frequencies are then used to calculate the weighted average response in the source population, averaging across the mean responses for each interval of Y. This adjusted mean response is then compared with the unadjusted mean for the target population to assess the consistency of response between the populations.

Nedelman et al. [15] develop a method comparing children and adults receiving a new drug as an add-on therapy, with the aim of using these data to support inferences about children receiving the drug as monotherapy. If the PK-Efficacy relationship is similar for adults and children receiving add-on therapy, this is taken to support an assumption of similar relationships for adults and children receiving monotherapy. Separate linear models are fitted to the PK-Efficacy data from adults and children, and model parameters are compared to establish whether there are differences between age-groups.

Chow et al. [38] apply the 'reproducibility probability' method [77] to bridging studies, calculating the reproducibility probability as the power of the bridging study to detect a treatment effect equal to the estimated effect from the reference study which itself produced a significant result. If the reproducibility probability exceeds a critical value (determined by a regulatory agency) then the bridging study may be considered unnecessary, that is, clinical data from the original region can be completely extrapolated to the new region to support claims of efficacy.

5 Discussion

This systematic review summarises statistical methods relevant for extrapolating data from a source population to a target population, and has captured a wide range of methodology. Several of the approaches identified are potentially applicable for making extrapolations to support paediatric drug development. In this context, adult data, pre-clinical data and data on children receiving treatment for related conditions may all be available at the time development of a medicine begins in children. Thus, methods which can harness existing data to derive informative prior distributions for key parameters in children are particularly appealing. However, we speculate that down-weighting existing data would be more acceptable in this setting to account for potential differences between adults and children. Therefore, the applicability of those eight methods which give comparable weight to historical and contemporary data is likely to be limited unless there is a strong prior rationale for believing that adults and children will respond similarly to treatment. Alternatively, the methods identified by this review for assessing the consistency of parameters of source and target populations may be used as objective criteria for determining when it is appropriate to pool data from adults and children, or indeed pool data across different age groups of children.

When there is some prior understanding of the factors that may explain differences between populations, a weight for the existing data may be pre-specified. Otherwise

Bayesian approaches such as the power prior, commensurate prior, mixture prior or MAP prior, which adaptively down-weight existing data, may be preferred. One criticism that has been made of MAP priors is that the posterior predictive distribution for θ_T given historical data must be typically derived using Markov Chain Monte Carlo. Therefore, since the prior is not available analytically, it cannot be easily reproduced by others unless they have access to the historical data combined in the meta-analysis. To overcome this challenge, Schmidli et al. [78] propose representing the MAP prior as a mixture of a small number of conjugate prior distributions which can be easily recorded and shared.

In Section 1 it was noted that there may be differences between age groups of children. Twenty-five methods [2, 15, 19, 56, 71, 76] identified by this review can accommodate a heterogeneous target population because key parameters are taken to be parameters of (semi-)parametric models capable of adjusting for baseline demographics. Several methods proposing a joint model for data from the source and target populations assume only that data from different populations are correlated. However, this is unlikely to be the case for paediatric drug development when source and target data will typically be observations on different patients. In this case multivariate meta-analytic models, as used by Arends et al. [48], are potentially more relevant since they can capture correlations between parameters of different populations. Future research will consider tailoring these models to support extrapolations in paediatric trials.

Several papers were identified by our literature search which, although they did not contain statistical methods, are relevant for discussion. Manolis et al. [9] discuss the role of modelling and simulation in paediatric investigation plans (PIPs), which are documents pre-specifying what studies will be conducted to support development of a medicine for children. The authors review positive PIP opinions (summarising key elements of PIPs supported by the EMA) and find that population PK models are the most frequently referenced modelling approach, while exposure-response and dose-response models are rarely cited: modelling and simulation, when proposed, is typically used to support dose predictions, study optimization and data analysis. Khalil and Läer [79] review physiologically based pharmacokinetic (PBPK) models as applied to paediatric drug development, where parameters of PBPK models for children may be extrapolated from another species or age group.

Other methods not included in the systematic review were found proposing other ways for using data from a source population to support inferences for a target population. Reif et al. [80] fit a population PK model to data from an adult Phase I trial and use this model to design clinical trial simulations needed to devise a sparse PK sampling schedule for children. De Santis [81] consider using a design prior borrowing information from historical data to plan a clinical trial, for instance to inform sample size selections. Additionally, 12 methods included in the review [18, 21, 28, 33, 37, 39, 47, 51, 53, 55, 73] use source data to inform the design (through sample size calculations) and analysis of

a prospective trial in the target population. In addition, four methods [20, 26, 32, 75] use source and target data to inform mid-study adaptations to the study in the target population.

Software was available for few of the 102 methods identified by this review. Computer syntax was included in a main paper or accompanying supplementary material for 9 methods [21, 27, 32, 33, 35, 36, 39]; code was stated as available upon request from the corresponding author of one method [70]; syntax for another method [17] was included in a related commentary article [82]. The strategy used to identify available software is described in Supplementary Appendix C, while the results are listed in Supplementary Appendix D.

This systematic review has aimed to be a comprehensive overview of methods for extrapolation. However, one limitation is that we chose to focus our literature searches on the four application areas listed in Section 2 and by doing so may have missed other relevant methods. Another limitation is that one author extracted the data so independent reviews of all papers were not performed.

Acknowledgements

We thank two reviewers for their thoughtful and supportive comments which helped to improve the manuscript. This work was supported in part by grants from the National Institute for Health Research [NIHR-RMOFS-2013-03-05, IW; NIHR-CDF-2010-03-32, TJ], the Medical Research Council [MR/J014079/1, LVH; MR/M013510/1, IW] and the Medical Research Council North-West Hub for Trials Methodology Research [MR/K025635/1].

References

1. European Medicines Agency. Concept paper on extrapolation of efficacy and safety in medicine development - EMA/129698/2012;. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2013/04/WC500142358.pdf (2013, accessed: 01/10/2014).

2. Ibrahim JG, Chen MH. Power prior distributions for regression models. Stat Sci. 2000;15(1):46–60.

3. Hobbs BP, Carlin BP, Mandrekar SJ, et al. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. Biometrics. 2011;67(3):1047–1056.

4. Neuenschwander B, Capkun-Niggli G, Branson M, et al. Summarizing historical information on controls in clinical trials. Clin Trials J. 2010;7(1):5–18.

5. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med. 1996;125(7):605–613.

6. International Conference on Harmonisation. Ethnic factors in the acceptability of foreign clinical data E5 (R1);. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E5_R1/Step4/E5_R1__Guideline.pdf (1998, accessed: 01/10/2014).

7. Dunne J, Rodriguez WJ, Murphy MD, et al. Extrapolation of adult data and other data in pediatric drug-development programs. Pediatrics. 2011;128(5):1242–1249.

8. Food and Drug Administration. Guidance for industry: exposure-response relationships - study design, data analysis, and regulatory applications; 2003. http://www.gmp-compliance.org/guidemgr/files/EXPOSURE.PDF (2003, accessed: 15/01/2015).

9. Manolis E, Osman TE, Herold R, et al. Role of modeling and simulation in pediatric investigation plans. Paediatr Anaesth. 2011;21(3):214–221.

10. International Conference on Harmonisation. E11: Note for guidance on clinical investigation of medicinal products in the paediatric population (CPMP/ICH/2711/99);. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002926.pdf (2001, accessed: 01/10/2014).

11. Joseph PD, Craig JC, Caldwell PHY. Clinical trials in children. Br J Clin Pharmacol. 2015;79(3):357–369.

12. Shilling V, Williamson PR, Hickey H, Sowden E, Smyth RL, Young B. Processes in recruitment to randomised controlled trials of medicines for children (RECRUIT): a qualitative study. Health Technol Assess. 2011;15(15):1–116.

13. Batchelor HK, Marriott JF. Paediatric pharmacokinetics: key considerations. Br J Clin Pharmacol. 2015;79(3):395–404.

14. Stephenson T. How childrens responses to drugs differ from adults. Br J Clin Pharmacol. 2005;59(6):670–673.

15. Nedelman JR, Rubin DB, Sheiner LB. Diagnostics for confounding in PK/PD models for oxcarbazepine. Stat Med. 2007;26(2):290–308.

16. Duan Y, Ye K, Smith EP. Evaluating water quality using power priors to incorporate historical information. Environmetrics. 2006;17(1):95–106.

17. Neuenschwander B, Branson M, Spiegelhalter DJ. A note on the power prior. Stat Med. 2009;28(28):3562–3566.

18. Chen MH, Ibrahim JG, Lam P, et al. Bayesian design of noninferiority trials for medical devices using historical data. Biometrics. 2011;67(3):1163–1170.

19. Hobbs BP, Sargent DJ, Carlin BP. Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. Bayesian Anal. 2012;7(3):639–674.

20. Hobbs BP, Carlin BP, Sargent DJ. Adaptive adjustment of the randomization ratio using historical control data. Clinl Trials J. 2013;10(3):430–440.

21. Gsteiger S, Neuenschwander B, Mercier F, et al. Using historical control information for the design and analysis of clinical trials with overdispersed count data. Stat Med. 2013;42(21):3609–3622.

22. Cuffe RL. The inclusion of historical control data may reduce the power of a confirmatory study. Stat Med. 2011;30(12):1329–1338.

23. Viele K, Berry S, Neuenschwander B, Amzal B, Chen F, Enas N, et al. Use of historical control data for assessing treatment effects in clinical trials. Pharm Stat. 2014;13(1):41–54.

24. Hsiao CF, Hsu YY, Tsou HH, et al. Use of prior information for Bayesian evaluation of bridging studies. J Biopharm Stat. 2007;17(1):109–121.

25. Gandhi M, Mukherjee B, Biswas D. A Bayesian approach for inference from a bridging study with binary outcomes. J Biopharm Stat. 2012;22(5):935–951.

26. De Santis F. Power priors and their use in clinical trials. Am Stat. 2006;60(2):122–129.

27. Rietbergen C, Klugkist I, Janssen KJM, et al. Incorporation of historical data in the analysis of randomized therapeutic trials. Contemp Clin Trials. 2011;32(6):848–855.

28. Schoenfeld DA, Zheng H, Finkelstein DM. Bayesian design using adult data to augment pediatric trials. Clin Trials J. 2009;6(4):297–304.

29. Chen YH, Wu YC, Wang M. A Bayesian approach to evaluating regional treatment effect in a multiregional trial. J Biopharm Stat. 2009;19(5):900–915.

30. Goodman SN, Sladky JT. A Bayesian approach to randomized controlled trials in children utilizing information from adults: the case of Guillain-Barre. Clin Trials J. 2005;2(4):305–310.

31. Greenhouse JB, Seltman H. Using prior distributions to synthesize historical evidence: comments on the Goodman–Sladky case study of IVIg in Guillain–Barré syndrome. Clin Trials J. 2005;2(4):311–318.

32. French JL, Thomas N, Wang C. Using historical data with Bayesian methods in early clinical trial monitoring. Stat Biopharm Res. 2012;4(4):384–394.

33. Whitehead J, Valdés-Márquez E, Johnson P, et al. Bayesian sample size for exploratory clinical trials incorporating historical data. Stat Med. 2008;27(13):2307–2327.

34. Gamalo MA, Wu R, Tiwari RC. Bayesian approach to noninferiority trials for proportions. J Biopharm Stat. 2011;21(5):902–919.

35. Kitsche A, Hothorn LA, Schaarschmidt F. The use of historical controls in estimating simultaneous confidence intervals for comparisons against a concurrent control. Comput Stat Data Anal. 2012;56(12):3865–3875.

36. Pei L, Hughes MD. A statistical framework for quantile equivalence clinical trials with application to pharmacokinetic studies that bridge from HIV-infected adults to children. Biometrics. 2008;64(4):1117–1125.

37. Tsou HH, Chien TY, Liu JP, et al. A consistency approach to evaluation of bridging studies and multi-regional trials. Stat Med. 2011;30(17):2171–2186.

38. Chow SC, Chiang C, Liu JP, et al. Statistical methods for bridging studies. J Biopharm Stat. 2012;22(5):903–915.

39. Gould AL, Jin T, Zhang LX, et al. A predictive Bayesian approach to the design and analysis of bridging studies. J Biopharm Stat. 2012;22(5):916–934.

40. Aitchison J, Dunsmore IR. Statistical prediction analysis. CUP Archive; 1975.

41. Hampson LV, Jennison C. Group sequential tests for delayed responses (with discussion). J Roy Stat Soc B. 2013;75(1):3–54.

42. Galbraith S, Marschner IC. Interim analysis of continuous long-term endpoints in clinical trials with longitudinal outcomes. Stat Med. 2003;22(11):1787–1805.

43. Marschner IC, Becker SL. Interim monitoring of clinical trials based on long-term binary endpoints. Stat Med. 2001;20(2):177–192.

44. Stallard N. A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. Stat Med. 2010;29(9):959–971.

45. Wüst K, Kieser M. Blinded sample size recalculation for normally distributed outcomes using long-and short-term data. Biometrical J. 2003;45(8):915–930.

46. Engel B, Walstra P. Increasing precision or reducing expense in regression experiments by using information from a concomitant variable. Biometrics. 1991;47(1):13–20.

47. Thall PF, Simon R. Incorporating historical control data in planning phase II clinical trials. Stat Med. 1990;9(3):215–228.

48. Arends LR, Vokó Z, Stijnen T. Combining multiple outcome measures in a meta-analysis: an application. Stat Med. 2003;22(8):1335–1353.

49. Chen CT, Hung HJ, Hsiao CF. Design and evaluation of multiregional trials with heterogeneous treatment effect across regions. J Biopharm Stat. 2012;22(5):1037–1050.

50. Ko FS. Approaches to allocate sample size rationally into individual regions for a multi-regional trial under heterogeneous effect size. Commun Stat Theory. 2012;41(20):3648–3665.

51. Liu JP, Hsueh H, Chen JJ. Sample size requirements for evaluation of bridging evidence. Biometrical J. 2002;44(8):969–981.

52. Ko FS. A design for survival studies: A multi-regional trial with unrecognized heterogeneity. Commun Stat Theory. 2013;42(13):2301–2323.

53. Lan KKG, Soo Y, Siu C, et al. The use of weighted Z-tests in medical research. J Biopharm Stat. 2005;15(4):625–639.

54. Huang Q, Chen G, Yuan Z, et al. Design and sample size considerations for simultaneous global drug development program. J Biopharm Stat. 2012;22(5):1060–1073.

55. Tsou HH, Tsong Y, Liu JT, et al. Weighted evidence approach of bridging study. J Biopharm Stat. 2012;22(5):952–965.

56. Chen J, Quan H, Binkowitz B, et al. Assessing consistent treatment effect in a multi-regional clinical trial: a systematic review. Pharm Stat. 2010;9(3):242–253.

57. Cochran WG. The combination of estimates from different experiments. Biometrics. 1954;10(1):101–129.

58. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. Stat Med. 1998;17(8):841–856.

59. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557–560.

60. Mittlböck M, Heinzl H. A simulation study comparing properties of heterogeneity measures in meta-analyses. Stat Med. 2006;25(24):4321–4333.

61. Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I2 in assessing heterogeneity may mislead. BMC Med Res Methodol. 2008;8(1):79.

62. Higgins JP, Thompson SG, et al. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21(11):1539–1558.

63. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. Biometrics. 1985;41(2):361–372.

64. Chow SC, Shao J, Hu OYP. Assessing sensitivity and similarity in bridging studies. J Biopharm Stat. 2002;12(3):385–400.

65. Ministry of Health, Labour and Welfare of Japan. Basic principles on global clinical trials;. `http://www.pmda.go.jp/kijunsakusei/file/guideline/new_drug/GlobalClinicalTrials_en.pdf` (2007, accessed: 01/10/2014).

66. Uesaka H. Sample size allocation to regions in a multiregional trial. J Biopharm Stat. 2009;19(4):580–594.

67. Tsou HH, Chow SC, Lan KK, et al. Proposals of statistical consideration to evaluation of results for a specific region in multi-regional trials–Asian perspective. Pharm Stat. 2010;9(3):201–206.

68. Quan H, Zhao PL, Zhang J, et al. Sample size considerations for Japanese patients in a multi-regional trial based on MHLW guidance. Pharm Stat. 2010;9(2):100–112.

69. Ko FS, Tsou HH, Liu JP, et al. Sample size determination for a specific region in a multiregional trial. J Biopharm Stat. 2010;20(4):870–885.

70. Luo X, Shih WJ, Ouyang SP, et al. An optimal adaptive design to address local regulations in global clinical trials. Pharm Stat. 2010;9(3):179–189.

71. Tsou HH, Hung HMJ, Chen YM, et al. Establishing consistency across all regions in a multi-regional clinical trial. Pharm Stat. 2012;11(4):295–299.

72. Tsong Y, Chang WJ, Dong X, et al. Assessment of regional treatment effect in a multiregional clinical trial. J Biopharm Stat. 2012;22(5):1019–1036.

73. Li G, Chen J, Quan H, et al. Consistency assessment with global and bridging development strategies in emerging markets. Contemp Clin Trials. 2013;36(2):687–696.

74. Chen J, Zheng H, Quan H, et al. Graphical assessment of consistency in treatment effect among countries in multi-regional clinical trials. Clin Trials J. 2013;10(6):842–851.

75. Hsiao CF, Xu JZ, Liu JP. A two-stage design for bridging studies. J Biopharm Stat. 2004;15(1):75–83.

76. Cai Z, Uesaka H, Tsujimoto M. Comparison of two ethnic populations using distribution adjusted mean. J Biopharm Stat. 2013;23(3):539–558.

77. Shao J, Chow SC. Reproducibility probability in clinical trials. Stat Med. 2002;21(12):1727–1742.

78. Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B. Robust meta-analytic-predictive priors in clinical trials with historical control information. Biometrics. 2014;70(4):1023–1032.

79. Khalil F, Läer S. Physiologically based pharmacokinetic modeling: methodology, applications, and limitations with a focus on its role in pediatric drug development. J Biomed Biotechnol. 2011;2011.

80. Reif S, Schultze-Mosgau M, Sutter G. From adults to children simulation-based choice of an appropriate sparse-sampling schedule. Pediatr Drugs. 2012;14(3):189–200.

81. De Santis F. Using historical data for Bayesian sample size determination. J Roy Stat Soc A Sta. 2007;170(1):95–113.

82. Gajewski BJ. Comments on A note on the power priorby Neuenschwander B, Branson M and Spiegelhalter DJ. Statistics in Medicine; DOI: 10.1002/sim. 3722. Stat Med. 2010;29(6):708–709.

Online supplementary material to accompany the manuscript "Extrapolation of efficacy and other data to support the development of new medicines for children: a systematic review of methods" by Wadsworth I, Hampson LV and Jaki T.

This document contains the following three appendices:

Appendix A: Systematic review search strategy.

Appendix B: Data extraction form used to record relevant information from articles identified by the systematic review

Appendix C: Search strategy for software implementing methods identified by the systematic review.

In addition, file "Appendix D Spreadsheet.xlsx" contains Appendix D.

Appendix D: For each method this file lists the following information: a) the citation number (as listed in the main text) and bibliographic details of the paper from which the method was extracted; b) a short description of the method; c) whether the method is Bayesian or frequentist; and d) whether software is available to implement the method and what statistical language this is written in (i.e., R, WinBUGS etc).

Appendix A: Systematic review search strategy

Below are listed the search terms used to perform three searches of the Web of Science SCI-EXPANDED database.

Search 1: Paediatric clinical trials and bridging trials extrapolating efficacy across ethnic groups or regions combined into one search

(TS=((bridging OR "borrow* strength" OR extrapolat* OR synthesize) AND (p$ediatric OR child* OR ethnic OR region* OR geotherapeutic* OR centre OR center) AND (trial* OR "bridging stud*"))) AND (WC=(Biology OR Mathematical & Computational Biology OR Mathematics, Applied OR Mathematics, Interdisciplinary Applications OR Medical Informatics OR Medicine, Research & Experimental OR Pediatrics OR Statistics & Probability))

Search 2: Historical controls in clinical trials

(TS=(("historical control*" OR "historical information" OR "historical data") AND (trial*))) AND (WC=(Biology OR Mathematical & Computational Biology OR Mathematics, Applied OR Mathematics, Interdisciplinary Applications OR Medical Informatics OR Medicine, Research & Experimental OR Pediatrics OR Statistics & Probability))

Search 3: Using short-term endpoints to support inferences about treatment effects on long-term endpoints

(TS=((short-term endpoint OR short-term end point OR biomarker OR surrogate endpoint OR surrogate end point) AND (long-term endpoint OR long-term end point) AND (trial*))) AND (WC=(Biology OR Mathematical & Computational Biology OR Mathematics, Applied OR Mathematics, Interdisciplinary Applications OR Medical Informatics OR Medicine, Research & Experimental OR Pediatrics OR Statistics & Probability))

Appendix B: Data extraction form used to record relevant information from articles identified by the systematic review

Citation:

DOI:

Repeat of Paper?

1. **What is the source population?**
   e.g. adult, original region.
2. **What is the target population?**
   e.g. paediatric, new region.
3. **Does the method assume a homogenous target population?**
4. **Is the question to be addressed based on,**
   - **4.1.** Comparison of interventions
   - **4.2.** Dose finding
   - **4.3.** Other

   Comments:
5. **Specific example of setting?**
   - **5.1.** Paediatric clinical trials.
   - **5.2.** Using short-term endpoints to support inferences about treatment effects on long-term endpoints.
   - **5.3.** Historical controls in clinical trials.
   - **5.4.** Bridging trials extrapolating efficacy across ethnic groups / regions / centres.
6. **Does the method require data from a source population?**
7. **Does the method require data from a target population?**

Comments:

8. What type of relevant data is required?

(can pick multiple)

	Source	Target
8.1. PK	☐	☐
8.2. Efficacy	☐	☐
8.3. Safety	☐	☐

9. What is the form of the required data?

(can pick multiple)

	Source	Target
9.1. Continuous outcome measure	☐	☐
9.2. Binary	☐	☐
9.3. Time-to-event	☐	☐
9.4. Ordered categorical	☐	☐
9.5. Unordered categorical	☐	☐
9.6. Count data	☐	☐

10. What quality of data does the method require / can the method accommodate?

	Source	Target
10.1. High (RCT)	☐	☐
10.2. Medium (observational studies)	☐	☐
10.3. Low (Case reports)	☐	☐
10.4. Not clear	☐	☐
10.5. Other	☐	☐

Comments: ☐

11. Is the method Bayesian and / or Frequentist?

11.1. Bayesian ☐

11.2. Frequentist ☐

12. Extrapolation process: Inferences regarding differences between source and target population (e.g. are Exposure-Response curves similar in adult and paediatric populations)

12.1. Are data collected to generate hypotheses about differences between the source and target populations? ☐ If **no**, go to **14**

12.2. Are these data from the source and target populations? ☐ If **no**, go to **12.3**

12.3. What data are collected for inference regarding differences between the source and target population?

12.4. How is the method exploring the differences between the source and target populations?

13. Details of the statistical model used for Q12.

13.1. Model used.

13.2. Model not known. ☐

13.3. NA ☐

14. Extrapolation process: Inferences about key parameter in target population (i.e. efficacy parameter in target population)

14.1. From the conclusion of **Q12**, are the source and target populations assumed to be similar? ☐ go to **14.2**.

14.2. For inference on the target population, are inferences:

14.2.1. made in the source population only? ☐ If **yes**, go to **14.3**

14.2.2. made in the target population only? ☐ If **yes**, leave comments, go to **15.**

14.2.3. made in both the source and target populations? ☐ If **yes**, go to **14.4**.

14.2.4. not clear? ☐ go to **16.**

Comments: ☐

14.3. Are key parameters of interest assumed to be the same in the source and target populations? ☐ Leave comments, go to **15**.

Comments: ☐

14.4. Are inferences about key parameters in the target population to be based on:

14.4.1. An overall model for the data from the source and target populations? ☐ If **yes**, go to **14.5**.

14.4.2. concurrent data from the target population? ☐ If **yes**, go to **14.6**.

14.4.3. Weighted test of source and target. ☐ If **yes**, go to **15.**

14.5. In the overall model,

14.5.1. Are key parameters in source and target populations assumed to be the same? ☐

14.5.2. Are nuisance parameters (e.g. variances) in source and target populations assumed to be the same? ☐

14.5.3. Other? ☐

Comments: ☐

14.6. How is the method borrowing strength from data in the source population?

14.6.1. Creation of informative prior? ☐
14.6.2. Use of point prior? ☐
14.6.3. Informal supportive analysis? ☐
14.6.4. Other? ☐

Comments:

15. Details of the statistical model used for Q14.

15.1. Model for the prior.

15.2. Model for the likelihood.

15.3. Model for the posterior.

15.4. Model not known. ☐

15.5. NA ☐

16. Has this method been devised with paediatric trials in mind? ☐

Appendix C: Search strategy for software implementing methods identified by the systematic review.

We searched for software implementing methods identified by the systematic review in the following ways:

1) By checking whether code was listed in the paper proposing the method (either in the main text, an Appendix, or on-line supplementary material). We also recorded whether it was stated in the paper that code is available from the authors upon request.
2) By checking the references of each paper for companion software papers;
3) By checking papers listed by Web of Science as having cited the original article to see whether these included companion software papers.

Groups of approaches	Main text reference	Paper reference
Bayesian method to create an informative prior for θ_T		
	2	Ibrahim JG, Chen MH. Power prior distributions for regression models. Stat Sci. 2000;15(1):46-60.
	3	Hobbs BP, Carlin BP, Mandrekar SJ, et al. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. Biometrics. 2011;67(3):1047-1056.
	4	Neuenschwander B, Capkun-Niggli G, Branson M, et al. Summarizing historical information on controls in clinical trials. Clin Trials J. 2010;7(1):5-18.
	17	Neuenschwander B, Branson M, Spiegelhalter DJ. A note on the power prior. Stat Med. 2009;28(28):3562-3566.
	18	Chen MH, Ibrahim JG, Lam P, et al. Bayesian design of noninferiority trials for medical devices using historical data. Biometrics. 2011;67(3):1163-1170.

19 Hobbs BP, Sargent DJ, Carlin BP. Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. Bayesian Anal. 2012;7(3):639-674.

20 Hobbs BP, Carlin BP, Sargent DJ. Adaptive adjustment of the randomization ratio using historical control data. Clinl Trials J. 2013;10(3):430-440.

21 Gsteiger S, Neuenschwander B, Mercier F, et al. Using historical control information for the design and analysis of clinical trials with overdispersed count data. Stat Med. 2013;42(21):3609-3622.

22 Cuffe RL. The inclusion of historical control data may reduce the power of a con rmatory study. Stat Med. 2011;30(12):1329-1338.

24 Hsiao CF, Hsu YY, Tsou HH, et al. Use of prior information for Bayesian evaluation of bridging studies. J Biopharm Stat. 2007;17(1):109-121.

25 Gandhi M, Mukherjee B, Biswas D. A Bayesian approach for inference from a bridging study with binary outcomes. J Biopharm Stat. 2012;22(5):935-951.

26 De Santis F. Power priors and their use in clinical trials. Am Stat. 2006;60(2):122-129.

27 Rietbergen C, Klugkist I, Janssen KJM, et al. Incorporation of historical data in the analysis of randomized therapeutic trials. Contemp Clin Trials. 2011;32(6):848-855.

28 Schoenfeld DA, Zheng H, Finkelstein DM. Bayesian design using adult data to augment pediatric trials. Clin Trials J. 2009;6(4):297-304.

29 Chen YH,Wu YC,Wang M. A Bayesian approach to evaluating regional treatment effect in a multiregional trial. J Biopharm Stat. 2009;19(5):900-915.

30 Goodman SN, Sladky JT. A Bayesian approach to randomized controlled trials in children utilizing information from adults: the case of Guillain-Barre. Clin Trials J. 2005;2(4):305-310.

31 Greenhouse JB, Seltman H. Using prior distributions to synthesize historical evidence: comments on the Goodman{Sladky case study of IVIg in Guillain-Barre syndrome. Clin Trials J. 2005;2(4):311-318.

32 French JL, Thomas N, Wang C. Using historical data with Bayesian methods in early clinical trial monitoring. Stat Biopharm Res. 2012;4(4):384-394.

33	Whitehead J, Vald es-M arquez E, Johnson P, et al. Bayesian sample size for exploratory clinical trials incorporating historical data. Stat Med. 2008;27(13):2307-2327.
34	Gamalo MA, Wu R, Tiwari RC. Bayesian approach to noninferiority trials for proportions. J Biopharm Stat. 2011;21(5):902-919.
35	Kitsche A, Hothorn LA, Schaarschmidt F. The use of historical controls in estimating simultaneous con dence intervals for comparisons against a concurrent control. Comput Stat Data Anal. 2012;56(12):3865-3875.
Bayesian method to assess consistency between θ_T and θ_S	
36	Pei L, Hughes MD. A statistical framework for quantile equivalence clinical trials with application to pharmacokinetic studies that bridge from HIV-infected adults to children. Biometrics. 2008;64(4):1117-1125.
37	Tsou HH, Chien TY, Liu JP, et al. A consistency approach to evaluation of bridging studies and multi-regional trials. Stat Med. 2011;30(17):2171-2186.
38	Chow SC, Chiang C, Liu JP, et al. Statistical methods for bridging studies. J Biopharm Stat. 2012;22(5):903-915.
39	Gould AL, Jin T, Zhang LX, et al. A predictive Bayesian approach to the design and analysis of bridging studies. J Biopharm Stat. 2012;22(5):916-934.
Frequentist method modelling data from source and target populations in a joint model	
41	Hampson LV, Jennison C. Group sequential tests for delayed responses (with discussion). J Roy Stat Soc B. 2013;75(1):3-54.

42	Galbraith S, Marschner IC. Interim analysis of continuous long-term endpoints in clinical trials with longitudinal outcomes. Stat Med. 2003;22(11):1787-1805.
43	Marschner IC, Becker SL. Interim monitoring of clinical trials based on long-term binary endpoints. Stat Med. 2001;20(2):177-192.
44	Stallard N. A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. Stat Med. 2010;29(9):959-971.
45	Wüst K, Kieser M. Blinded sample size recalculation for normally distributed outcomes using long-and short-term data. Biometrical J. 2003;45(8):915-930.
47	Thall PF, Simon R. Incorporating historical control data in planning phase II clinical trials. Stat Med. 1990;9(3):215-228.
48	Arends LR, Vok o Z, Stijnen T. Combining multiple outcome measures in a meta-analysis: an application. Stat Med. 2003;22(8):1335-1353.
49	Chen CT, Hung HJ, Hsiao CF. Design and evaluation of multiregional trials with heterogeneous treatment e ect across regions. J Biopharm Stat. 2012;22(5):1037-1050.
50	Ko FS. Approaches to allocate sample size rationally into individual regions for a multi-regional trial under heterogeneous e ect size. Commun Stat Theory. 2012;41(20):3648-3665.
51	Liu JP, Hsueh H, Chen JJ. Sample size requirements for evaluation of bridging evidence. Biometrical J. 2002;44(8):969-981.
52	Ko FS. A design for survival studies: A multi-regional trial with unrecognized heterogeneity. Commun Stat Theory. 2013;42(13):2301-2323.
Frequentist method combining source and target population data in a weighted test statistic	
53	Lan KKG, Soo Y, Siu C, et al. The use of weighted Z-tests in medical research. J Biopharm Stat. 2005;15(4):625-639.
54	Huang Q, Chen G, Yuan Z, et al. Design and sample size considerations for simultaneous global drug development program. J Biopharm Stat. 2012;22(5):1060-1073.
55	Tsou HH, Tsong Y, Liu JT, et al. Weighted evidence approach of bridging study. J Biopharm Stat. 2012;22(5):952-965.

Frequentist method to assess consistency between θ_T and θ_S

Nedelman JR, Rubin DB, Sheiner LB. Diagnostics for confounding in PK/PD models for oxcarbazepine. Stat Med. 2007;26(2):290-308.

Pei L, Hughes MD. A statistical framework for quantile equivalence clinical trials with application to pharmacokinetic studies that bridge from HIV-infected adults to children. Biometrics. 2008;64(4):1117-1125.

Tsou HH, Chien TY, Liu JP, et al. A consistency approach to evaluation of bridging studies and multi-regional trials. Stat Med. 2011;30(17):2171-2186.

Chow SC, Chiang C, Liu JP, et al. Statistical methods for bridging studies. J Biopharm Stat. 2012;22(5):903-915.

Chen CT, Hung HJ, Hsiao CF. Design and evaluation of multiregional trials with heterogeneous treatment e ect across regions. J Biopharm Stat. 2012;22(5):1037-1050.

Ko FS. Approaches to allocate sample size rationally into individual regions for a multi-regional trial under heterogeneous e ect size. Commun Stat Theory. 2012;41(20):3648-3665.

Liu JP, Hsueh H, Chen JJ. Sample size requirements for evaluation of bridging evidence. Biometrical J. 2002;44(8):969-981.

Ko FS. A design for survival studies: A multi-regional trial with unrecognized heterogeneity. Commun Stat Theory. 2013;42(13):2301-2323.

Chen J, Quan H, Binkowitz B, et al. Assessing consistent treatment effect in a multi-regional clinical trial: a systematic review. Pharm Stat. 2010;9(3):242-253.

66 Uesaka H. Sample size allocation to regions in a multiregional trial. J Biopharm Stat. 2009;19(4):580-594.

67 Tsou HH, Chow SC, Lan KK, et al. Proposals of statistical consideration to evaluation of results for a speci c region in multi-regional trials - Asian perspective. Pharm Stat. 2010;9(3):201-206.

68 Quan H, Zhao PL, Zhang J, et al. Sample size considerations for Japanese patients in a multi-regional trial based on MHLW guidance. Pharm Stat. 2010;9(2):100-112.

69 Ko FS, Tsou HH, Liu JP, et al. Sample size determination for a specific region in a multiregional trial. J Biopharm Stat. 2010;20(4):870-885.

70 Luo X, Shih WJ, Ouyang SP, et al. An optimal adaptive design to address local regulations in global clinical trials. Pharm Stat. 2010;9(3):179-189.

71 Tsou HH, Hung HMJ, Chen YM, et al. Establishing consistency across all regions in a multi-regional clinical trial. Pharm Stat. 2012;11(4):295-299.

72 Tsong Y, Chang WJ, Dong X, et al. Assessment of regional treatment effect in a multiregional clinical trial. J Biopharm Stat. 2012;22(5):1019-1036.

73 Li G, Chen J, Quan H, et al. Consistency assessment with global and bridging development strategies in emerging markets. Contemp Clin Trials. 2013;36(2):687-696.

74 Chen J, Zheng H, Quan H, et al. Graphical assessment of consistency in treatment e ect among countries in multi-regional clinical trials. Clin Trials J. 2013;10(6):842-851.

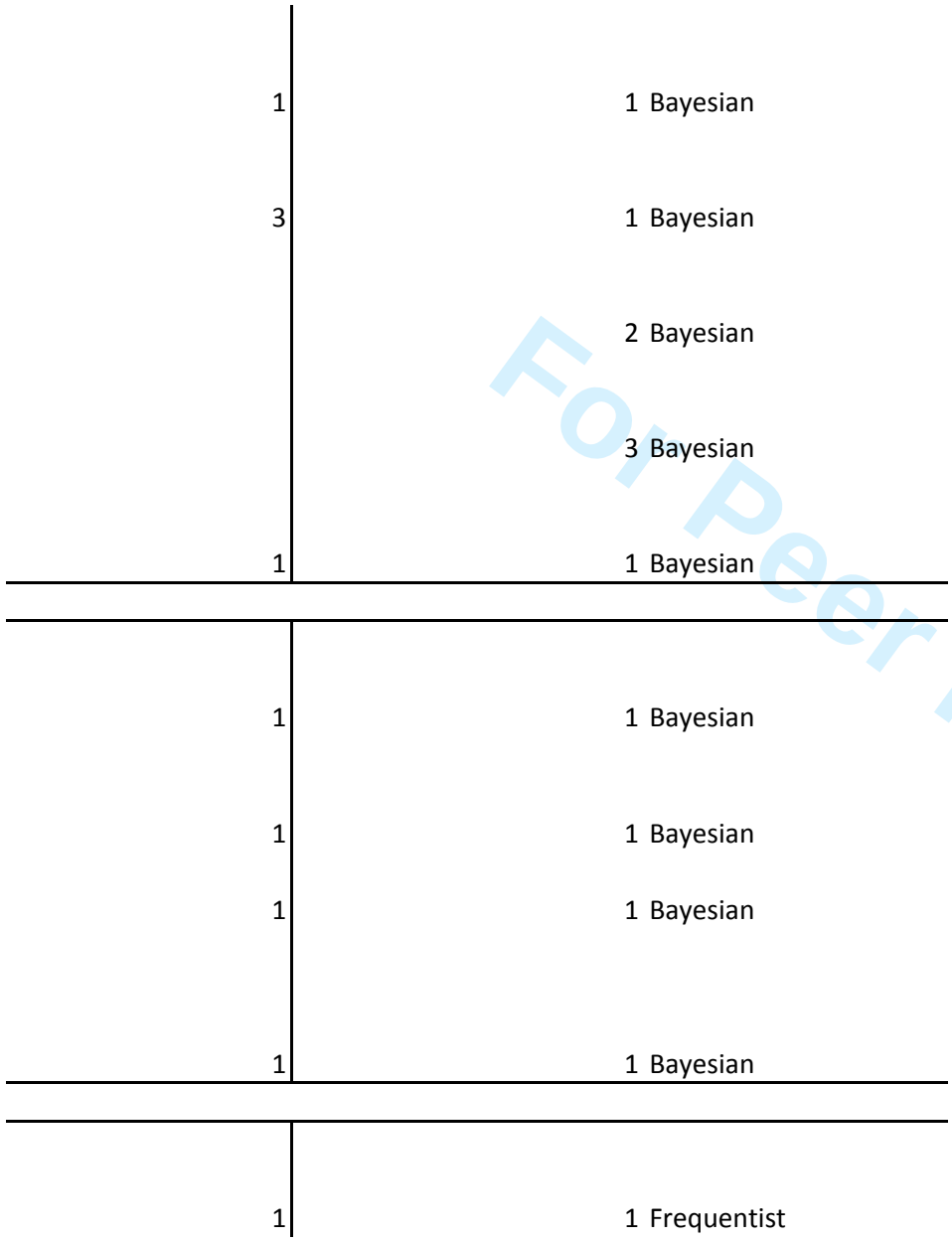
75 Hsiao CF, Xu JZ, Liu JP. A two-stage design for bridging studies. J Biopharm Stat. 2004;15(1):75-83.

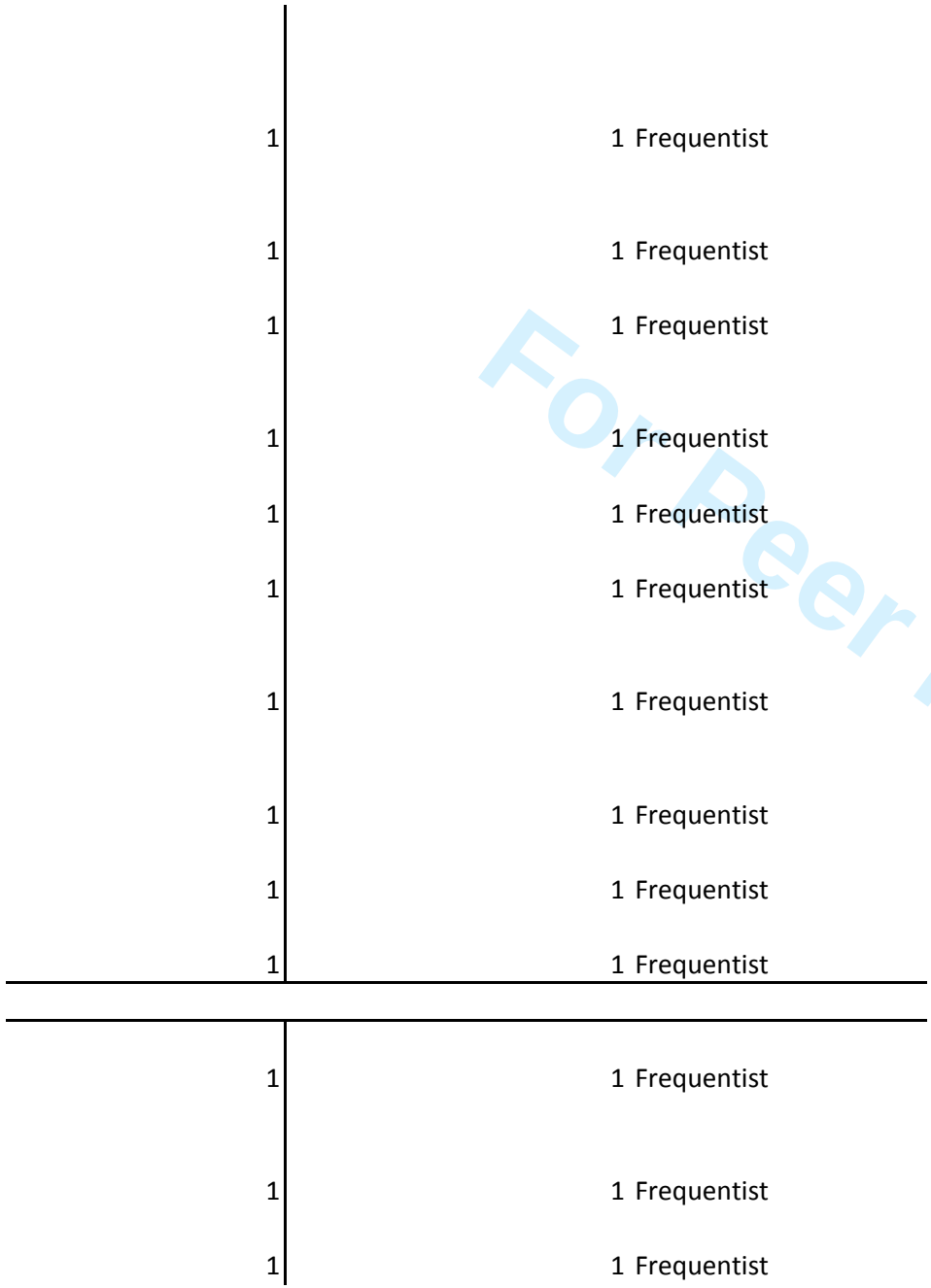
76 Cai Z, Uesaka H, Tsujimoto M. Comparison of two ethnic populations using distribution adjusted mean. J Biopharm Stat. 2013;23(3):539-558.

Number of methods	Method number in paper	Frequentist/Bayesian
7	1	Bayesian
	2	Bayesian
	3	Bayesian
	4	Bayesian
	5	Bayesian
	6	Bayesian
	7	Bayesian
8	1	Bayesian
	2	Bayesian
	3	Bayesian
	4	Bayesian
	5	Bayesian
	6	Bayesian
	7	Bayesian
	8	Bayesian
1	1	Bayesian
1	1	Bayesian
2	1	Bayesian
	2	Bayesian

10	1	Bayesian
	2	Bayesian
	3	Bayesian
	4	Bayesian
	5	Bayesian
	6	Bayesian
	7	Bayesian
	8	Bayesian
	9	Bayesian
	10	Bayesian
1	1	Bayesian
1	1	Bayesian
1	1	Bayesian
1	1	Bayesian
3	1	Bayesian
	2	Bayesian
	3	Bayesian
2	1	Bayesian

		2	Bayesian
	1	1	Bayesian
	1	1	Bayesian
	1	1	Bayesian
	1	1	Bayesian
	4	1	Bayesian
		2	Bayesian
		3	Bayesian
		4	Bayesian
	3	1	Bayesian
		2	Bayesian
		3	Bayesian





1	1 Frequentist
1	1 Frequentist
1	1 Frequentist
1	1 Frequentist
1	1 Frequentist
2	1 Frequentist 2 Frequentist
1	1 Frequentist
2	1 Frequentist 2 Frequentist
8	1 Frequentist 2 Frequentist 3 Frequentist 4 Frequentist 5 Frequentist 6 Frequentist 7 Frequentist 8 Frequentist

1	1 Frequentist
1	1 Frequentist
1	1 Frequentist
1	1 Frequentist
1	1 Frequentist
1	1 Frequentist
1	1 Frequentist
2	1 Frequentist
	2 Frequentist
1	1 Frequentist
1	1 Frequentist
1	1 Frequentist

Short description
Power prior for an arbitrary regression model.
Hierarchical power prior, specifying a prior distribution for the power parameter.
Hierarchical power prior with multiple historical studies.
Power prior for generalized linear models.
Power prior for generalized linear mixed models.
Power prior for proportional hazards model.
Power prior for cure rate model.
Conditional power prior, with fixed power paramter.
Modified power prior, specifying a prior distribution for the power parameter and including the normalising constant.
Location commensurate power prior, borrowing strength from the historical study depends upon the evidence in the data for commensurability between the location parameters
Commensurate prior, estimation of commensurability among the historical and current data in a hierarchical model, fixed commensurability paramter.
Commensurate prior, estimation of commensurability among the historical and current data in a hierarchical model, specifying prior for the commensurability paramter.
Location commensurate prior, in the case of Guassian data.
Location-scale commensurate prior, in the case of Gaussian data.
Location-scale commensurate mixture priors, in the case of Gaussian data.
Meta-analytic-predictive approach to find the predictive distribution of the control parameter in the new study, to be used as a formal prior to be incorporated in the final analysis.
Power prior, prior distribution placed on the power parameter, normalising constant included.
Synthesises historical and contemporary data within a Bayesian random-effects meta-analytic model, considers both normal and binary data.
Power prior, prior distribution placed on the power parameter, normalising constant included, considers both normal and binary data.

Commensurate prior with one historical study.
Commensurate prior with multiple historical studies.
Commensurate prior for general linear models with fixed effects.
Commensurate prior for general linear models with mixed effects.
Commensurate prior for generalized linear models with fixed effects.
Commensurate prior for generalized linear models with fixed effects, considering binary data.

Commensurate prior for generalized linear models with fixed effects, considering survival data.
Commensurate prior for generalized linear models with mixed effects.
Commensurate prior for generalized linear models with mixed effects, considering binary data.
Commensurate prior for generalized linear models with mixed effects, considering count data.

Commensurate prior, with spike and slab prior considered for the commensurability paramter.

Meta-analaytic predictive prior for the log mean count on control in a new trial when count data are overdispersed and follow a negative binomial model.

Bayesian random-effects model incorporating data from historical and contemporary controls.
Mixture prior to augment data from a future clinical trial in a new geographic region with data from an area that has previously been studied.
Empirical Bayes approach, used existing binary data from a geographic region in which a drug has been shown to be effective to create an informative prior distribution for the treatment effect of a bridging trial in a new region, without any down-weighting.

Mixture prior approach, used existing binary data from a geographic region in which a drug has been shown to be effective to create a mixture prior for the treatment effect of a bridging trial in a new region

Power prior with fixed power paramter, incorporating existing binary data from a geographic region in which a drug has been shown to be effective into the analysis of a bridging trial conducted in a new region.
Power prior, fixed power parameter equal to a value specified by the analyst divided by the number of subjects in the source population data.

Power prior, fixed power parameter equal to a value specified by the analyst divided by the number of subjects in the source population data, modified to weight different historical datasets by different fractions when they differ in their relevance to the new trial.

Power prior with fixed power paramter, incorporating data from several historical studies, assigning data from each study a weight elicited from expert opinion.
Hierarchical model, augments data from a clinical trial in children with data from a completed adult trial, assuming parameters of adult and paediatric data are samples from a normal population distribution

A Bayesian empirical prior distribution for a treatment effect in a specific local region of a MRCT which borrows strength from data from other trial sites.

Creation of informative prior for the treatment effect in the target population, based on shifting the location and/or inflating the standard error of an estimate of the treatment effect in the source population.

Used data from a source population to create an informative prior distribution for the treatment effect in the target population, without any down-weighting.
Creation of informative prior for the treatment effect in the target population, based on shifting the location and/or inflating the standard error of an estimate of the treatment effect in the source population - Equal but discounted prior.
Creation of informative prior for the treatment effect in the target population, based on shifting the location and/or inflating the standard error of an estimate of the treatment effect in the source population - Skeptical prior.
Creation of informative prior for the treatment effect in the target population, based on shifting the location and/or inflating the standard error of an estimate of the treatment effect in the source population - Enthusiastic prior.
Used data from a source population to create an informative prior distribution for the treatment effect in the target population, without any down-weighting.

Used data from a source population to create an informative prior distribution for the treatment effect in the target population, standard deviation inflated to account for the potential between study variability.
Prior distribution for hazard rates based on several historical studies.

Uses historical placebo data to create an informative prior distribution for the expected response on placebo in the new trial, in the context of sample size calculation.

Used data from a source population to create an informative prior distribution for the treatment effect in the target population, without any down-weighting - Fully Bayesian approach.

Used data from a source population to create an informative prior distribution for the treatment effect in the target population, without any down-weighting - Normal Approximation to Beta approach.

Used data from a source population to create an informative prior distribution for the treatment effect in the target population, without any down-weighting - Normal approximation to likelihood approach.

Used data from a source population to create an informative prior distribution for the treatment effect in the target population, without any down-weighting.

In a Bayesian framework, assess whether candidate doses for adults and children result in similar percentages of patients experiencing low levels of a drug.
Uses Bayesian most plausible prediction to assess the consistency of treatment effect estimates generated by a new clinical trial comparing an experimental treatment (E) with control (C), and reference studies which have demonstrated the advantage of E versus C.
Uses posterior predictive probabilities to assess the consistency of data from a bridging trial and reference studies.

Bayesian method to assess the consistency of parameters in source and target populations, results of a bridging study are judged to be consistent with those of the reference studies if they fall within contours or regions of the posterior predictive distribution derived from the reference confirmatory trials.

In group sequential tests, monitor a long-term outcome by incorporating data on a correlated short-term endpoint so as to increase the Fisher information available for the treatment effect at each interim analysis.

In group sequential tests, incorporate repeated measurements of a continuous endpoint taken at an arbitrary number of follow-up time, assuming a multivariate normal distribution, with correlations between the measurements being exploited to improve estimation and inference associated with the long-term measurement.

Increase the interim information available for a long-term response probability by incorporating data on a short-term binary endpoint.
Uses observations on short- and long-term endpoints to support early stopping decisions and treatment selection decisions of a seamless Phase II/III clinical trial.

Derive a more precise estimator of the variance of a long-term outcome incorporating short- and long-term data, using this improved estimator to inform blinded sample size adjustments at an interim analysis.

Combines historical and contemporary control data via a univariate random-effects meta-analysis.

Models short-term and long-term outcomes from trials using a multivariate random effects model.

Random effects model to deal with heterogeneity between regions.

Random effects model to deal with heterogeneity between regions.
Hierarchical model, incorporate data from original completed studies to evaluate the sample size required for the analysis of a bridging study.
Frailty model for survival data, different regions have different frailties (like random effects within a hazard function).

Final inferences about the efficacy of a new treatment in a new geographic region on the basis of a weighted test-statistic combining information from the source and target populations.

Final inferences about the efficacy of a new treatment in a new geographic region on the basis of a weighted test-statistic combining information from the source and target populations.
Final inferences about the efficacy of a new treatment in a new geographic region on the basis of a weighted test-statistic combining information from the source and target populations.

Compares PK-PD relationships (linear models) in adults and children receiving a new drug as an add-on therapy, if found to be similar this is taken as support for an assumption of similar relationships for adults and children receiving monotherapy.

In a frequentist framework, assess whether candidate doses for adults and children result in similar percentages of patients experiencing low levels of a drug.
Statistical criterion to assess the consistency between the region of interest and overall results in a multi-regional trial.
`Reproducibility probability' method, assessing whether clinical data from the original region can be completely extrapolated to the new region to support claims of efficacy.

Consistency criteria.

Consistency criteria.
Assessing consistent trend.

Non-inferiority hypothesis tests, one study in several regions.

Consistency criteria.
Assessing consistent trend.

Cochran’s Q heterogeneity statistic.
The Gail-Simon qualitative test.
Higgins I-squared, derived from the Cochran’s Q statistic.
Declare treatment effects as consistent if there are no significant pairwise differences between effect estimates.
Equivalence hypothesis tests.
PMDA method 1, consistency if a positive trend is observed.
PMDA method 2, consistency if a positive trend is observed.
Method based on confidence interval coverage of a target value.

Consistency criteria.

Consistency criteria.

Consistency criteria.

Consistency criteria.

Consistency criteria.

Consistency criteria.

Consistency criteria.

Qualitative consistency criteria.
Quantitative consistency criteria.

Derive standardised weighted least squares residuals, used to create Q-Q plots for assessing consistency between regional treatment effects.
Two-stage design for bridging trials: The trial begins recruiting patients from the original region. If efficacy in this region is confirmed at the interim analysis, the trial proceeds to recruit patients from the new region in Stage 2. On conclusion of the trial, data accumulated from both regions are pooled. If the result of Stage 1 is similar to the pooled result of Stage 2, the result from the new region is declared consistent with that from the original region.
Evaluates the similarity of data from clinical trials performed in different ethnic populations using a 'distribution adjusted mean'.

Software used / code available
None provided.
None provided.
None provided.
None provided.
None provided.
None provided.
None provided.
None provided.
None provided.
None provided.
None provided.
None provided.
None provided.
None provided.
None provided.
None provided.
None provided, though there is a comment paper (ref [82]) containing WinBugs code.
None provided.
None provided.

None provided.
None provided.
None provided.
None provided.
None provided.
None provided.

None provided.
None provided.
None provided.
None provided.

None provided.

Includes WinBugs code.

None provided.

None provided.

None provided.

None provided.

None provided.

None provided.

None provided.

Includes WinBugs code.

None provided.

None provided.

None provided.

None provided.

None provided.

None provided.

None provided.

Includes OpenBugs code.

Includes OpenBugs code.

Includes OpenBugs code.

SAS code included for the sample size calculation.
None provided.
None provided.
None provided. R code to conduct the analysis is provided online (link http://www.biostat.uni-hannover.de/software)
WinBugs code provided in supplementary material, appendix B.
None provided.
None provided. R code provided in online supplementary material, appendix 2 (link http://www.tandfonline.com/doi/suppl/10.1080/10543406.2012.701579#tabModule)
None provided.

None provided.

None provided.

None provided.

None provided.

None provided.

None provided.

None provided.

None provided.

None provided.

None provided.

None provided.

None provided.

None provided.

None provided.

None provided.
None provided.
None provided.
None provided.
None provided.
None provided.
None provided.
None provided.
None provided.
None provided.
None provided.
None provided.
None provided.
None provided.
None provided.
None provided.
None provided.
None provided.

None provided.

None provided.

None provided.

None provided.

R code for the examples within the paper are available from the corresponding author.

None provided.

None provided.

None provided.

None provided.

None provided.

None provided.

None provided.