Bridging the Gap: Using Extrapolation to Reduce the Experimental Burden in Children

Ian Wadsworth, BSc., MSc.

Submitted for the degree of Doctor of Philosophy

at Lancaster University,

February 2018.

Bridging the Gap: Using Extrapolation to Reduce the Experimental Burden in Children

Ian Wadsworth, BSc., MSc.

Department of Mathematics and Statistics, Lancaster University

Submitted for the degree of Doctor of Philosophy at Lancaster University, February 2018.

Abstract

When developing a new drug product for children, it is important to provide safe and effective medicines whilst minimising the experimental burden where possible. Extrapolating data from clinical trials in adults or other relevant populations has the potential to reduce the number or size of clinical trials required to obtain a licence for a new drug in the paediatric population. The ethical and practical benefits extrapolation can provide, such as speeding up access to medicines, reducing drug development costs and avoiding replication of existing information, have to be balanced against potential risks i.e. if extrapolation from adults is incorrectly deemed to be appropriate, children could be exposed to a harmful or ineffective treatment. Extrapolation is therefore a challenging but important aspect for future paediatric medicine development.

The work presented in this thesis is broadly interested in approaches for reducing the experimental burden in the paediatric population by leveraging external information, such as existing adult data or the opinion of clinical experts. The concept of extrapolation, especially in the context of paediatric clinical trials, is explored over the thesis, beginning with a literature review of extrapolation methodology in Chapter 3, particularly aiming to identify potentially relevant methods for performing extrapolations to children.

Extrapolation can be a potentially controversial approach to take in drug development, so the opinion of experts within a target clinical setting is incredibly valuable. Chapters 4 and 5 describe the process and outcome of seeking the opinion of clinical experts regarding extrapolation in epilepsy drug development. Based on this expert opinion, an outline for a new drug development paradigm is presented allowing for simultaneous recruitment of adults and paediatric patients aged 2 years and older, from Phase II onwards.

In order for the extrapolation of efficacy data from adult trials to the paediatric population to be plausible and appropriate, strong assumptions regarding similarity between these populations are required. One important assumption is whether adults and children can be said to have similar pharmacokinetic-pharmacodynamic (PK-PD) relationships. In Chapter 6, an approach is developed to use data from existing studies of adults and adolescents, along with expert opinion, to quantify prior uncertainty regarding the similarity of PK-PD relationships in adults and younger children. A bias-adjusted meta-analysis of existing adult and adolescent data allows the derivation of prior distributions quantifying our uncertainty about the extrapolation assumption and calculation of the prior probability that the extrapolation assumption holds. This approach could be extended to quantifying prior uncertainty in other contexts; here we consider PK-PD relationships in adults and younger children to provide a clear focus.

Within the paediatric population itself, there may exist distinct age groups with different PK-PD relationships requiring separate dosing rules to account for pharmacological differences. Chapter 7 considers model-based approaches to quantify how parameters of PK-PD models differ over age. Based on this, an approach for deriving optimal dosing rules accounting for differences between age groups is developed.

Acknowledgements

I would like to thank Dr. Lisa Hampson and Prof. Thomas Jaki for the invaluable supervision, advice and encouragement throughout my PhD and for giving me such a fantastic opportunity. I have learned a great deal from both over the last four years, not only with what I have found a very interesting PhD topic, but also generally about research, academia and that I am more capable than I thought!

The work presented in this thesis was funded by two grants: the National Institute for Health Research (NIHR) grant NIHR-RMOFS-2013-03-05; and the Medical Research Council (MRC) grant MR/M013510/1. I am thankful for this funding and all of the hard work Lisa (as PI) and everyone else involved put into those grant applications. Thanks also to Prof. Anne Whitehead who, along with Jack, gave me my first research job and convinced me that I could do a PhD at Lancaster. Lancaster University has given me a great deal over the last decade and I have achieved much more here than I ever thought I could.

Thanks also to Dr. Graeme Sills, whose clinical input and discussions have been a huge help, as has the support from Prof. Tony Marson, Dr. Richard Appleton and all epilepsy experts that have taken time to contribute their knowledge and experience. Visiting Dr. Björn Bornkamp at the Novartis campus in Basel was a very valuable experience and his support has been a great help. Thanks also to the useful discussions had with Dr. Linh Van, also of Novartis.

I would also like to thank all of my friends for their support (especially Dr. Amanda Minter who read over part of this thesis) and my parents for always being there whenever I need them. Finally, thanks to my wife Lucy for all of her love and

support over the last seven years.

Declaration

I declare that this thesis is my own work and has not been submitted elsewhere for the award of any other degree.

Ian Wadsworth

List of papers

This thesis includes the following four papers:

- I. Wadsworth, L. V. Hampson, and T. Jaki. Extrapolation of efficacy and other data to support the development of new medicines for children: A systematic review of methods. Statistical Methods in Medical Research, 2016. *Published online ahead of print.* DOI: 10.1177/0962280216631359.

- I. Wadsworth, T. Jaki, G. J. Sills, R. Appleton, J. H. Cross, A. G. Marson, T. Martland, A. McLellan, P. E. Smith, J. M. Pellock, and L. V. Hampson. Clinical drug development in epilepsy revisited: a proposal for a new paradigm streamlined using extrapolation. CNS Drugs, vol. 30, no. 11, pp. 1011–1017, 2016.

- I. Wadsworth, L. V. Hampson, T. Jaki, G. J. Sills, A. G. Marson, R. Appleton. A quantitative framework to inform extrapolation decisions in children. *Under review*

- I. Wadsworth, L. V. Hampson, B. Bornkamp, T. Jaki. Exposure-response modelling approaches for determining optimal dosing rules in children. *Under review*

The Statistical Methods in Medical Research paper is reproduced under SAGE's Green Open Access policy which allows up to one full article in an unpublished dissertation or thesis. The CNS Drugs paper is reproduced under Springer's Open Choice policy, the paper is published open access under the Creative Commons

Attribution Non-Commercial 4.0 International (CC BY-NC) license. The first two papers are as similar as possible to their original printed form and are given in Chapters 3 and 5, respectively. The final two papers are as similar as possible to the manuscript submission and are given in Chapters 6 and 7, respectively.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Decision making in clinical research

In order for new medicines to be approved for use in humans, evidence must be accrued demonstrating a medicine's efficacy, safety and correct dosage before a decision regarding licencing (and subsequent use in the general population) can be made. The primary approach to accrue this evidence is to perform clinical trials. Clinical trials are defined as planned investigations of a treatment in human subjects with the aim of assessing one or more of the following: safety; efficacy; clinical effects; pharmacokinetics (PK; what effect the body has on an administered drug i.e. absorption, distribution, metabolism and excretion); pharmacodynamics (PD; the effect an administered drug has on the body); and adverse reactions.[4,5] Clinical research is often separated into phases, with trials in each phase designed to address specific questions, though for some trials the distinction between phases may not be as clear and there are different opinions on the exact details of each phase.[6] The ICH E8 guidelines suggest that classifying trials by study objectives could be preferred and that, ideally, information from small early studies in drug development should be used to support and plan larger and more definitive studies.[7]

Before clinical research can begin, there is a preclinical phase of testing the treatment in vitro and in vivo on animal subjects which allows for a preliminary assessment of the safety of a treatment for the proposed investigation in humans.[7] Early phase clinical research begins with first-in-man Phase I studies which are designed to address the initial safety, PK and PD of a treatment. These are usually performed in healthy volunteers, unless the test treatment is likely to be unsafe for healthy subjects, such as in oncology studies. Phase II studies seek to assess whether the treatment may be potentially efficacious in a larger group of subjects with the condition to be treated, providing a threshold to be crossed when moving to Phase III, though not with enough certainty to provide definitive conclusions. Additionally, Phase II studies may inform the design of future Phase III studies, such as the identification of an appropriate dose range. Phase III studies are large confirmatory studies which aim to confirm whether a treatment can be considered efficacious and intend to provide adequate evidence in favour of the treatment being licensed for use. Phase IV studies take place after the treatment has been approved and marketed and explore the use of the treatment in practice. Safety is an important aspect and is monitored throughout all phases of clinical research. Seamless Phase II/III studies aim to combine Phase II and III studies into one trial consisting of two stages: the first stage focuses on treatment, dose or subgroup selection; the second stage then aims to definitively compare the selected treatment/dose/subgroup with control.[8,9]

Clinical trials are specifically designed such that if a conclusion of a causal treatment effect is claimed there is confidence in the results. One such design choice is for the clinical trials to be controlled. This is where a separate group of participants do not receive the test treatment but instead a placebo or active control (such as the current standard of care) for comparison to the test treatment. Participants may be randomised to test treatment or control in an attempt to remove bias associated with treatment allocation and to enable causal inferences to be

made. If participants are not made aware of what treatment they have been assigned, then the trial is said to be blinded; this is used in an attempt to reduce the bias associated with the placebo effect. Researchers may also be blinded from treatment allocation in order to eliminate the risk of bias from any preferential care which may arise or blinded from the assessment of outcomes to reduce chance of bias in estimating treatment effects.[10] There are often situations where blinding is not possible, for example where treatment involves physical therapy, surgery or qualitatively different interventions. The gold standard for clinical trials are randomised controlled trials (RCT) which allow for causal inferences to be made regarding the effects of a treatment relative to control.[11]

Clinical trials are designed in such a way that inferences made regarding a parameter, say θ (e.g. a treatment effect), can be considered reliable whilst looking for a clinically meaningful result. For frequentist studies, they are designed to control type I and II error rates at chosen levels whilst testing a null hypothesis, H_0 (for example, no treatment effect, $\theta = 0$) against an alternative hypothesis, H_1 (e.g. positive treatment effect exists, $\theta > 0$, or some treatment effect exists, $\theta \neq 0$). The type I error rate, α, is the probability of rejecting H_0 given that it is true. The type II error rate, β, is the probability of failing to reject H_0 when it is in fact false; the power of a trial is $1 - \beta$. When a hypothesis test is performed, a p-value and confidence interval for the parameter of interest can be calculated. The p-value is the probability of obtaining a test statistic at least as extreme as the result observed, under the null hypothesis; that is, how much probability is contained in the tail(s) of the distribution of the test statistic under H_0, beyond the test result observed. If the alternative hypothesis is only in one direction from H_0 (as with the $\theta > 0$ example above), a one-sided test is performed, otherwise a two-sided test is performed and both tails of the test statistic distribution must be considered. When this p-value is less than the significance level of the test, the result of the hypothesis test is said to be statistically significant. A $(1 - \alpha)100\%$ confidence

interval is defined such that if one were to repeat an experiment according to some protocol many times, $(1-\alpha)100\%$ of the confidence intervals calculated according to the same method would contain the true value of the parameter of interest.

In contrast to the frequentist approach, in a Bayesian study model parameters are also considered to be random variables and suitable prior distributions for the parameters of interest (such as treatment effects) are updated with accumulated data from the clinical trial using Bayes Theorem to produce posterior distributions for parameters.[12] These posterior distributions can then be used to make probability statements about the parameters of interest, e.g. the true value of the treatment effect has a 95% probability of lying between a and b. Such an interval is called a credibility interval, and has a much more simple interpretation than the frequentist confidence interval. Choice of prior distributions for parameters in a Bayesian study is an important step. Options include specifying priors as: non-informative (such as flat priors placing equal weight on all possibilities); vague/diffuse (to provide minimal information on specific parameters and have little influence on the posterior); operational (chosen to ensure certain operating characteristics under different scenarios); or informative (based on existing data or expert opinion). For informative priors we discuss the use of expert opinion in Section 1.6 and the paper presented in Chapter 3 provides details of some approaches where prior distributions are formed from existing data. Many Bayesian models can be complex and analytically intractable, meaning that explicit evaluation of posterior distributions is not possible and numeric approaches to sample from the posterior must be taken, for example, using Markov chain Monte Carlo (MCMC) methods. Modern statistical computing software such as OpenBUGS[13] and Stan[14] (e.g. using the 'RStan' package[15] run through R[16]) have made analysing complex models in a Bayesian framework much more simple. OpenBUGS performs Bayesian modelling using Gibbs sampling,[17] whilst RStan uses Hamiltonian Monte Carlo,[17] both are flexible and user-friendly approaches to implementing Bayesian inference.

When small populations would mean recruiting to both treatment and control arms would be difficult, existing data may be taken as evidence of the control to compare to a single arm trial. Eichler et al.[18] define a threshold crossing approach for situations where RCTs are not feasible. Defining the factual as the average outcome on an experimental treatment and the counterfactual as how patients would have responded on average if not given the new treatment (i.e. given nothing or another treatment known to be effective), Eichler et al. state that a RCT is the standard way to assess the factual over the counterfactual. Eichler et al. propose the following threshold crossing framework: use existing data to decide on an efficacy threshold that a new single arm trial must successfully cross for the new treatment to be deemed effective and a futility threshold that if not crossed would mean the new treatment is deemed ineffective, otherwise if between these thresholds the treatment would be deemed to be 'promising' with either a second single-arm trial (where success is only possible by exceeding the efficacy threshold) or a RCT (if feasible) being performed.[18]

In this thesis, emphasis will be on how to inform the development of medicines for children.

1.2 Paediatric clinical trials

Paediatric clinical trials are important for the development of safe and effective medicines for children. However, Bourgeois et al.[19] looked at the proportion of trials, which were performed in children, of drugs intended to treat diseases with a large burden in the paediatric population between 2006 and 2011 listed on ClinicalTrials.gov. The authors found that within the diseases considered, only 12% of trials were in paediatrics even though 59.9% of the disease burden was in the paediatric population. An important issue in paediatric trials is the potential

heterogeneity of the paediatric population.[20] Batchelor and Marriott[21] state that anatomical and physiological differences between younger and older children and adults can create differences for drug PK. Furthermore, there may be age-related changes in the safety profile and pharmacodynamics of a drug, although Stephenson[22] states that adults' and children's responses to many drugs have much in common. To capture this heterogeneity, the ICH E11 guideline[23] suggests one possible set of age groupings to categorise paediatric patients: preterm newborn infants, term newborn infants (0 to 27 days), infants and toddlers (28 days to 23 months), children (2 to 11 years), adolescents (12 to 16/18 years, dependent on region); though it is emphasised that any classification of the paediatric population into age categories is somewhat arbitrary and that there is considerable overlap in terms of physical, cognitive, and psychosocial developmental issues across those categories suggested.

Within paediatric trials there can be an issue of gaining consent, with parents and guardians having to give consent on behalf of the child in most cases; in most EU countries the legal age for independent consent in paediatric clinical trials is 18, below this often only assent is sought for differing age ranges.[24] Clearly, there are no healthy volunteers in paediatric clinical trials, so trials are always in the population with the condition to be treated. There is also a common perception that recruitment into paediatric trials will be challenging, with clinicians, parents and guardians being reluctant to expose children to experimental treatments.[25] However, recent research has shown that there is a willingness from both parents and practitioners to enter children into trials, so barriers to recruitment are perhaps not as great as perceived.[26] Funding of paediatric trials has also had challenges, with Bourgeois et al.[19] identifying that between 2006 and 2011 the primary funding sources for trials in the paediatric population were government and nonprofit organizations.

In 2006 the EU Paediatric Regulation (EC 1901/2006) came into effect, with the objective of motivating the development of new medicinal products in children aged 0 to 17 years, and to ensure that new medicines are appropriately licensed for use across the paediatric population.[27,28] Following this regulation, for all new medicines an agreed paediatric investigation plan (PIP) is required, with results of studies as described in the PIP needed for marketing authorisation. PIPs are defined as prospectively agreed documents stipulating how the development of a new medicine should proceed in children and outline all of the studies that are to be conducted, detailing the conditions, indications and age groups concerned. Proposals for PIPs are submitted to the Paediatric Committee (PDCO) of the European Medicines Agency (EMA) for consideration; the Paediatric Regulation states submission should be no later than completion of adult PK studies.[27] The PDCO is responsible for agreeing or refusing the PIP, with positive PIP opinion (adopted by the PDCO) summarising the binding elements of an agreed development plan. In order to get the approval to market a medicine, all applications must include the results of any studies described in the agreed PIP, unless the medicine is exempt because of a deferral or waiver.[29] Prior to the introduction of the EU Paediatric Regulation, it has been estimated that at least 50% of drugs prescribed for children have never been tested in the paediatric population.[28,30,31] However, Weda et al.[32] concluded that the introduction of the Paediatric Regulation (1901/2006/EC) does not seem to have led to a lower prevalence of off-label use, where they define off-label use as "intentional use of an authorised product not covered by the terms of its marketing authorisation". Based on data from 16 EU Member States, 32 studies of hospital based paediatric off-label drug prescription showed a range of 13-69% of investigated drugs being used off-label.[32]

Hampson et al.[33] reviewed 73 PIP opinions to explore strategies that were adopted to support dose recommendations in the paediatric population and suggested that there are opportunities for using Bayesian methods to quantify uncertainty regard-

ing extrapolation assumptions in paediatric medicine development.

1.3 Extrapolation

The European Medicines Agency (EMA) defines extrapolation as:

> 'extending information and conclusions available from studies in one or more subgroups of the patient population (source population(s)), or in related conditions or with related medicinal products, in order to make inferences for another subgroup of the population (target population), or condition or product, thus reducing the amount of, or general need for, additional information (types of studies, design modifications, number of patients required) needed to reach conclusions'.[34,35]

Examples of extrapolation in practice include extrapolating from historical data to predict drug effects in contemporary patients, extrapolating from one geographic region to another to predict clinical benefits, or extrapolating from adults to support licensing decisions in the paediatric population. In order for extrapolations to be appropriate, strong assumptions on similarity are required between the source and target populations; a big challenge is deciding whether and to what extent extrapolation is appropriate. The US FDA[1] and International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) E11[23] guidelines outline an algorithmic, assumptions based approach to deciding on the level of extrapolation appropriate in paediatric medicine development and for determining which data are needed to support paediatric licensing of a medicine. For a particular treatment and indication, the choice of the level of extrapolation depends on whether it is reasonable to make the following assumptions between adults and children: similar disease progression; similar response to the intervention; similar PK-PD relationships; and whether there is a PD response that can be used to predict efficacy. Depending on which of these assumptions

are deemed reasonable, extrapolations can range from no extrapolation, through partial extrapolation, to complete extrapolation. According to Dunne et al.,[36] for each level of extrapolation the evidence required in the paediatric population is as follows: for no extrapolation a full development programme is required; for partial extrapolation the evidence needed ranges from a single adequate, well-controlled trial to confirm efficacy to a PK-PD study to confirm response; for complete extrapolation only PK studies, to establish dosing, and safety studies are needed (in certain cases only safety data may be needed). Hampson et al.[33] describe this type of extrapolation, where uncertainty regarding extrapolation assumptions is not accommodated, as 'deterministic extrapolation'.

An alternative framework for the use of extrapolation in the development of medicines for paediatrics has more recently been proposed by the EMA.[34,35] This framework supports the use of quantitative methodology to help understand (in terms of disease, drug pharmacology and clinical response) how relevant existing source population (e.g. adults) information is to the target paediatric population and identify any important assumptions and uncertainties about the relation between dose, PK, PD and clinical efficacy which should be documented as an 'extrapolation concept'. Conditional on these assumptions, the question of whether clinical efficacy can be predicted in the paediatric population, from the source population, can be assessed and a specific extrapolation plan can be developed to address any identified gaps in knowledge (and also identify where large confidence exists, to avoid unnecessary duplication of information). This extrapolation plan would detail any trials and study objectives needed to fill in knowledge gaps and provide evidence which could validate the extrapolation concept. After any planned studies, if the extrapolation concept is deemed to be valid, relevant evidence from the source population and evidence generated in the paediatric population could contribute to regulatory decision making for marketing authorisation. In order to mitigate uncertainty and risk in any regulatory decisions made, addi-

tional post-authorisation data may need to be gathered.[35] Such an approach to extrapolation, where knowledge regarding extrapolation assumptions can be updated and verified, is referred to by Hampson et al.[33] as 'stochastic extrapolation'.

A benefit of extrapolation in the setting of paediatric trials is the potential to reduce the number and size of studies required to demonstrate efficacy of a new medicine in children. This is important as it may be difficult to recruit children into clinical trials, there may not be many children meeting inclusion criteria for recruitment (e.g. having already been exposed to the drug off-label) and there may be constraints on the number and type of clinical measurements which can be taken from children. Additionally, there is a tradeoff between risk and benefit for testing medicines in children: if a highly effective treatment already exists there may be a reluctance to expose children to a potentially harmful new treatment; however, if there is currently an unmet medical need in children, it would be considered ethical to trial a drug without proven efficacy, even with the potential for adverse events. Conversely, the consequences of extrapolating when the assumptions do not hold could include exposing children to an ineffective medicine or exposing children to an unacceptably toxic dose.

Examples of extrapolation in practice can be seen in the cases of Zmax (azithromycin extended-release) for treating Community-Acquired Pneumonia, where use in the paediatric population (6 months and older) is based on extrapolation of adult efficacy data with additional safety and PK data in paediatric patients;[37] and oxcarbazepine (an anti epileptic drug), where the efficacy data from trials of adjunctive therapy (the test treatment in conjunction with another treatment) were used to inform the approval of oxcarbazepine as monotherapy.[38] In patients with focal epilepsies, adjunctive therapy data from adults and paediatrics were used to support the extrapolation of efficacy data from adults on oxcarbazepine monotherapy to paediatric patients in order to gain FDA approval. However, it should be

noted that this extrapolation was made with the knowledge that oxcarbazepine was already approved as monotherapy for paediatric patients in the EU.

1.4 Epilepsy

Much of the work in this thesis will be placed in the context of epilepsy clinical trials as there is a consensus of expert groups that extrapolation of efficacy from adults to younger children can be appropriate for focal epilepsies. Epilepsy is defined as "a tendency to suffer recurrent epileptic seizures" and is one of the most common groups of neurological disorders.[39] Epilepsy may occur at any age, but most commonly starts in childhood and old age.[40] There are many types of epilepsy (called epilepsy syndromes) and many types of seizures. Seizures can be defined as "a disturbance of movement, feeling or consciousness occasioned by sudden, inappropriate and excessive electrical discharges in the grey matter of the brain".[41] Epileptic seizures are usually grouped into either generalised or focal (also called partial) seizures. A seizure is called a focal seizure if the discharge remains in one part of the brain; there are a range of possible focal seizures depending on the location of the discharge. Generalised seizures involve all parts of the brain. A primary generalised seizure involves all parts of the brain at the seizure's onset. The discharge of a focal seizure can also lead to a focal seizure with secondary generalisation, where the focal discharge can spread through the brain, initiating a generalised seizure discharge.

With regard to focal epilepsies, whilst there is evidence to suggest that differences between treatment effects in adults and children would be quantitative rather than qualitative[42] there has been some disagreement between expert groups regarding what age is acceptable to extrapolate from adults down to. French et al.,[43] a group of US experts, suggest 2 years of age could be appropriate, whilst an EMA paediatric epilepsy experts group meeting[44] suggested 4 years of age could be

acceptable. The FDA recently concluded that extrapolation of efficacy from adult to paediatric patients aged greater than 4 years of age with focal epilepsies is acceptable.[45]

1.5 Meta-analysis

Meta-analytic methods involve quantitatively synthesising evidence across multiple related but independent studies in order to make inference based on a body of relevant research and potentially allow for an increase in power to detect a treatment effect over all included studies.[46,47] The term 'meta-analysis' was first used by Glass[48] to mean "The statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings"; since then, more modern meta-analyses do not necessarily rely on a "large collection" of results, with Davey et al.[49] reporting that from 2011 and earlier, 36% of the 22453 meta-analyses listed in the Cochrane Database of Systematic Reviews were based on two studies, whilst 75% were based on five or fewer studies.

Meta-analyses may synthesise summary measures reported from completed studies in an aggregate data meta-analysis or make use of the original individual participant data (IPD) from each included study. An IPD meta-analysis may be performed in either a one-step or two-step approach, described by Riley et al.[50] as follows. In a one-step approach, the IPD over all studies are fitted together in one overall model, whilst ensuring that the clustering of each study is handled appropriately (e.g. by inclusion of study as a factor in the model[47]). In a two-step approach, first the IPD in each trial are analysed in an appropriate manner producing summary measures for each included study, with the second step being an aggregate data meta-analysis performed using these aggregate data. Of course, if the original study results can be exactly reproduced from the IPD in this first step, an aggregate data meta-analysis using reported summary results and the results

of a two-step IPD meta-analysis are equivalent. There are many benefits to IPD meta-analyses such as greater consistency in analysis across studies (inclusion / exclusion criteria, analysis methodology, summary measures used, consistency in handling missing data); potential verification of original study results and modelling assumptions; possible long-term follow-up results may be available for inclusion; and publication bias can be reduced by inclusion of unpublished studies, where available.[50] The primary disadvantage of IPD meta-analyses is how heavily resource intensive they can be; expert statistical input will likely be required for the more advanced meta-analytic methods used, whilst obtaining, understanding and cleaning each set of IPD can take a considerable amount of time and contact with original study investigators, with no guarantee that the original data will always be completely available.[50] A disadvantage of any meta-analysis is that if included studies are not of high quality then any biases contributing from individual studies will result in inferior overall results. Though, there are approaches available for measuring (and adjusting for) the risk of bias in studies to be included in a meta-analysis.[51,52]

When performing a meta-analysis there are two ways to model between-study variability: fixed-effects, where study-specific treatment effects are considered to be the same across all studies; and random-effects, where study-specific treatment effects are allowed to be different between studies, allowing for the incorporation of between-study heterogeneity.[47] Focusing on the case of aggregate data for simplicity of presentation, a fixed-effects meta-analysis can be described as follows. Let the parameter of interest, θ, be the overall population treatment effect and summary measures Y_j, for $j = 1, \ldots, H$ trials, be estimates of this treatment effect. The Y_j can be modelled as:

$$Y_j = \theta + \epsilon_j,$$

where the ϵ_j's are random error terms and follow a normal distribution with mean 0 and variance σ_j^2; typically the variances σ_j^2 are taken to be the estimated variance of the Y_j summary measure, $s_j^2 = \widehat{\text{var}}(Y_j)$. As such, the Y_j can be assumed to be taken from a distribution:

$$Y_j \sim N(\theta, s_j^2),$$

with mean θ;[46,47] the Y_j and s_j^2 are summaries that could be calculated in a two-step IPD meta-analysis.

For a random-effects meta-analysis, summary measures Y_j are taken to be estimates of study-specific treatment effects θ_j and can be modelled as realisations of a study-specific distribution:

$$Y_j \sim N(\theta_j, s_j^2).$$

That is, the random-effects model assumes that treatment effects in each study can be different and that these θ_j effects are themselves realisations from a population distribution:

$$\theta_j \sim N(\theta, \tau^2),$$

with mean θ, the overall treatment effect, and τ^2 which measures the between-study variance. As such, summary measures Y_j can be modelled as:

$$Y_j = \theta + \xi_j + \epsilon_j,$$

where the ϵ_j terms are defined as in the fixed-effects approach and model the variability within study, and the ξ_j are random error terms which are normally distributed with mean 0 and variance τ^2 and model the variability between studies.

The between-trial heterogeneity term τ^2 is often unknown and will be estimated as part of the random-effects meta-analysis.[46,47]

Deciding between fixed and random effects meta-analyses can be based on a prior understanding of the underlying population. For example, if studies included in the meta-analysis were spaced apart in time, standard of care could have changed slightly resulting in each study giving a different (though still relevant) estimate of treatment effect. Alternatively, a test of whether there exists significant between-study heterogeneity can also be performed. One such approach is to consider Cochran's Q statistic, which can be used to test for heterogeneity of study treatment effects.[46,53] The Q statistic is found as a weighted sum of squared deviations between the summary measures in each trial and the fixed-effect overall treatment effect estimate:

$$Q = \sum_{j=1}^{H} w_j (Y_j - \widehat{\theta})^2,$$

where $\widehat{\theta} = \frac{\sum_{j=1}^{H} w_j Y_j}{\sum_{j=1}^{H} w_j}$ is an estimate of the overall treatment effect and the weights, $w_j = 1/s_j^2$, are the reciprocal of the estimated variances of the Y_j summary measures. Under the null hypothesis of no between-trial heterogeneity this Q statistic is approximately Chi-square distributed with $H - 1$ degrees of freedom, so at a specified significance level a standard hypothesis test can be performed to test for heterogeneity between trials.[46,54] However, Hardy and Thompson[54] show that using Cochran's Q as a test for heterogeneity has low power, especially when one study in the meta-analysis contributes a large proportion of the total information of all studies and do not recommend basing the choice of a fixed or random effects meta-analysis solely on such tests of heterogeneity, instead suggesting clinical insight may be more relevant.

In cases where there may be several potentially correlated variables to be modelled,

a multivariate meta-analysis is possible.[55] In the bivariate random-effects case the following illustrates this idea. Summary measures $Y_{1,j}$ and $Y_{2,j}$ are taken to be estimates of study-specific treatment effects $\theta_{1,j}$ and $\theta_{2,j}$, respectively, and can be modelled as realisations of a study-specific bivariate normal distribution:

$$\begin{pmatrix} Y_{1,j} \\ Y_{2,j} \end{pmatrix} \sim N \left(\begin{pmatrix} \theta_{1,j} \\ \theta_{2,j} \end{pmatrix}, \begin{pmatrix} s_{1,j}^2 & \rho_j s_{1,j} s_{2,j} \\ \rho_j s_{1,j} s_{2,j} & s_{2,j}^2 \end{pmatrix} \right),$$

where $s_{1,j}^2$ and $s_{2,j}^2$ are estimated variances of the $Y_{1,j}$ and $Y_{2,j}$ summary measures and ρ_j is the within-study correlation.[56] The $\theta_{1,j}$ and $\theta_{2,j}$ treatment effects are themselves realisations from a population bivariate normal distribution:

$$\begin{pmatrix} \theta_{1,j} \\ \theta_{2,j} \end{pmatrix} \sim N \left(\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \tau_{12} \\ \tau_{12} & \tau_2^2 \end{pmatrix} \right),$$

This multivariate meta-analysis allows borrowing of information between the $\theta_{1,j}$ and $\theta_{2,j}$ parameters and is used in this thesis in Chapter 6 in a Bayesian framework.

Compared to the frequentist approaches already discussed in this section, Bayesian methods differ in that the model parameters, such as the between-study variance, are also considered to be random variables and must be assigned prior distributions. One benefit of the Bayesian approach is the potential to approximate the information contained in the existing studies of a meta-analysis as a prior distribution, which could be used for the design or analysis of a future related study, see for example the meta-analytic predictive priors of Schmidli et al.[57] Sutton and Abrams[58] describe several advantages and disadvantages of the Bayesian approach being used in a meta-analysis. Disadvantages suggested include sensitivity to the choice of prior distributions, subjectivity involved when using informative priors and computational complexity, all of which are issues in most Bayesian analyses. Advantages suggested by Sutton and Abrams[58] include the fact that parameter

uncertainty can be accounted for in the analysis, the potential for easier extension to more complex models and the ability to incorporate external evidence (including expert opinion) in informative prior distributions.

1.6 Elicitation of expert opinion

The elicitation of expert prior opinion is the process of constructing a probability distribution that represents the knowledge and uncertainty extracted from an expert regarding one or more unknown quantities, such as a probability or treatment effect. Elicitation of expert opinion can be a useful way to quantify the state of knowledge about an unknown quantity prior to data collection. For rare diseases or small populations, where a conventional, well powered frequentist trial may not be feasible, elicited expert prior distributions can be used in the design and analysis of a future Bayesian clinical trial to improve the understanding of a treatment.

Based on a thorough review of literature, O'Hagan et al.[59] suggest the following steps to a model for the whole elicitation process: (a) background and preparation, including identifying quantities of interest and planning the elicitation session; (b) identifying and recruiting experts; (c) motivating and training experts, including explaining probability, probability distributions, common rules and instinctive processes people use when forming judgments (heuristics, such as anchor-and-adjustment), possible biases and trialling practice elicitation questions; (d) structuring and decomposition, including considering any dependencies and reviewing the available evidence base with experts; and (e) the elicitation itself, eliciting specific summaries from experts, constructing the probability distribution to represent these summaries, assessing adequacy and having the expert iteratively update if needed. Of course, many alternate approaches are possible.[59] As mentioned in (c), when people need to provide judgments where uncertainty is involved there are several simple strategies that are often employed called heuristics.[59,60] For

example, the anchor-and-adjustment heuristic is where a judgment is based on the adjustment from an initial starting value, called an anchor. This can be an issue as when making such quantitative judgments, experts often stay too close to the anchor value and fail to adjust far enough in either direction. It is important to keep such heuristics in mind when developing elicitation schemes to try to at least reduce the risk of associated biases.

The aim of prior elicitation is to quantify expert opinion as a useful probability distribution, X. When eliciting opinion, the choice of probability distribution used to quantify expert knowledge will depend on the unknown quantity of interest. For example, opinion on a probability may be appropriately modelled by a beta distribution, whilst opinion on a continuous variable which can take any real number could be modelled by a normal distribution. Clearly, if the expert opinion is to be quantified by a continuous distribution, it is impossible to elicit their opinion for the infinite number of probabilities contained within that distribution. Instead, several probabilities or specific summary measures of the distribution can be elicited and the probability distribution can be inferred from these elicited values.[59] Specific summaries of the distribution may include location measures (e.g. mean, median or mode) and measures of variability (e.g. standard deviation). However, measures of variability are difficult to elicit directly,[59] instead the spread of the distribution is often inferred by quantiles or credibility intervals, which can be used to give a value for the desired parameter by taking into consideration the mathematical properties of the distribution in question.

The bisection method begins by the statistical facilitator eliciting from the expert the median of X, that is, the value $x_{0.5}$ such that $P(X < x_{0.5}) = 0.5$; an expert may be asked to give a value such that X is equally likely to be less than or greater than the value given.[59] Following this, the expert is then asked to bisect the distribution above and below the median, i.e. provide quantiles $x_{0.25}$ and $x_{0.75}$ such

that $P(X < x_{0.25}) = 0.25$ and $P(X < x_{0.75}) = 0.75$. For a normal distribution this provides more points along the distribution than the number of parameters, meaning that unless $x_{0.25}$ and $x_{0.75}$ are symmetrical around $x_{0.5}$, a perfect fit to the expert's values will not be possible. The minimum number of questions needing answers to be elicited is equal to the number of parameters in the chosen probability distribution. Eliciting more quantities than the number of parameters means that it is unlikely a set of parameter values will be found that gives a distribution matching the elicited opinion exactly, instead the parameters of the distribution could be chosen in an optimal way; for the normal case, if eliciting more than one quantile for the measure of variability, the variance parameter could be chosen to minimise the sum of the absolute or squared differences between elicited and fitted quantiles of the distribution. Compared to absolute differences, squared differences give larger weight to large differences when optimising. For particularly complex elicitation problems, eliciting the answers to more questions will help with stability when fitting the elicited distribution. Additionally, the redundant information can be useful for assessing modelling assumptions; if the quantiles are dramatically asymmetric, and one is confident that the expert has indeed understood the questions, a symmetric distribution may not appropriately capture expert belief.

An alternative method could be to elicit X as a histogram.[61,62] In this approach, X is split into bins and the expert is asked to place a certain number of 'chips' (representing probability) amongst the bins to represent their belief regarding the probability distribution X. Placing all chips in one bin suggests that the expert is completely certain that X lies within that bin and placing an equal number of chips in every bin would suggest complete uncertainty regarding X.[62] With a small number of chips it may be difficult to adequately reflect X, for example, an expert may want to spread at least some probability across all bins to properly reflect their uncertainty, yet still focus more chips in bins where they wish to place most probability; a small number of chips may make this challenging. However, if

an expert finds it difficult to adequately assign their probability to bins, a small number of chips may make the task more manageable to reflect their opinion. The elicited histogram could then be used to approximate a parametric distribution for X, for example, by considering quantiles of the histogram.

When eliciting opinion from a group of experts there are several approaches to combining opinion in to one overall prior distribution.[59] One approach is to elicit individual prior distributions from each expert and mathematically combine them, for example, as a weighted sum. Such a weighted sum could either assign equal weight to all experts or could weight individual experts differently, say by clinical experience, though such weightings would be highly subjective. Alternatively, a consensus opinion over all experts could be quantified by having experts give their individual opinions, then join together to agree on one final prior distribution through constructive and open discussion; care must be made by the statistical facilitator to ensure that experts equally share opinion, come to one agreed prior and no one expert dominates the discussion.[63] Advantages of this type of behavioural aggregation include the automatic averaging over all experts opinion; drawing on every experts opinion at once to form one prior without the need for a complex mathematical aggregation model; and all experts must agree on a final prior distribution (compared to mathematical aggregation, where a weighted combination of priors could result in a distribution no expert agrees with). However, the statistical facilitators job is more challenging, strong personalities could heavily influence the final results, more reserved personalities may not contribute their valuable experience and it may not be clear how to proceed if experts are unable to agree on a final prior.

An example of a user-friendly software for eliciting probability distributions is contained in the Sheffield Elicitation Framework (SHELF).[64] SHELF is a formal procedure comprising of advice and tools for a facilitator with expertise in prior

elicitation to elicit expert opinion on an uncertain quantity and to quantify this opinion as a probability distribution. SHELF offers templates to structure and record elicitation sessions (called SHELF workshops) and an R[16] package which allows interactive elicitation, fitting and visualisation of expert opinion as probability distributions; the approaches to prior elicitation can either be performed in an interactive Shiny[65] application or by the facilitator manually inputting elicited quantities. SHELF allows the use of several methods to elicit opinion: quartile method (bisection approach described earlier); tertile approach (similar to bisection, though rather than splitting the distribution into four parts with equal probability, the distribution is split into three parts); and the roulette method (the histogram approach described earlier). SHELF proposes first eliciting opinion from individuals, then having experts regroup to share their expertise and opinions, before eliciting a group consensus distribution. To answer the issue of experts not all coming to a perfect agreement regarding the consensus prior distribution, SHELF asks the experts to consider the opinion of a rational impartial observer (RIO); after the RIO has seen all individual expert opinion and heard all discussions, experts must agree on a prior distribution that the RIO might reasonably believe to be true about the uncertain quantity of interest.

There are numerous examples of prior elicitation in practice across many fields, such as engineering, physics, psychology, agriculture, economics and medicine.[59,61,66] One example of elicitation being used in a clinical trials context is that of the MYPAN study (mycophenolate mofetil for childhood polyarteritis nodosa) where expert prior opinion was elicited for a future Bayesian RCT testing mycophenolate mofetil (MMF) against cyclophosphamide (CYC) for the treatment of children aged 4–18 years old with polyarteritis nodosa (PAN).[67,68] A Bayesian trial design was chosen to improve understanding about treatments for PAN, as recruitment for a definitive frequentist trial would not have been feasible; a frequentist non-inferiority trial with 90% power, 2.5% one-sided significance level and remission

rates on both treatments assumed to be 70% would have required 513 patients on each treatment arm, with previous studies of PAN suggesting recruitment would have taken over 30 years.[68] The Bayesian approach taken in the MYPAN trial began by first eliciting expert opinion on the 6-month remission rate on CYC (p_C) and the log-odd ratio, θ, between the 6-month remission rate on PAN (p_E) and p_C. Expert opinion on p_C and θ was elicited by asking six questions about different probabilities and proportions, marking answers on a visual analogue scale ranging from 0 to 1, with answers rounded to the nearest 0.05 probability. The prior for p_C was modelled as a beta distribution and experts were asked questions to establish the mode and lower quartile to infer the distribution. The prior for θ was taken to be a normal distribution, asking experts questions to establish the prior probability that $p_E > p_C$ and $p_E - p_C < -0.1$; answers to these questions were used to infer values for the mean and variance of the prior distribution for θ. Redundant questions regarding p_E were also asked in order to assess goodness of fit of the model and the consistency of expert opinion. Individual experts had plots of the probability density functions for p_C and p_E, fitted based on their elicited opinion, presented to them and were allowed to make changes to previously answered questions until they felt the plots represented their prior belief. All experts were then brought together to discuss their individual opinions and had a further opportunity to revise their answers. By taking the means and medians of the expert's final answers to the elicitation questions, a set of consensus prior distributions were determined that all experts agreed upon. Further to this, Hampson et al.[67] elicited expert opinion regarding the relevance of a related trial of MMF and CYC treating a different (but related) condition to PAN. The 6-month remission rates for both treatments between both study populations were related by log-odds ratios λ_C (for CYC) and λ_E (for MMF). These λ_C and λ_E parameters were used to measure the difference in treatment effect of CYC and MMF, respectively, between the trials. Expert opinion regarding probability distributions for these log-odds ratios was elicited in a similar way to the previous log-odds prior. Again, experts gave

individual opinions to begin with and then came together to reach a consensus opinion and agree on a single set of answers to the elicitation questions, giving consensus prior distributions. Existing data from the relevant trial were then revealed, updating the prior probability densities for p_C, p_E and θ to be shared with the experts. Ultimately, the experts agreed on these updated prior distributions (incorporating the relevant trial data) as the consensus prior distribution to be used for the Bayesian trial.[67] This example highlights the complexity of expert opinion that can be elicited and the worth of eliciting opinion in rare diseases or small populations.

Chapter 2

Thesis Summary

This thesis aims to explore extrapolation in paediatric medicine development and consists of four substantive pieces of work: a systematic review of methods for extrapolating between populations; conclusions of a focus group of epilepsy experts eliciting their opinions regarding extrapolation in paediatrics; a quantitative framework to inform extrapolation decisions in younger children using existing data from trials of adults and adolescents; and approaches to quantify how parameters of an E-R model vary over age to derive dosing rules.

Chapter 3 contains the systematic review paper entitled "Extrapolation of efficacy and other data to support the development of new medicines for children: A systematic review of methods".[69]

Chapter 4 gives some background to the focus group of epilepsy experts. This leads to the second published paper included in this thesis which is contained in Chapter 5, entitled "Clinical drug development in epilepsy revisited: A proposal for a new paradigm streamlined using extrapolation".[70]

Chapter 6 contains the third piece of work entitled "A quantitative framework to inform extrapolation decisions in children". This chapter details a framework using existing data from trials of adults and adolescents, along with expert opinion on external biases, to quantify prior uncertainty regarding the similarity of PK-PD

relationships between adults and younger children.

The final piece of work contained in this thesis is given in Chapter 7 and is entitled "Exposure-response modelling approaches for determining optimal dosing rules in children". This chapter considers approaches to quantify how parameters of a PK-PD model vary over a continuous age range and, given this, an approach for deriving optimal dosing rules which account for pharmacological differences between paediatric age groups.

Finally, Chapter 8 ends with a discussion overview of the thesis, considering some limitations and potential future work.

Chapter 3

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

3.1 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'.[34] Extrapolating from existing data, also commonly referred to as bridging or borrowing strength, is common in drug development.[36,71] Examples include incorporating historical data into the analysis of contemporary clinical trials[72–74] and, more controversially, using information on a drug's short-term effect to draw conclusions about its long-term effect.[75] 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.[76]

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.[1,23,36,77] 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.[25] 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.[26] Dunne et al.[36] discuss the paediatric study decision tree[1,23] shown in Figure 3.1, 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[23] 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[21] 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[22] notes that adults' and children's responses to many drugs have much in common. The European Medicines Agency (EMA)[34] 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

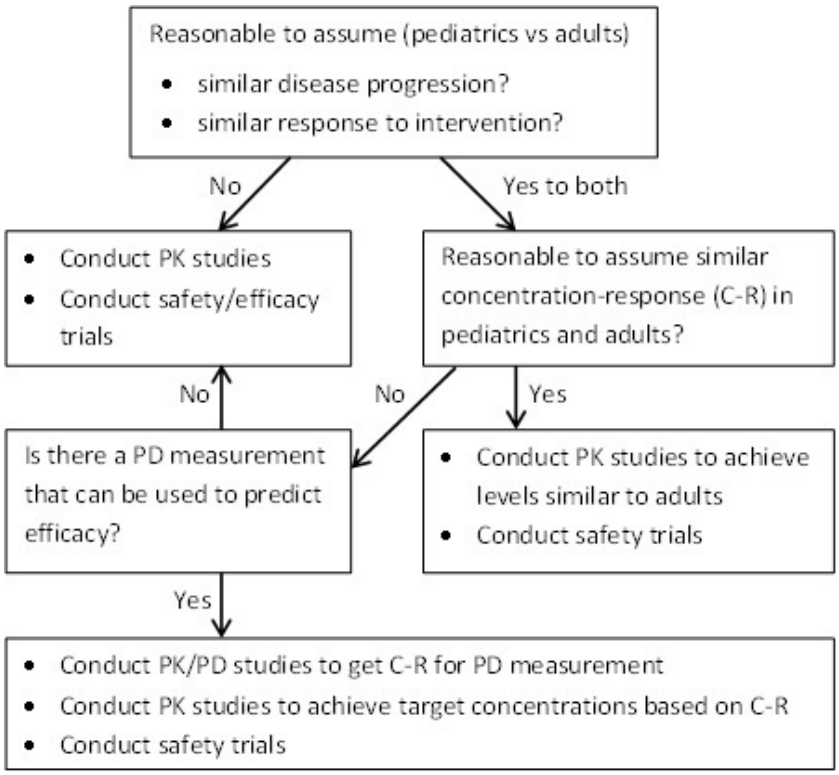


Figure 3.1: Paediatric Study Decision Tree: Image reproduced from Food and Drug Administration[1]

data will be generated in the target population, and these data should, in turn, be used to verify the extrapolation concept.

This chapter 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 chapter proceeds as follows. Section 3.2 outlines the strategy used to identify relevant papers and methods which are briefly summarised in Section 3.3.

In Section 3.4, we give a detailed account of the methods found, grouped according to four common approaches. We conclude in Section 3.5 with a discussion of the suitability of these methods for making extrapolations in paediatric drug development.

3.2 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 categories: 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 (Appendix A.1). 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 (Appendix A.2) 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.

3.3 Results

Searches identified 52 papers satisfying the stated inclusion/exclusion criteria as summarised in Figure 3.2, 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 3.3. All

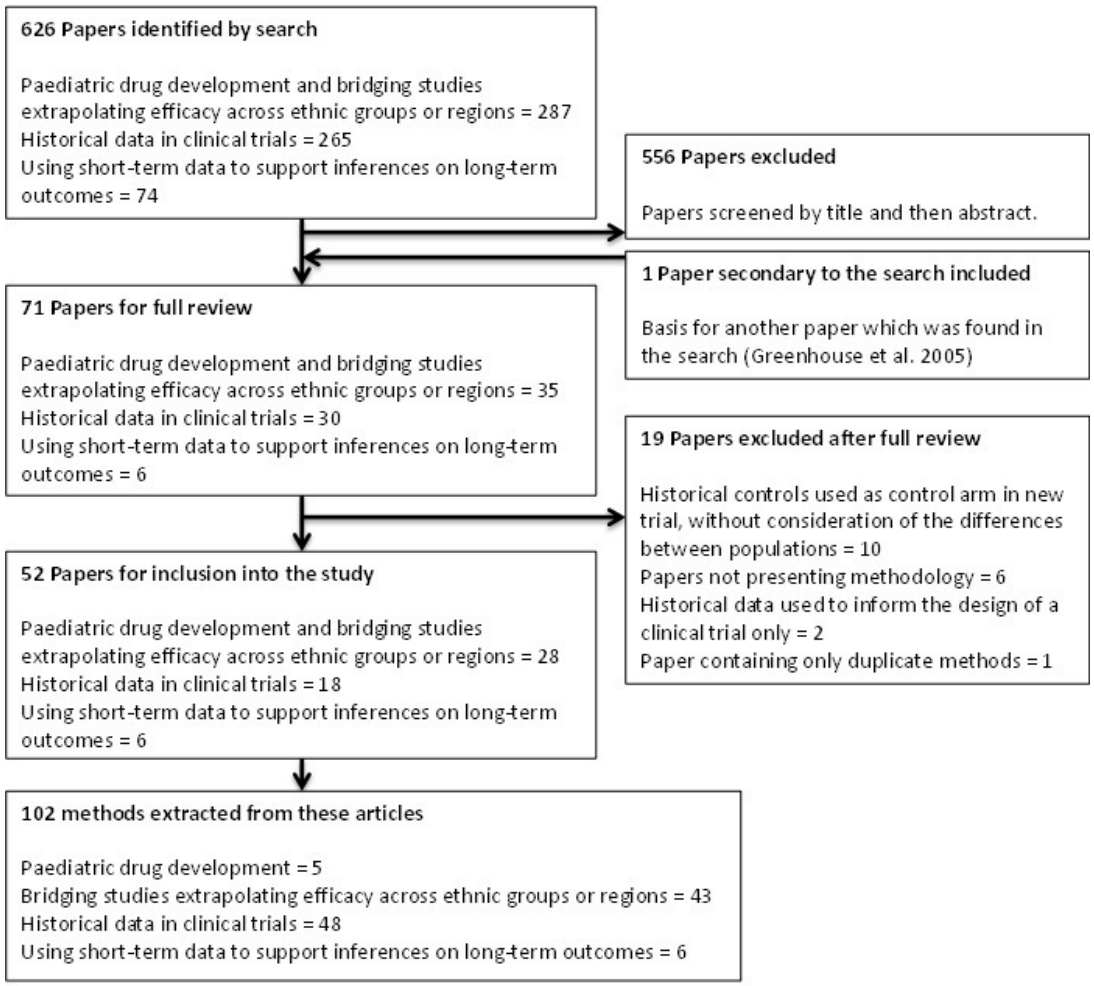


Figure 3.2: Flow diagram of systematic review results

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).

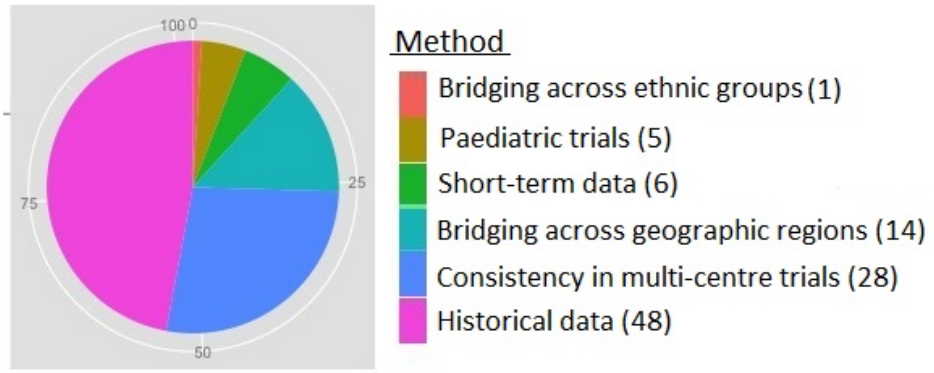


Figure 3.3: Plot showing distribution of methods across four main areas

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.[38] suggest that a necessary condition for using adult efficacy data to support conclusions about the efficacy of oxcarbazepine 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.

3.4 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 approaches, 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).

3.4.1 Bayesian methods

Searches identified 58 Bayesian methods from 25 papers.[72–74,78–99] Of these, 54 methods[72–74,78–95] sought to create an informative prior for θ_T while four[96–99] assessed the consistency of treatment effects or PK responses between the source and target populations.

3.4.1.1 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[72–74,78–85] 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.[72,73,78,79] 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 is:[72]

$$\pi^{PP}(\theta, a_0|x_S) \propto L(\theta|x_S)^{a_0}\pi_0(\theta)\pi(a_0). \tag{3.1}$$

The prior for a_0 captures prior uncertainty about the commensurability of parameters of the historical and contemporary data. Ibrahim and Chen[72] 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 equation (3.1) 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[72] extend π^{PP} in equation (3.1) 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 equation (3.1) violates the likelihood principle since it omits the normalising constant for a_0.[78,100] Modifying equation (3.1) 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

$$\pi^{MPP}(\theta, a_0|x_S) = C(a_0)L(\theta|x_S)^{a_0}\pi_0(\theta)\pi(a_0), \tag{3.2}$$

which Hobbs et al.[73] refer to as the modified power prior (MPP). Chen et al.[79] extend the MPP to accommodate several historical datasets, as well as binary and normally distributed data. Hobbs et al.[73] modify the MPP in equation (3.2)

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

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

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.[73] 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

$$\pi^{CP}(\theta_T, \tau | x_S, \theta_S) \propto L(\theta_S|x_S) \times \sqrt{\tau}\phi((\theta_T - \theta_S)\sqrt{\tau})\,\pi_0(\theta_T)\,\pi(\tau). \qquad (3.3)$$

Once data from the new trial become available, the posterior density for (θ_T, τ) given x_T and x_S is proportional to equation (3.3) 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.[80,81] suggest defining $\pi(\tau)$ in (3.3)

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).[80] Hobbs et al.[81] use the CP[73] 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.[80] illustrate this approach with applications to binary, survival and count data.

Hobbs et al.[73] adapt the CP in equation (3.3) 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.[73] 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,[74,82] 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. The MAP prior is then updated using Bayes theorem when data from the new trial become available. We classify methods[74,82] as adaptive approaches to down-weighting data from the source population since the MAP prior can be approximated as a mixture of conjugate distributions[57] and have heavier tails than a simple conjugate prior. Thus, in the event of a prior-data conflict, the historical data will eventually be discarded from the posterior analysis of the new trial.

When deriving the MAP prior, meta-analytic 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.[74] assume parameter estimates are normally distributed with known standard errors $s_{S1}, \ldots, s_{SH}$. A Bayesian random-effects

meta-analytic model is:

$$\begin{aligned} X_{Sh} \mid \theta_{Sh} &\sim N(\theta_{Sh},\, s_{Sh}^2), \quad \text{for } h = 1, \ldots, H, \\ \theta_{S1}, \ldots, \theta_{SH}, \theta_T | \theta^\star, \nu^2 &\sim N(\theta^\star,\, \nu^2), \\ \theta^\star &\sim \pi(\theta^\star), \\ \nu^2 &\sim \pi(\nu^2). \end{aligned} \tag{3.4}$$

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.[74] recommend using priors for ν to check the sensitivity of conclusions in a fully Bayesian meta-analysis. Gsteiger et al.[82] 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.[79] 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.[80] state that when $H = 1$, there is a one-to-one relationship between the commensurability parameter τ in equation (3.3) and the between-study variance ν in model (3.4).

Cuffe[83] 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

$$\begin{gathered} X_S \mid \theta_S \sim N(\theta_S, \sigma^2/n_S) \quad \text{and} \quad X_T \mid \theta_T \sim N(\theta_T, \sigma^2/n_T), \\ \theta_S, \theta_T \mid \theta^\star \sim N(\theta^\star, \sigma^2/n_b), \\ \theta^\star \sim N(0, \sigma_1^2), \end{gathered} \tag{3.5}$$

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

$$\lim_{\sigma_1 \to \infty} \mathbb{E}(\theta_T | x_T, x_S) = \frac{n_b n_S}{2n_S n_T + n_b n_T + n_b n_S} x_S + \frac{2n_S n_T + n_b n_T}{2n_S n_T + n_b n_T + n_b n_S} x_T. \tag{3.6}$$

Model (3.5) 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

$$\hat{n}_b = (n_m/d_m) \max\{d_m - |x_T - x_S|, 0_+\}, \tag{3.7}$$

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 equation (3.7) 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 equation (3.6). 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.[101] 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[84,85] 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.[84] 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.[73] 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[72,73,85–93] 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 equation (3.1) is taken to be a fixed constant and Hobbs et al.[73] refer to this approach as the conditional power prior (CPP). Six methods[72,73,85–87] propose power priors with fixed a_0. Ibrahim and Chen[72] 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.[72,73] De Santis[86] 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[86] illustrates how the geometric prior can be used to inform early stopping decisions in a new Bayesian clinical trial. Rietbergen et al.[87] consider the CPP incorporating data from several historical studies, assigning data from each study a weight elicited from expert opinion. Gandhi et al.[85] 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.[73] also provide a variation on the commensurate prior described in the previous subsection which treats τ as fixed.

Schoenfeld et al.[88] 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 CPP 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.[88] 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.[89] 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.[89] 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.[90–93] For example, French et al.[92] 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.[93]

consider Bayesian sample size calculations, using historical placebo data to create an informative 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[85,91,92,94,95] 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.

3.4.1.2 Assessing consistency between source and target populations

Four Bayesian methods were proposed to assess the consistency of parameters in source and target populations.[96–99] Pei and Hughes[96] 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.[97] use Bayesian most plausible prediction[102] 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, $\hat{\boldsymbol{\theta}}_{S\setminus h}$ 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.[98] also use posterior predictive probabilities to assess the consistency of data from a bridging trial and reference studies. Gould et al.[99] 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.

3.4.2 Frequentist methods

Forty-four frequentist methods were identified[9,38,96–98,103–127] of which 11 methods[9,103–112] synthesised data from source and target populations in a joint model, three methods[113–115] combined data across populations through a weighted test statistic, and 30 methods[38,96–98,109–112,116–127] proposed criteria to assess the consistency of estimates of key parameters in different populations.

3.4.2.1 Joint model incorporating data from source and target populations

Five methods[9,103–106] proposed using short-term data to support inferences about a long-term endpoint assuming simple models to relate observations on differ-

ent 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[103] 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,[104] 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[105] 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[9] 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 distribu-

tion, fitted using the double regression method of Engel and Walstra.[128] Wüst and Kieser[106] 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[107–112] synthesize data from source and target populations using a frequentist random-effects model. Thall and Simon[107] combine historical and contemporary control data via a univariate random-effects meta-analysis while Arends et al.[108] model short-term and long-term outcomes from trials using a multivariate random effects model. Chen et al.[109] and Ko[110] use a random effects model to accommodate heterogeneity between regions and test for an overall treatment effect. Liu et al.[111] use a random effects model to test for similarity or non-inferiority between treatment effects in different regions. Ko[112] 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.

3.4.2.2 Combining data across populations in a weighted test-statistic

Three methods[113–115] 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.[113] 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.[98] 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.[113] suggest that this weight may be based on evidence of efficacy established in the original region.

3.4.2.3 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.[116] 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[53] 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[54] 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,[129] 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.[130] This statistic also depends on the within-centre precision[131] and the number of centres, s, such that under H_0, $\mathbb{E}(I^2) = -200/(s-3)$ if $s > 3$.[130] An alternative measure of consistency not found by this review but pointed out by a reviewer is $H^2 = Q/(s-1)$, which Higgins and Thompson[132] state does not intrinsically depend on the number of studies.

Global test statistics can also be used to test for a qualitative interaction between the treatment effect and trial regions. The Gail-Simon test[133] 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.[116] 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 $|\Delta_s - \Delta| > m$,[134] 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.[116] 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.[111] 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 $j = 1, \ldots, s$.[135] 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[97,109,110,112,117–124] 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.[120] 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.[125] 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[96] propose a frequentist version of their method described in Section 2.4.1.2 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.[126] 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.[127] 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 population 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.[38] 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.[98] apply the 'reproducibility probability' method[136] 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.

3.5 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 similarities. 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.[57] 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 3.1 it was noted that there may be differences between age groups of children. Twenty-five methods[38,72,80,116,122,127] 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.,[108] 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.[77] 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[137] 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.[138] 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[139] 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[79,82,88,93,97,99,107,111,113,115,124] 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[81,86,92,126] 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;[82,87,92,93,95,96,99] code was stated as available upon request from

the corresponding author of one method;[121] syntax for another method[78] was included in a related commentary article.[140] The strategy used to identify available software is described in Appendix A.3, while the results are listed in online Supplementary Appendix D which can be found at (`http://www.research.lancs.ac.uk/portal/en/publications/-(8911844e-2638-4dec-a844-8b842f034168).html`) and is described in more detail in Appendix A of this thesis.

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 3.2 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.

Chapter 4

Background to 'Clinical drug development in epilepsy revisited' paper

4.1 Introduction

After the systematic review presented in Chapter 3, we then sought to elicit the opinion of UK based epilepsy experts regarding the role and acceptability of extrapolation in paediatric epilepsy treatment research. This meeting placed our extrapolation research into a clear context and resulted in a publication in the journal CNS Drugs given in Chapter 5. In Chapter 5, we discuss the opinions of experts regarding extrapolation and propose a new paradigm for the development of epilepsy medicines streamlined using the partial extrapolation assumption which the experts felt appropriate for focal epilepsies. Section 4.2 below discusses the organisation of the epilepsy experts meeting.

4.2 One-day meeting outline

An initial plan for the meeting was laid out in the proposal for the NIHR grant (NIHR-RMOFS-2013-03-05): expert opinion regarding the role of extrapolation in paediatric epilepsy drug development would be elicited during a 1-day meeting. The plan was for the meeting to elicit opinion regarding perceived sources of heterogeneity in the paediatric population, such as between different age groups; the appropriateness of source populations (e.g. adults) from which efficacy data can be extrapolated to younger children; and the evidence base needed in such source populations to support extrapolations.

Experts were defined as neurologists with experience in adult and/or paediatric epilepsy and the aim was to recruit 6 to 8 experts who would be identified through the International League Against Epilepsy, Association of British Neurologists and the British Paediatric Neurology Association; our collaborator Dr. Graeme Sills, a Senior Lecturer in the Department of Molecular and Clinical Pharmacology at the University of Liverpool, was instrumental in identifying the experts to be invited and informally contacted each expert to gauge their interest. Initially, 12 experts were invited, with seven experts attending the meeting: two adult neurologists (Anthony G. Marson; Philip E. M. Smith) and five paediatric neurologists (Richard Appleton; J. Helen Cross; Tim Martland; Ailsa McLellan; Chris Rittey).

Before the meeting could take place, Dr. Lisa Hampson, as principal investigator for the project, had to gain ethical approval. This involved filling in an ethics self assessment form, confirming that any recordings taken would be deleted as soon as possible after transcription and ensuring that all experts read through and signed a consent form. Experts were sent this consent form prior to the meeting, along with a participant information sheet which provided details about the purpose of the meeting, what would be expected of experts, what data would be collected

and what they could do if they changed their mind about participating.

As the experts were based across the UK, the Manchester Meeting Place (part of the University of Manchester) was chosen as a central location and the one-day meeting was held on the 6th November 2014. The meeting was split into morning and afternoon sessions, and on arrival the experts were given an information booklet which contained the participant information sheet, a list of attendees, an agenda, and the morning and afternoon session slides.

During the morning session, a brief talk was given outlining the project work, project team, the aims of the meeting, and an explanation of extrapolation (definition; use; issues specific to paediatric epilepsy medicine development; examples of extrapolation). A discussion session was held in the afternoon where several questions were asked:

- Are paediatric patients just small adults?
- Are all paediatric patients the same?
- If a drug is safe in adults, is it safe in paediatrics?
- Can we extrapolate efficacy data from adjunctive therapy to monotherapy?
- When is it reasonable to use therapies in paediatrics that are licensed only for adults?

This discussion session was recorded and the company UK Transcription were hired to transcribe the recordings. Based on these discussions, an opinion paper was written which discussed the experts thoughts on the place of extrapolation in paediatric epilepsy medicine development and a new paradigm for having phase II and III trials recruit both adults and children aged 2 years and above, justified by the experts belief about an assumption of partial extrapolation between adults and children being appropriate.

Chapter 5

Clinical drug development in epilepsy revisited: A proposal for a new paradigm streamlined using extrapolation

5.1 Introduction

The European Medicines Agency (EMA) defines extrapolation as '...extending information and conclusions available from studies in one or more subgroups of the patient population (source population)... to make inferences for another subgroup of the population (target population)...'[34,141] There are several examples of how this definition can be applied. Using the terminology of Dunne et al.,[36] extrapolation can range from complete (no additional data needed in the target population) to partial (supporting data needed) to none. Extrapolation can be used to streamline drug development. Avoiding unnecessary studies in populations whose response to therapy is well understood enables sponsors to focus research on patient groups about which least is known. This chapter considers how the extrapolation

of adult efficacy and safety data can be used to streamline the development of drugs for use in paediatric epilepsies.

Off-label prescribing in paediatrics is prevalent in the US[142] and EU.[143] In routine clinical practice, informal extrapolation from adult data increases the confidence of doctors and families about off-label prescribing in children. When developing new medicines, it is reasonable practice to extrapolate from adult data to predict the clinical benefits of a new medicine in paediatrics such that smaller trials may suffice to demonstrate efficacy in this age group. However, extrapolations only have value if robust assumptions on similarity hold when applied to the adult and paediatric populations. The US Food & Drug Administration (FDA)[1] and ICH E11[23] guidelines outline an algorithmic approach for determining which data are needed to support paediatric licensing of a medicine depending upon whether it is reasonable to assume that disease progression, drug pharmacology, and pharmacokinetic-pharmacodynamic (PK-PD) relationships are consistent across adults and paediatric patients. An alternative framework has recently been proposed which stipulates that emerging and cumulative data in the target population should be used to confirm extrapolation assumptions.[34,141]

In the context of epilepsy research, it is not always possible to predict clinical benefits in paediatric patients using adult data due to disparities in the different types (syndromes) of epilepsy and in their specific natural histories. The acceptability of extrapolation will depend on several factors, including age, seizure type and epilepsy syndrome, treatment regimen, and the individual antiepileptic drug (AED). Whilst there is broad agreement that efficacy in adults with focal epilepsies can be extrapolated to paediatric patients with focal epilepsies, there is disagreement about the boundary of certainty, with different expert groups supporting extrapolation down to the ages of either 2[43] or 4 years.[44] The FDA has recently suggested that complete extrapolation of efficacy from adult to paediatric patients

aged 4 years and older with partial onset seizures is acceptable.[45] This is a major development, and one that is consistent with our view, but the potential of extrapolation goes much further.

This chapter explores these issues and provides recommendations on the role of extrapolation in drug development for epilepsy and identifies opportunities to improve current practice. It reflects work conducted within a project funded by the National Institute for Health Research (UK) on extrapolation approaches in paediatric trials.

5.2 Considerations

5.2.1 Are paediatric patients just small adults?

In the case of common focal epilepsies, the answer may well be yes. The aetiology of extra-temporal focal epilepsy in both adults and children is predominated by vascular lesions, trauma, and, most frequently, cortical dysplasias. Dysplasias are present from birth and while the time to seizure onset may vary widely, the underlying pathology is the same which suggests that this is likely to reflect a single pathophysiological process independent of age.

All patients aged 2 years and above with focal epilepsy would be expected to respond similarly to drug treatment in terms of seizure frequency reduction, provided that dosing led to an equivalent serum concentration-time profile. Although there are some subtle differences in semiology of focal seizures in the youngest age groups (i.e. paucity of automatisms, predominance of bilateral motor signs, etc.), these rapidly disappear with age and there is no evidence that these seizure types are differentially responsive to first-line therapies for focal epilepsy.[144] Consequently, it should be possible to extrapolate efficacy data obtained in adults with focal

epilepsy to patients aged 2 years and above. It would be inappropriate, however, to extrapolate efficacy to patients below 2 years of age primarily because of greater variability in aetiology and difficulties in diagnosis.

While the natural history of epilepsies may differ between adults and paediatrics, any differences in treatment effect between adult and paediatric patients with focal epilepsies are likely to be quantitative rather than qualitative.[42,43,145,146] However, this does not obviate the continued need for trials of new AEDs in paediatrics, particularly in the case of the rarer epilepsy syndromes.

5.2.2 Are all paediatric patients the same?

For focal epilepsies, the older age groups proposed in the ICH E11 guidance (Table 1)[23] could in theory be merged to create a single group that encompasses children and adolescents aged 2 to 16/18 years. However, there would be less confidence regarding the younger age groups and discussions with neonatologists would be required.

There is no doubt that preterm and term infants are relatively under-investigated with minimal Phase I or randomised controlled trial (RCT) data. Extrapolation of efficacy data from adults or older paediatric patients to these groups is not possible because of differences in the pathophysiology of the epilepsy as well as brain biochemistry, brain development, and drug-metabolism. Drug-clearance is low in preterm and term newborn infants, subsequently increases rapidly until around 2 years, and then declines steadily until around 12 years at which point it is considered to have reached adult levels, such that adult dosing can be considered for adolescents aged $>$ 12 years;[147] this is well-illustrated by carbamazepine.[148] However, sufficient variability exists that PK studies are likely to be required to support dose choices for paediatric patients aged 2 to 12 years even when efficacy

is extrapolated.

In general, the behaviour of AEDs in patients aged 2 years and above is usually predictable. There is a need for more robust studies in patients under 2 years with both focal and generalised epilepsies. This is acknowledged to be challenging, especially for patients less than 1 month old in whom the study design would be critical.

5.2.3 If a drug is safe in adults, is it safe in paediatrics?

There are a variety of adverse outcomes associated with AED use, including those that are acute and dose-related, those that are chronic and exposure-related, and those that are idiosyncratic and likely to be immune-mediated. For the purposes of this article we group them all under the term "safety". Most safety issues are considered to be essentially similar in adults and paediatrics at equivalent doses. There is anecdotal evidence suggesting that some idiosyncratic reactions occur at differing frequencies in adults and children (i.e. lamotrigine-induced Stevens-Johnson syndrome) but this may simply reflect differences in drug disposition and in systemic exposure to the drugs or their reactive metabolites. Those aside, it is possible, with appropriate caution, to extrapolate most adult safety data to paediatric patients aged 2 years and above.

Important safety issues that are specific to paediatrics include effects on growth and on pubertal, motor, speech and language, and cognitive development. These paediatric safety signals cannot be reliably identified from an adult population. Effects on learning and on social and educational development are also important and in paediatric patients with severe epilepsies it may be difficult to distinguish the influence of the epilepsy and its underlying aetiology from the effects of the medications used to treat it. Nevertheless, improvements in attention, memory,

cognition, and behaviour can be observed during AED withdrawal in paediatric patients with challenging epilepsy, suggesting a strong influence of drug treatment.

Seizure aggravation is another important safety issue, particularly in rare idiopathic focal epilepsies that are typically diagnosed in childhood only. Standard treatments can occasionally exacerbate seizures in these children but their low prevalence in the focal epilepsy population means that they might evade detection in controlled trials of short duration. Inclusion of EEG follow-up in the Phase II/III trial protocol for paediatric participants would improve detection of these paradoxical effects.

5.2.4 When is it reasonable to use therapies in paediatrics that are licensed only for adults?

For drugs licensed for use in adults but not in paediatrics, this would depend on the clinical situation with a risk-benefit trade-off determining the acceptability of off-label prescribing. When prescribing off-label in paediatrics, a drug will often be tried initially in adolescents before then being used in younger patients.

There would likely be greater confidence to enter patients in clinical trials rather than prescribe an AED off-label, particularly because of the detailed monitoring performed within a trial. There is a clear need for paediatric RCTs to be conducted earlier than at present and in parallel or in conjunction with adult trials. This would incentivise the recruitment of children into trials since accrual can be challenging when a trial treatment licensed in adults is available off-label in children. Improving enrolment will improve the quality of paediatric RCTs since inadequate accrual currently obliges many trialists to recruit from small, inexperienced centres, increasing patient heterogeneity and the risk of internal biases. Earlier paediatric RCTs would also widen participation in trials to include children with refractory

epilepsy who are often excluded from new drug studies on the basis that they have already been prescribed the drug off-label after failing all other licensed medicines.

It is important to acknowledge that there may be paediatric-specific issues for any RCTs undertaken in the idiopathic focal epilepsies of childhood and particularly benign partial epilepsy with centro-temporal spikes (BECTS) and benign epilepsy of childhood with occipital paroxysms (BECOP; Panayiotopoulos syndrome). The natural history of BECTS (and probably BECOP), is such that a spontaneous remission may occur any time, including soon after its onset or diagnosis. Consequently, any apparent efficacy of a drug in RCT participants with BECTS or BECOP may be due to the drug itself or to the natural history of the syndrome. This might risk assay sensitivity in a non-inferiority trial but would be of less concern if the trial was designed to detect differences and found them.

5.2.5 Can we extrapolate efficacy data from adjunctive therapy to monotherapy?

Extensive trial data and clinical experience with existing AEDs has failed to find any instance where a drug behaves differently in terms of its spectrum of efficacy and adverse effects when administered alone or as adjunctive therapy, except in circumstances where drug interactions might be expected. Consequently, it would be reasonable to extrapolate efficacy data from adjunctive trials to inform the use of an AED as monotherapy.

Mintzer et al (2015) state that the need for separate monotherapy and adjunctive therapy licenses in epilepsy is "unnecessarily restrictive" and that AEDs should be approved for specific seizure types or epilepsy syndromes only.[149]

5.3 An alternative paradigm for developing medicines for focal epilepsies

This section outlines our proposal for future clinical development of drugs for focal epilepsies. This proposal uses a partial extrapolation of adult efficacy data, generating only supportive efficacy data in children aged 2 years and above, and a limited extrapolation of adult safety data to justify joint Phase II and III studies recruiting adult and paediatric patients aged 2 years and above.

5.3.1 Phase I trials

The primary purpose of these trials remains the identification of a safe range of doses of a new compound to be used in the subsequent clinical development programme. Such studies should continue to be undertaken in healthy male adults only in an effort to reduce variability, limit confounding influences and minimize the likelihood of unexpected adverse events.

5.3.2 Phase IIa and IIb trials

The primary purpose remains determination of the effective dose range and a preliminary assessment of safety and efficacy. Trials should be randomised, placebo-controlled, adjunctive therapy studies following current guidelines for adjunctive trials but now recruiting patients with focal epilepsy aged 2 years and above, obviating the current requirement for a separate development programme in paediatrics. Using partial extrapolation of adult efficacy data, power calculations should be based on the entire study population but the final analysis should be stratified by age. Long-term extension will allow provisional assessment of safety in adults, paediatrics or both. There would not be a requirement to complete the long-term extension before progressing to Phase III. PK investigations will reveal the dose-concentration relationship in adults, paediatrics or both. Wherever pos-

sible, PK data should be analysed using population-PK models to accommodate sparse sampling schedules. Inclusion of mandatory EEG follow-up for paediatric participants would allow detection of seizure aggravation.

5.3.3 Phase III trials

The primary purpose remains the identification of efficacy in comparison to placebo. Traditional approaches are appropriate; i.e. randomised and placebo-controlled trials of adjunctive therapy. Efficient adaptive[150] and/or Bayesian[151] strategies to the design and analysis of trials should be considered if appropriate. Studies should again recruit patients with focal epilepsy aged 2 years and above and should be powered to detect treatment effects based on the total sample size accumulated across adults and paediatrics but should also include potential for stratified analysis. Minimum sample sizes in each age group might be pre-specified to ensure that reliable (but not necessarily definitive) conclusions can be drawn from the paediatric data. If a significant treatment effect was demonstrated in adults but not in paediatric patients and the differences could be attributed to sample size alone, then the treatment would still be acceptable for paediatric use provided there were no qualitative differences in the effects between adults and paediatric patients.

Long-term extension will allow additional open-label assessment of safety and efficacy in adults, paediatrics or both. Further PK investigations and EEG follow-up may be required, particularly in paediatric patients.

5.3.4 Licensing of treatments

Under this new paradigm, since all pivotal trials would be conducted in both paediatrics and adults, licensing should also apply to all age groups from 2 years upwards. Licenses should be granted for a general indication of focal epilepsy,

allowing the discretionary use as either adjunctive therapy or monotherapy unless there is reason to impose a restriction. Approval for paediatric use (2 to 16/18 years) should be conditional on a prospective, time-limited commitment to collect safety data from paediatric patients on growth and on neurological and cognitive development. Ideally, these neuro-developmental safety data would be collected within a randomised, placebo-controlled design, but this is likely to pose significant logistical issues. Consequently, it would be appropriate and sufficient to collate multiple audit and observational data. This is a pragmatic solution since the challenges of deducing unbiased estimates of causal effects from observational data in the presence of unmeasured confounders are well documented.[152] Caution should also be exercised to prevent or at least monitor the use of drugs licensed for focal epilepsy in more complex epilepsies that express multiple seizure types; the focal component may be improved but other seizure types may be simultaneously exacerbated.

The approach proposed above has been used to develop rufinamide for Lennox-Gastaut syndrome.[153,154] However, this was a syndrome-specific development programme for a relatively rare epilepsy. The paradigm suggested here is a more general framework for common epilepsies which considers other factors such as adjunctive therapy and monotherapy. It dispenses with the need for a separate paediatric development programme and a separate monotherapy trial programme, neither of which have clear additional benefits in focal epilepsies.

5.3.5 Shift in research culture

The adoption of this proposed framework for drug development in epilepsy may require a shift in culture. A network of specialist paediatric epilepsy centres is needed to coordinate recruitment of patients into regulatory trials of AEDs, in a manner similar to the common practice in paediatric oncology. Rather than

specialists making third or fourth line treatment decisions for paediatric patients, they should randomise those patients into trials; this would advance knowledge much more rapidly. Those anxious about undertaking combined trials in adults and paediatrics should consider the SANAD studies, which remain the largest ever randomised trials in epilepsy and which successfully recruited across the age spectrum from 5 years upwards.[155,156] Fears over inclusion of paediatric patients in randomised trials should be tempered with examples of paediatric epilepsy studies that have successfully hit their recruitment targets in a timely manner[157–160] and with evidence regarding parents opinions on enrolment of their children into RCTs.[26] Finally, improved interaction with neonatologists would help to ensure that treatments for epileptic seizures in the very youngest age groups do not lag behind those for others.

5.4 Conclusions

This proposed paradigm for drug development in epilepsy has many potential benefits for epilepsy and epilepsy research; paediatric patients gain from immediate access to new treatments, trialists have access to a broader patient population, fewer trials and less restrictive licensing will incentivise sponsors, broaden their market and re-invigorate drug development for epilepsy, and R&D savings can be expected to have knock-on effects for medication costs and the allocation of healthcare resources.

It is acknowledged that there are potential dangers in a condensed AED trial programme because of the volume of data and number of patient exposures. There may also be additional complexities to conducting trials in adults and children if, for example, drug formulations or dosing rules vary across age groups, although several successful trials show these barriers are not insurmountable.[160] The next step in this process is to seek the opinion of patients, parents and guardians,

regulatory authorities, and sponsors on the risks, benefits and feasibility of the proposed paradigm.

This commentary is written within the context of growing international interest in the place of extrapolation in the development of medicines for paediatric epilepsies. Following the publication of robust evidence demonstrating that efficacy in RCTs recruiting adults with focal epilepsies can similarly predict efficacy in children,[42] a US consortium from academia, industry, the FDA and the Epilepsy Foundation was formed to further explore and develop this concept. The Pediatric Epilepsy Academic Consortium for Extrapolation (PEACE) has since drafted a white paper establishing disease similarity in adults and children. Additional pharmacometric analyses are also currently underway at the FDA to further evaluate pharmacokinetic and pharmacodynamic properties of AEDs. The PEACE group has shown that ever since a 1994 National Institute of Neurological Disorders and Stroke (NINDS) workshop agreed that most children with focal epilepsies would respond to a drug that was also efficacious in adults with focal epilepsies,[161] further clinical and basic science data have served to strengthen this viewpoint. After excluding children below four years and those with focal seizures associated with epileptic encephalopathy, such as Lennox-Gastaut syndrome, the pathophysiology of focal epilepsies is similar in children and adults. The PEACE white paper will therefore recommend that AEDs shown to be effective in adults with focal epilepsies should be considered as effective in children aged four years and above. This proposal will be limited to efficacy, noting that safety and PK may not necessarily be extrapolated.

There are subtle differences in the proposals being developed by the PEACE group in the US and those presented here. Nonetheless, it is encouraging to note that these discussions are taking place, simultaneously and independently, on both sides of the Atlantic. Extrapolation is clearly high on the agenda of those interested in

expediting the development of new medications for epilepsy.

Chapter 6

A quantitative framework to inform extrapolation decisions in children

6.1 Chapter background

Our systematic review in Chapter 3 highlighted many approaches taken in the literature to extrapolating between source and target populations, across both frequentist and Bayesian methodology. In this chapter, we aim to develop a quantitative framework to inform extrapolation decisions in children, after data has been collected in adults and adolescents, but before data in younger children. With wanting to utilise existing data to inform extrapolation decisions in an unstudied population, a Bayesian framework seems appropriate. Within a Bayesian framework, we can model exposure-response relationships in existing studies, borrowing strength across multiple studies in a Bayesian meta-analytic model, quantifying this as a prior distribution. The focus group meeting of epilepsy experts highlighted how useful the elicitation of expert opinion can be. We quantify subjective expert opinion to learn about the similarity of studied and unstudied populations. We consider adjusting the results of our Bayesian meta-analytic model for elicited

expert opinion on whether differences between E-R relationships in adults and adolescents are representative of differences between adults and younger children. Ultimately, in this chapter we propose that this bias adjusted meta-analysis can be used to inform extrapolation decisions in younger children.

6.2 Introduction

An experimental medicine must pass through several phases of experimentation, and only once its safety and efficacy have been confirmed can it be approved for general use. At the time of planning a drug development programme, relevant information may already be available from routine clinical practice; clinical trials of the drug performed in related diseases or different age groups; or studies of similar medicines. The design of the development programme can then be optimised in light of this so that any new studies fill in the gaps in our existing knowledge base without replicating information. Leveraging existing data in this way is particularly desirable when our aim is to develop medicines for small or vulnerable populations such as children. The European Medicines Agency (EMA) defines extrapolation as "... extending information and conclusions available from studies in one or more subgroups of the patient population (source population) ... to make inferences for another subgroup of the population (target population)..."[34,35] In many cases, we may seek to extrapolate adult efficacy data to children. Wadsworth et al.[69] report the findings of a systematic review of statistical methods relevant for extrapolating efficacy and other data from adults to children. The authors identify methods originally proposed in a variety of contexts, ranging from incorporating historical controls in new studies to evaluating the consistency of results across sites in a multi-centre trial, reflecting the wide ranging applications of extrapolation.

To justify the extrapolation of adult efficacy data to children, we must often make strong assumptions about the similarity of age groups in terms of disease progression, response to intervention and pharmacokinetic-pharmacodynamic (PK-PD) relationships. These assumptions are made explicit in the paediatric decision tree (see FDA[1] and Figure 6.1) where, in the terminology of Dunne et al.,[36] judgements about the plausibility of each aspect of similarity determine whether a 'complete', 'partial' or 'no' extrapolation strategy is adopted. Dunne et al.[36] reviewed 370 paediatric studies submitted to the FDA between 1998 and 2008 to identify cases in which efficacy data were extrapolated: of the 166 drug products considered, 14.5% followed a complete extrapolation strategy, 68% a partial extrapolation strategy and 17.5% did not extrapolate. Sun et al.,[71] in an update on the review by Dunne et al.,[36] reviewed 388 paediatric studies between 2009 and 2014. The proportion of products using partial extrapolation fell to 29%, whilst the use of no and complete extrapolation both rose to 37% and 34%, respectively. There is likely to be prior uncertainty about the plausibility of different assumptions. Hlavin et al.[162] use a scepticism factor to represent uncertainty about the plausibility of a complete extrapolation approach, whereby the full weight of evidence supporting drug efficacy in adults is taken to support a claim of efficacy in children. This factor could be established from historical data or expert opinion. The EMA extrapolation framework stipulates that data which are subsequently collected in the target population should be used to confirm extrapolation assumptions.[34,35]

Since 2006, the EU paediatric regulation[27] has mandated that the programme of studies intended to support licensing of a medicine for children in the EU must follow a Paediatric Investigation Plan (PIP), which itself must be agreed ahead of time with the EMA's Paediatric Committee (PDCO). When selecting (approving) an extrapolation strategy, sponsors (regulators) must first ask themselves how plausible needed assumptions are given the data currently to hand, where extrapolation strategies relying on more plausible assumptions are to be preferred.

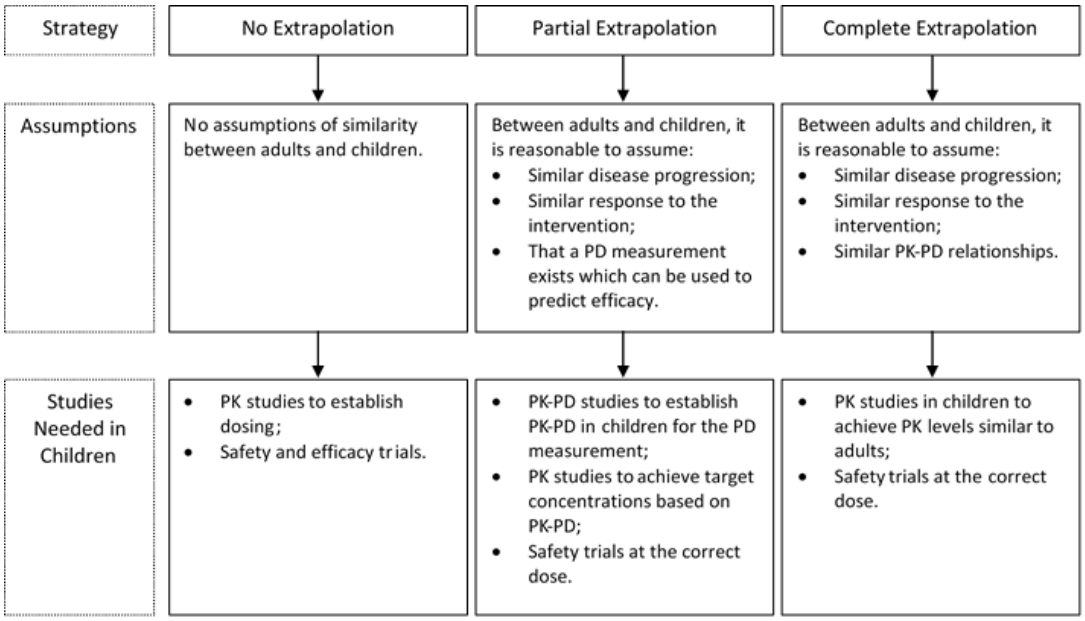


Figure 6.1: Extrapolation strategies, assumptions made and required studies, based on the FDA paediatric decision tree.[1]

This chapter presents a framework for using existing data to inform a decision on whether to perform a complete extrapolation of efficacy data from adults to children or a partial extrapolation instead. This decision will determine whether the sponsor will collect only PK data in children to support dose-finding, or both PK and PD data. The proposed framework requires pre-specification of a numerical criterion which PK-PD curves in adults and children must satisfy in order to be considered 'similar'. The sponsor can then use historical data or expert opinion to quantify the prior plausibility of the stated degree of similarity. This process enables sponsors and regulators to define transparent success criteria which emerging data in the target population must satisfy in order to be judged as verifying the assumption.

We propose that the process of choosing between complete and partial extrapolation strategies should begin by performing a Bayesian random-effects meta-analysis of existing PK-PD data to derive priors for parameters representing differences between PK-PD relationships in adults and children. When studying small populations it is likely that few historical studies will be available for synthesis. The

methodological challenges associated with performing meta-analyses of few trials have been noted in Friede et al. and Turner et al.[163,164] In this setting, using a frequentist approach, we lack power to detect between-trial heterogeneity,[163] while the results of a Bayesian random-effects meta-analysis are sensitive to the choice of prior for the between-trial heterogeneity parameter.[163,165] Furthermore, 'external biases'[51] may be inherent in the existing data if there are differences between the source and target populations, for example, if existing data are measurements on adults and adolescents but our question is whether PK-PD relationships in adults and children aged 2-11 years are similar. This data scenario often arises in practice because drug development in adults and children is typically staggered, starting in adults first. Furthermore, older adolescents are also often recruited into adult trials in therapeutic areas such as epilepsy[155,156,166,167] and asthma.[168,169] To derive prior distributions for key parameters accounting for external biases, existing data may be down-weighted according to either a pre-specified weight (see, for example, Ibrahim and Chen;[72] Tan et al.;[170] Rietbergen et al.[87]) or a dynamic weight reflecting their commensurability with new data collected in the target population.[72–74] The challenges of dynamic downweighting are noted in Galwey.[171] Alternatively, we can model the external biases and either define empirical priors[51,172] or priors elicited from expert opinion[51] on the bias parameters. We adopt the latter approach here.

To make things consistent, throughout this chapter we illustrate the proposed extrapolation framework with applications to anti-epileptic drug (AED) development in mind. In this setting, there is broad agreement about the acceptability of extrapolating efficacy data in adults with partial-onset seizures (POS) to older children with POS, although there is some uncertainty about what age we can extrapolate down to.[43–45,70] This chapter proceeds as follows. In Section 6.3 we introduce our framework, define a Bayesian bias-adjusted multivariate meta-analytic model for existing PK-PD data and propose a quantitative criterion to justify complete

extrapolation of the efficacy data. Section 6.4 describes a scheme for eliciting expert opinion on external biases that may be inherent in the existing data. In Sections 6.5 and 6.6, we describe the simulation study used to evaluate properties of our framework in a range of scenarios before concluding in Section 6.7 with a discussion.

6.3 Using existing data to inform an extrapolation decision

6.3.1 Motivation

Our aim is to use existing adult and adolescent data to inform our decision of whether to adopt a complete or partial extrapolation approach for younger children. Suppose PK-PD data are available from H historical trials which recruited both adults and adolescents. Let Y_{ij} represent the response of the i^{th} subject in historical study j, for $i = 1, \ldots, N_j$, and $j = 1, \ldots, H$. Then, writing A_{ij} as a binary indicator of age which takes the value 1 if the i^{th} subject in study j is an adolescent and 0 otherwise, let

$$\mathbb{E}[Y_{ij}] = g^{-1}\left(\gamma_0 + \sum_{k=1}^{K}\gamma_k x_{kij} + \gamma_C C_{ij} + \gamma_A A_{ij} + \gamma_I C_{ij} A_{ij}\right), \tag{6.1}$$

where C_{ij} is a measure of drug exposure; $x_{1ij}, \ldots, x_{Kij}$ are baseline covariates (such as weight) influencing response; and g is the link function of the generalised linear model.

For present purposes, we assume that regression parameters remain constant across studies. In this case, the relationship between exposure and the expected PD response, hereafter referred to as the PK-PD 'relationship' or 'curve', is said to be identical in adults and adolescents in each study if $\gamma_A = \gamma_I = 0$. The assumption of between-trial homogeneity is relaxed in Section 6.3.3, in which case γ_A and γ_I

can be interpreted as population means.

To simplify the presentation of our methods, we will assume throughout that the PD response of interest is normally distributed and that a linear model is an adequate description of the underlying relationship between exposure and response, so that

$$Y_{ij} = \gamma_0 + \sum_{k=1}^{K} \gamma_k x_{kij} + \gamma_C C_{ij} + \gamma_A A_{ij} + \gamma_I C_{ij} A_{ij} + \epsilon_{ij}, \tag{6.2}$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$ is a random error term. Linear models have been used to analyse PK-PD data for the AEDs oxcarbazepine[38] and topiramate[166] setting $Y = \log\{Z+110\}$, where Z is the percent change from baseline in seizure frequency and C represents the steady state trough concentration under repeated dosing.

Consider now the data that we would accumulate if we performed a PK-PD study, indexed by T, in adults and younger children, that is, the target population. If we made a complete extrapolation of efficacy data from adults to younger children, we would not need to perform this study but it is useful to consider the data that it would generate. Suppose we measure PD responses Y_{iT}, for $i = 1, \ldots, N_T$, following the linear model

$$Y_{iT} = \beta_0 + \sum_{k=1}^{K} \beta_k z_{kiT} + \beta_C C_{iT} + \beta_A A_{iT} + \beta_I C_{iT} A_{iT} + \epsilon_{iT}, \tag{6.3}$$

where $\epsilon_{iT} \sim N(0, \sigma^2)$; $z_{1iT}, \ldots, z_{KiT}$ are baseline prognostic covariates defined analogously to $x_{1ij}, \ldots, x_{Kij}$; C_{iT} is a measure of exposure for subject i defined similar to C_{ij}; and A_{iT} is a binary age covariate taking the value 1 if subject i is a younger child and 0 otherwise. PK-PD relationships in adults and younger children would be said to be identical if $\beta_A = \beta_I = 0$. The age boundary separating adolescents from younger children can be chosen based on the ICH E11 guidance[23] which stipulates that children are aged 2 to 11 years and adolescents 12 to 16-18 years (dependent on region). We relate parameters in the source and target

populations described by models (6.2) and (6.3) via the relations:

$$\beta_A = \gamma_A + \delta_A \qquad \text{and} \qquad \beta_I = \gamma_I + \delta_I. \tag{6.4}$$

This makes the assumption that the PK-PD relationship in younger children only differs from adolescents by additive bias terms δ_A and δ_I. Here δ_A and δ_I represent external biases arising because PK-PD curves in adolescents and younger children may differ due to the effects of maturation and physical development on drug absorption, distribution, metabolism and elimination, and on the action of and response to a drug.[173] Stephenson[22] also notes that the responses of adults and children to many drugs have much in common, although there are examples, such as warfarin[174] and cyclosporine,[175] where there are true pharmacodynamic differences between age groups. An alternative to the additive bias model (6.4) is a proportional model stipulating $\beta_A = \delta_A\gamma_A$ and $\beta_I = \delta_I\gamma_I$.[51] However, we prefer an additive model in this setting since there may be differences between adults and younger children even if there are no differences between adults and adolescents. Furthermore, we can think of differences between adults and younger children as the sum of differences between adults and adolescents and differences between adolescents and younger children.

The existing data, $\mathcal{D}_E$, are said to be relevant for learning about likely differences between PK-PD relationships in adults and younger children if δ_A and δ_I are both close to 0. In the following sections, we outline how priors for (β_A, β_I) can be derived by first performing a Bayesian meta-analysis of $\mathcal{D}_E$ to obtain posterior distributions for (γ_A, γ_I), and then adjusting for our prior opinion on (δ_A, δ_I) according to model (6.4). Based on the priors for (β_A, β_I) thus obtained, we can calculate the prior probability that PK-PD curves in adults and younger children are similar enough to satisfy the criterion for complete extrapolation that will be proposed in the following section. If this probability is high, it may be used to

support a decision not to perform a PK-PD study in younger children and instead leverage the adult data to support a claim of efficacy in this age group.

6.3.2 Extrapolation criterion

We define how similar PK-PD curves in adults and younger children must be in order to justify a complete extrapolation of efficacy data. Let $\mathcal{C}^\star$ denote a level of exposure known to be effective in adults, for example, the adult EC_{90}, the exposure at which the expected adult response is 90% of the maximum. We propose a similarity criterion evaluating whether a summary measure of the distribution of responses in adults and younger children on placebo and at an effective level of exposure are equivalent. Then, PK-PD curves are said to be similar if

$$\mathcal{M}[h\{Y(C_T = 0; A_T = 1)\}] - \mathcal{M}[h\{Y(C_T = 0; A_T = 0)\}] \in (-\eta_1,\, \eta_1) \qquad (6.5)$$

and

$$\mathcal{M}[h\{Y(C_T = \mathcal{C}^\star; A_T = 1)\}] - \mathcal{M}[h\{Y(C_T = \mathcal{C}^\star; A_T = 0)\}] \in (-\eta_2,\, \eta_2), \qquad (6.6)$$

where $h(Y(C_T,\, A_T))$ is a function of the PD response of a subject randomised to exposure C_T in age group A_T; and $\mathcal{M}$ is a measure of location such as the mean or median. It may be more straightforward to specify equivalence margins with differences between a transformed outcome in mind. For example, in the context of AEDs, criteria (6.5)-(6.6) would be written setting $h(Y(C_T; A_T)) = \exp(Y(C_T; A_T)) - 110$. In this case, it would be more appropriate to use the median as a summary measure since the log-normal distribution is asymmetric. In practice, bounds η_1 and η_2 would be set by regulators based on clinical judgement. Larger bounds imply that larger differences between the average PD responses of adults and younger children will be tolerated if we incorrectly perform a complete extrapolation and dose younger children targeting the adult effective concentration. Whilst different equivalence bounds can be applied at exposures $C_T = 0$ and $C_T = \mathcal{C}^\star$, to simplify, we set $\eta_1 = \eta_2$.

Working in a Bayesian framework, the prior probability of criteria (6.5)-(6.6) can be taken as a measure of the prior plausibility of an assumption that PK-PD curves are similar enough in adults and children to justify a complete extrapolation of efficacy data across these subgroups. We speculate that a prior probability in excess of 0.8 or 0.9 would be sufficient to support the immediate adoption of a complete extrapolation strategy. Lower probabilities would prompt a sponsor to collect additional PK-PD data in younger children, where the exact sample size may be determined according to an expected value of information calculation.[176,177] However, a very low prior probability could be consistent either with extreme uncertainty about the relevance of the existing data or a strong degree of scepticism about the similarity of PK-PD curves in adults and younger children. In both cases the most appropriate testing strategy would be to plan a PK-PD study in younger children sized to support independent dose-finding in this age group.

6.3.3 Bayesian bias adjusted meta-analytic model for existing data

We begin the process of quantifying what is known about differences between adults and younger children by using a Bayesian meta-analysis to learn about γ_A and γ_I. We assume individual patient data are available but summary measures could be used if maximum likelihood estimates and corresponding standard errors are available for all parameters in the linear predictor of model (6.2). At the first level of the meta-analytic model, existing data from study j, $j = 1, \ldots, H$, enrolling adults and adolescents, are modelled as

$$Y_{ij} = \gamma_{0j} + \gamma_{Cj} C_{ij} + \gamma_{Aj} A_{ij} + \gamma_{Ij} C_{ij} A_{ij} + \epsilon_{ij}, \tag{6.7}$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$ and for ease of presentation, we assume that the only baseline covariate prognostic of outcome is age. To limit model complexity we regard the

study-specific intercepts and effects of exposure as fixed effects, with population parameters γ_0 and γ_C estimated as $\sum_{j=1}^{H} \gamma_{0j}/H$ and $\sum_{j=1}^{H} \gamma_{Cj}/H$.

To accommodate between-trial heterogeneity in the remaining study-specific parameters in model (6.7), we model $(\gamma_{Aj}, \gamma_{Ij})$ as samples from a bivariate Normal (BVN) population distribution with mean $\boldsymbol{\mu} = (\gamma_A, \gamma_I)$ and covariance matrix Σ. Considering the prior distributions for Σ, one approach would be to place an inverse-Wishart prior on the covariance matrix. However, our investigations found the result of the meta-analysis to be very sensitive to the choice of the inverse-Wishart scale matrix; decreasing the diagonal elements of this matrix reduces the variances of the marginal posterior distributions of γ_A and γ_I. Gelman[178] shows that inverse-Gamma(ε, ε) priors with $\varepsilon \approx 0$ are informative for variance parameters in hierarchical models and suggests that inverse-Wishart prior distributions for covariance matrices incur similar issues. To avoid this sensitivity, we adopt an alternative parameterisation[179] for the BVN population distribution which gives the analyst more flexibility in how they specify priors for variance parameters. For $j = 1, \ldots, H$, define

$$\begin{aligned} \gamma_{Aj} &\sim N(\gamma_A, \xi_1^2), \\ \gamma_{Ij} \mid \gamma_{Aj} &\sim N(\lambda_0 + \lambda_1(\gamma_{Aj} - \bar{\gamma}_A), \xi_2^2), \end{aligned}$$

where $\bar{\gamma}_A = (1/H) \sum \gamma_{Aj}$, which implies that

$$\begin{pmatrix} \gamma_{Aj} \\ \gamma_{Ij} \end{pmatrix} \sim N\left(\begin{pmatrix} \gamma_A \\ \gamma_I \end{pmatrix}, \begin{pmatrix} \xi_1^2 & \lambda_1 \xi_1^2 \\ \lambda_1 \xi_1^2 & \xi_2^2 + \xi_1^2 \lambda_1^2 \end{pmatrix} \right), \tag{6.8}$$

where, under this representation, $\gamma_I = \lambda_0 + \lambda_1(\gamma_A - \bar{\gamma}_A)$. It is clear therefore that the proposed parameterisation allows for a correlation between γ_{Aj} and γ_{Ij}, for each $j = 1, \ldots, H$.

The definition of the meta-analytic model is completed by defining priors for all unknown parameters. For each j, $j = 1, \ldots, H$, the study-specific intercept and effect of exposure, γ_{0j} and γ_{Cj}, are assigned independent Normal$(0, \zeta^2)$ prior distributions. For the residual precision we stipulate $\sigma^{-2} \sim$ Gamma(a, b), with a and b chosen to define a diffuse prior. For parameters of population distribution (6.8), we place a Normal$(0, 100)$ prior on γ_A and specify priors $\xi_1 \sim$ Gamma(a_1, b_1), $\xi_2 \sim$ Gamma(a_2, b_2), $\lambda_0 \sim$ t(μ_t, σ_t, ν_t), and $\lambda_1 \sim$ Normal(μ_1, σ_1^2). In the examples we have considered, we have chosen parameters of these prior distributions to ensure the prior for the correlation between each pair $(\gamma_{Aj}, \gamma_{Ij})$ has a bucket shape placing probability mass at -1 and 1 and furthermore that prior probability mass is placed on a range of plausible values for the between-trial standard deviations.[74]

The Bayesian meta-analytic model can be fitted using Markov Chain Monte Carlo (MCMC). The joint posterior distribution of $(\gamma_0, \gamma_C, \gamma_A, \gamma_I)$ will not be of a standard form. Similar to Schmidli et al.,[57] to facilitate communication and reproducibility of the joint posterior we approximate it as a mixture of K 4-dimensional multivariate Normal (MVN) distributions using the 'flexmix' package[180–182] in R:[16]

$$f(\gamma_0, \gamma_C, \gamma_A, \gamma_I \mid \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_H) \approx \Sigma_{i=1}^{K} \omega_i \, \phi_4(\boldsymbol{\mu}_i, \Sigma_i), \tag{6.9}$$

where $\phi_4(\boldsymbol{\mu}, \Sigma)$ is a four-dimensional MVN probability density function (pdf) with mean $\boldsymbol{\mu}$ and variance Σ; and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_H$ are vectors representing the adult and adolescent data from existing studies $1, \ldots, H$. Increasing K in (6.9) increases the accuracy of the finite mixture approximation as measured by the Kullback-Leibler (KL) divergence;[57,183] however, these increases diminish as K increases and must be balanced against increases in model complexity. In our investigations, we have found setting $K = 2$ in (6.9) to be adequate and in the simulations described in Section 6.5, we occasionally found that some two-component model fits actually returned a one-component solution, that is, the posterior could be accurately ap-

proximated by a single 4-dimensional MVN distribution.

If we consider (γ_A, γ_I) to be systematically biased for the parameters (β_A, β_I) in model (6.3), then we can elicit expert opinion on the size of these external biases. We assume that prior opinion on the vector of bias parameters can be modelled as a bivariate normal distribution, written as $\boldsymbol{\delta} \sim N_2(\boldsymbol{\nu}, \Pi)$, where $\boldsymbol{\nu} = (\nu_A, \nu_I)$ are the prior modal values of the biases. Our protocol for eliciting $\boldsymbol{\nu}$ and Π is described in Section 6.4. Then, by sampling pairs (γ_A, γ_I) and (δ_A, δ_I) from $f(\gamma_0, \gamma_C, \gamma_A, \gamma_I | \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_H)$ and $\phi_2(\boldsymbol{\nu}, \Pi)$, respectively, we generate samples from the prior distribution of (β_A, β_I) given the existing data. Fitting these Monte Carlo samples using maximum likelihood estimation, we obtain the approximate prior

$$g(\beta_A,\, \beta_I \mid \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_H) = \Sigma_{i=1}^{2} \omega_i \phi_2(\boldsymbol{\mu_i}, \Sigma_i). \tag{6.10}$$

6.3.4 Effective sample size

It is useful to quantify how much we have learnt about differences between adults and younger children from the existing data in adults and adolescents. From this, we can infer to what degree the existing data are down-weighted due to our uncertainty about their relevance. We measure our uncertainty by the effective sample size (ESS)[184] of the (β_A, β_I) prior shown in (6.10). The ESS is defined as the total number of subjects (adults and younger children) that would be required to participate in a future PK-PD study to provide the same amount of information about β_A and β_I as is represented by prior (6.10).

To calculate the ESS we follow the approach of Morita et al.[184] For each sample size $m = 1, \ldots, M$, we compare the trace of the Fishers information matrix of $g(\beta_A, \beta_I \mid \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_H)$ with that of the posterior $g_\epsilon^m(\beta_A, \beta_I | \boldsymbol{D})$ formed by updating an ϵ-information prior, $g_\epsilon(\beta_A, \beta_I)$, with hypothetical new data $\boldsymbol{D}$, a vector listing

the PD responses, ages and exposures of m subjects recruited into a new PK-PD study of adults and younger children. The ϵ-information prior is of the same form as $g(\beta_A, \beta_I \mid \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_H)$, with the same number of mixture components, mixture weights and component-wise means and correlations, but with the component-wise variances inflated such that the ϵ-information prior contains minimal information. The ESS is obtained by interpolating between values of m to find the sample size corresponding to the smallest distance between the traces of the information matrices of $g(\beta_A, \beta_I \mid \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_H)$ and $g_\epsilon^m(\beta_A, \beta_I|\boldsymbol{D})$.

Assuming data from a new PK-PD study in adults and younger children would follow model (6.3), the trace of the information matrix of $g_\epsilon^m(\beta_A, \beta_I|\boldsymbol{D})$ is a function of unknown model parameters and the unobserved PD responses, ages and exposures contained in $\boldsymbol{D}$, and therefore must be estimated using Monte Carlo simulation. In practice we recommend simulating PD responses from model (6.3) assuming $\beta_0 = \gamma_0$ and $\beta_C = \gamma_C$, where γ_0 and γ_C are as defined in Section 6.3.3; the remaining model parameters β_A and β_I should be set equal to their expectations defined according to $g(\beta_A, \beta_I|\boldsymbol{Y}_1, ..., \boldsymbol{Y}_H)$.[184] This means that if there is a prior-data conflict, the ESS may not be an accurate indicator of the contribution of the prior (based on existing data and prior opinion) to the posterior (combining existing data and opinion with data from the new PK-PD trial). Age and exposure covariates $(A_{1T}, C_{1T}), \ldots$ can be sampled from models fitted to the existing data; in particular, this would assume that adults and younger children would be recruited into a new PK-PD study in the same proportion as adults and adolescents were enrolled into the existing studies. Following these steps, for each $m = 1, ..., M$, one can simulate 10,000 datasets from a new PK-PD study in adults and younger children with sample size m, and calculate the average difference between the traces of $g(\beta_A, \beta_I|\boldsymbol{Y}_1, ..., \boldsymbol{Y}_H)$ and $g_\epsilon^m(\beta_A, \beta_I|\boldsymbol{D})$. When evaluating the properties of our procedure in Section 6.5, to increase the accuracy of calculations we computed the ESS of $g(\beta_A, \beta_I \mid \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_H)$ simulating $\boldsymbol{D}$ setting β_0 and β_C equal to the

values of γ_0 and γ_C used to simulate the adult and adolescent data. Simulated values of $A_{1T}, A_{2,T}, \ldots$ and $C_{1T}, C_{2T}, \ldots$ were sampled from the same models that were used to generate the existing adult and adolescent data. We simulate 10000 new data sets for the Monte Carlo simulations and evaluate the difference between the traces of $g(\beta_A, \beta_I \mid \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_H)$ and $g_\epsilon^m(\beta_A, \beta_I|\boldsymbol{D})$ by calculating the trace of $g_\epsilon^m(\beta_A, \beta_I|\boldsymbol{D})$ for sample sizes $m = 1, \ldots, M$ and averaging over the simulations, so we have M trace averages.

6.4 Eliciting prior opinion on external biases

6.4.1 Overview

In this section we describe our proposal for eliciting an individual expert's opinion on the additive biases δ_A and δ_I. Experts should be subject matter specialists, such as consultant-level clinicians with a relevant specialism. Our full elicitation scheme can be broken down into four main components:

Part 1 Present to each expert fitted dose-response curves for adults and adolescents derived from existing data;

Part 2 Elicit each expert's prior modal guess at the dose-response curve in younger children;

Part 3 Elicit from the expert their uncertainty about their answer to Part 2 as a 90% credibility interval;

Part 4 Use the expert's answers from Parts 2 and 3 to derive a fitted prior for δ_A and δ_I. Feedback to the expert the consequences of their opinions by presenting summaries of their fitted priors for the dose-response relationship in younger children. Give the expert the opportunity to revise their earlier answers until they are happy that their fitted prior reflects their beliefs.

Note that we frame elicitation questions in terms of the dose-response, rather than PK-PD, relationship since clinicians are likely to be more familiar expressing beliefs about the former; in our experience, serum concentrations of AEDs (and other drugs) are not typically measured in routine clinical practice, so clinicians tend to be more familiar with dose than with concentration. Answers to elicitation questions can then be translated to opinions on PK-PD parameters assuming a certain relationship between dose and exposure which might be derived using existing PK-PD data or through a further elicitation exercise with pharmacometricians or clinical pharmacologists. In the examples we have considered, we have assumed dose-proportionality holds over the dose range of interest, with a known constant of proportionality. Below we give more detail on each aspect of the elicitation scheme.

6.4.2 Rationale for the elicitation scheme

It is challenging to elicit opinion directly on the biases δ_A and δ_I. Instead, we propose an indirect approach, whereby an expert is first presented with fitted adult and adolescent dose-response curves derived from a meta-analysis of historical trials and then conditional on this information, is asked for their opinion on the dose-response curve in younger children. We show below that under certain assumptions, we can deduce from the expert's answers their joint prior distribution for (δ_A, δ_I). In some cases, it may be necessary to present plots and frame elicitation questions in terms of a transformed PD response which is easier for the expert to express opinion on. For example, in clinical trials for adjunctive therapies in epilepsy, the primary endpoint is often log-transformed percent change in seizure frequency from baseline,[38,166] which would be challenging to express opinions on; instead, the percent change scale would be more interpretable for experts. We will discuss this issue in more detail in the following subsection and start by assuming that Y is on a scale that experts can express opinions on directly.

Let d denote dose. Assuming dose-proportionality, we can write exposure as $\mathcal{C} = \kappa d$. It seems reasonable to suppose that, when presented with the fitted adult and adolescent dose-response curves, the expert will take these to be the true response curves for these age groups, disregarding any estimation error. Ignoring for the moment between-study heterogeneity in PK-PD parameters, model (6.2) stipulates that at dose level $d^\star$, the fitted average PD response in adults is $\mathcal{F}^1_{d^\star} = \gamma_0 + \gamma_C \kappa d^\star$ and the fitted average response in adolescents is $\mathcal{F}^1_{d^\star} + \mathcal{F}^2_{d^\star} = \gamma_0 + \gamma_C \kappa d^\star + \gamma_A + \gamma_I \kappa d^\star$. Therefore, from the presentation of the existing data, at dose level $d^\star$ an expert can deduce: a) the average response in adults, $\mathcal{F}^1_{d^\star}$; and b) the difference between the adult and adolescent expected responses $\mathcal{F}^2_{d^\star} = \gamma_A + \gamma_I \kappa d^\star$.

Assuming bias model (6.4) and assuming no drift in the parameters of the adult PK-PD relationship so that in a future PK-PD study enrolling adults and younger children we would have $\beta_0 = \gamma_0$ and $\beta_C = \gamma_C$, model (6.3) stipulates that the expected response of a younger child in such a study given dose $d^\star$ would be

$$\mathbb{E}[Y_T \mid A_T = 1, \mathcal{C} = \kappa d^\star] = \gamma_0 + \gamma_C \kappa d^\star + (\gamma_A + \delta_A) + (\gamma_I + \delta_I)\kappa d^\star.$$

Conditioning on what has been learnt from the historical data, we have

$$\mathbb{E}[Y_T \mid A_T = 1, \mathcal{C} = \kappa d^\star] \mid \mathcal{F}^1_{d^\star}, \mathcal{F}^2_{d^\star} = \mathcal{F}^1_{d^\star} + \mathcal{F}^2_{d^\star} + \delta_A + \delta_I \kappa d^\star.$$

Assuming that prior opinion on (δ_A, δ_I) is independent of opinion on other PK-PD model parameters and can be modelled as $N_2(\boldsymbol{\nu}, \Pi)$, with $\boldsymbol{\nu} = (\nu_A, \nu_I)$ and

$$\Pi = \begin{pmatrix} \pi_A^2 & \pi_{AI} \\ \pi_{AI} & \pi_I^2 \end{pmatrix},$$

then at dose level $d^\star$ we obtain

$$\mathbb{E}[Y_T \mid A_T = 1, \mathcal{C} = \kappa d^\star] \mid \mathcal{F}^1_{d^\star}, \mathcal{F}^2_{d^\star}$$
$$\sim N(\mathcal{F}^1_{d^\star} + \mathcal{F}^2_{d^\star} + \nu_A + \nu_I \kappa d^\star, \pi_A^2 + 2\pi_{AI}\kappa d^\star + \pi_I^2(\kappa d^\star)^2).$$

The parameters of an expert's (unconditional) bias prior can therefore be identified by asking for their conditional beliefs, given the existing data about the average response of younger children on placebo, a 'medium' or a 'high' dose, where these doses are denoted by d_0, d_M and d_H, respectively. The proposed wording of the elicitation questions is given in the following section. In practice, d_M and d_H could be chosen on the basis of adult dose-finding studies or, if the drug is already licensed in adults, using WHO lists of defined daily doses.[185] The assumption that opinion on (δ_A, δ_I) is independent of opinion on other PK-PD model parameters is a pragmatic one which ensures that elicitation questions have a direct interpretation and can be answered by non-statisticians. If an expert expresses the consistent opinion that the average response in younger children is similar to the fitted average in adolescents, this suggests they believe that δ_A and δ_I are small, that is, that the existing data are highly relevant for informing our understanding of likely differences between adults and younger children. Note that the proposed scheme asks for opinions on the relevance of the existing data after seeing how supportive they are of a complete extrapolation of efficacy data from adults to adolescents. To increase the credibility of beliefs elicited in this way, one could interview independent experts not directly involved with the drug development programme.

By asking an expert for their best guesses at the average PD responses in children on placebo and a 'high' dose, we deduce fitted values of ν_A and ν_I. To find ν_A, one can subtract $(\gamma_0 + \gamma_A)$ from the expert's best guess at the average response on placebo; ν_I is obtained by subtracting $(\gamma_C + \gamma_I)$ from the slope calculated by divid-

ing the difference between the expert's best guesses at the expected PD responses on a high dose and placebo by κd_H. An expert's uncertainty about the average PD response in younger children on a dose $d^\star$ is established by asking questions to establish the 5^{th}, 25^{th}, 75^{th} and 95^{th} percentiles of their prior distribution for this quantity. Given values of ν_A and ν_I, we can then adapt the approach of[186] to search over configurations of π_A^2, π_I^2 and π_{AI} to find the triplet which defines a positive definite variance matrix and minimises the absolute difference between percentiles of the fitted prior and the expert's stated percentiles. To ensure positive definiteness, Π is represented in the optimisation routine using the Cholesky decomposition.

6.4.3 Elicitation protocol

In order to capture expert opinion we developed an interactive web application written in R[16] using the 'Shiny' package.[65] Screen shots from the app can be found in Appendix B.1. The app expects the statistical facilitator to summarise the protocols of the historical trials (eligibility criteria, outcomes, treatments, etc.) and then directs them through the following elicitation protocol:

Step 1 Display the fitted dose-response curves for adults and adolescents derived from the historical data. If individual patient data (IPD) are available, overlay these as a scatter plot to provide a more complete description of the existing data, as illustrated in Supplementary Figure B.1. Conditional on these existing data, elicit opinion on how the average PD response in younger children varies with dose according to Steps 2-5. When interacting with clinicians, the dose-response curve should be referred to throughout as the line of best fit that would be plotted if we were able to randomise a large number of younger children to each of a range of dose levels.

Step 2 To elicit the expert's prior modal guess at the dose-response curve in younger children given existing data, first show them a range of different shapes for

the dose-response curve in this age group and ask them to select the one which most closely reflects their current best guess, as shown in Supplementary Figure B.2. The range of shapes should include curves which: a) lie above the fitted adult curve; b) between the fitted adult and adolescent curves; c) lie below the fitted adolescent curve; d) are identical to either the fitted adult or adolescent curves. Guide the expert to iteratively refine their selected 'shape' until they find a line which more closely reflects their current best guess at the dose-response curve in younger children, as illustrated in Supplementary Figure B.3. Refinements are driven by the expert's answer to the following question, which is repeated for placebo and dose d_H: "Given the existing data, give your best guess at the average response amongst children aged 2-11 years on dose d of the test treatment". From this step of the elicitation process, the prior modal values of the bias parameters ν_A and ν_I can be deduced.

Step 3 To establish the expert's uncertainty about the dose-response relationship in younger children, given the existing data in adults and adolescents, explain the concept of a credibility interval. Then show the expert a range of shapes which might be formed by interpolating between the limits of their 90% credibility intervals for the average PD response in younger children on placebo and doses d_M and d_H. Four options are shown in Supplementary Figure B.4, which have been informed by discussions with clinical experts. The statistical facilitator should interpret each shape and ask the expert to select the one most closely describing their prior uncertainty. For each dose in turn, the expert is then asked: "Given the existing data, state a value which you are 95% sure the average response amongst younger children on dose d of the test treatment will lie below". The expert's answer is the 95^{th} percentile of their prior distribution for the expected PD response at the dose in question. Only the upper limits of credibility intervals need to be elicited to deduce fitted values for elements of the variance-covariance matrix

of $\boldsymbol{\delta}$ (due to symmetry). Despite this, the fitted upper and lower limits of credibility intervals for expected PD responses are shown to the expert. This stage of the elicitation procedure is illustrated in Supplementary Figure B.5.

Step 4 To further refine and validate the expert's priors, elicit three histograms representing their prior distributions for the average PD response of younger children on placebo, dose d_M and dose d_H. This information is collected to check the consistency of the expert's opinion, to increase the stability of the prior fitting routine (by also using the 25^{th} and 75^{th} percentiles) and to provide a more accurate quantification of the expert's opinion. Each histogram is elicited by asking the expert to allocate a prior weight to different response intervals, where their weights must sum to 1. For reference, coloured lines marking the mode and 95^{th} percentiles elicited from Steps 2 and 3 are shown, along with the 5^{th} percentile implied by symmetry. Supplementary Figure B.6 is a screen shot of this process. The elicited histograms provide the 5^{th}, 25^{th}, 75^{th} and 95^{th} percentiles of the expert's prior distributions for the average PD response of younger children on placebo, dose d_M and dose d_H, which can be used to allow parameters of the prior variance matrix of $\boldsymbol{\delta}$ to be established.

Step 5 Once all three histograms have been elicited, feed back to the expert summaries of the prior distribution for the dose-response relationship in younger children that would be implied by their expressed opinions. In particular, allow the expert to compare the credibility bound established from Step 3 with that from Step 4, to ensure that they are confident the histograms reflect their belief. The average response line will be that given in Step 2, whilst the uncertainty around that line will be taken from the opinions elicited in Step 4. If the fitted prior lacks face validity, the expert should be allowed to re-evaluate their answers to Step 4 until they feel confident that their beliefs have been adequately captured. Supplementary Figures B.7 and B.8 illustrate this step of the scheme.

6.4.3.1 Example: Application to AED development

The prior elicitation protocol and software described above has been tested and refined through several rounds of applications asking clinicians about dose-response curves for an AED. In that context, we defined an expert as a paediatric or adult neurologist with practical experience of treating adult and/or paediatric epilepsy. Testing included face-to-face pilot runs with eight neurologists (adult and paediatric) attending the International League Against Epilepsy British and Irish Chapters Meeting (Dublin, October 2016), which suggested that adult neurologists may feel unable to provide confident answers regarding the paediatric population. However, the population of experts may be biased if adult neurologists are neglected completely. The final version of the protocol was also piloted on three neurologists via web conference. The figures in Supplementary Appendix B.1 are of the app tailored to the AED development application.

A few comments on the application of our elicitation scheme to the AED example are needed. In this context, the PD response, $Y = \log\{Z + 110\}$ is the log-transformed percent change in seizure frequency from baseline. Since a log-transformed percent change is difficult to give opinions on, it makes sense to elicit beliefs on the percent change in seizure frequency, Z, instead. It seems natural to think that if we took an expert's best guess at the relationship between dose and $\mathbb{E}(Z)$ and then transformed it, we would obtain their best guess at the relationship between dose and $\mathbb{E}(Y)$. Therefore, the prior mode for $\mathbb{E}(Y)$ at a particular dose is obtained by transforming the prior mode of $\mathbb{E}(Z)$. The expert's quantiles of their prior distribution for $\mathbb{E}(Z)$ can similarly be transformed to give corresponding quantiles for $\mathbb{E}(Y)$, as quantiles are invariant under a monotonic transformation.

Using the proposed elicitation procedure and Shiny app, bias priors (E1) and (E2) below were elicited from two epileptologists who were presented with simulated

IPD on a licensed AED shown in Figure 6.2 which were generated from the fitted models presented in Girgis et al.[166] We assumed the dose-proportionality constant, κ, was equal to 1; the PK-PD relationship is then equivalent to the relationship between dose and the PD response.

$$\text{Prior E1:} \quad \begin{pmatrix} \delta_A \\ \delta_I \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.101^2 & 3.898 \times 10^{-5} \\ 3.898 \times 10^{-5} & 0.016^2 \end{pmatrix} \right). \quad \text{(E1)}$$

$$\text{Prior E2:} \quad \begin{pmatrix} \delta_A \\ \delta_I \end{pmatrix} \sim N\left(\begin{pmatrix} 0.050 \\ 0.003 \end{pmatrix}, \begin{pmatrix} 0.101^2 & 3.898 \times 10^{-5} \\ 3.898 \times 10^{-5} & 0.016^2 \end{pmatrix} \right). \quad \text{(E2)}$$

Prior (E1) is consistent with the opinion that it is most likely that the average PD responses of adolescents and younger children are the same. Prior (E2) is consistent with the best prior guess that younger children have an average PD response slightly worse than that for adolescents, so that differences between PK-PD curves in adults and younger children are larger than those between adults and adolescents. However, as can be seen from Figure 6.2(b), both experts were uncertain about the dose-response curve in younger children given the existing data which indicates that extrapolation decisions are likely to be made in the presence of significant uncertainty. The implications of this are explored further in the subsequent section.

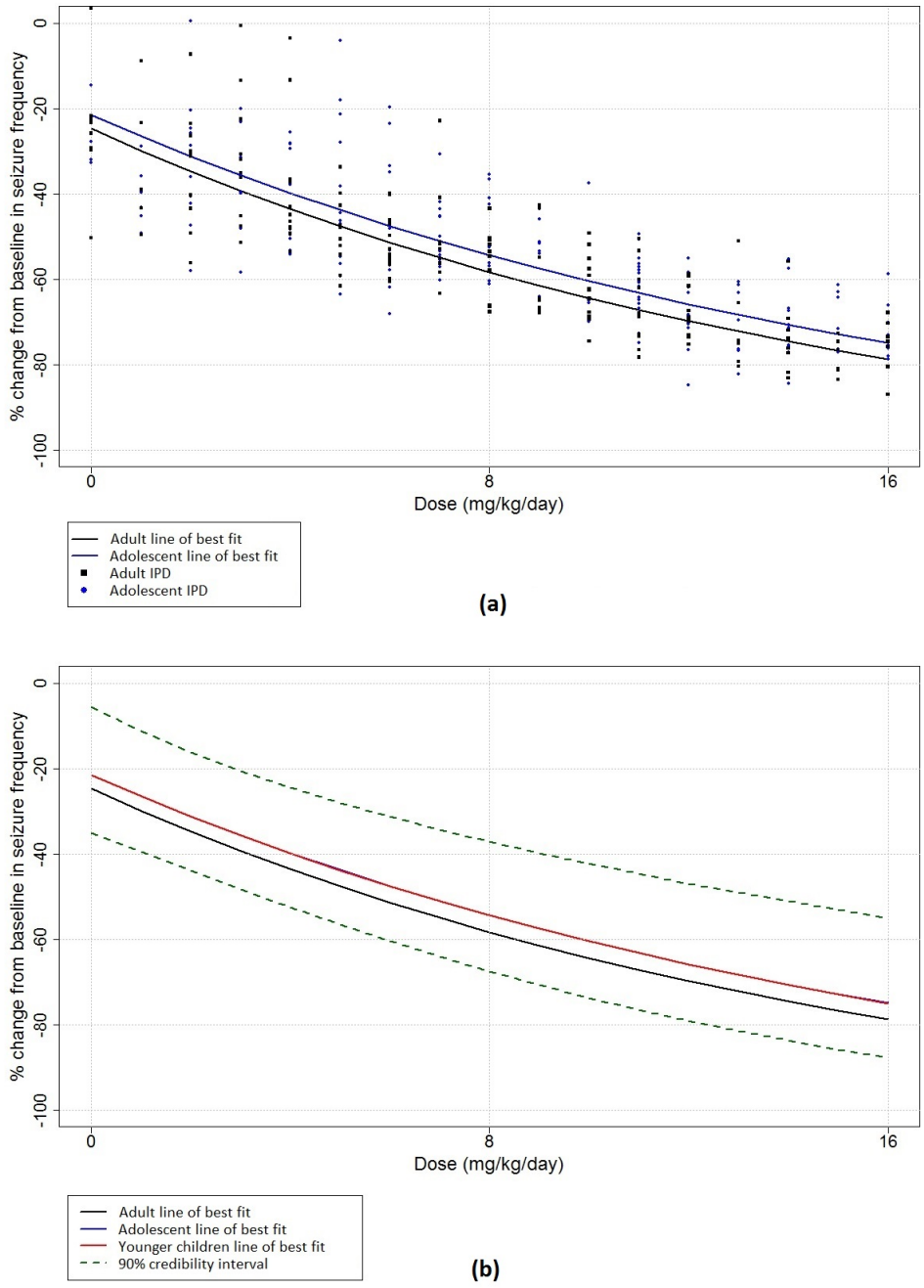


Figure 6.2: (a) Simulated adult and adolescent IPD and lines of best fit. (b) Elicited modal values of the dose-response relationship in younger children, captured in expert prior E1, with corresponding 90% credibility interval. Also plotted are adult and adolescent lines of best fit obtained from the simulated IPD.

6.5 Simulation study

To assess the operating characteristics of our proposed extrapolation framework, we performed a simulation study considering a wide range of realistic scenarios. Scenarios were informed by applications to AED development for partial onset seizures. Motivated by the PK-PD studies described in Girgis et al.[166] and Nedelman et al.,[38] in what follows we took the PD response to be $Y = \log(Z + 110)$, where Z represents the percent change from baseline seizure frequency, and exposure to be C_{min}, the steady state trough concentration under repeated dosing.

6.5.1 Epilepsy application extrapolation criterion

In all simulation scenarios, PK-PD curves were said to be similar in two age groups if the difference between median percent changes from baseline in seizure frequency was less than 10%:

$$\mathcal{M}[\exp\{Y(C_T = 0; A_T = 1)\} - 110] - \mathcal{M}[\exp\{Y(C_T = 0; A_T = 0)\} - 110] \in (-10, 10) \quad \text{and} \tag{6.11}$$

$$\mathcal{M}[\exp\{Y(C_T = \mathcal{C}^\star; A_T = 1)\} - 110] - \mathcal{M}[\exp\{Y(C_T = \mathcal{C}^\star; A_T = 0)\} - 110] \in (-10, 10), \tag{6.12}$$

where $\mathcal{M}$ represents the median and $\mathcal{C}^\star$ is the adult EC_{90}, the exposure at which the expected adult response is 90% of the maximum. Our choices for η_1 and η_2 were based on clinical feedback on acceptable differences in average responses. We wrote the similarity criteria in terms of the transformed PD endpoint to make it easier to elicit similarity bounds. We chose the median as our summary measure of response since if Y follows a log-normal distribution with median m_Y, the median of $Z = \exp\{Y\} - 110$ is given by $m_Z = \exp\{m_Y\} - 110$, thus simplifying the mapping of properties from Z to Y.

6.5.2 Simulating historical PK-PD data in adults and adolescents

We simulated the PD responses of adults and adolescents as normally distributed with residual variance σ^2 according to model (6.7). Setting $\sigma^2 = 0.0243$ ensured that for each $i = 1, \ldots, N_j$, $j = 1, \ldots, H$, each transformed response Z_{ij} lay within $\pm 10\%$ of its median given the patient's age group and level of exposure with probability 0.95. We simulated age group indicators $A_{ij} \sim Bern(0.15)$ so that on average 15% of historical trial participants were adolescents. This proportion appears reasonable based on the studies cited in Girgis et al.[166] Furthermore, we set 10% of patients in each study to be assigned to placebo. For patients allocated to the drug, we generate $\log(C_{min})$ values as samples from a normal distribution with mean $\log(2.94)$ and variance 0.921, truncating samples above by $\log(17.27)$ to avoid excessively high concentrations. In this way, we generated C_{min} values with quartiles and 1st and 99th percentiles similar to those reported by studies cited in Girgis et al.[166] where C_{min} values ranged between $0.19 - 17.27$ μg/ml.

For each simulated historical study, study-specific parameters of PK-PD model (6.7) were generated by sampling $\gamma_{0j} \sim N(\gamma_0, \sigma_0^2)$, $\gamma_{Cj} \sim N(\gamma_C, \sigma_C^2)$ and

$$\begin{pmatrix} \gamma_{Aj} \\ \gamma_{Ij} \end{pmatrix} \sim N\left(\begin{pmatrix} \gamma_A \\ \gamma_I \end{pmatrix}, \begin{pmatrix} \sigma_A^2 & 0 \\ 0 & \sigma_I^2 \end{pmatrix} \right)$$

setting $\gamma_0 = 4.4469$ and $\gamma_C = -0.0627$ which are maximum likelihood estimates of these model parameters taken from Girgis et al.[166] Let ϵ_P and ϵ_C represent the difference between $\mathcal{M}(Z \mid A = 1, \mathcal{C})$ and $\mathcal{M}(Z \mid A = 0, \mathcal{C})$ when $\mathcal{C} = 0$ and $\mathcal{C} = \mathcal{C}^\star$, respectively. We chose values for γ_A and γ_I such that values of ϵ_P and ϵ_C, when evaluated under the population mean parameters, spanned a realistic range of differences. We considered pairs $(\epsilon_P, \epsilon_C) \in \{(0,0), (5,5), (10,10), (20,20), (5,10), (5,20)\}$ corresponding to the six pairs of (γ_A, γ_I) labelled in Table 6.1 as PK-PD Models

Table 6.1: Population means of the effects of age (γ_A) and the interaction between age and exposure (γ_I) for adults and adolescents in the six PK-PD simulation models, with the interpretation of each model. The population median PD response refers to $\exp\{m_Y\} - 110$, where $m_Y = \mathcal{M}(Y)$ is calculated setting PK-PD model parameters equal to their population means.

	γ_A	γ_I	Model interpretation
Model S1	0	0	Population median PD response identical in adults and adolescents
Model S2	0.057	0.006	Small differences between population median PD responses satisfy (6.11) and (6.12)
Model S3	0.111	0.010	Moderate differences between population median PD responses satisfy (6.11) and (6.12)
Model S4	0.211	0.018	Large differences between population median PD responses do not satisfy (6.11) and (6.12)
Model S5	0.057	0.014	Small differences between population median PD responses on placebo; moderate differences at EC_{90} satisfy (6.11) and (6.12)
Model S6	0.057	0.027	Small differences between population median PD responses on placebo; large differences at EC_{90} do not satisfy (6.12)

1-6.

The variances of study-specific PK-PD parameters were chosen to characterise low, moderate, high and very high levels of between-trial heterogeneity. Parameters σ_0^2 and σ_C^2 stipulate the level of between-trial heterogeneity in the adult PK-PD relationship: σ_0^2 was therefore chosen so that for each historical study, the median percent change from baseline in seizure frequency for adults on placebo lay within $\pm 10\%$ of the median of Z calculated setting the PK-PD model parameters equal to their population means with probability 0.6 (very high heterogeneity), 0.7 (high), 0.8 (moderate) or 0.95 (low). Fixing σ_0^2, σ_C^2 was then set to ensure the study-specific median percent change in seizure frequency on the EC_{90} lay within $\pm 10\%$ of the median of Z calculated setting the PK-PD model parameters equal to their population means with the same probability. We chose σ_A^2 and σ_I^2 to fix the prob-

Table 6.2: Standard deviations for the intercept (σ_0) and effects of exposure (σ_C), age (σ_A) and the interaction between age and exposure (σ_I) chosen to reflect low, moderate, high and very high levels of between-trial heterogeneity.

Level of heterogeneity	σ_0	σ_C	σ_A	σ_I
Low	0.057	0.008	0.059	0.009
Moderate	0.086	0.012	0.091	0.014
High	0.107	0.015	0.113	0.018
Very high	0.132	0.019	0.140	0.023

ability that an individual historical trial will be consistent with an assumption of similar PK-PD curves in adults and adolescents according to criteria (6.11)-(6.12). Specifically, we chose σ_A^2 such that with probability 0.6, 0.7, 0.8 or 0.95, the true difference between $\mathcal{M}\{Z \mid A = 0, \mathcal{C} = 0\}$ and $\mathcal{M}\{Z \mid A = 1, \mathcal{C} = 0\}$ lay within $\pm 10\%$. For a particular choice of σ_A^2, we then fixed σ_I^2 such that with probability 0.6, 0.7, 0.8 or 0.95, the true difference between $\mathcal{M}\{Z \mid A = 0, \mathcal{C} = EC_{90}\}$ and $\mathcal{M}\{Z \mid A = 1, \mathcal{C} = EC_{90}\}$ lay within $\pm 10\%$. Different configurations of the heterogeneity parameters are listed in Table 6.2.

Simulation scenarios considered different numbers of historical trials ($H = 2, 3, 4, 5, 10, 20$) and numbers of subjects per trial ($N = 30, 170$). Numbers of historical trials were chosen to explore a plausible range: Davey et al.[49] report that 36% of the 22453 meta-analyses listed in the Cochrane Database of Systematic Reviews as of 2011 were based on two studies, 75% were based on five or fewer studies, and 1% were based on 28 or more studies. Numbers of subjects per trial were selected to explore the impact of smaller and larger trials; we had access to data from four industry-sponsored trials of an anti-epileptic drug, the average sample size of which was 168 patients. Overall, this gave 288 different simulation scenarios to consider.

6.5.3 Meta-analysis of simulated historical PK-PD studies

For each of the 288 simulation scenarios, we simulated 1000 trials and fitted the Bayesian multivariate meta-analytic model described in Section 6.3.3 to each dataset. All simulations were performed in R[16] fitting the meta-analytic model by calling OpenBUGS version 3.2.3[13] using the 'R2OpenBUGS' package.[187] We fitted the Bayesian model by running three chains using a thinning rate of 5, running the chain for 30000 iterations including a burn-in of 10000 iterations. The 'coda' package[188] was then used to extract posterior samples from the OpenBUGS output.

The meta-analytic model was fitted placing prior distributions on the unknown parameters of the form described in Section 6.3.3, with parameters given in Table 6.3. Parameters defining the bivariate normal population distribution variability, ξ_1 and ξ_2, were chosen so that the prior means of ξ_1 and ξ_2 were equal to our choices for the moderate between-trial standard deviation for γ_A and γ_I; 95% of the probability mass for the ξ_1 prior was between (0.0119, 0.2495); and 95% of the probability mass for the ξ_2 prior was between (0.0005, 0.0503). Therefore, low weight was given to very low and high between-trial variances. We chose this range based on the heterogeneity scenarios we explored; in practice one may have some historical information about the range the between-trial variability is likely to lie in. If not, a more diffuse prior, spreading probability mass over a wider interval could be used. We tested the use of a more vague Gamma(1.5, 3) prior on both ξ_1 and ξ_2, which resulted in a slight reduction in the probability of extrapolation; perhaps as a more conservative approach, this could be appropriate. Our choice of priors for λ_0 and λ_1 implies a reasonable spread of probability mass between -1 and 1 for the prior correlation coefficient.

Table 6.3: Prior distributions placed on unknown model parameters defined in Section 6.3.3.

Parameter	Prior Distribution
$\gamma_{01}, \ldots, \gamma_{0H}$	Normal(0, 100)
$\gamma_{C1}, \ldots, \gamma_{CH}$	Normal(0, 100)
ξ_1	Gamma(2.097, 23.003)
ξ_2	Gamma(1.118, 78.149)
λ_0	$t(0, 0.03, \text{df} = 3)$
λ_1	Normal(0, 1)

6.6 Results

In this section we summarise the results of the simulation study. Some tables and figures are listed in the Supplementary Materials.

6.6.1 Meta-analysis of historical data

As can be seen from supplementary tables B.1 - B.8 in Appendix B.3, the Bayesian multivariate meta-analysis of the existing adult and adolescent data accurately estimates the effects of age and the interaction between age and exposure, with low bias, empirical standard deviations and mean squared error in most scenarios. In all cases, accuracy increases with the sample size per study. Empirical standard deviations are highest under the highest level of between-trial heterogeneity, but the bias remains small. The intercept and effect of exposure are also estimated with small bias and high precision (results not presented). This suggests that the meta analysis is capturing the true values underlying our simulated data well.

6.6.2 ESS of the approximate joint posterior for β_A and β_I

Figure 6.3 plots average ESSs of the BVN mixture approximation to the joint posterior for (β_A, β_I), which are the target parameters describing differences between adults and younger children. As only the parameter population means are changing between simulation models S1-S6, ESSs are almost identical across the different underlying models. We therefore only present results for model S1; the means and

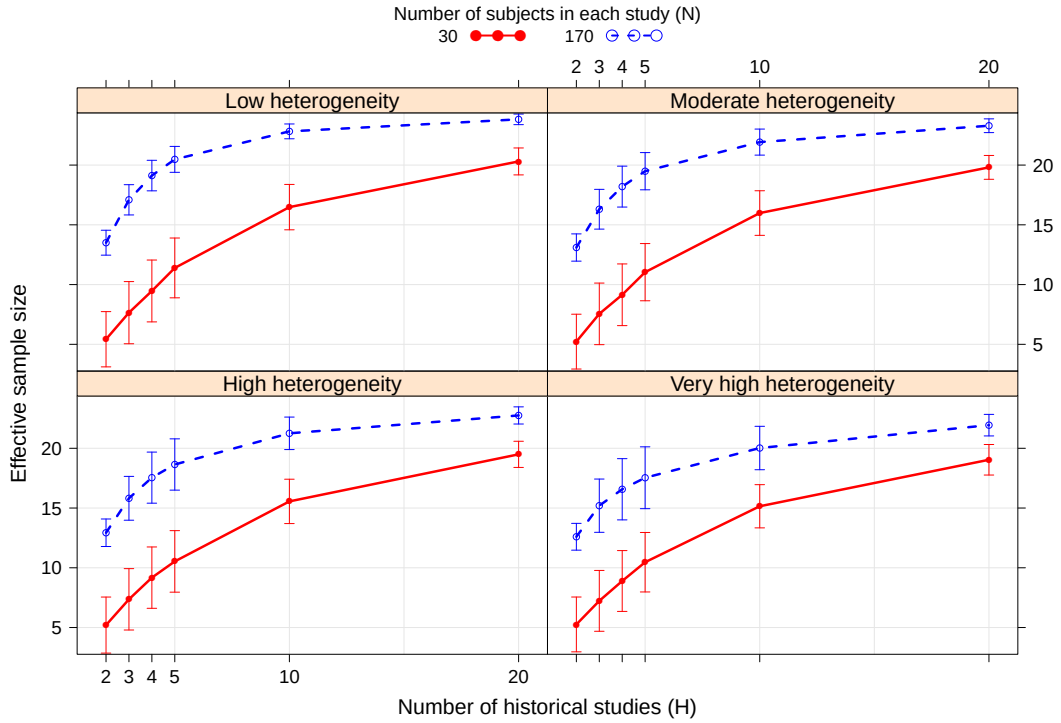


Figure 6.3: Average ESS under bias prior E1 when the true PK-PD relationships in adults and adolescents follows Model 1, with bars of ± 1 standard deviations of the mean

empirical standard deviations are listed in Supplementary Table ST13. We note that the between-trial heterogeneity appears to have little impact on ESSs. Additionally, if we consider the case where we have existing data from 20 studies, with each study containing 170 subjects, it is interesting to note that for low between-trial heterogeneity the ESS is 23.83; this demonstrates the downweighting of the existing data due to the bias-adjusted meta-analysis. In particular, the expert's uncertainty captured by the variability in the elicited prior for the bias parameters results in a large downweighting of the existing data. Were the expert more confident in their answers to the elicitation questions, less downweighting would be observed and a larger ESS would be obtained.

For moderate between-trial heterogeneity and 30 subjects per trial, Figure 6.4 shows how the ESS changes as the variance-covariance matrix of the bias prior (E1) is scaled; the means and correlations of the bias prior remain the same but

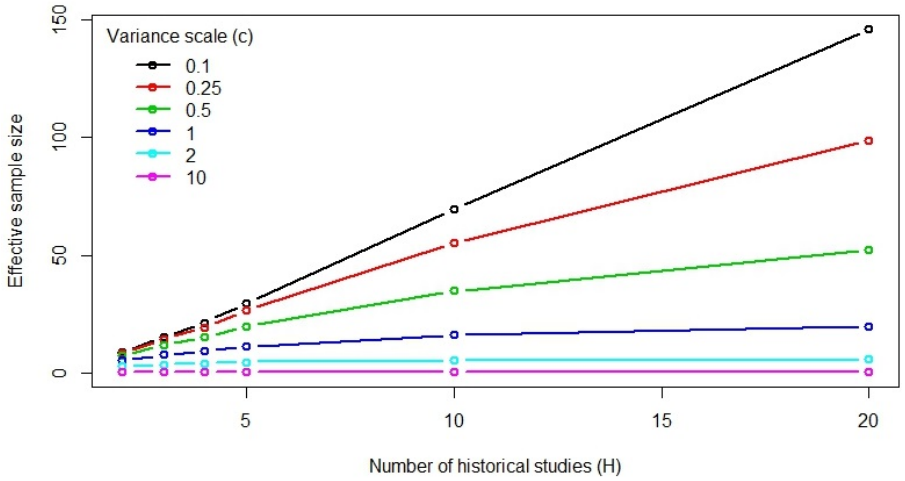


Figure 6.4: Average ESS when the true PK-PD relationships in adults and adolescents follows Model 1 and bias prior E1 has the covariance matrix scaled by 0.1, 0.25, 0.5, 1, 2, or 10.

the variance is scaled by a factor c. Comparing the black line ($c = 0.1$) and the pink line ($c = 10$), we see that the ESS of the existing data increases as prior uncertainty about the external biases decreases.

6.6.3 Prior probability that PK-PD curves are similar in adults and younger children

First we look at how the prior probability of similar PK-PD curves in adults and younger children (hereafter referred to as the extrapolation probability) varies with the true PK-PD relationship in adults and adolescents. Supplementary tables B.9 - B.12 in Appendix B.3 present the means and empirical standard deviations of extrapolation probabilities for a range of scenarios. Figures 6.5(b), 6.6(b), and 6.7(b) illustrate PK-PD relationships in adults, adolescents and younger children under simulation models S1, S3 and S4 and bias prior E1. Figures 6.5(a), 6.6(a) and 6.7(a) illustrate how our confidence in the extrapolation assumption changes as differences between adult and adolescent PK-PD relationships increase from none (model S1), to moderate (S3) to large (S4). A general trend seen in our results is that larger prior probabilities of extrapolation criteria (6.11)-(6.12) are

recorded in scenarios where the true PK-PD curves in adults and adolescents are more closely aligned. This makes intuitive sense when working with bias prior E1, since this prior is consistent with the best guess that PK-PD relationships in adolescents and younger children are identical. For example, comparing Figure 6.5(a) with Figure 6.6(a), we see that when there are moderate differences between the true adult and adolescent PK-PD curves (model S3), prior extrapolation probabilities are lower than under model S1. This is because observed differences between adults and adolescents increase our scepticism that PK-PD curves will be similar in adults and younger children.

Scaling the variance matrix of the bias prior influences the extrapolation probability. Under model S1 and bias prior E1, the probability of extrapolation reaches a maximum value of 0.572 when the between-trial heterogeneity is low and data are available from $H = 20$ historical studies, each having recruited 170 subjects. This maximum probability is far from 1. Lines representing cases when the bias prior variance matrix is scaled by a factor of 0.5, 0.01 or 0 show that if uncertainty about external biases were to be significantly reduced, the prior extrapolation probability would increase. For example, when the bias prior variance matrix is multiplied by 0.01, the extrapolation probability does tend towards 1 as H increases. For bias prior E1, a scale factor of 0 would reflect the opinion that we are certain that differences between adult and adolescent PK-PD curves reflect differences between curves for adults and younger children.

There is a question of whether it is plausible that an expert would be confident enough in their beliefs for us to attain a high prior extrapolation probability. Suppose a probability of 0.8 would be sufficient to support a complete extrapolation strategy. Looking at Figure 6.5(a), we see that under model S1 with low between-trial heterogeneity and 170 subjects per trial, if we scale the bias prior variance matrix by 0.5, the probability of extrapolation reaches 0.8 when the number of ex-

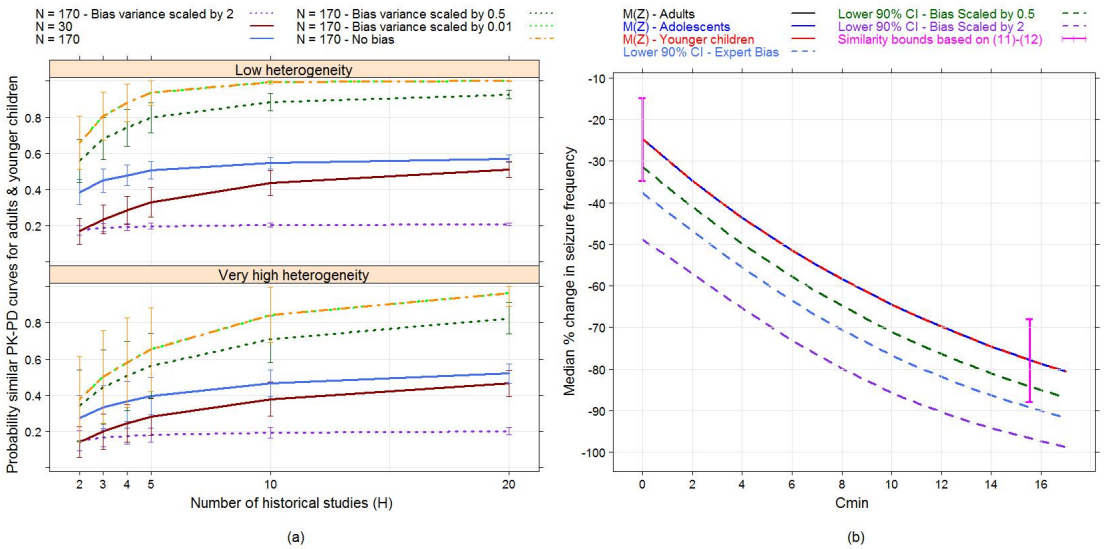


Figure 6.5: (a) Average probability of the extrapolation assumption under bias prior E1, with bars of ± 1 standard deviations of the mean, when the true PK-PD relationships in adults and adolescents follows model S1 (when no differences between PK-PD curves in adults and adolescents). (b) Median PK-PD curves for adults, adolescents and younger children following model S1, along with lower bounds of 90% empirical credibility intervals for the median response in younger children resulting from the expert bias prior E1 when its variance matrix is unscaled; or scaled by a factor of 0.5 or 2. Credibility intervals are calculated conditioning on the true values of the adult and adolescent PK-PD parameters. Similarity bounds at placebo and the EC_{90} given by criteria (6.11) and (6.12) are also shown.

isting studies exceeds five. Looking at Figure 6.5(b), we can see what this scaling factor would correspond to in terms of the level of confidence an expert must have in the location of the PK-PD curve in younger children. We speculate that experts could possess this level of confidence in practice. Furthermore, from Figure 6.5(a) we see that with at least 10 existing studies, low between-trial heterogeneity and 170 subjects per study, if an expert's confidence in the relevance of the existing data was consistent with bias prior E1 with variance matrix scaled by a factor between 0.5 and 1, then the prior probability of extrapolation would still exceed 0.8; with enough existing data and strong, but still feasible, expert opinion, a high probability of extrapolation is plausible.

Focusing on model S3, when there is low between-trial heterogeneity we notice that smaller variances for the bias parameters lead to smaller extrapolation probabilities

due to our increased confidence in the relevance of differences observed between adults and adolescents. In contrast, under very high levels of between-trial heterogeneity, smaller bias prior variances lead to larger extrapolation probabilities. We see from Figure 6.7(a) (model S4) that when there are large differences between the PK-PD relationships for adults and adolescents, the prior extrapolation probability approaches 0 as H increases since we have greater precision to identify differences in the source population. However, in this setting, the extrapolation probability increases with the variances of the bias parameters. This is because as our uncertainty about the relevance of the existing data increases, at each dose our 90% prior credibility interval for $\mathcal{M}(Z)$ in younger children widens and we place more probability mass on values consistent with criteria (6.11)-(6.12), as can be seen in Figure 6.7(b) when the variance matrix of the bias prior is scaled by a factor of two.

Clearly the prior extrapolation probability is heavily influenced by the variability of the elicited bias prior. It seems natural that the prior extrapolation probability should be a combination of quantitative evidence and subjective opinion. For example, if existing data suggest that adult and adolescent curves are similar and experts are confident that the relationship between these curves reflects that between adults and younger children, then complete extrapolation of efficacy data from adults to younger children should be recommended. However, as can be seen from Figure 6.5(a), if there is uncertainty about the relevance of the existing data, the prior probability of extrapolation should be lowered to reflect this, perhaps to the extent where we would caution against adopting a complete extrapolation strategy in favour of collecting a reduced amount of PK-PD data in the youngest age group. In a similar manner, if we observe differences between adult and adolescent PK-PD curves, it seems appropriate that the extrapolation probability should increase slightly as the variance of the bias parameters increases since the existing data are discounted from our decision making, as shown in Figure 6.7(a).

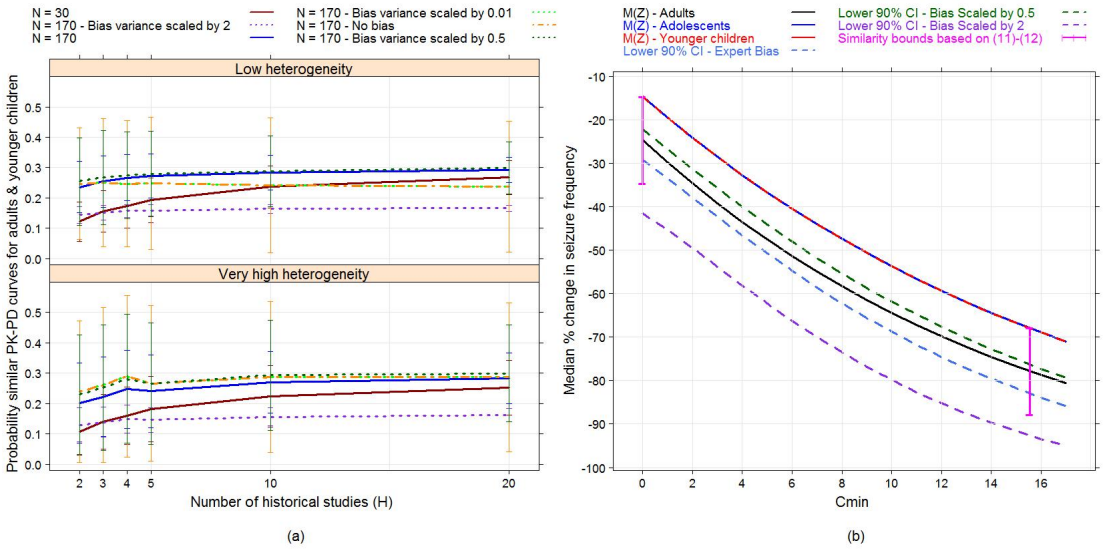


Figure 6.6: (a) Average prior probability that PK-PD relationships in adults and younger children satisfy criteria (6.11)-(6.12), with bars of $\pm$ 1 standard deviations from the mean. The true PK-PD relationships in adults and adolescents follow model S3. (b) Median PK-PD curves for adults, adolescents and younger children following model S3, along with lower bounds of 90% empirical credibility intervals for the median response in younger children resulting from the expert bias prior E1 when its variance matrix is unscaled; or scaled by a factor of 0.5 or 2. Credibility intervals are calculated conditioning on the true values of the adult and adolescent PK-PD parameters. Similarity bounds at placebo and the EC_{90} given by criteria (6.11) and (6.12) are also shown.

Comparing results derived under bias prior E1 setting $N_1 = \ldots = N_H = N$, we see that the extrapolation probability increases as N increases from 30 to 170, although differences diminish with larger values of H. This trend occurs under models S1 and S3 when differences between adults and adolescents are small enough to satisfy criteria (6.11)-(6.12) and our best prior guess is that external biases are equal to 0. It arises because as the study-specific sample size and H increase, we are able to estimate the population mean effects of age and age by exposure interaction on the PD response more accurately. For a similar reason, from Figure 6.7(a) based on model S4, for configurations of the bias prior with small variances the prior extrapolation probability decreases with H as we are able to deduce with more certainty that there are important differences between the adult and adolescent PK-PD curves and because we are confident these differences reflect those between

adults and younger children. However, Figure 6.7(a) shows that in scenarios where we have 30 subjects per study, or 170 subjects but the bias prior variance matrix is scaled by a factor of 2, the extrapolation probability increases slightly with H.

Prior probabilities of extrapolation under models S3, S5 and S6 are provided in Appendix B.2 as Figures B.9–B.11. Similar patterns are seen in the results generated under models S2 and S5 as under model S1, although prior probabilities tend to be lower overall reflecting increased differences between adult and adolescent PK-PD curves. A similar comment applies to results generated under models S4 and S6, and the extrapolation probability is even lower under model S6 when scaling the bias prior variance matrix by 2.

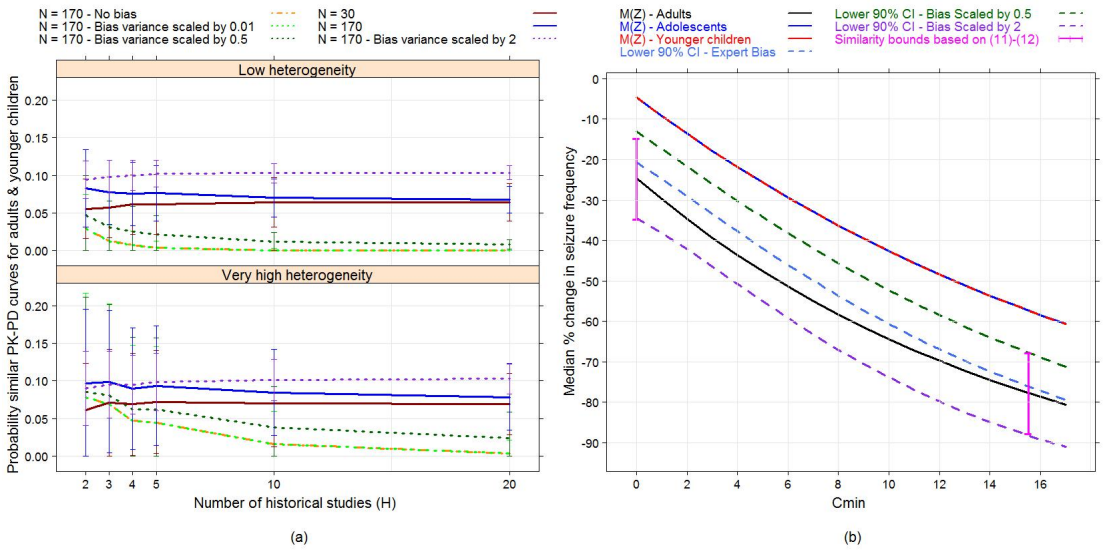


Figure 6.7: (a) Average probability of the extrapolation assumption under bias prior E1, with bars of $\pm$ 1 standard deviations of the mean, when the true PK-PD relationships in adults and adolescents follows model S4 (when clearly large differences between PK-PD curves in adults and adolescents). (b) Median PK-PD curves for adults, adolescents and younger children following model S4, along with lower bounds of 90% empirical credibility intervals for the median response in younger children resulting from the expert bias prior E1 when its variance matrix is unscaled; or scaled by a factor of 0.5 or 2. Credibility intervals are calculated conditioning on the true values of the adult and adolescent PK-PD parameters. Similarity bounds at placebo and the EC_{90} given by criteria (6.11) and (6.12) are also shown.

We have repeated our investigations using bias prior E2 in place of prior E1.

Prior E2 is consistent with the belief that the PK-PD curve in younger children lies above that for adolescents (indicating a worse average response). Comparing results generated under the two priors, we see that the probability of extrapolation under E2 is lower in all scenarios, demonstrating that it is not only an expert's uncertainty about external biases which influences the probability of extrapolation but also the expert's opinion on the direction of differences between PK-PD curves in adolescents and younger children. Bias prior E2 has the same covariance matrix as prior E1 meaning that ESSs do not vary substantially since the ESS is driven by the variability in the meta-analysis results and the prior variance of the bias parameters but not the location of the bias prior.

6.7 Discussion

This chapter proposes a quantitative framework for using existing pharmacological data to inform our understanding of likely differences between PK-PD relationships in adults and younger children. The prior probability of acceptably small differences between these relationships is used to inform a decision of whether to perform a complete or partial extrapolation of adult efficacy data to younger children. The elicitation of expert opinion is an essential yet challenging aspect of our approach. Elicitation questions must be precisely tailored to the disease area and comparison of interest, meaning that close collaborations with clinical colleagues are necessary to ensure we construct an appropriate elicitation scheme. This is especially important given the influence that the elicited bias prior has on the prior extrapolation probability. Currently, we propose that extrapolation probabilities in excess of 0.8 or 0.9 would support a decision to adopt a complete extrapolation strategy, although further work will explore whether the choice of this cut-off can be refined and formalised through use of a decision-theoretic argument. A decision theoretic approach would consider various risks and costs including: the risk to children of incorrectly adopting a complete extrapolation strategy when in fact

PK-PD curves are different in adults and younger children, in which case dosing younger children to match exposures seen in adults would lead to the former age group being treated at toxic or ineffective doses; the costs to patients and the sponsor of failing to perform a complete extrapolation when this is appropriate, in which case younger children are recruited into an unnecessary PK-PD study and the wider population outside this study is delayed access to the new medicine.

A low prior probability of similar PK-PD curves in adults and younger children could be interpreted in different ways. For instance, the low probability could be due to marked uncertainty about the magnitude of external biases or high levels of between-study heterogeneity in the existing data. Another possibility is that the extrapolation probability is low because there is a clear signal that there exist large differences between PK-PD curves in different age groups. When it is unclear whether a complete extrapolation strategy should be adopted or not, one could use an expected value of information analysis[189,190] to quantify the value, in terms of improved decision making, of collecting varying numbers of additional PK-PD data in younger children given the risks outlined above of incorrectly making a complete extrapolation of efficacy data from adults to younger children or missing an opportunity to do so when this is appropriate. Further work will explore using a decision theoretic approach to set the extrapolation decision rule.

When performing the bias-adjusted meta-analysis upon which the prior probability of extrapolation is based, it is essential that included studies should have been identified through a process of systematic review according to a pre-specified protocol.[191,192] The eligibility criteria for the systematic review should support inclusion of studies that were not performed in the target population but are considered relevant on the basis of our current understanding of the disease of interest, the drug's mechanism of action and our understanding of the effects of baseline covariates.

Our current approach assumes exposure-response relationships can be captured by models which represent age as a categorical variable, that is, assuming there are no important differences within an age group. While this assumption will never hold exactly, we do expect it to hold approximately for suitably defined age groups: if important differences were expected to occur within an age group (for example, in the setting of our motivating example, if children aged from 2 to 4 years were expected to respond differently to those aged from 5 to 11 years), then a more suitable approach would be to consider each homogeneous age group in turn and select an extrapolation strategy for each by application of the methods described in Sections 6.3 and 6.4. While the motivating example for this work has been extrapolating across age groups, a similar framework could be used to inform the extrapolation of efficacy data across ethnic groups or geographic regions, where subgroups in this setting are naturally discrete.

In this chapter, we have considered the relatively simple case of linear PK-PD models. It would be interesting to extend our approach to more complex cases such as non-linear PK-PD models. With careful thought this could be possible, although one would need to consider: a) how to parameterise the more complex PK-PD models for adults, adolescents and younger children; b) how to represent differences between the various PK-PD relationships and define decision criteria governing extrapolation decisions; c) how one would devise a scheme to elicit opinion on biases affecting parameters governing the similarity of PK-PD relationships. Furthermore, depending on the choice of models for PK-PD relationships and bias parameters, the resulting posterior distributions for key parameters in the extrapolation decision criteria may be more complex, so that larger mixtures of (Normal) distributions may be required to obtain accurate approximations.

6.8 Appendix: OpenBUGS code

```
model {
  for(j in 1:H) {
  for(i in 1:N) {
     mu[i,j] <- gamma_0j[j] + gamma_Ej[j]*X1[i,j] +
           gamma_Aj[j]*X2[i,j] + gamma_Ij[j]*X1[i,j]*X2[i,j]
     y[i,j] ~ dnorm(mu[i,j], tau)
  }
    gamma_0j[j] ~ dnorm(0,0.01)
    gamma_Ej[j] ~ dnorm(0,0.01)
    gamma_Aj[j] ~ dnorm(gamma_A,tau1)
    gamma_Ij[j] ~ dnorm(gamma_I[j], tau2)
    gamma_I[j] <- lambda0 + lambda1*(gamma_Aj[j] - mean(gamma_Aj[]))
  }
  tau ~ dgamma(0.5,0.5)
  tau1 <- 1/pow(sig1,2)
  sig1 ~ dgamma(2.086339,22.88263)
  tau2 <- 1/pow(sig2,2)
  sig2 ~ dgamma(1.135401,79.3581)
  gamma_A ~ dnorm(0,0.01)
  lambda0 ~ dt(0,0.03,3)
  lambda1 ~ dnorm(0,1)
  gamma0mean <- mean(gamma_0j[])
  gammaEmean <- mean(gamma_Ej[])
  gammaImean <- mean(gamma_I[])
}
```

Chapter 7

Exposure-response modelling approaches for determining optimal dosing rules in children

7.1 Chapter background

Our quantitative approach to informing extrapolation decisions in the previous chapter made the assumption of homogenous age groups across adults, adolescents and younger children. Homogeneity of age groups in terms of exposure-response (E-R) relationships may not necessarily be entirely appropriate. As such, approaches to quantify how E-R model parameters change over age and a way to utilise this information to identify distinct age groups for practical dosing rules would be of great use.

7.2 Introduction

Children of different ages given a new medicine may be characterised by different dose-exposure and E-R relationships due to age related differences in growth, de-

velopment and physiological differences.[23] Several regulatory guidance documents have suggested general age groupings, such as the International Conference on Harmonisation E11 document,[23] which suggests one possible categorisation: 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, depending on region). The National Institute of Child Health and Human Development (NICHD) guideline, suggests similar age groups, but with extra splits at 1 and 6 years. This chapter aims to estimate the E-R relationship in children over a chosen age range and to identify age groupings which define practical and effective dosing rules.

An understanding of how the E-R relationship of a drug varies with age will inform whether and how we leverage adult data to support drug development in children. Hampson et al.[33] reviewed paediatric investigation plans (PIPs) and found that it was common to plan to identify paediatric doses by matching exposures in children with target adult exposures. This is an appropriate dose-finding strategy if E-R relationships are similar in adults and children. An assumption of similar E-R relationships might be justified for some paediatric subgroups, but not others. For example, Takahashi et al.[174] concluded that whilst pubertal (12 to 18 years) and adult patients had similar response to long-term warfarin therapy, differences existed in the pharmacodynamic response between pre-pubertal (1 to 11 years) patients versus pubertal and adult patients. If E-R relationships can be assumed to be similar across age groups, it may be appropriate to make a complete extrapolation of efficacy data from one age group to another, so that only dose-exposure data are needed in the unstudied age group to identify doses producing exposures efficacious in the studied age group.[33,36] However, if E-R relationships cannot be considered similar, a partial extrapolation approach[36] may be considered, where dose-exposure and E-R data may be accrued in specified age groups to establish differences in E-R relationships. Parkinson et al.[193] developed a sig-

moid Emax model for the relationship between dapagliflozin exposure and urinary glucose excretion for adult and paediatric patents with type 2 diabetes mellitus. After accounting for significant covariates (e.g. sex, race, baseline fasting plasma glucose), further covariates were included for paediatric patients which failed to improve model fit. The authors took this as evidence that adult and paediatric patients had similar E-R relationships. Earp et al.[194] used E-R modelling and exposure matching analyses to estimate paediatric doses for esomeprazole for the treatment of gastroesophageal reflux disease. The authors modelled E-R relationships of intragastric pH for adults and children separately and concluded similarity of E-R based on a visual inspection of fitted E-R relationships. In this chapter, a more quantitative approach to evaluating differences between E-R relationships is taken using sophisticated modelling approaches.

Age groups characterised by different E-R relationships can be considered as distinct subgroups. Lipkovich et al.[195] review methods for the identification and analysis of subgroups in clinical trials. Ondra et al.[196] reviewed methods for designing or analysing clinical trials that aim to investigate differences in treatment effects across subgroups. In this chapter, we consider two model-based approaches to quantify how E-R model parameters vary over a continuous age range: Bayesian penalised B-splines,[197] and model-based recursive partitioning (MOB)[3,198] which is used to fit model-based trees to bootstrapped samples of the E-R data. Based on estimates of how E-R model parameters vary with age, we propose an approach to identify the age groups and exposure levels that define a dosing rule which is optimal for targeting a certain level of response; definition of the dosing rule is then completed by using the exposure levels and estimated dose-exposure relationship to make dosing recommendations for each age group. The estimated dose-exposure relationship is not considered in this thesis. Thomas et al. use MOB to estimate patient subgroups with different dose-response curves, and apply this method to data from a dose-finding trial. In this chapter, we focus on estimating

age groups with different E-R relationships since in practice, when seeking to relate dose to response, a two-step process relating dose to exposure then exposure to response is often adopted. For example, the ICH E4 guidance[199] states that E-R information can help identify a range of concentrations which likely lead to a satisfactory response, which can in turn inform dose selection. While parameters of the dose-exposure relationship are expected to depend on age, for some medicines parameters of the E-R relationship are expected to remain stable across age groups. In such cases, the two-step modelling process can be advantageous because it enables separate modelling of the dose-exposure and E-R relationships, which allows for changes due to age to be captured in each relationship separately. In a simulation study to compare the performance of the two-step and single stage (dose-response) approaches to dose finding, Berges and Chen[200] found that the two-step approach resulted in more precise E-R model parameter estimation and more accurate dose selection, though the authors show that the gain with the two-step approach depends on properties of the drug, trial design features and the response level being targeted. Hsu[201] found that in scenarios with increased intrinsic PK variability, E-R modelling has advantages for dose selection over dose-response modelling, provided measurement error for exposures is small. As an example of a two-stage approach to selecting a dosing rule, Schoemaker et al.[202] developed a population PK model to describe the relationship between brivaracetam dose and plasma concentration in adults with partial onset seizures, and a population pharmacokinetic-pharmacodynamic model to describe the relationship between brivaracetam plasma concentration and daily seizure counts. The authors then simulated from these models to estimate the relationship between dose and response, enabling them to identify a dose range producing the maximum response.

This chapter proceeds as follows. Section 7.3 gives a motivating example while Section 7.4 defines two E-R models. In Section 7.5, we introduce the methods that will be used to estimate parameters of E-R relationships. Section 7.6 proposes an

approach for using fitted E-R models to identify practical dosing rules for children. We use simulation to evaluate the performance of E-R modelling approaches and the operating characteristics of the dosing rule algorithm. The design of the simulation study is described in Section 7.7 and the results are presented in Section 7.8. An example illustrating how the E-R modelling approaches can be applied to non-linear models is given in Section 7.9. The chapter concludes with a discussion in Section 7.10.

7.3 Motivating example

We motivate the work that follows by considering the development of epilepsy medicines for paediatric patients with partial onset seizures. Girgis et al.[166] study both monotherapy and adjunctive therapy with the anti-epileptic drug topiramate, whilst Nedelman et al.[38] consider adjunctive therapy with oxcarbazepine. For adjunctive therapy, Girgis et al.[166] and Nedelman et al.[38] take response, $Y = \log\{Z + 110\}$, to be the log-transformed percent change from baseline in seizure frequency (where Z is the percent change from baseline in seizure frequency). The response, Y, is assumed to be normally distributed and a linear function of exposure, measured by the average steady-state trough concentration (C_{min}). Girgis et al.[166] and Nedelman et al.[38] evaluate the similarity of E-R relationships in adults and children on adjunctive therapy with the aim of justifying the use of extrapolation to support the approval of monotherapy in children. The models the authors use and parameter estimates provided by Girgis et al.[166] will be used to inform the design of realistic simulation scenarios.

7.4 Exposure-Response models

We start by considering a linear model for the relationship between exposure and response. Suppose E-R data are available from a single study which recruited

children aged 0 to 18 years and let Y_i represent the response of subject i, for $i = 1, \ldots, N$. If the E-R relationship does not depend upon age, we could model it as:

$$Y_i = \gamma_0 + \sum_{p=1}^{P} \gamma_p x_{pi} + \gamma_C C_i + \epsilon_i,$$

where C_i is a measure of drug exposure (such as C_{min}), $x_{1i}, \ldots, x_{Pi}$ are other covariates influencing response (such as body weight), and $\epsilon_i \sim N(0, \sigma^2)$ is a random error term. We consider the situation where the E-R relationship may differ between age groups, that is, γ_0 and γ_C are functions of age (A):

$$Y_i = \gamma_0(A_i) + \sum_{p=1}^{P} \gamma_p x_{pi} + \gamma_C(A_i) C_i + \epsilon_i. \tag{7.1}$$

In Section 6.3 we will consider different approaches for parameterising $\gamma_0(A_i)$ and $\gamma_C(A_i)$.

The non-linear sigmoid Emax model is often used to represent the relationship between exposure and response:

$$Y_i = \gamma_0(A_i) + \sum_{p=1}^{P} \gamma_p x_{pi} + \frac{E_{max}(A_i) C_i^{\delta(A_i)}}{EC_{50}(A_i)^{\delta(A_i)} + C_i^{\delta(A_i)}} + \epsilon_i, \tag{7.2}$$

where for subject i, aged A_i years old, $\gamma_0(A_i)$ is the intercept, $E_{max}(A_i)$ is the maximum effect attributable to the drug, $EC_{50}(A_i)$ is the concentration of the drug that produces half of the maximum effect, and $\delta(A_i)$ (the Hill parameter) governs slope steepness. Here, four of the model parameters may potentially depend upon age.

7.5 Estimating the exposure-response relationship

In this Section we describe three E-R modelling approaches that can be applied when we assume that the E-R relationship follows model (7.1) with age-dependent intercept and slope. These methods are: linear regression with categorical covariates for age groups; MOB and partially additive linear model (PALM) trees; and Bayesian penalised B-splines. We also highlight where methods can be applied more generally with non-linear E-R models. A worked example illustrating how each method can be applied to estimate a linear E-R relationship is given in Appendix C.1.

7.5.1 Linear model fit with categorical age covariates

If we knew that the age groups defined by different E-R relationships were, $(a_0 = 0, a_1], (a_1, a_2], \ldots, (a_{H-1}, a_H = 18]$, we could define a linear model for the E-R relationship which adjusts for a categorical age covariate:

$$Y_i = \gamma_0 + \sum_{p=1}^{P} \gamma_p x_{pi} + \gamma_C C_i + \sum_{h=2}^{H} \mathbb{1}_{\mathcal{A}_h}(A_i) \left\{ \gamma_{A,h} + \gamma_{I,h} C_i \right\} + \epsilon_i, \tag{7.3}$$

where $\mathcal{A}_h$ is the interval $(a_{h-1}, a_h]$; $\mathbb{1}_{\mathcal{A}_h}(A_i)$ is an indicator function (1 if $A_i \in \mathcal{A}_h$, 0 otherwise); $\gamma_{A,2}, \ldots, \gamma_{A,H}$ are the main effects of the age groups; and $\gamma_{I,2}, \ldots, \gamma_{I,H}$ are the interactions between age group and exposure. Fitting this model permits estimation of age group specific intercepts and slopes. We include this simple model as a benchmark for comparison with other more complex modelling approaches. Unlike the other methods we consider, this approach requires that age groups be pre-specified rather than estimating them from the data.

7.5.2 MOB and PALM trees

Building on this simple model, MOB allows data to be split into groups based on partitioning variables, with each subgroup characterised by its own parametric model.[198] We implement MOB using age as the only partitioning variable. The MOB algorithm we use comprises the following steps:[198] Fit a parametric model to the dataset, estimating model parameters by minimising the objective function; test for whether the intercept and slope parameters significantly change over age by using a generalized M-fluctuation test,[198,203] which assesses whether the scores of the model systematically deviate from 0 over age; partition the model into two subgroups with respect to age by finding the value of age which minimises an objective function segmented at this age split point; repeat the fitting, testing and splitting procedure in each identified age group until no significant changes are found in the intercept and slope parameters over age within each group. The MOB algorithm[198] can be implemented using the 'mob' function found in the 'partykit' package[2,198] in R.[16] As MOB allows subgroups with any parametric model, non-linear models (such as Emax models) are possible.

PALM trees are a variation of MOB, allowing for global parameters which remain constant across subgroups, although PALM trees are restricted to generalised linear models (GLM) rather than any parametric model.[3] For our linear model example with outcome Y_i and partitioning age variable A_i, PALM trees can contain globally fixed linear effects $\gamma_1, \ldots, \gamma_P$ for covariates $x_{1i}, \ldots, x_{Pi}$ and subgroup-wise varying linear effects $\gamma_0(A_i)$ and $\gamma_C(A_i)$, as in equation (7.1). PALM trees use the MOB algorithm described above to identify age groups with separate GLMs. In order to allow for global parameters which do not change over the partitioned age subgroups, an EM-type algorithm is used. This iterates between estimating the global effects, $\gamma_1, \ldots, \gamma_P$, for the currently estimated PALM tree and estimating the PALM tree (using the above algorithm) for a given set of global effect estimates, $\hat{\gamma}_1, \ldots, \hat{\gamma}_P$. The algorithm can be implemented in R[16] using the 'palmtree'

function found in the 'partykit' package.[2,3] We implement PALM trees with the default tuning parameters, i.e. a significance level of 0.05 and no maximum tree depth. An advantage of tree based methods is the easy to understand output: each final partitioned subgroup of the tree represents an age group, with model parameter estimates given for each group.

We implement MOB and PALM tree approaches using bootstrap aggregating[204] to improve the accuracy and precision of age-specific E-R model parameters and reduce overfitting. The E-R data are bootstrapped and each bootstrap sample is used to fit a MOB or PALM tree. From each bootstrap tree fit, estimates of age-specific model parameters (intercept and slope) can be evaluated for a grid of ages covering the interval [0, 18] years. For each grid point in turn, we then aggregate across the bootstrap samples and, applying linear interpolation to the average age-specific parameter estimates, can thus obtain an estimate of the E-R intercept or slope for any given age. The important aspect to note here is that no parametric assumptions are made about the form of the relationship between each model parameter and age. One disadvantage of this is that these relationships cannot then be easily recorded in a closed form for future reference.

We fit linear E-R models using PALM trees in Section 7.7 because we also consider the case of having an additional global covariate whose effect is independent of age, which we present in Appendix C.3. In Section 7.9, we fit non-linear E-R models using MOB.

7.5.3 Bayesian penalised B-splines

Splines define flexible regression models by joining smooth curves (differentiable at every point) together at knot points.[205] An E-R model parameter that can be written as a smooth function of A, $f(A)$, can be modelled as a spline. Here, we

will consider the penalised B-splines developed by Eilers and Marx.[197] B-splines can be written as a linear combination of B-spline basis functions of degree d, that is, $B_1(A;d),\ldots,B_J(A;d)$:

$$f(A)=\sum_{j=1}^{J}\beta_j B_j(A;d). \tag{7.4}$$

A B-spline basis function of degree d consists of $d+1$ polynomial curves of degree d, each joined in sequence.[197] The degree of the B-spline basis controls how differentiable the spline is and can influence the smoothness of the spline. We implement B-splines of degree 2 as in the examples we have considered we gain little in terms of smoothness for the added complexity of using degree 3 B-splines. We therefore fit linear E-R models defining the intercept and slope as B-splines of degree 2:

$$\gamma_0(A_i)=\sum_{j=1}^{J}\beta_{0j}B_j(A_i;d=2),$$
$$\gamma_C(A_i)=\sum_{j=1}^{J}\beta_{Cj}B_j(A_i;d=2).$$

$J=26$ given our choice of degree and number of knots; five knots equally spaced knots within each of the four ICH E11 age groups (not including pre-term newborn infants), knots at each age group boundary, along with two external knots below age zero and two above age 18. We use the function 'splineDesign' in the R package 'splines'[16] to construct our 26 B-spline basis functions. Further details of how the B-spline basis functions are constructed can be found in Bowman and Evers.[205] Note that for penalised B-splines, Eilers and Marx[197] recommend using equidistant knots and suggest that there are no gains to be made from using unequally spaced knots, as the penalty smooths any sparse areas. However, we specify knots using the prior information on potential age groupings that is contained in the ICH E11 guidance document.[23] By specifying an equal number of knot points on each ICH E11 age group, knots are more densely spread across age ranges where

model parameters are expected to change most rapidly with age.

For penalised B-splines, a roughness penalty is used to control the smoothness of the estimated spline, rather than the choice of knot location and number.[197] In a Bayesian context, penalised B-splines are implemented placing random walk priors on the B-spline coefficients.[205,206] For example, to penalise differences between adjacent B-spline coefficients, first-order random walk priors are used:

$$\begin{aligned} \beta_{0,j}|\beta_{0,j-1} &\sim N(\beta_{0,j-1}, \tau_0^2), \qquad \text{for } j = 2, \ldots, J \\ \beta_{C,j}|\beta_{C,j-1} &\sim N(\beta_{C,j-1}, \tau_C^2), \end{aligned}$$

with $\beta_{0,1} \sim N(0, 100)$ and $\beta_{C,1} \sim N(0, 100)$. This penalises B-spline coefficients by shrinking towards a common constant,[205] which is desirable in our context since we anticipate that there may be age ranges on which a model parameter is fairly stable followed by periods of rapid change. We stipulate diffuse Inverse-Gamma(1, 0.005) priors for τ_0 and τ_C, similar to Lang and Brezger (2004)[206] who place an Inverse-Gamma prior on the variance of the random walk prior. We do not weight τ_0^2 and τ_C^2 by the distance between successive knot points, as suggested by Kneib et al.,[207] to allow larger prior variation when there are larger steps between knots. This is because in our setting, we have purposefully placed knots closer together over age intervals where the most rapid changes with age are anticipated.

We fit the Bayesian penalised B-splines model using Hamiltonian Monte Carlo, calling Stan[14] from R[16] using the RStan package[15] running three chains with a default thinning rate of one for 3000 iterations, 1500 of which are discarded as burn-in samples. Following equation (7.4), the posterior means of the B-spline coefficients are multiplied by the B-spline basis functions to estimate the B-spline for the respective E-R model parameter.

Bayesian penalised B-splines are a very flexible modelling approach, with the capacity to be used to represent the E-R parameters of any parametric model for the E-R relationship. The ability to write the relationship between E-R parameters and age in a simple form, as in equation (7.4), means it is easy to record and share the estimated relationship. However, Bayesian penalised B-spline models can comprise many parameters which can make them computationally expensive to fit.

7.6 Dosing recommendations

7.6.1 Optimisation criterion

We could use the modelling approaches described in Section 7.5 to derive personalised dosing recommendations tailored to a patient's exact age. However, for practical reasons, we seek to identify dosing rules based on wider age subgroups. First, we derive target exposure levels for up to K age groups of children. For practical reasons, K would likely be small, e.g. $K = 5$ in the ICH E11 guideline.[23] When defining the target exposure for each age group, we would like to minimise the difference between the expected response and a target response denoted by Y^*. For the epilepsy example, a 50% change in seizure frequency from baseline would be an appropriate target response, so that $Y^* = \log(-50 + 110)$.

We derive dosing rules assuming the E-R model and parameter estimates are identical to the true model and parameter values. Given a proposed age grouping, let C_k denote the target exposure for the kth age group $(a_{k-1}, a_k]$ needed for a patient aged $(a_{k-1} + a_k)/2$ years to have expected PD response equal to Y^*. Furthermore, define $D_a = |\mathbb{E}[Y \mid A = a, C = C_k] - Y^*|$. One approach would be to find the dosing rule minimising the objective function $F = \int_0^{18} D_a \, da$, where rules minimising F minimise the total absolute difference between the expected response

and Y^*. F weights equally the performance of the dosing rule at every age. This is undesirable in our context since if E-R model parameters do depend on age, it may be reasonable to expect parameters to change rapidly over short intervals (i.e. between 0 - 2 years) and remain fairly stable across the adolescent age range. Minimising F would favour rules which dose most ages effectively, where inaccurate dosing over narrow age intervals would not be seriously penalised. However, our aim is to ensure all ages are dosed appropriately. With this in mind, we choose dosing rules to minimise:

$$G = \frac{1}{a_1^*}\int_0^{a_1^*} D_a \; da \; + \; \frac{1}{a_2^* - a_1^*}\int_{a_1^*}^{a_2^*} D_a \; da \; + \; \ldots \; + \; \frac{1}{a_P^* - a_{P-1}^*}\int_{a_{P-1}^*}^{a_P^*} D_a \; da, \tag{7.5}$$

where $a_1^* < a_2^* < \ldots < a_P^*$ are fixed and pre-specified age boundaries and may be based on regulatory guidance, such as the ICH E11 guideline[23] or the NICHD guideline.[208] We define these boundaries in line with the NICHD guidelines. Finding dosing rules which minimise G means that we give equal weight to the performance of the dosing rule in a number of paediatric age groups considered as our best prior guesses.

7.6.2 Identifying an optimal number of age groups in our dosing rule

Define $\mathbb{a}_K = (a_0, \ldots, a_K)$ as the vector of age boundaries defining the optimal dosing rule with K groups; $\mathcal{C}_K$ as the vector of target exposures; and G_K^* as the minima of G for K age groups. Furthermore, let K_{max} denote the maximum number of age groups considered to be plausible or workable in practice, which would be pre-specified based on feedback from clinicians. We use the following algorithm to define a paediatric dosing rule:

1. Begin with $K = 1$ age group;

2. For K age groups, search over configurations of a_K to find the dosing rule minimising G_K;

3. Save G_K^*, a_K^*, and $\mathcal{C}_K^*$;

4. Repeat steps (2) and (3), successively increasing K by one until $K = K_{max}$.

The minima $G_1^*, ..., G_{K_{max}}^*$ can be compared to see if increasing K always produces a worthwhile increase in the accuracy of the dosing rule. The optimum value of K, balancing the trade-off between complexity and accuracy, is denoted by K^*. In some scenarios, a more automated approach to selecting K^* is possible. In these cases, for each $K = 1, ..., 5$, we propose calculating the percentage difference between G_{K+1}^* and G_K^*. The value of K where the percentage change is less than $c = 25\%$ is taken as K^*. The arbitrary choice of c used here is intended to illustrate one possible approach and will be adopted in the simulation study described in the next section.

7.7 Design of the simulation study

We performed a simulation study to explore the performance of the modelling approaches described in Section 7.5 and the approach of Section 7.6 for defining dosing rules. We consider a range of data generation scenarios for the linear model described in Section 7.4. For the categorical age covariates model, we follow the ICH E11 age groups to fix the age intervals $\mathcal{A}_1 = (0, 28/365], \mathcal{A}_2 = (28/365, 2], \mathcal{A}_3 = (2, 12]$ and $\mathcal{A}_4 = (12, 18]$ of equation (7.3), across all scenarios of the simulation study.

We simulate studies enrolling 25 subjects into each of four ICH E11 age groups, $(0, 28/365], (28/365, 2], (2, 12], (12, 18]$, excluding preterm newborn infants. Within age group $(a_{i-1}, a_i]$, the age of patient i is sampled from a Uniform(a_{i-1}, a_i) distribution. We consider 11 scenarios, as illustrated in Figures 7.1 and 7.2, for how

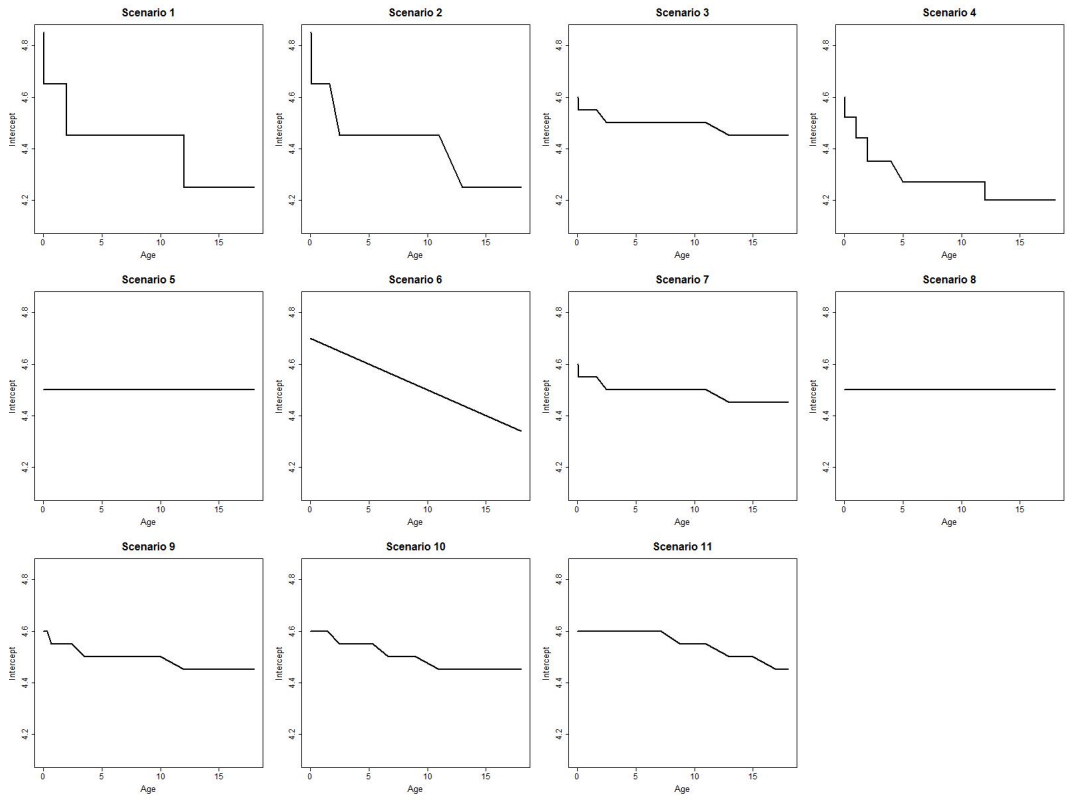


Figure 7.1: Plot showing how the intercept of the E-R model changes with age in simulation scenarios 1-11.

E-R model parameters vary with age. More detail on these scenarios is provided in Supplementary Tables C.1 and C.2, Supplementary Figure C.9 and Appendix C.2. We only consider scenarios where the E-R intercept and slope change monotonically with age, since these differences are most realistic in the context of the epilepsy example.

We measure exposure by C_{min}. Following Wadsworth et al.,[209] we sample $\log(C_{min})$ from a $N(\log(2.94), 0.921)$ distribution, truncating samples above by $\log(17.27)$ to avoid excessively high concentrations. We sample random errors from a $N(0, 0.02)$ distribution. These simulated values are used to generate patient responses, Y_i, according to equation (7.1). We simulate 1000 data sets for each scenario and approach using the statistical software R.[16]

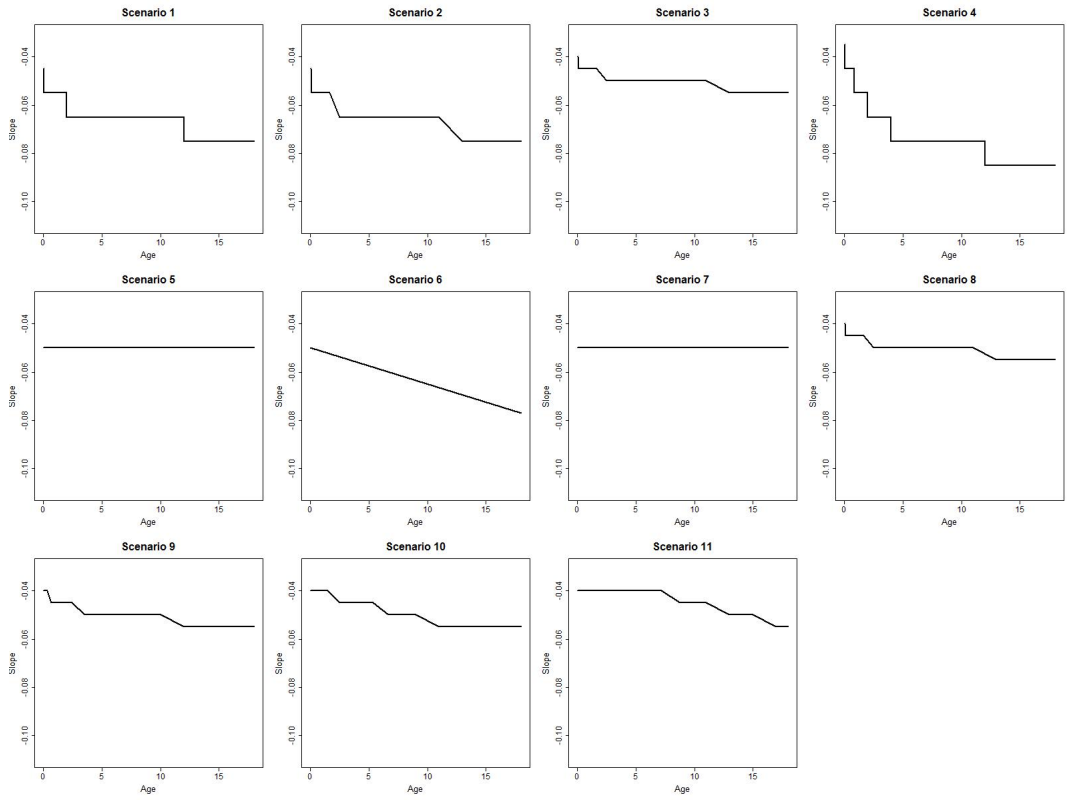


Figure 7.2: Plot showing how the slope of E-R model changes with age in simulation scenarios 1-11.

7.7.1 Evaluating different approaches to modelling the E-R relationship

We use the following measures to compare the modelling approaches described in Section 7.5. Define $\mathcal{A}$ as a grid of $Q = 40000$ equally spaced ages between 0 and 18 years. For each age, $A_q \in \mathcal{A}$, we first measure how well each of the methods has estimated the true intercept and slope parameters. We do this by comparing the true parameters, $\gamma_0(A_q)$ and $\gamma_C(A_q)$, with our estimates of the parameters, $\hat{\gamma}_0^{(m)}(A_q)$ and $\hat{\gamma}_C^{(m)}(A_q)$, based on simulated dataset m, for $m = 1, \ldots, 1000$. For simplicity, henceforth we will refer to a general E-R model parameter $\gamma^{(m)}(A_q)$ and corresponding estimate $\hat{\gamma}^{(m)}(A_q)$.

Let $\widehat{\mathbb{E}[\hat{\gamma}(A_q)]} = \frac{1}{M}\sum_{m=1}^{M} \hat{\gamma}^{(m)}(A_q)$. We evaluate the average absolute bias (AAB), Empirical Standard Deviation (ESD) and Empirical Mean Squared Error (EMSE)

of an estimated parameter at age A_q as:

$$\begin{aligned}
\text{AAB}(\hat{\gamma}(A_q)) &= \frac{1}{M}\sum_{m=1}^{M}\left|\hat{\gamma}^{(m)}(A_q) - \gamma(A_q)\right|, && \text{for } q = 1,\ldots,Q,\\
\text{ESD}(\hat{\gamma}(A_q)) &= \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\gamma}^{(m)}(A_q) - \widehat{\mathbb{E}[\hat{\gamma}(A_q)]}\right)^2}, && \text{for } q = 1,\ldots,Q,\\
\text{EMSE}(\hat{\gamma}(A_q)) &= \frac{1}{M}\sum_{m=1}^{M}\left(\hat{\gamma}^{(m)}(A_q) - \gamma(A_q)\right)^2, && \text{for } q = 1,\ldots,Q.
\end{aligned}$$

Using the grids of AAB, ESD and EMSE values thus produced, we use Simpson's rule[210,211] to calculate the integrated absolute bias, integrated empirical SD and integrated empirical MSE for the E-R model parameter. These metrics can be interpreted as overall measures of the accuracy, precision and MSE of an estimate of the functional relationship between an E-R model parameter and age.

Similarly, let Y_{qj} denote the response at age, A_q, and exposure, $C_j \in \mathcal{C}$, where $\mathcal{C}$ is a grid of $J = 40000$ equally spaced exposures between 0 and 18. We wish to compare the estimated expected response at exposure level C_j, $\hat{\mathbb{E}}^{(m)}[Y_{qj}] = \hat{\gamma}_0^{(m)}(A_q) + \hat{\gamma}_C^{(m)}(A_q)C_j$, with the true expected response at C_j given by $\mathbb{E}[Y_{qj}] = \gamma_0(A_q) + \gamma_C(A_q)C_j$.

Let $\widehat{\mathbb{E}[Y_{qj}]} = \frac{1}{M}\sum_{m=1}^{M}\hat{\mathbb{E}}^{(m)}[Y_{qj}]$. For each $j = 1,\ldots,J$, and $q = 1,\ldots,Q$ calculate:

$$\begin{aligned}
\text{AAB}(\hat{\mathbb{E}}[Y_{qj}]) &= \frac{1}{M}\sum_{m=1}^{M}\left|\hat{\mathbb{E}}^{(m)}[Y_{qj}] - \mathbb{E}[Y_{qj}]\right|,\\
\text{ESD}(\hat{\mathbb{E}}[Y_{qj}]) &= \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\mathbb{E}}^{(m)}[Y_{qj}] - \widehat{\mathbb{E}[Y_{qj}]}\right)^2},\\
\text{EMSE}(\hat{\mathbb{E}}[Y_{qj}]) &= \frac{1}{M}\sum_{m=1}^{M}\left(\hat{\mathbb{E}}^{(m)}[Y_{qj}] - \mathbb{E}[Y_{qj}]\right)^2.
\end{aligned}$$

These evaluations produce $Q \times J$ matrices of values for AAB, ESD, EMSE. For each C_j, for $j = 1,\ldots,J$, we then numerically integrate over age using Simpson's

rule, and then apply Simpson's rule again to integrate over exposure to obtain the integrated absolute bias, integrated empirical SD and integrated empirical MSE for a patient's expected response. These can be interpreted as overall measures of the accuracy, precision and MSE of our estimate of the E-R relationship across a continuum of ages.

7.7.2 Measuring the accuracy of dosing rules

Following the algorithm of Section 7.6, we find dosing rules comprising $K = 1, \ldots, 6$ age groups, with associated target exposures and minimum objective function values. We want to assess the performance of this dosing rule identification process in our simulation study. For each simulated dataset, m, we first take the derived K 'optimal' age groups, $(a_0^{(m)} = 0, a_1^{(m)}], \ldots, (a_{K-1}^{(m)}, a_K^{(m)} = 18]$, and estimates of corresponding target exposure levels, $\hat{C}_1^{(m)}, \ldots, \hat{C}_K^{(m)}$, and evaluate the true expected response, at the target exposure levels, according to the simulation model. That is, at age $A_q \in \mathcal{A}$, we define

$$\hat{\mathbb{E}}^{(m)}\left[Y_{qK}\right] = \sum_{k=1}^{K} \mathbb{1}_{\mathcal{A}_k^{(m)}}(A_q)\left[\gamma_0(A_q) + \gamma_C(A_q)\hat{C}_k^{(m)}\right], \qquad \text{for } q = 1, \ldots, Q,$$

where $\mathcal{A}_k^{(m)}$ is the interval $(a_{k-1}^{(m)}, a_k^{(m)}]$ and $\mathbb{1}_{\mathcal{A}_k^{(m)}}(A_q)$ is an indicator function, which takes the value 1 if $A_q \in \mathcal{A}_k^{(m)}$ and 0 otherwise. This measure is the true expected response, under the simulation model, implied by the estimated dosing rule. Comparing this to the target response will allow us to measure the accuracy of our dosing rule. For each $q = 1, \ldots, Q$ and $K = 1, \ldots, K_{max}$ we find $Y_{qK,\text{diff}}$, the absolute difference between $\hat{\mathbb{E}}^{(m)}\left[Y_{qK}\right]$ and Y^* averaged over the 1000 simulated datasets:

$$Y_{qK,\text{diff}} = \frac{1}{M}\sum_{m=1}^{M}\left|\hat{\mathbb{E}}\left[Y_{qK}^{(m)}\right] - Y^*\right|.$$

This measure can be interpreted as the accuracy of the K-group optimal dosing rule at age A_q. As with Section 7.7.1, we calculate the integral of $Y_{qK,\text{diff}}$ over

age using Simpson's integration. This measure gives an overall measure of the accuracy of the K-group optimal dosing rule and allows us to evaluate how close the true expected response (derived from the simulation model) is to the target response when children are dosed according to the estimated optimal dosing rule. We also consider how many of the simulated datasets would lead us to select a dosing rule with $K^* = 1, \ldots, K_{max}$ groups according to the algorithm described in Section 7.6.2, in order to evaluate the typical complexity of optimal dosing rules and how this varies with the extent of differences between E-R model parameters across age groups.

7.8 Results

Figures 7.3–7.5 plot the integrated absolute bias and integrated empirical SD of E-R model parameter estimators for each modelling approach in each simulation scenario. For estimates obtained fitting Bayesian penalised B-splines, bootstrapped PALM trees, a single PALM tree and the linear model with categorical age covariate, Supplementary Tables C.3–C.6, in Appendix C, present the integrated average absolute bias, empirical SD (as shown in Figures 7.3–7.5) and empirical MSE (not included in the chapter) of the estimated intercepts, slopes and expected response.

Comparing different modelling approaches within a scenario, Figures 7.3–7.5 suggest that, in general, estimates of the functional relationship between the E-R model intercept and slope parameters obtained via Bayesian penalised B-splines are more accurate than estimates obtained using bootstrapped PALM trees. The single PALM tree fit is outperformed by the bootstrapped PALM tree approach in terms of both integrated absolute bias and empirical SD across most scenarios and both parameters, suggesting that bootstrapping is a refinement to the single PALM tree approach. As would be expected, the categorical covariate fit performs best in terms of the accuracy and precision of the estimates of the E-R relationship pa-

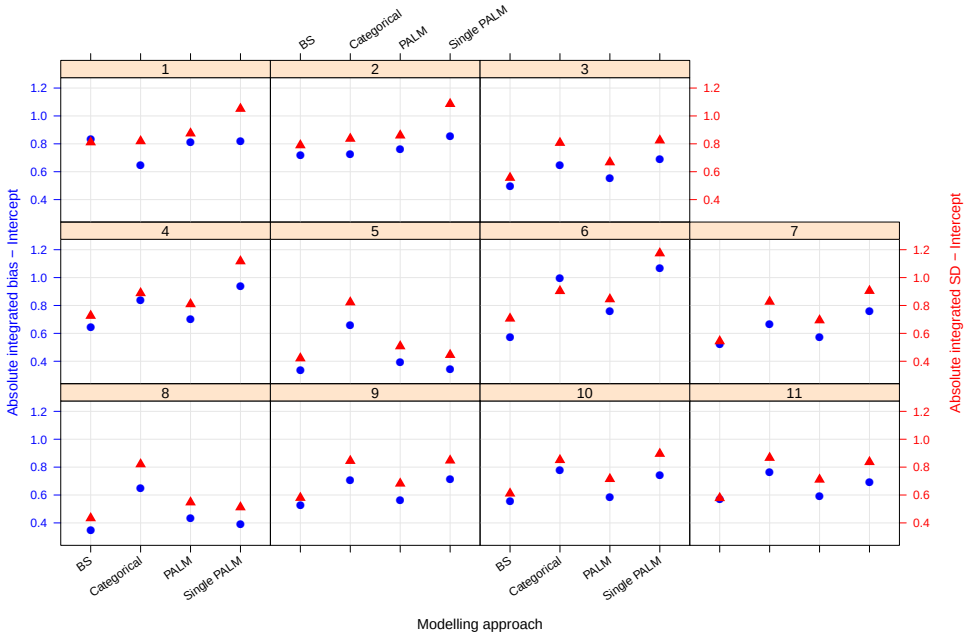


Figure 7.3: Integrated absolute bias (blue circles) and integrated empirical SD (red triangles) for the E-R model intercept. On the horizontal axis, 'BS' refers to the Bayesian penalised B-splines approach, 'Categorical' the linear model adjusted for a categorical age covariate, and 'PALM' and 'singlePALM' label the bootstrapped PALM tree approach and single PALM tree, respectively.

rameters and expected response in scenario 1, where age groups E-R relationships are most distinct and follow the categories suggested by the ICH E11 guidance, excluding pre-term newborns.

Figure 7.6 compares the performance of dosing rules minimising G_K under different values of K, derived from E-R models fitted using different modelling approaches. As the linear model adjusting for a categorical age covariate approach has fixed age groups and the single PALM tree approach estimates specific age groupings, results of the dosing rules optimisation are only presented for the Bayesian penalised B-splines and bootstrapped PALM tree approaches. Figure 7.6 shows that overall both Bayesian penalised B-splines and bootstrapped PALM trees define K-group dosing rules with a similar performance in getting the expected response close to the target response under the true simulation scenarios. In most simulation scenarios, there comes a point at which there is little to be gained in terms of

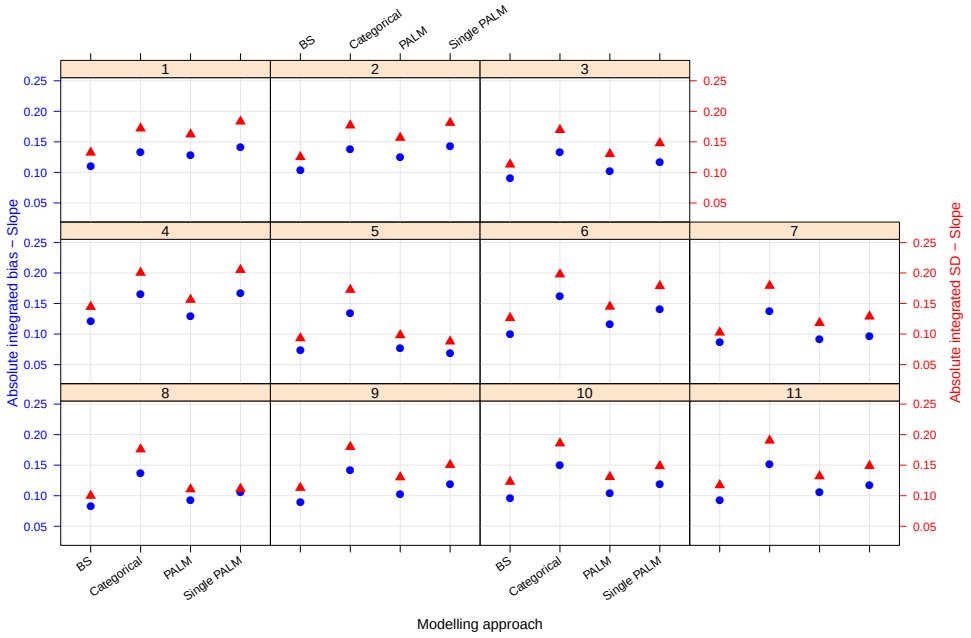


Figure 7.4: Integrated absolute bias (blue circles) and integrated empirical SD (red triangles) for the slope of the E-R model. On the horizontal axis, 'BS' refers to the Bayesian penalised B-splines approach, 'Categorical' the linear model adjusting for a categorical age covariate, and 'PALM' and 'singlePALM' label the bootstrapped PALM tree approach and single PALM tree, respectively.

accuracy by refining the dosing rule further by allowing for additional age groups. In most scenarios, as K increases either the true expected response (under the simulation model and implied by the estimated dosing rule) better matches the target response or there is little difference in the K-group dosing rules of both modelling approaches.

Figure 7.7 shows the percentage of the simulations where the global optimum dosing rule comprises K^* age groups for various values of K^* when dosing rules are derived modelling the E-R relationship using Bayesian penalised B-splines or bootstrapped PALM trees. It is important to note, however, that the 'true' dosing rule age groups determined using the algorithm described in Section 7.6.2 may not necessarily be the same as the true underlying E-R age groups, as if there are large differences in expected response between underlying E-R age groups, a more optimal fit may be found by dosing rules splitting age groups around big changes and

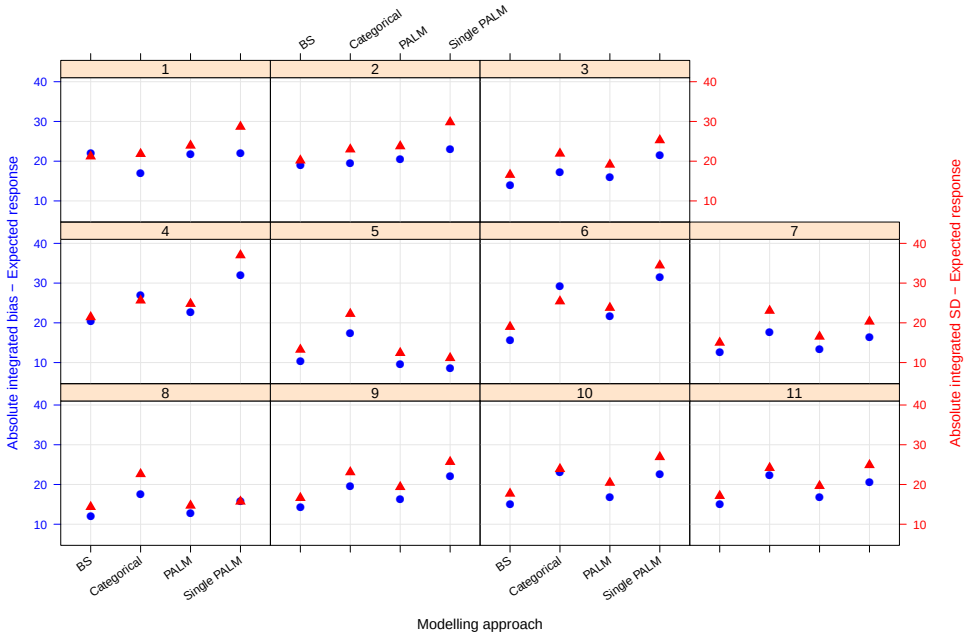


Figure 7.5: Integrated absolute bias (blue circles) and integrated empirical SD (red triangles) for the expected response. On the horizontal axis, 'BS' refers to the Bayesian penalised B-splines approach, 'Categorical' the linear model adjusting for a categorical age covariate, and 'PALM' and 'singlePALM' label the bootstrapped PALM tree approach and single PALM tree, respectively.

combining age groups with smaller changes. However, it is interesting to explore the values of K^* defining the global optimal dosing rules to assess their complexity. Additionally, the complexity of the derived dosing rules will depend on the quantitative threshold used to identify K^* as described in Section 7.7.2; with a different threshold, c, dosing rules with different K^* may be chosen as optimal. Optimal dosing rules minimising G are cautious, forming slightly more age groups than the underlying E-R age groups.

Focusing on Bayesian penalised B-splines, we see from Figure 7.7 that in scenario 1, where larger differences are present in the underlying E-R model parameters, the large majority (81.6%) of simulated datasets would lead to the investigator selecting a global optimum dosing rule with $K^* = 4$, as would the majority (54%) of simulated datasets in scenario 3. This suggests that when underlying E-R relationships across age groups become less distinct, dosing rules with smaller K^*

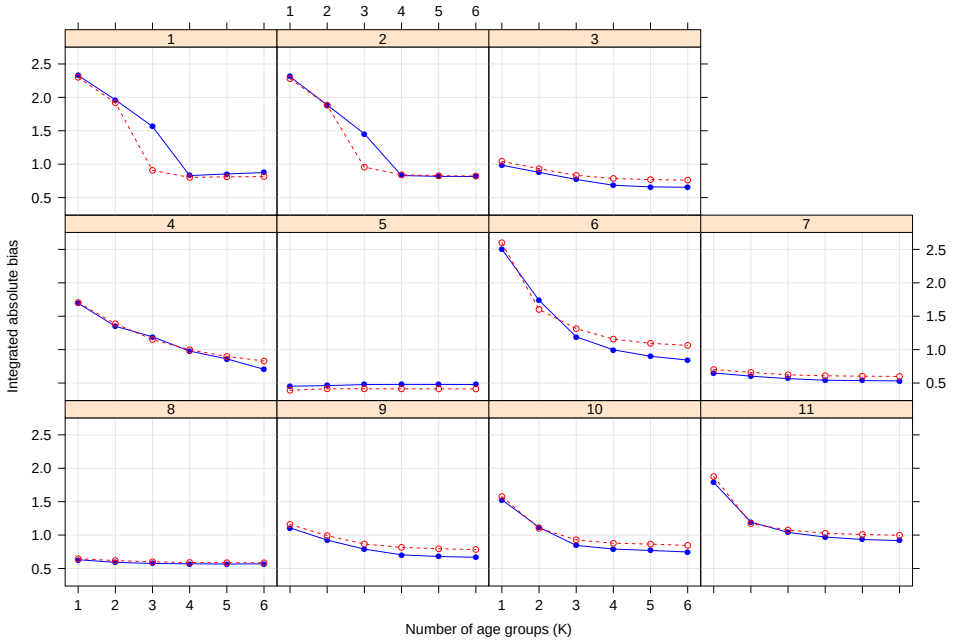


Figure 7.6: Integrated absolute difference between the target response and true expected response when children are dosed according to the K group optimal dosing rule. Results are shown for dosing rules obtained modelling the E-R relationship using Bayesian penalised B-splines (solid blue line) and bootstrapped PALM trees (dashed red line).

are selected. In scenario 4, the majority of simulated datasets would lead to the investigator selecting global optimum dosing rules with $K^* = 4$, although there is a trend to larger K^* compared with other scenarios. In scenario 5, where underlying E-R model parameters do not depend on age a higher percentage of datasets lead to the selection of a dosing rule defined by a smaller value of K^* chosen.

Similar patterns are seen for the bootstrapped PALM trees approach in Figure 7.7. It seems that both bootstrapped PALM trees and the Bayesian penalised B-splines approach are capable of identification of dosing rules with multiple age groups when differences in the underlying E-R relationships across age groups are large, but fewer are identified as differences diminish. For the single PALM tree fit, for scenarios where larger differences are present in the underlying E-R model parameters, as in scenarios 1 and 2, a single PALM tree often identifies dosing rules with four groups; 96.9% and 94.6% would choose four groups, respectively.

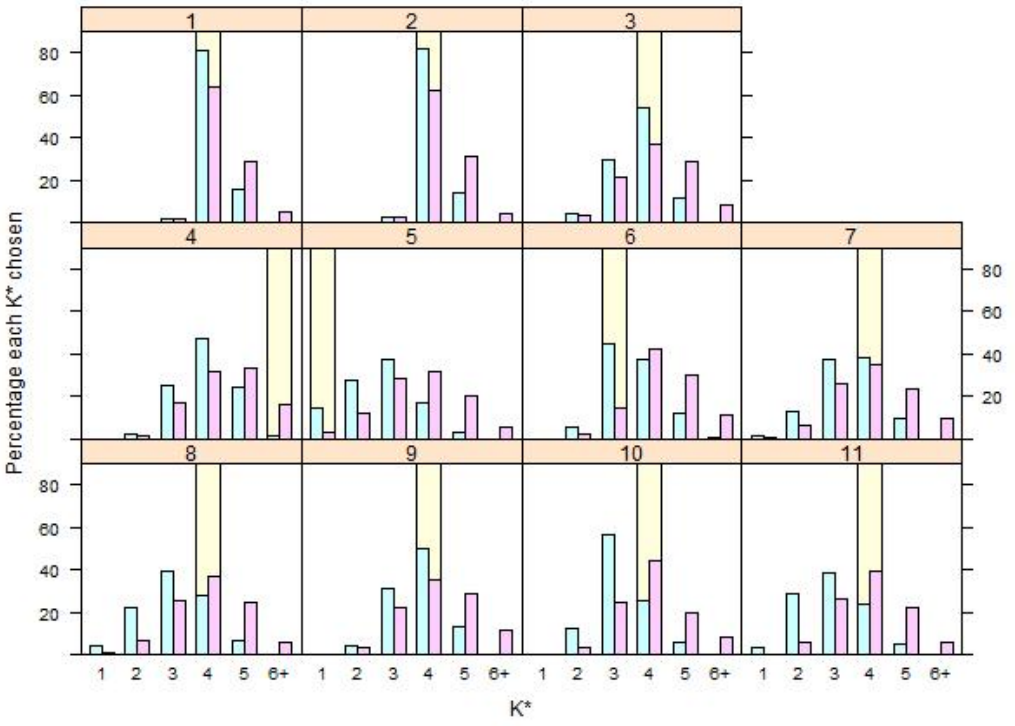


Figure 7.7: Percentage of 1000 simulations in which K^*, the optimal number of age groups in the dosing rule, takes each value shown. K^* is selected according to the algorithm described in Section 7.6.2 for Bayesian penalised B-spline (blue) and bootstrapped PALM tree (pink) approaches. The values of K^* chosen by applying the algorithm in Section 7.6.2 to the true underlying E-R relationships in each scenario are shown by the yellow bars.

It is interesting to note, in not one scenario did a single PALM tree select a dosing rule with $K^* > 4$.

7.9 Extension to Emax model

We consider a simulated example informed by the data presented in Marshall and Kearns,[175] who model the relationship between cyclosporine concentration and in vitro inhibitory effect on peripheral blood monocyte (PBM) proliferation as a sigmoid Emax curve (7.2). We simulate responses for 41 subjects assigned to one of four age groups: 10 infants (0–1 year); 12 children (1–4 years); 9 pre-adolescents (4–12 years); and 10 adults (12–18 years). Data are generated such that for each of the following concentrations of cyclosporine (6.25; 12.5; 25; 50;

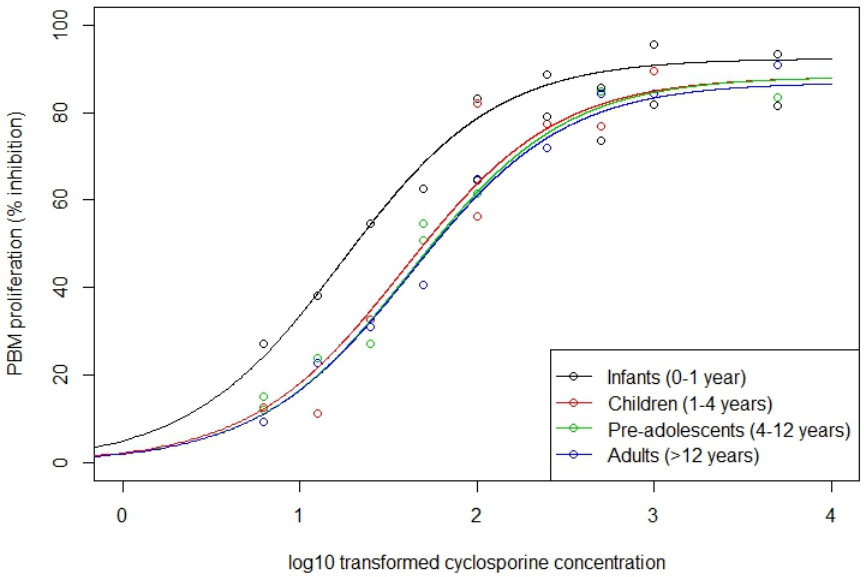


Figure 7.8: Fitted curves of the relationship between log base-10 transformed cyclosporine concentrations and PBM proliferation based on frequentist two parameter Emax model fit for each of the four age groups considered. Fitted curves are the solid lines and the points are simulated data.

100; 250; 500; 1000; and 5000 ng/mL) a patient was recruited from each age group and the remaining patients in each age group (1 infant; 3 children; 1 adult) were randomly assigned a concentration from this set. Within an age group, patients' ages are assumed to follow a uniform distribution. In a deviation from Marshall and Kearns,[175] patient responses are simulated according to a hyperbolic Emax model (setting $\delta(A) \equiv 1$), although we follow the original publication to force a zero intercept ($\gamma_0(A) \equiv 0$). Patient responses are simulated setting the remaining EC50 and Emax model parameters equal to the age group specific parameter estimates provided by Marshall and Kearns,[175] and we assume a normally distributed random error with mean zero and variance 15. We restrict attention to a hyperbolic Emax model because estimates of age group specific Hill parameters are not reported by Marshall and Kearns. Using these simulated data, we fitted a two parameter Emax model separately to each age group. The four fitted curves are shown in Figure 7.8.

7.9.1 Bayesian penalised B-splines

We implement the Bayesian penalised B-splines model by running three Markov chains (as in our simulation study), although now using a thinning rate of 3 and 9000 iterations, 4500 of which are discarded as burn-in samples. We adopt the first-order random walk prior defined in Section 7.5.3 for the penalisation. We found a great deal of sensitivity, in terms of convergence, to the choice of prior for the standard deviation parameters of the random walk priors on the B-spline coefficients of the Emax and EC50 parameters. This sensitivity was found when using the Inverse-Gamma priors as used in Section 7.7. We would advise caution and appropriate checks to ensure posterior results are reliable. One should check a priori the plausible range of values for these standard deviation parameters, which would depend on the magnitude of the Emax and EC50 parameters. Gamma$(2, 1/A)$ priors, with A large (such as $A = 10$) are recommended by Chung et al. (2013)[212] and the Stan user guide[213] as boundary-avoiding priors in hierarchical models for hierarchical standard deviation parameters. Placing Gamma$(2, 0.1)$ priors on these random walk prior standard deviation parameters allowed the two parameter Emax model to fit well to the simulated data shown in Figure 7.8, with the chains converging with Gelman-Rubin convergence diagnostic < 1.011 for all parameters.

Figure 7.9 shows the fitted Bayesian penalised B-spline for the Emax and EC50 parameters over age, showing the median, 2.5th and 97.5th quantiles and the parameter values reported by Marshall and Kearns[175] (green dotted lines). The fitted B-splines for both the EC50 and Emax parameters seem to follow closely to the true underlying parameter values and, as can be seen from Figure 7.10a, the underlying E-R relationships are accurately estimated. Figure 7.10a plots fitted expected response against concentration in each of the four age groups. For each age group, the fitted expected response is calculated by setting the Emax and EC50 parameters at values gained by evaluating the Emax and EC50 fitted B-splines at

the mid-points of each age group.

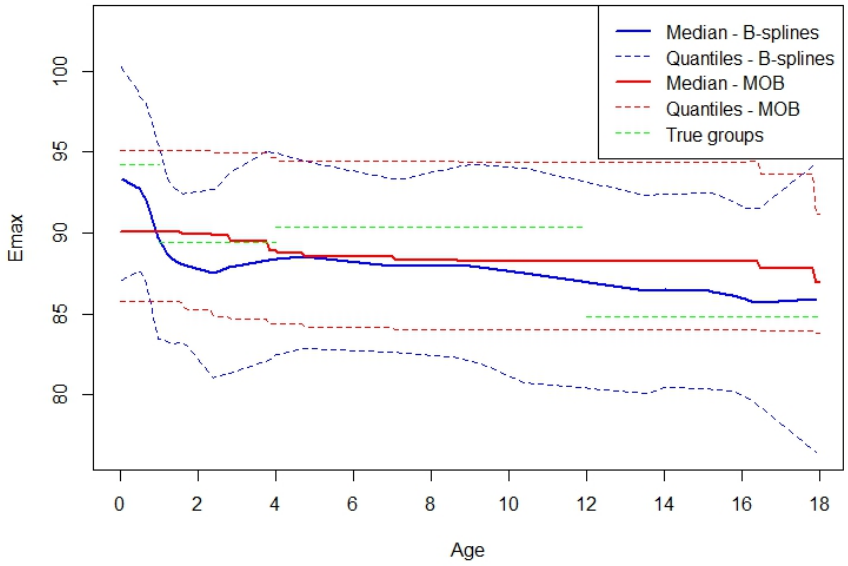
7.9.2 Bootstrapped MOB

To implement the bootstrapped MOB approach, we used the 'mob' function in R with a two parameter Emax model. Otherwise, the approach proceeds exactly as the bootstrapped PALM trees approach described in Section 7.5.2. To incorporate a two parameter Emax model in the 'mob' function, we built on code provided by Thomas and Bornkamp,[214] using the 'nls' function in R[16] to specify the two parameter Emax model.

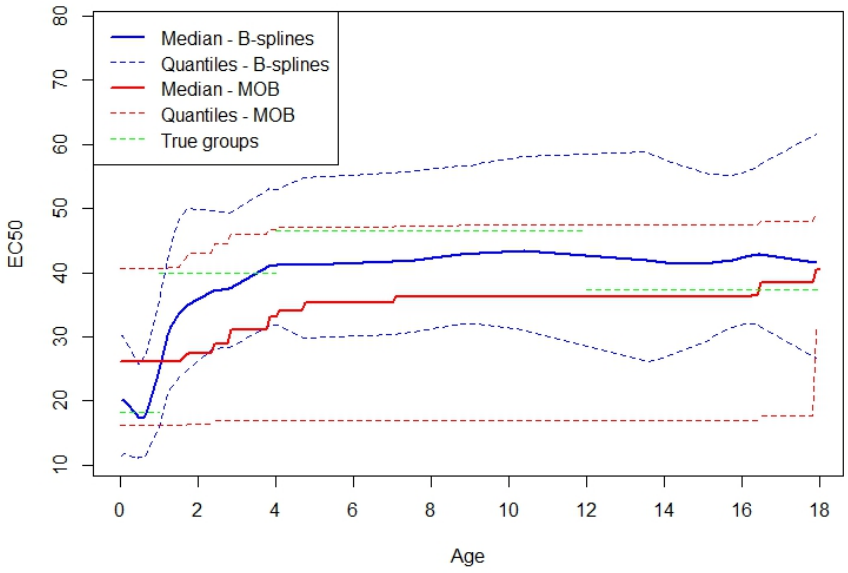
Figure 7.9 shows the fitted bootstrapped MOB to the Emax and EC50 parameters over age, showing the median, 2.5th and 97.5th quantiles over the bootstrapped samples and the true parameter (green dotted lines). The fitted Emax and EC50 parameters do change with age. However, they are both quite far from the true underlying values. When looking at Figure 7.10b we see that the model still fits fairly well to the general shape of the data. However, in comparison to the Bayesian penalised B-splines, Figure 7.10b highlights that there is worse separation between the fitted E-R curves for different age groups across the whole concentration range when using the bootstrapped MOB approach.

7.9.3 Deriving dosing rules

Following the procedure to derive optimal dosing rules described in Section 7.6, Figure 7.11 provides a plot of the objective function values for dosing rules based on both the Bayesian penalised B-splines and bootstrapped MOB approaches. Overall, the bootstrapped MOB approach has lower objective function values than the Bayesian penalised B-splines approach. For both the bootstrapped MOB and Bayesian penalised B-splines approaches two groups would almost certainly be recommended by visual inspection.

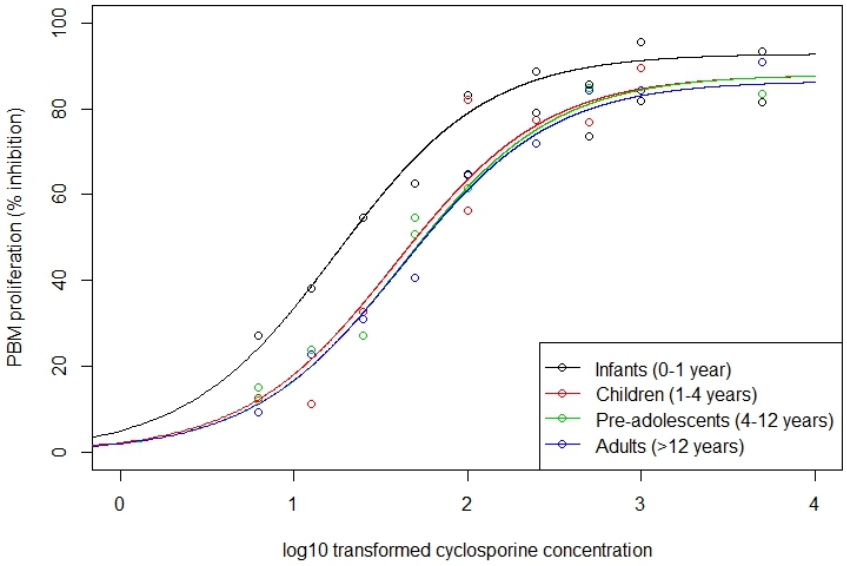


(a) Emax

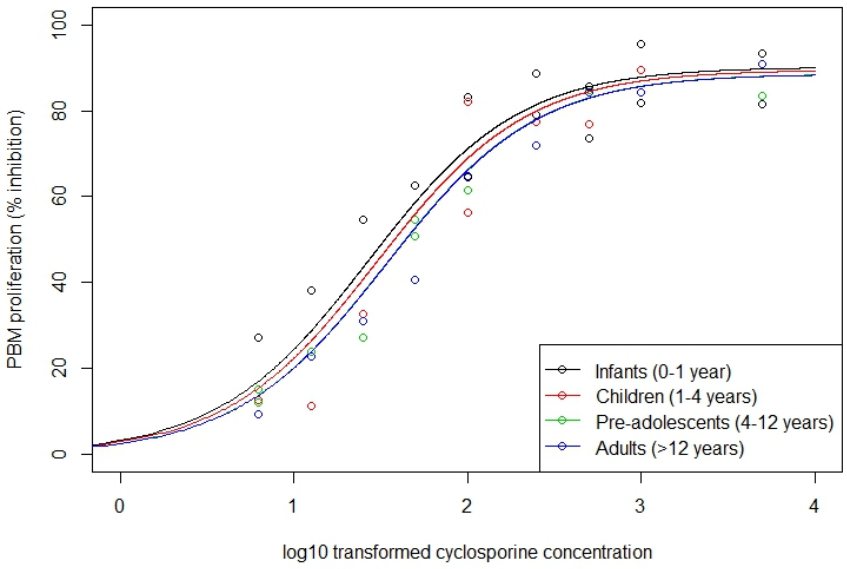


(b) EC50

Figure 7.9: Plots of the Bayesian penalised B-spline and bootstrapped MOB fits of (a) the Emax parameter and (b) the EC50 parameter. The median of each parameter, with 2.5th and 97.5th quantiles, over the 1000 simulated bootstrap samples and true parameter values given by the green dotted lines are also shown.



(a) B-splines



(b) MOB

Figure 7.10: Fitted relationships between log base-10 transformed cyclosporine concentrations and PBM proliferation based on parameter estimates for the four age groups obtained with (a) the Bayesian penalised B-spline approach and (b) the bootstrapped MOB approach.

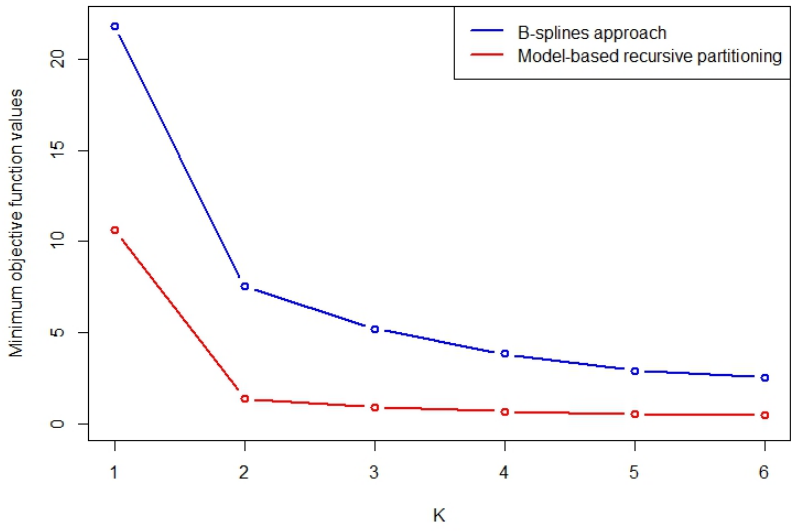


Figure 7.11: Plot of the objective function values from the optimisation procedure used to identify age groups for the Bayesian penalised B-splines approach (blue line) and bootstrapped MOB (red line).

For two age groups, the optimal age groups defining the bootstrapped MOB dosing rule would be 0 to 3.33 years and 3.33 to 18 years, with target exposures of 191.95 and 294.87, respectively. For the optimal age groups defining the Bayesian penalised B-splines dosing rule would be 0 to 0.84 years and 0.84 to 18 years, with target exposures of 110.36 and 446.04, respectively. It is interesting to note how different the dosing rules are for these two methods: the bootstrapped MOB rule stipulates a wider youngest age group, with larger target exposure levels than the Bayesian penalised B-splines rule. However, overall the bootstrapped MOB dosing rule has a lower maximum target exposure than the Bayesian penalised B-splines dosing rule. This seems to be indicative of the larger differences for the E-R relationships found when using the Bayesian penalised B-splines approach.

7.10 Discussion

In this chapter we have considered several approaches to estimating if and how E-R model parameters change over age in order to determine practical dosing rules for distinct paediatric age groups. Our approaches concentrate on the relationship between exposure and response, deriving target exposures for age groups. These target exposures can then be used to identify dosing rules based on a separate relationship between dose and exposure. We do not develop PK models relating dose and exposure in this chapter, many methods exist to do this.[215]

We derive the target exposures of each age group by taking each age group mid-point and finding the exposure level at which the expected response would be equal to the target response. In reality, this may not actually be the optimal exposure level over the whole age group. A more appropriate method may be to search for age specific exposure levels at which the expected response would be equal to the target response for each age over the whole age group, then calculate the expected response over the whole age group using each of these potential target exposures. The target exposure which minimises the absolute difference between the expected response and the target response integrated over the age group would be a more optimum target exposure level for that age group. This approach is computationally more demanding making it unsuitable for our simulation study, but should be quickly implemented for one dataset in practice.

Results of our simulation study of linear model scenarios suggest that the Bayesian penalised B-splines and bootstrapped PALM tree approaches perform similarly in terms of estimating changes in E-R model parameters over age, though the integrated absolute bias and empirical SD is consistently lower in the Bayesian penalised B-splines approach. Plots of the absolute difference between the true expected response implied by proposed target exposures and the target response

also suggest that for most scenarios both approaches perform similarly well, though in some scenarios Bayesian penalised B-splines perform better than bootstrapped PALM trees and vice versa. In fact, the Bayesian penalised B-splines approach appears to outperform all other approaches in most scenarios; only the approach using categorical covariates sometimes has lower integrated absolute bias, and even then, only in some scenarios where the true underlying E-R models contain four age groups matching ICH E11 guidance (as is assumed in the categorical covariates approach).

Chapter 8

Discussion

8.1 Summary

This thesis contains four papers placed in the area of extrapolating between populations, with a particular focus on paediatric clinical trials in an epilepsy context. Chapter 3 detailed a systematic review of methodology for extrapolating between 'source' and 'target' populations and provides a detailed overview of extrapolation approaches. The expert group meeting, discussed in Chapter 5, gave an insight into the opinions of leading UK epilepsy experts regarding extrapolation in the development of drugs to treat epilepsy. This led to the conclusion that these experts felt a partial extrapolation of adult efficacy data and a limited extrapolation of adult safety data would be appropriate for drugs treating paediatric focal epilepsies, and that a single combined drug development program in adults and children 2 years and older could be suitable.

Our discussion with clinical experts regarding the acceptability of extrapolation in epilepsy medicine development motivated our development of a quantitative framework for informing extrapolation decisions in paediatric medicine development given in Chapter 6 and Chapter 7 investigating how PK-PD relationships

change with age in order to define practical dosing rules.

Our quantitative framework to inform extrapolation decisions in the paediatric population highlights the potential gains from utilising existing data, along with expert prior opinion, for the decision making process regarding appropriate levels of extrapolation. The prior probability of an extrapolation assumption provides valuable information which can be leveraged to determine a future course of action with regards to extrapolation: a low probability would be strong evidence that a separate drug development programme is required; a high probability may be sufficient evidence to state that PK-PD relationships between adults, adolescents and younger children are similar enough that complete extrapolation of efficacy data is suitable; whilst a moderate probability could be incorporated into some decision theoretic approach to determine how worthwhile a future study in younger children would be, given existing information.

When trialling the prior elicitation schemes that lead to our final elicitation protocol in Chapter 6, several interesting issues were found. Some of our trial schemes began by asking for histograms of the average response (rather than the expert's best guess at the average response), which experts found to be particularly difficult. Experts had difficulty determining where they wanted to centre the histogram, with all four experts who trialled such a scheme underestimating what they felt the average response should be in children; all elicited histograms were centred closer to zero response than experts expected. This led to the elicitation of the expert's best guess at the average response line first, using this to mark on the histograms where experts would expect to centre. We also had a version of the histograms where experts clicked the plot to add blocks to chosen bins. This was a very simple and more interactive approach which was well received, however, choice of block size and quantity (in terms of how much probability each block was worth) became an issue; one expert wanted to spread their probability further

than was possible given the number of chips available. As such, we changed to allowing experts to place specific quantities in each bin.

Chapter 7 describes model-based approaches to deriving practical dosing rules for paediatric age groups. The Bayesian penalised B-splines and bootstrapped MOB or PALM tree approaches estimate the relationship between PK-PD model parameters and a continuous age variable. This essentially gives us the potential for completely personalised dosing recommendations over the whole age range. In practice, however, such personalised dosing would likely be impossible; the practical difficulties of clinicians needing to have access to an algorithm to give dosing recommendations for a specific age and the investment in terms of time, money and the need for a wide (perhaps even continuous) range of dosing formulations are just a few reasons. Hence, the benefit of forming practical dosing rules for a few distinct age groups, deriving age groups by optimising based on a sum of weighted integrals (over age) of the difference between expected and target response.

8.2 Limitations

As mentioned in Chapter 3, one limitation of the systematic review is that we focussed on just four application areas: paediatric drug development; use of historical data in contemporary clinical trials; bridging trials extrapolating efficacy data between ethnic groups or geographic regions; and the use of short-term data to support inferences on long-term outcomes. As a result, we may have missed other relevant methods, such as in approaches for borrowing information across species.

With our focus group meeting of epilepsy experts, we perhaps could have gained a broader view of the place of extrapolation in paediatric medicine development for epilepsy by speaking to experts with different backgrounds. For example, perhaps

pharmacologists and pharmacometricians could have offered a different perspective on the appropriateness of extrapolation.

One limitation of our approach for the prior elicitation scheme in Chapter 6, is that only certain shapes are possible for the credibility intervals around the expert's opinion regarding the distribution of the average response of younger children. This is due to the BVN assumption for the bias priors; in Section 6.4.2 it can clearly be seen that the variance of the distribution of the expected response for younger children, conditional on the existing adult and adolescent data, is a quadratic in dose where the quadratic coefficient is a variance and can therefore only be a non-negative real number. This stops us from being able to achieve a shape for the credibility interval which could be described to an expert as "more confidence in the response at low and high doses, but more uncertainty at moderate doses"; this is a belief that one clinical collaborator suggested might be how experts would feel, as they suggested that the middle part of the dose range may not be as well explored or understood.

For our Bayesian penalised B-spline approach to estimate the relationship between PK-PD model parameters and age, there can be a great deal of sensitivity to the choice of prior distribution for the random walk prior standard deviation parameter. In practice, one would want to explore sensitivity to the choice of prior to ensure confidence in the results.

8.3 Further work

An interesting extension of our systematic review in Chapter 3 would be to have a single consistent extrapolation example to compare as many of the methods identified as possible. It may be that several examples would be needed within groupings of methods i.e. where specific data types and distributional assumptions

have been made which would conflict between methods. Additionally, given the recent increased interest in extrapolation approaches, such as the EMA extrapolation framework,[34,35] it may be worthwhile to update this review with new methodology proposed in the literature since our paper (beyond 31st January 2014).

With regards to the proposed joint drug development program in Chapter 5, it may be interesting to quantify the potential gains (such as in terms of patient recruitment, cost and development time) from considering joint recruitment of adults and children, with standard practice in paediatric trials of epilepsy.

Future work from the paper in Chapter 6 could be to consider extensions to more complex PK-PD models, such as the non-linear Emax model. However, this could be quite a challenging endeavour as the increased complexity would cascade throughout the entire approach. For example, considering extending to the Emax model, the comparison between adults and younger children would then be based on parameters of the Emax model (intercept, Emax, ED50; Hill parameter if considering Sigmoid Emax model) and corresponding bias parameters. With potentially four bias parameters, a bivariate Normal distribution would no longer be an appropriate model for these parameters. Careful thought and clinical collaboration would be required to develop an elicitation scheme that would best capture expert belief under this more complex model. Even if an appropriate prior elicitation scheme could be developed to elicit expert opinion for the expert's best guess at the average response in younger children, and some measure of uncertainty, it would be more difficult to derive estimates for the underlying bias parameters and would require careful thought on modelling assumptions. Additionally, our prior probability of an extrapolation assumption would likely need to consider more exposure levels than placebo and some other level; with the non-linear nature of the Emax model, matching similar PK-PD relationships between adults and younger children may want to take in to account some moderate exposure level, such as

the EC50, along with placebo and a higher exposure level.

As discussed, the prior probability of the extrapolation assumption holding could be incorporated into some Bayesian decision theoretic approach. Future work to develop such an approach to determine how worthwhile a future study would be, would be valuable. Future studies of adults and younger children could also be conducted according to an adaptive design, updating this prior probability of an extrapolation assumption holding at interim analyses to potentially provide early stopping in favour of extrapolation.

At present, we have simply assumed dose-proportionality and a direct one-to-one relationship between dose and exposure when translating the elicited expert opinion to a prior distribution for the bias parameters of our PK-PD model for a future study of adults and younger children. As mentioned in Chapter 6, existing PK studies may be available to determine whether an assumption of dose proportionality is appropriate and to estimate the proportionality constant, or to assess whether some other relationship holds. If existing data are not available, it could be interesting further work to establish a scheme to elicit the expert opinion of pharmacologists (see the approach of Whitehead et al.[216]) regarding the relationship between dose and exposure.

For future work, a more comprehensive testing of the prior elicitation scheme would be ideal. Clearly, to really demonstrate how well this approach could work, a real-world practical application would be desirable, though perhaps a well designed simulation study with strong clinical input would suffice. At present, although the scheme has been trialled with experts, each testing session only had a short window of time (between 10 to 30 minutes). In practice, one would want to have a full day session with initial elicitation training, detailed presentation of the problem requiring elicitation and plenty of time to run through the elicitation scheme

with individuals. Additionally, if behavioural aggregation of the elicited opinion was chosen to derive a consensus prior, time would need to be allocated for discussion of individual opinion amongst the group of experts to arrive at a final joint bias prior distribution. When elicitation schemes were being trialled, some adult neurologists felt unable to answer the questions due to feeling they did not have enough understanding or experience with the paediatric population to give meaningful answers. This may make behavioural aggregation an ideal approach to deriving consensus opinions; adult experts can still bring the knowledge they do have, whilst taking on board the expertise of their paediatric neurology peers when agreeing to a consensus prior. However, given the amount of information we elicit (especially with the three histograms), this behavioural aggregation could be a very time consuming process. If this approach were not felt to be appropriate due to the complexity of the elicitation scheme, mathematical aggregation could be considered as an alternative.

An extension to the work in Chapter 7, could be to consider other objective functions for the optimisation procedure, perhaps based on eliciting expert opinion on where age groupings are most likely to exist. At present, our objective function uses age groupings suggested by guidance documents. Whilst using guidance documents appears to be a very reasonable approach, it may be possible that age groups could stray from this and if this difference may be suggested by experts in advance, gains could be made when optimising groups to establish dosing rules. Currently, the approaches considered in Chapter 7 derive optimal age groups based on an assumption that the estimated relationship between PK-PD parameter and age is true; future work could perhaps look at incorporating the uncertainty in the estimated relationship.

Further, in our approach we find practical dosing rules by using a continuous optimisation method over age, however, we could instead search for optimal age

groupings over a grid of possible age boundaries. This would allow the incorporation of practical restrictions into the design of the search grid, such as setting a minimum width of any age groups or ensuring age boundaries are in terms of practical units of age (e.g. months or years). Additionally, this would ensure the identification of the global optimum across the age grid chosen. We tested this approach, however, the computation time dramatically increases as the number of age groups increases (even over a small age grid of equally spaced years), making this infeasible for our simulation study. It may be possible to find a computationally quicker approach to evaluate this grid search optimisation in a simulation study, but given time constraints we felt the continuous optimisation approach was sufficient. For future work it may be worthwhile to explore this grid optimisation in more detail.

8.4 Wider application

Whilst the focus of this thesis has been the extrapolation of efficacy data from adults to the paediatric population, many of the approaches could generalise to other areas. This is clearly demonstrated by the systematic review of methods for extrapolating between 'source' and 'target' populations, where extrapolation from adults to children was only one of four areas explored.

For example, our quantitative framework for informing extrapolation decisions in Chapter 6 could be extended to any homogenous populations where existing data already exists for two groups and expert opinion can be sought for the similarity between an existing group and some new 'target' population. This approach would work across finer paediatric age groupings, extending to geriatrics, considering differences across related conditions, indications or drugs with similar mechanisms of action, and informing extrapolation decisions across geographic regions.

The focus group meeting should also provide clear evidence that seeking the opinion of subject leaders regarding the place of extrapolation in a specific area of drug development can be worthwhile. Expert opinion can potentially lead to pragmatic approaches to extrapolate across populations, as is the case with our joint drug development paradigm.

This thesis has aimed to demonstrate how utilising existing historical data and expert opinion can be useful to focus and prioritise paediatric drug development research, and hopefully more broadly across other areas of drug development.

Bibliography

[1] FDA, Guidance for industry: Exposure-response relationships - Study design, data analysis, and regulatory applications, 2003. Accessed: 15/01/2015. Available at: http://www.gmp-compliance.org/guidemgr/files/EXPOSURE.PDF.

[2] T. Hothorn and A. Zeileis, Partykit: A modular toolkit for recursive partytioning in R., *Journal of Machine Learning Research*, vol. 16, pp. 3905–3909, 2015.

[3] H. Seibold, T. Hothorn, and A. Zeileis, Generalised linear model trees with global additive effects, *arXiv preprint arXiv:1612.07498*, 2016.

[4] C. L. Meinert, *Clinical Trials: Design, Conduct and Analysis*. Oxford University Press, 2012.

[5] ICH, Integrated Addendum to ICH E6 (R1): Guideline for Good Clinical Practice E6 (R2), *ICH Guideline*, 2015. Accessed: 19/12/2017. Available at:http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6/E6_R2__Step_4_2016_1109.pdf.

[6] World Health Organization, Guidelines for good clinical practice (GCP) for trials on pharmaceutical products, *WHO Technical Report Series*, vol. 850, pp. 97–137, 1995. Accessed: 19/12/2017.

[7] ICH, ICH Topic E8 - General considerations for clinical trials, *ICH Guideline*, 1998. Accessed: 07/02/2018. Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002877.pdf.

[8] F. Bretz, H. Schmidli, F. König, A. Racine, and W. Maurer, Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: General concepts, *Biometrical Journal*, vol. 48, no. 4, pp. 623–634, 2006.

[9] N. Stallard, A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information, *Statistics in Medicine*, vol. 29, no. 9, pp. 959–971, 2010.

[10] P. J. Karanicolas, F. Farrokhyar, and M. Bhandari, Blinding: Who, what, when, why, how?, *Canadian journal of surgery*, vol. 53, no. 5, p. 345, 2010.

[11] S. J. Pocock, *Clinical Trials: A Practical Approach*. John Wiley & Sons, 2013.

[12] D. J. Spiegelhalter, K. R. Abrams, and J. P. Myles, *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. John Wiley & Sons, 2004.

[13] D. Lunn, D. Spiegelhalter, A. Thomas, and N. Best, The BUGS project: Evolution, critique and future directions, *Statistics in Medicine*, vol. 28, no. 25, pp. 3049–3067, 2009.

[14] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, A. Riddell, *et al.*, Stan: A probabilistic programming language, *Journal of Statistical Software*, vol. 20, no. 2, pp. 1–37, 2016.

[15] Stan Development Team, RStan: the R interface to Stan, 2016. R package version 2.14.1.

[16] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. Available from: `https://www.R-project.org/`.

[17] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*. CRC press Boca Raton, FL, 2014.

[18] H.-G. Eichler, B. Bloechl-Daum, P. Bauer, F. Bretz, J. Brown, L. V. Hampson, P. Honig, M. Krams, H. Leufkens, R. Lim, *et al.*, Threshold-crossing: A useful way to establish the counterfactual in clinical trials?, *Clinical Pharmacology & Therapeutics*, vol. 100, no. 6, pp. 699–712, 2016.

[19] F. T. Bourgeois, S. Murthy, C. Pinto, K. L. Olson, J. P. Ioannidis, and K. D. Mandl, Pediatric versus adult drug trials for conditions with high pediatric disease burden, *Pediatrics*, vol. 130, no. 2, pp. 285–292, 2012.

[20] J. P. Karlberg, *Clinical Trials in Paediatrics. In: Machin D, Day S, Green S (eds) Textbook of Clinical Trials*. Wiley Online Library, 2004.

[21] H. K. Batchelor and J. F. Marriott, Paediatric pharmacokinetics: Key considerations, *British Journal of Clinical Pharmacology*, vol. 79, no. 3, pp. 395–404, 2015.

[22] T. Stephenson, How childrens responses to drugs differ from adults, *British Journal of Clinical Pharmacology*, vol. 59, no. 6, pp. 670–673, 2005.

[23] ICH, E11: Note for guidance on clinical investigation of medicinal products in the paediatric population, 2001. Accessed: 29/12/2017. Available at: `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002926.pdf`.

[24] P. Lepola, A. Needham, J. Mendum, P. Sallabank, D. Neubauer, and S. de Wildt, Informed consent for paediatric clinical trials in europe, *Archives of Disease in Childhood*, vol. 101, pp. 1017–1025, 2016.

[25] P. D. Joseph, J. C. Craig, and P. H. Caldwell, Clinical trials in children, *British Journal of Clinical Pharmacology*, vol. 79, no. 3, pp. 357–369, 2015.

[26] V. Shilling, P. Williamson, H. Hickey, E. Sowden, R. Smyth, and B. Young, *Processes in Recruitment to Randomised Controlled Trials of Medicines for Children (RECRUIT): A Qualitative Study*. Prepress Projects Limited, 2011.

[27] European Parliament, No 1901/2006 of the European Parliament and of the Council of 12 December 2006, *Official Journal of the European Union*, 2006. Accessed: 19/12/2017. Available at: `http://ec.europa.eu/health/documents/eudralex`.

[28] B. Lehmann, Regulation (EC) No. 1901/2006 on medicinal products for paediatric use & clinical research in vulnerable populations, *Child and Adolescent Psychiatry and Mental Health*, vol. 2, no. 1, p. 37, 2008.

[29] European Medicines Agency, EMA website - Paediatric investigation plans. Accessed: 15/02/2017. Available at: `http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000608.jsp&mid=WC0b01ac0580925b1b`.

[30] S. Conroy, I. Choonara, P. Impicciatore, A. Mohn, H. Arnell, A. Rane, C. Knoeppel, H. Seyberth, C. Pandolfini, M. P. Raffaelli, *et al.*, Survey of unlicensed and off label drug use in paediatric wards in European countries, *The BMJ*, vol. 320, no. 7227, pp. 79–82, 2000.

[31] A. Saint-Raymond and R. Herold, *Five Years of Pediatric Legislation in the European Union, in Pediatric Drug Development: Concepts and Applications (eds A. E. Mulberg, D. Murphy, J. Dunne and L. L. Mathis)*. John Wiley & Sons Ltd, Chichester, UK, 2013.

[32] M. Weda, J. Hoebert, M. Vervloet, C. Moltó Puigmarti, N. Damen, S. Marchange, J. Langedijk, J. Lisman, and L. v. Dijk, Study on off-label use of medicinal products in the European Union, 2017. Accessed: 20/12/2017. Available at: `https://ec.europa.eu/health/sites/health/files/files/documents/2017_02_28_final_study_report_on_off-label_use_.pdf`.

[33] L. V. Hampson, R. Herold, M. Posch, J. Saperia, and A. Whitehead, Bridging the gap: a review of dose investigations in paediatric investigation plans, *British Journal of Clinical Pharmacology*, vol. 78, no. 4, pp. 898–907, 2014.

[34] European Medicines Agency, Concept paper on extrapolation of efficacy and safety in medicine development (Final) - EMA/129698/2012, 2013. Accessed: 25/01/2017. Available at: `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2013/04/WC500142358.pdf`.

[35] European Medicines Agency, Reflection paper on the use of extrapolation in the development of medicines for paediatrics (Draft) - EMA/199678/2016, 2017. Accessed: 01/12/2017. Available at: `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2017/10/WC500236640.pdf`.

[36] J. Dunne, W. J. Rodriguez, M. D. Murphy, *et al.*, Extrapolation of adult data and other data in pediatric drug-development programs, *Pediatrics*, vol. 128, no. 5, pp. 1242–1249, 2011.

[37] Pfizer Laboratories Div. Pfizer Inc., Zmax- azithromycin dihydrate powder, for suspension - Highlights of prescribing information., 2017. Accessed: 20/12/2017. Available at: http://labeling.pfizer.com/ShowLabeling.aspx?id=650#S12.3.

[38] J. R. Nedelman, D. B. Rubin, and L. B. Sheiner, Diagnostics for confounding in PK/PD models for oxcarbazepine, *Statistics in Medicine*, vol. 26, no. 2, pp. 290–308, 2007.

[39] G. Alarcón and A. Valentín, *Introduction to Epilepsy*. Cambridge University Press, 2012.

[40] R. Appleton and A. G. Marson, *Epilepsy: The Facts (Facts)*. Oxford University Press, 2009.

[41] M. J. Brodie and S. C. Schachter, *Epilepsy Fast Facts Series*. Health Press Ltd., 1999.

[42] J. M. Pellock, W. J. Carman, V. Thyagarajan, T. Daniels, D. L. Morris, and D. O'Neill, Efficacy of antiepileptic drugs in adults predicts efficacy in children - A systematic review, *Neurology*, vol. 79, no. 14, pp. 1482–1489, 2012.

[43] J. A. French, A. M. Kanner, J. Bautista, B. Abou-Khalil, T. Browne, C. L. Harden, W. H. Theodore, C. Bazil, J. Stern, S. C. Schachter, *et al.*, Efficacy and tolerability of the new antiepileptic drugs, II: Treatment of refractory epilepsy - Report of the TTA and QSS subcommittees of the American Academy of Neurology and the American Epilepsy Society, *Epilepsia*, vol. 45, no. 5, pp. 410–423, 2004.

[44] European Medicines Agency, Conclusions of the paediatric epilepsy experts group meeting, held in London 1 September 2009 - EMA/153272/2010, 2010. Accessed: 25/01/2017. Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Other/2010/05/WC500090218.pdf.

[45] Pediatric News., FDA conducts analysis to assess acceptability of extrapolation of antiepileptic drug (AED) effectiveness in adults to children four years of age and older with partial onset seizures (POS)., *The Journal of Pediatric Pharmacology and Therapeutics*, vol. 21, no. 1, p. 98, 2016. Accessed: 25/01/2017. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4778704/pdf/i1551-6776-21-1-98.pdf.

[46] S.-L. T. Normand, Tutorial in biostatistics - Meta-analysis: Formulating, evaluating, combining, and reporting, *Statistics in Medicine*, vol. 18, no. 3, pp. 321–359, 1999.

[47] A. Whitehead, *Meta-Analysis of Controlled Clinical Trials*. John Wiley & Sons, 2002.

[48] G. V. Glass, Primary, secondary, and meta-analysis of research, *Educational Researcher*, vol. 5, no. 10, pp. 3–8, 1976.

[49] J. Davey, R. M. Turner, M. J. Clarke, and J. P. Higgins, Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: A cross-sectional, descriptive analysis, *BMC Medical Research Methodology*, vol. 11, no. 160, pp. 1–11, 2011.

[50] R. D. Riley, P. C. Lambert, and G. Abo-Zaid, Meta-analysis of individual participant data: Rationale, conduct, and reporting, *The BMJ*, vol. 340, p. c221, 2010.

[51] R. M. Turner, D. J. Spiegelhalter, G. Smith, and S. G. Thompson, Bias modelling in evidence synthesis, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 172, no. 1, pp. 21–47, 2009.

[52] S. Thompson, U. Ekelund, S. Jebb, A. K. Lindroos, A. Mander, S. Sharp, R. Turner, and D. Wilks, A proposed method of bias adjustment for meta-analyses of published observational studies, *International Journal of Epidemiology*, vol. 40, no. 3, pp. 765–777, 2010.

[53] W. G. Cochran, The combination of estimates from different experiments, *Biometrics*, vol. 10, no. 1, pp. 101–129, 1954.

[54] R. J. Hardy and S. G. Thompson, Detecting and describing heterogeneity in meta-analysis, *Statistics in Medicine*, vol. 17, no. 8, pp. 841–856, 1998.

[55] H. C. Van Houwelingen, L. R. Arends, and T. Stijnen, Tutorial in biostatistics - Advanced methods in meta-analysis: Multivariate approach and meta-regression, *Statistics in Medicine*, vol. 21, no. 4, pp. 589–624, 2002.

[56] R. D. Riley, Multivariate meta-analysis: The effect of ignoring within-study correlation, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 172, no. 4, pp. 789–811, 2009.

[57] H. Schmidli, S. Gsteiger, S. Roychoudhury, A. O'Hagan, D. Spiegelhalter, and B. Neuenschwander, Robust meta-analytic-predictive priors in clinical trials with historical control information, *Biometrics*, vol. 70, no. 4, pp. 1023–1032, 2014.

[58] A. J. Sutton and K. R. Abrams, Bayesian methods in meta-analysis and evidence synthesis, *Statistical Methods in Medical Research*, vol. 10, no. 4, pp. 277–303, 2001.

[59] A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow, *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons, 2006.

[60] A. Tversky and D. Kahneman, Judgment under uncertainty: Heuristics and biases, *Science*, vol. 185, no. 4157, pp. 1124–1131, 1974.

[61] S. R. Johnson, G. A. Tomlinson, G. A. Hawker, J. T. Granton, and B. M. Feldman, Methods to elicit beliefs for Bayesian priors: A systematic review, *Journal of Clinical Epidemiology*, vol. 63, no. 4, pp. 355–369, 2010.

[62] B. Grigore, J. Peters, C. Hyde, and K. Stein, A comparison of two methods for expert elicitation in health technology assessments, *BMC Medical Research Methodology*, vol. 16, no. 1, p. 85, 2016.

[63] R. T. Clemen and R. L. Winkler, Combining probability distributions from experts in risk analysis, *Risk Analysis*, vol. 19, no. 2, pp. 187–203, 1999.

[64] J. E. Oakley and A. O'Hagan, SHELF: the Sheffield Elicitation Framework (version 3.0), 2016. Accessed: 28/12/2017. Available at: `http://tonyohagan.co.uk/shelf`.

[65] W. Chang, J. Cheng, J. Allaire, Y. Xie, and J. McPherson, *shiny: Web Application Framework for R*, 2017. R package version 1.0.0. Available from: `https://CRAN.R-project.org/package=shiny`.

[66] B. Grigore, J. Peters, C. Hyde, and K. Stein, Methods to elicit probability distributions from experts: A systematic review of reported practice in health technology assessment, *Pharmacoeconomics*, vol. 31, no. 11, pp. 991–1003, 2013.

[67] L. V. Hampson, J. Whitehead, D. Eleftheriou, and P. Brogan, Bayesian methods for the design and interpretation of clinical trials in very rare diseases, *Statistics in medicine*, vol. 33, no. 24, pp. 4186–4201, 2014.

[68] L. V. Hampson, J. Whitehead, D. Eleftheriou, C. Tudur-Smith, R. Jones, D. Jayne, H. Hickey, M. W. Beresford, C. Bracaglia, A. Caldas, *et al.*, Elicitation of expert prior opinion: Application to the MYPAN trial in childhood polyarteritis nodosa, *PloS One*, vol. 10, no. 3, pp. 1–14, 2015.

[69] I. Wadsworth, L. V. Hampson, and T. Jaki, Extrapolation of efficacy and other data to support the development of new medicines for children: A systematic review of methods, *Statistical Methods in Medical Research*, 2016. Epub ahead of print.

[70] I. Wadsworth, T. Jaki, G. J. Sills, R. Appleton, J. H. Cross, A. G. Marson, T. Martland, A. McLellan, P. E. Smith, J. M. Pellock, and L. V. Hampson, Clinical drug development in epilepsy revisited: A proposal for a new paradigm streamlined using extrapolation, *CNS Drugs*, vol. 30, no. 11, pp. 1011–1017, 2016.

[71] H. Sun, J. W. Temeck, W. Chambers, G. Perkins, R. Bonnel, and D. Murphy, Extrapolation of efficacy in pediatric drug development and evidence-based medicine: Progress and lessons learned, *Therapeutic Innovation & Regulatory Science*, pp. 1–7, 2017.

[72] J. G. Ibrahim and M.-H. Chen, Power prior distributions for regression models, *Statistical Science*, vol. 15, no. 1, pp. 46–60, 2000.

[73] B. P. Hobbs, B. P. Carlin, S. J. Mandrekar, *et al.*, Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials, *Biometrics*, vol. 67, no. 3, pp. 1047–1056, 2011.

[74] B. Neuenschwander, G. Capkun-Niggli, M. Branson, *et al.*, Summarizing historical information on controls in clinical trials, *Clinical Trials*, vol. 7, no. 1, pp. 5–18, 2010.

[75] T. R. Fleming and D. L. DeMets, Surrogate end points in clinical trials: Are we being misled?, *Annals of Internal Medicine*, vol. 125, no. 7, pp. 605–613, 1996.

[76] ICH, Ethnic factors in the acceptability of foreign clinical data E5 (R1). Accessed: 01/10/2014. Available at: `http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E5_R1/Step4/E5_R1__Guideline.pdf`.

[77] E. Manolis, T. E. Osman, R. Herold, *et al.*, Role of modeling and simulation in pediatric investigation plans, *Pediatric Anesthesia*, vol. 21, no. 3, pp. 214–221, 2011.

[78] B. Neuenschwander, M. Branson, and D. J. Spiegelhalter, A note on the power prior, *Statistics in Medicine*, vol. 28, no. 28, pp. 3562–3566, 2009.

[79] M.-H. Chen, J. G. Ibrahim, P. Lam, *et al.*, Bayesian design of noninferiority trials for medical devices using historical data, *Biometrics*, vol. 67, no. 3, pp. 1163–1170, 2011.

[80] B. P. Hobbs, D. J. Sargent, and B. P. Carlin, Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models, *Bayesian Analysis*, vol. 7, no. 3, pp. 639–674, 2012.

[81] B. P. Hobbs, B. P. Carlin, and D. J. Sargent, Adaptive adjustment of the randomization ratio using historical control data, *Clinical Trials*, vol. 10, no. 3, pp. 430–440, 2013.

[82] S. Gsteiger, B. Neuenschwander, F. Mercier, *et al.*, Using historical control information for the design and analysis of clinical trials with overdispersed count data, *Statistics in Medicine*, vol. 42, no. 21, 2013.

[83] R. L. Cuffe, The inclusion of historical control data may reduce the power of a confirmatory study, *Statistics in Medicine*, vol. 30, no. 12, pp. 1329–1338, 2011.

[84] C.-F. Hsiao, Y.-Y. Hsu, H.-H. Tsou, *et al.*, Use of prior information for Bayesian evaluation of bridging studies, *Journal of Biopharmaceutical Statistics*, vol. 17, no. 1, pp. 109–121, 2007.

[85] M. Gandhi, B. Mukherjee, and D. Biswas, A Bayesian approach for inference from a bridging study with binary outcomes, *Journal of Biopharmaceutical Statistics*, vol. 22, no. 5, pp. 935–951, 2012.

[86] F. De Santis, Power priors and their use in clinical trials, *The American Statistician*, vol. 60, no. 2, pp. 122–129, 2006.

[87] C. Rietbergen, I. Klugkist, K. J. Janssen, *et al.*, Incorporation of historical data in the analysis of randomized therapeutic trials, *Contemporary Clinical Trials*, vol. 32, no. 6, pp. 848–855, 2011.

[88] D. A. Schoenfeld, H. Zheng, and D. M. Finkelstein, Bayesian design using adult data to augment pediatric trials, *Clinical Trials*, vol. 6, no. 4, pp. 297–304, 2009.

[89] Y.-H. Chen, Y.-C. Wu, and M. Wang, A Bayesian approach to evaluating regional treatment effect in a multiregional trial, *Journal of Biopharmaceutical Statistics*, vol. 19, no. 5, pp. 900–915, 2009.

[90] S. N. Goodman and J. T. Sladky, A Bayesian approach to randomized controlled trials in children utilizing information from adults: The case of Guillain-Barre, *Clinical Trials*, vol. 2, no. 4, pp. 305–310, 2005.

[91] J. B. Greenhouse and H. Seltman, Using prior distributions to synthesize historical evidence: Comments on the Goodman–Sladky case study of IVIg in Guillain–Barré syndrome, *Clinical Trials*, vol. 2, no. 4, pp. 311–318, 2005.

[92] J. L. French, N. Thomas, and C. Wang, Using historical data with Bayesian methods in early clinical trial monitoring, *Statistics in Biopharmaceutical Research*, vol. 4, no. 4, pp. 384–394, 2012.

[93] J. Whitehead, E. Valdés-Márquez, P. Johnson, *et al.*, Bayesian sample size for exploratory clinical trials incorporating historical data, *Statistics in Medicine*, vol. 27, no. 13, pp. 2307–2327, 2008.

[94] M. A. Gamalo, R. Wu, and R. C. Tiwari, Bayesian approach to noninferiority trials for proportions, *Journal of Biopharmaceutical Statistics*, vol. 21, no. 5, pp. 902–919, 2011.

[95] A. Kitsche, L. Hothorn, and F. Schaarschmidt, The use of historical controls in estimating simultaneous confidence intervals for comparisons against a concurrent control, *Computational Statistics & Data Analysis*, vol. 56, no. 12, pp. 3865–3875, 2012.

[96] L. Pei and M. D. Hughes, A statistical framework for quantile equivalence clinical trials with application to pharmacokinetic studies that bridge from HIV-infected adults to children, *Biometrics*, vol. 64, no. 4, pp. 1117–1125, 2008.

[97] H.-H. Tsou, T.-Y. Chien, J.-P. Liu, *et al.*, A consistency approach to evaluation of bridging studies and multi-regional trials, *Statistics in Medicine*, vol. 30, no. 17, pp. 2171–2186, 2011.

[98] S.-C. Chow, C. Chiang, J.-P. Liu, *et al.*, Statistical methods for bridging studies, *Journal of Biopharmaceutical Statistics*, vol. 22, no. 5, pp. 903–915, 2012.

[99] A. L. Gould, T. Jin, L. X. Zhang, *et al.*, A predictive Bayesian approach to the design and analysis of bridging studies, *Journal of Biopharmaceutical Statistics*, vol. 22, no. 5, pp. 916–934, 2012.

[100] Y. Duan, K. Ye, and E. P. Smith, Evaluating water quality using power priors to incorporate historical information, *Environmetrics*, vol. 17, no. 1, pp. 95–106, 2006.

[101] K. Viele, S. Berry, B. Neuenschwander, B. Amzal, F. Chen, N. Enas, B. Hobbs, J. G. Ibrahim, N. Kinnersley, S. Lindborg, *et al.*, Use of historical control data for assessing treatment effects in clinical trials, *Pharmaceutical Statistics*, vol. 13, no. 1, pp. 41–54, 2014.

[102] J. Aitchison and I. R. Dunsmore, *Statistical Prediction Analysis*. CUP Archive, 1975.

[103] L. V. Hampson and C. Jennison, Group sequential tests for delayed responses (with discussion), *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 75, no. 1, pp. 3–54, 2013.

[104] S. Galbraith and I. C. Marschner, Interim analysis of continuous long-term endpoints in clinical trials with longitudinal outcomes, *Statistics in Medicine*, vol. 22, no. 11, pp. 1787–1805, 2003.

[105] I. C. Marschner and S. L. Becker, Interim monitoring of clinical trials based on long-term binary endpoints, *Statistics in Medicine*, vol. 20, no. 2, pp. 177–192, 2001.

[106] K. Wüst and M. Kieser, Blinded sample size recalculation for normally distributed outcomes using long-and short-term data, *Biometrical Journal*, vol. 45, no. 8, pp. 915–930, 2003.

[107] P. F. Thall and R. Simon, Incorporating historical control data in planning phase II clinical trials, *Statistics in Medicine*, vol. 9, no. 3, pp. 215–228, 1990.

[108] L. R. Arends, Z. Vokó, and T. Stijnen, Combining multiple outcome measures in a meta-analysis: An application, *Statistics in Medicine*, vol. 22, no. 8, pp. 1335–1353, 2003.

[109] C.-T. Chen, H. J. Hung, and C.-F. Hsiao, Design and evaluation of multi-regional trials with heterogeneous treatment effect across regions, *Journal of Biopharmaceutical Statistics*, vol. 22, no. 5, pp. 1037–1050, 2012.

[110] F.-S. Ko, Approaches to allocate sample size rationally into individual regions for a multi-regional trial under heterogeneous effect size, *Communications in Statistics - Theory and Methods*, vol. 41, no. 20, pp. 3648–3665, 2012.

[111] J.-P. Liu, H. Hsueh, and J. J. Chen, Sample size requirements for evaluation of bridging evidence, *Biometrical Journal*, vol. 44, no. 8, pp. 969–981, 2002.

[112] F.-S. Ko, A design for survival studies: A multi-regional trial with unrecognized heterogeneity, *Communications in Statistics - Theory and Methods*, vol. 42, no. 13, pp. 2301–2323, 2013.

[113] K. Lan, Y. Soo, C. Siu, *et al.*, The use of weighted Z-tests in medical research, *Journal of Biopharmaceutical Statistics*, vol. 15, no. 4, pp. 625–639, 2005.

[114] Q. Huang, G. Chen, Z. Yuan, *et al.*, Design and sample size considerations for simultaneous global drug development program, *Journal of Biopharmaceutical Statistics*, vol. 22, no. 5, pp. 1060–1073, 2012.

[115] H.-H. Tsou, Y. Tsong, J.-T. Liu, *et al.*, Weighted evidence approach of bridging study, *Journal of Biopharmaceutical Statistics*, vol. 22, no. 5, pp. 952–965, 2012.

[116] J. Chen, H. Quan, B. Binkowitz, *et al.*, Assessing consistent treatment effect in a multi-regional clinical trial: A systematic review, *Pharmaceutical Statistics*, vol. 9, no. 3, pp. 242–253, 2010.

[117] H. Uesaka, Sample size allocation to regions in a multiregional trial, *Journal of Biopharmaceutical Statistics*, vol. 19, no. 4, pp. 580–594, 2009.

[118] H.-H. Tsou, S.-C. Chow, K. Lan, *et al.*, Proposals of statistical consideration to evaluation of results for a specific region in multi-regional trials – Asian perspective, *Pharmaceutical Statistics*, vol. 9, no. 3, pp. 201–206, 2010.

[119] H. Quan, P.-L. Zhao, J. Zhang, *et al.*, Sample size considerations for Japanese patients in a multi-regional trial based on MHLW guidance, *Pharmaceutical Statistics*, vol. 9, no. 2, pp. 100–112, 2010.

[120] F.-S. Ko, H.-H. Tsou, J.-P. Liu, *et al.*, Sample size determination for a specific region in a multiregional trial, *Journal of Biopharmaceutical Statistics*, vol. 20, no. 4, pp. 870–885, 2010.

[121] X. Luo, W. J. Shih, S. P. Ouyang, *et al.*, An optimal adaptive design to address local regulations in global clinical trials, *Pharmaceutical Statistics*, vol. 9, no. 3, pp. 179–189, 2010.

[122] H.-H. Tsou, H. James Hung, Y.-M. Chen, *et al.*, Establishing consistency across all regions in a multi-regional clinical trial, *Pharmaceutical Statistics*, vol. 11, no. 4, pp. 295–299, 2012.

[123] Y. Tsong, W.-J. Chang, X. Dong, *et al.*, Assessment of regional treatment effect in a multiregional clinical trial, *Journal of Biopharmaceutical Statistics*, vol. 22, no. 5, pp. 1019–1036, 2012.

[124] G. Li, J. Chen, H. Quan, *et al.*, Consistency assessment with global and bridging development strategies in emerging markets, *Contemporary Clinical Trials*, vol. 36, no. 2, 2013.

[125] J. Chen, H. Zheng, H. Quan, *et al.*, Graphical assessment of consistency in treatment effect among countries in multi-regional clinical trials, *Clinical Trials*, vol. 10, no. 6, pp. 842–851, 2013.

[126] C.-F. Hsiao, J.-Z. Xu, and J.-P. Liu, A two-stage design for bridging studies, *Journal of Biopharmaceutical Statistics*, vol. 15, no. 1, pp. 75–83, 2004.

[127] Z. Cai, H. Uesaka, and M. Tsujimoto, Comparison of two ethnic populations using distribution adjusted mean, *Journal of Biopharmaceutical Statistics*, vol. 23, no. 3, pp. 539–558, 2013.

[128] B. Engel and P. Walstra, Increasing precision or reducing expense in regression experiments by using information from a concomitant variable, *Biometrics*, pp. 13–20, 1991.

[129] J. P. Higgins, S. G. Thompson, J. J. Deeks, and D. G. Altman, Measuring inconsistency in meta-analyses, *The BMJ*, vol. 327, no. 7414, p. 557, 2003.

[130] M. Mittlböck and H. Heinzl, A simulation study comparing properties of heterogeneity measures in meta-analyses, *Statistics in Medicine*, vol. 25, no. 24, pp. 4321–4333, 2006.

[131] G. Rücker, G. Schwarzer, J. R. Carpenter, and M. Schumacher, Undue reliance on I2 in assessing heterogeneity may mislead, *BMC Medical Research Methodology*, vol. 8, no. 1, p. 79, 2008.

[132] J. P. Higgins and S. G. Thompson, Quantifying heterogeneity in a meta-analysis, *Statistics in Medicine*, vol. 21, no. 11, pp. 1539–1558, 2002.

[133] M. Gail and R. Simon, Testing for qualitative interactions between treatment effects and patient subsets, *Biometrics*, vol. 41, no. 2, pp. 361–372, 1985.

[134] S.-C. Chow, J. Shao, and O. Y.-P. Hu, Assessing sensitivity and similarity in bridging studies, *Journal of Biopharmaceutical Statistics*, vol. 12, no. 3, pp. 385–400, 2002.

[135] Ministry of Health, Labour and Welfare of Japan, Basic principles on global clinical trials. Accessed: 01/10/2014. Available at: `http://www.pmda.go.jp/kijunsakusei/file/guideline/new_drug/GlobalClinicalTrials_en.pdf`.

[136] J. Shao and S.-C. Chow, Reproducibility probability in clinical trials, *Statistics in Medicine*, vol. 21, no. 12, pp. 1727–1742, 2002.

[137] F. Khalil and S. Läer, Physiologically based pharmacokinetic modeling: Methodology, applications, and limitations with a focus on its role in pediatric drug development, *Journal of Biomedicine and Biotechnology*, vol. 2011, 2011.

[138] S. Reif, M. Schultze-Mosgau, and G. Sutter, From adults to children simulation-based choice of an appropriate sparse-sampling schedule., *Pediatric Drugs*, vol. 14, no. 3, pp. 189–200, 2012.

[139] F. De Santis, Using historical data for Bayesian sample size determination, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 170, no. 1, pp. 95–113, 2007.

[140] B. J. Gajewski, Comments on 'A note on the power prior by Neuenschwander B, Branson M and Spiegelhalter DJ. Statistics in Medicine; DOI: 10.1002/sim.3722, *Statistics in Medicine*, vol. 29, no. 6, pp. 708–709, 2010.

[141] European Medicines Agency, Reflection paper on extrapolation of efficacy and safety in paediatric medicine development (Draft) - EMA/199678/2016, 2016. Accessed: 24/05/2016. Available at: `http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2016/04/WC500204187.pdf`.

[142] A. T. Bazzano, R. Mangione-Smith, M. Schonlau, M. J. Suttorp, and R. H. Brook, Off-label prescribing to children in the United States outpatient setting, *Academic Pediatrics*, vol. 9, no. 2, pp. 81–88, 2009.

[143] A. Palmaro, R. Bissuel, N. Renaud, G. Durrieu, B. Escourrou, S. Oustric, J.-L. Montastruc, and M. Lapeyre-Mestre, Off-label prescribing in pediatric outpatients, *Pediatrics*, vol. 135, no. 1, pp. 49–58, 2014.

[144] A. Fogarasi, H. Jokeit, E. Faveret, J. Janszky, and I. Tuxhorn, The effect of age on seizure semiology in childhood temporal lobe epilepsy, *Epilepsia*, vol. 43, no. 6, pp. 638–643, 2002.

[145] B. F. Bourgeois and H. P. Goodkin, Efficacy of antiepileptic drugs in adults vs children - Does one size fit all?, *Neurology*, vol. 79, pp. 1420–1421, 2012.

[146] S. Rheims, M. Cucherat, A. Arzimanoglou, and P. Ryvlin, Greater response to placebo in children than in adults: A systematic review and meta-analysis in drug-resistant partial epilepsy, *PLoS medicine*, vol. 5, no. 8, p. e166, 2008.

[147] G. Ginsberg, D. Hattis, B. Sonawane, A. Russ, P. Banati, M. Kozlak, S. Smolenski, and R. Goble, Evaluation of child/adult pharmacokinetic differences from a database derived from the therapeutic drug literature, *Toxicological Sciences*, vol. 66, no. 2, pp. 185–200, 2002.

[148] D. Battino, M. Estienne, and G. Avanzini, Clinical pharmacokinetics of antiepileptic drugs in paediatric patients. Part II: Phenytoin, carbamazepine, sulthiame, lamotrigine, vigabatrin, oxcarbazepine and felbamate., *Clinical Pharmacokinetics*, vol. 29, no. 5, pp. 341–369, 1995.

[149] S. Mintzer, J. A. French, E. Perucca, J. A. Cramer, J. A. Messenheimer, D. E. Blum, M. A. Rogawski, and M. Baulac, Is a separate monotherapy indication warranted for antiepileptic drugs?, *The Lancet Neurology*, vol. 14, no. 12, pp. 1229–1240, 2015.

[150] S.-C. Chow and M. Chang, *Adaptive Design Methods in Clinical Trials*. CRC press, 2011.

[151] S. M. Berry, B. P. Carlin, J. J. Lee, and P. Muller, *Bayesian Adaptive Methods for Clinical Trials*. CRC press, 2010.

[152] T. J. VanderWeele and O. A. Arah, Unmeasured confounding for general outcomes, treatments, and confounders: Bias formulas for sensitivity analysis, *Epidemiology*, vol. 22, no. 1, p. 42, 2011.

[153] Eisai Inc., Highlights of prescribing information (Banzel)., 2015. Accessed: 13/01/2016. Available at: `https://www.banzel.com/pdfs/BanzelPI.pdf`.

[154] T. Glauser, G. Kluger, R. Sachedo, G. Krauss, C. Perdomo, and S. Arroyo, Efficacy and safety of rufinamide adjunctive therapy in patients with Lennox–Gastaut Syndrome (LGS): A multicenter, randomized, double-blind, placebo-controlled, parallel trial [abstract no. LBS. 001], *Neurology*, vol. 64, no. 10, p. 1826, 2005.

[155] A. G. Marson, A. M. Al-Kharusi, M. Alwaidh, R. Appleton, G. A. Baker, D. W. Chadwick, C. Cramp, O. C. Cockerell, P. N. Cooper, J. Doughty, *et al.*, The SANAD study of effectiveness of carbamazepine, gabapentin, lamotrigine, oxcarbazepine, or topiramate for treatment of partial epilepsy: An unblinded randomised controlled trial, *The Lancet*, vol. 369, no. 9566, pp. 1000–1015, 2007.

[156] A. G. Marson, A. M. Al-Kharusi, M. Alwaidh, R. Appleton, G. A. Baker, D. W. Chadwick, C. Cramp, O. C. Cockerell, P. N. Cooper, J. Doughty, *et al.*, The SANAD study of effectiveness of valproate, lamotrigine, or topiramate for generalised and unclassifiable epilepsy: An unblinded randomised controlled trial, *The Lancet*, vol. 369, no. 9566, pp. 1016–1026, 2007.

[157] J. McIntyre, S. Robertson, E. Norris, R. Appleton, W. P. Whitehouse, B. Phillips, T. Martland, K. Berry, J. Collier, S. Smith, *et al.*, Safety and efficacy of buccal midazolam versus rectal diazepam for emergency treatment of seizures in children: A randomised controlled trial, *The Lancet*, vol. 366, no. 9481, pp. 205–210, 2005.

[158] A. L. Lux, S. W. Edwards, E. Hancock, A. L. Johnson, C. R. Kennedy, R. W. Newton, F. J. O'Callaghan, C. M. Verity, J. P. Osborne, *et al.*, The United Kingdom Infantile Spasms Study comparing vigabatrin with prednisolone or tetracosactide at 14 days: A multicentre, randomised controlled trial, *The Lancet*, vol. 364, no. 9447, pp. 1773–1778, 2004.

[159] T. A. Glauser, A. Cnaan, S. Shinnar, D. G. Hirtz, D. Dlugos, D. Masur, P. O. Clark, E. V. Capparelli, and P. C. Adamson, Ethosuximide, valproic acid, and lamotrigine in childhood absence epilepsy, *New England Journal of Medicine*, vol. 362, no. 9, pp. 790–799, 2010.

[160] Novartis Pharmaceuticals, A placebo-controlled study of efficacy & safety of 2 trough-ranges of everolimus as adjunctive therapy in patients with tuberous sclerosis complex (TSC) & refractory partial-onset seizures (EXIST-3) [ClinicalTrials.gov identifier NCT01713946], 2000. Accessed: 05/03/2015. Available at: `https://clinicaltrials.gov/ct2/show/NCT01713946`.

[161] P. H. Sheridan and M. P. Jacobs, The development of antiepileptic drugs for children. Report from the NIH workshop, Bethesda, Maryland, February 17–18, 1994, *Epilepsy research*, vol. 23, no. 1, pp. 87–92, 1996.

[162] G. Hlavin, F. Koenig, C. Male, M. Posch, and P. Bauer, Evidence, eminence and extrapolation, *Statistics in Medicine*, vol. 35, no. 13, pp. 2117–2132, 2016.

[163] T. Friede, C. Röver, S. Wandel, and B. Neuenschwander, Meta-analysis of few small studies in orphan diseases, *Research Synthesis Methods*, vol. 8, no. 1, pp. 79–91, 2016.

[164] R. M. Turner, D. Jackson, Y. Wei, S. G. Thompson, and J. Higgins, Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis, *Statistics in Medicine*, vol. 34, no. 6, pp. 984–998, 2015.

[165] T. Friede, C. Röver, S. Wandel, and B. Neuenschwander, Meta-analysis of two studies in the presence of heterogeneity with applications in rare diseases, *Biometrical Journal*, vol. 59, no. 4, pp. 658–671, 2016.

[166] I. G. Girgis, P. Nandy, J. S. Nye, L. Ford, S. Mohanty, S. Wang, S. Ochalski, M. Eerdekens, and E. Cox, Pharmacokinetic–pharmacodynamic assessment of topiramate dosing regimens for children with epilepsy 2 to $<$ 10 years of age, *Epilepsia*, vol. 51, no. 10, pp. 1954–1962, 2010.

[167] J. A. French, G. L. Krauss, V. Biton, D. Squillacote, H. Yang, A. Laurenza, D. Kumar, M. A. Rogawski, V. Campanille, J. Floridia, *et al.*, Adjunctive perampanel for refractory partial-onset seizures randomized phase III study 304, *Neurology*, vol. 79, no. 6, pp. 589–596, 2012.

[168] J. M. FitzGerald, E. R. Bleecker, P. Nair, S. Korn, K. Ohta, M. Lommatzsch, G. T. Ferguson, W. W. Busse, P. Barker, S. Sproule, *et al.*, Benralizumab, an anti-interleukin-5 receptor α monoclonal antibody, as add-on treatment for patients with severe, uncontrolled, eosinophilic asthma (CALIMA): a randomised, double-blind, placebo-controlled phase III trial, *The Lancet*, vol. 388, no. 10056, pp. 2128–2141, 2016.

[169] P. M. O'Byrne, J. M. FitzGerald, N. Zhong, E. Bateman, P. J. Barnes, C. Keen, G. Almqvist, K. Pemberton, C. Jorup, S. Ivanov, *et al.*, The SYGMA programme of phase III trials to evaluate the efficacy and safety of budesonide/formoterol given 'as needed' in mild asthma: study protocols for two randomised controlled trials, *Trials*, vol. 18, no. 1, p. 12, 2017.

[170] S.-B. Tan, K. B. Dear, P. Bruzzi, and D. Machin, Strategy for randomised clinical trials in rare cancers, *The BMJ*, vol. 327, no. 7405, p. 47, 2003.

[171] N. Galwey, Supplementation of a clinical trial by historical control data: Is the prospect of dynamic borrowing an illusion?, *Statistics in Medicine*, vol. 36, no. 6, pp. 899–916, 2017.

[172] N. Welton, A. Ades, J. Carlin, D. Altman, and J. Sterne, Models for potentially biased evidence in meta-analysis using empirically based priors, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 172, no. 1, pp. 119–136, 2009.

[173] G. L. Kearns, S. M. Abdel-Rahman, S. W. Alander, D. L. Blowey, J. S. Leeder, and R. E. Kauffman, Developmental pharmacology – Drug disposition, action, and therapy in infants and children, *New England Journal of Medicine*, vol. 349, no. 12, pp. 1157–1167, 2003.

[174] H. Takahashi, S. Ishikawa, S. Nomoto, Y. Nishigaki, F. Ando, T. Kashima, S. Kimura, M. Kanamori, and H. Echizen, Developmental changes in pharmacokinetics and pharmacodynamics of warfarin enantiomers in Japanese children, *Clinical Pharmacology & Therapeutics*, vol. 68, no. 5, pp. 541–555, 2000.

[175] J. D. Marshall and G. L. Kearns, Developmental pharmacodynamics of cyclosporine, *Clinical Pharmacology & Therapeutics*, vol. 66, no. 1, pp. 66–75, 1999.

[176] A. R. Willan and E. M. Pinto, The value of information and optimal clinical trial design, *Statistics in Medicine*, vol. 24, no. 12, pp. 1791–1806, 2005.

[177] E. C. Wilson, A practical guide to value of information analysis, *PharmacoEconomics*, vol. 33, no. 2, pp. 105–121, 2015.

[178] A. Gelman, Prior distributions for variance parameters in hierarchical models, *Bayesian Analysis*, vol. 1, no. 3, pp. 515–534, 2006.

[179] MRC Biostatistics Unit, The BUGS project FAQs, 2017. Accessed 06/01/2017. Available at: `http://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-faqs/`.

[180] F. Leisch, FlexMix: A general framework for finite mixture models and latent class regression in R, *Journal of Statistical Software*, vol. 11, no. 8, pp. 1–18, 2004.

[181] B. Grün and F. Leisch, Fitting finite mixtures of generalized linear regressions in R, *Computational Statistics & Data Analysis*, vol. 51, no. 11, pp. 5247–5252, 2007.

[182] B. Grün and F. Leisch, FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters, *Journal of Statistical Software*, vol. 28, no. 4, pp. 1–35, 2008.

[183] S. Kullback and R. A. Leibler, On information and sufficiency, *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[184] S. Morita, P. F. Thall, and P. Müller, Determining the effective sample size of a parametric prior, *Biometrics*, vol. 64, no. 2, pp. 595–602, 2008.

[185] WHO Collaborating Centre for Drug Statistics Methodology, Defined Daily Dosing - Definition and general considerations, 2016. Accessed: 01/03/2017. Available at: `https://www.whocc.no/ddd/definition_and_general_considera/`.

[186] B. Neuenschwander, M. Branson, and T. Gsponer, Critical aspects of the Bayesian approach to phase I cancer trials, *Statistics in Medicine*, vol. 27, no. 13, pp. 2420–2439, 2008.

[187] S. Sturtz, U. Ligges, and A. Gelman, R2WinBUGS: A package for running WinBUGS from R, *Journal of Statistical Software*, vol. 12, no. 3, pp. 1–16, 2005. Available from: `http://www.jstatsoft.org`.

[188] M. Plummer, N. Best, K. Cowles, and K. Vines, CODA: Convergence Diagnosis and Output Analysis for MCMC, *R News*, vol. 6, no. 1, pp. 7–11, 2006. Available from `http://CRAN.R-project.org/doc/Rnews/Rnews_2006-1.pdf`.

[189] A. H. Briggs, K. Claxton, and M. J. Sculpher, *Decision Modelling for Health Economic Evaluation*. Handbooks in Health Economic Evaluation, 2006.

[190] A. Heath, I. Manolopoulou, and G. Baio, A review of methods for analysis of the expected value of information, *Medical Decision Making*, vol. 37, no. 7, pp. 747–758, 2017.

[191] Higgins, Julian PT and Green, Sally, Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011], The Cochrane Collaboration, 2011. Available from www.handbook.cochrane.org.

[192] K. Khan, R. Kunz, J. Kleijnen, and G. Antes, *Systematic Reviews to Support Evidence-Based Medicine*. Crc Press, 2011.

[193] J. Parkinson, W. Tang, C.-C. Johansson, D. Boulton, and B. Hamrén, Comparison of the exposure–response relationship of dapagliflozin in adult and paediatric patients with type 2 diabetes mellitus, *Diabetes, Obesity and Metabolism*, vol. 18, no. 7, pp. 685–692, 2016.

[194] J. C. Earp, N. Mehrotra, K. E. Peters, R. P. Fiorentino, D. Griebel, S. C. Lee, A. Mulberg, K. Röhss, M. Sandström, A. Taylor, *et al.*, Esomeprazole FDA approval in children with GERD: Exposure-matching and exposure-response, *Journal of Pediatric Gastroenterology and Nutrition*, vol. 65, no. 3, pp. 272–277, 2017.

[195] I. Lipkovich, A. Dmitrienko, *et al.*, Tutorial in biostatistics: Data-driven subgroup identification and analysis in clinical trials, *Statistics in Medicine*, vol. 36, no. 1, pp. 136–196, 2017.

[196] T. Ondra, A. Dmitrienko, T. Friede, A. Graf, F. Miller, N. Stallard, and M. Posch, Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review, *Journal of Biopharmaceutical Statistics*, vol. 26, no. 1, pp. 99–119, 2016.

[197] P. H. Eilers and B. D. Marx, Flexible smoothing with B-splines and penalties, *Statistical Science*, pp. 89–102, 1996.

[198] A. Zeileis, T. Hothorn, and K. Hornik, Model-based recursive partitioning, *Journal of Computational and Graphical Statistics*, vol. 17, no. 2, pp. 492–514, 2008.

[199] ICH, E4: Dose-response information to support drug registration, 1994. Accessed: 26/01/2018. Available at: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E4/Step4/E4_Guideline.pdf.

[200] A. Berges and C. Chen, Dose finding by concentration-response versus dose-response: A simulation-based comparison, *European Journal of Clinical Pharmacology*, vol. 69, no. 7, pp. 1391–1399, 2013.

[201] C.-H. Hsu, Evaluating potential benefits of dose–exposure–response modeling for dose finding, *Pharmaceutical Statistics*, vol. 8, no. 3, pp. 203–215, 2009.

[202] R. Schoemaker, J. R. Wade, and A. Stockis, Brivaracetam population pharmacokinetics and exposure-response modeling in adult subjects with partial-onset seizures, *The Journal of Clinical Pharmacology*, vol. 56, no. 12, pp. 1591–1602, 2016.

[203] A. Zeileis and K. Hornik, Generalized m-fluctuation tests for parameter instability, *Statistica Neerlandica*, vol. 61, no. 4, pp. 488–508, 2007.

[204] L. Breiman, Bagging predictors, *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[205] A. Bowman and L. Evers, Nonparametric smoothing, 2016. APTS Lecture notes. Accessed: 07/02/2018. Available at: `http://www2.warwick.ac.uk/fac/sci/statistics/apts/students/resources-1617/smoothing-notes.pdf`.

[206] S. Lang and A. Brezger, Bayesian P-splines, *Journal of Computational and Graphical Statistics*, vol. 13, no. 1, pp. 183–212, 2004.

[207] T. Kneib, A. Brezger, and C. M. Crainiceanu, *Generalized Semiparametric Regression with Covariates Measured with Error. In: Kneib T, Tutz G (eds) Statistical Modelling and Regression Structures - Festschrift in Honour of Ludwig Fahrmeir*. Springer, 2010.

[208] K. Williams, D. Thomson, I. Seto, D. G. Contopoulos-Ioannidis, J. P. Ioannidis, S. Curtis, E. Constantin, G. Batmanabane, L. Hartling, and T. Klassen, Standard 6: Age groups for pediatric trials, *Pediatrics*, vol. 129, no. Supplement 3, pp. S153–S160, 2012.

[209] I. Wadsworth, L. V. Hampson, T. Jaki, G. J. Sills, A. G. Marson, and R. Appleton, A quantitative framework to inform extrapolation decisions in children, *Under review*, 2018.

[210] C. Jennison and B. W. Turnbull, *Group Sequential Methods with Applications to Clinical Trials*. CRC Press, 1999.

[211] R. L. Burden and J. D. Faires, *Numerical Analysis*. Cengage Learning, 2011.

[212] Y. Chung, S. Rabe-Hesketh, V. Dorie, A. Gelman, and J. Liu, A nondegenerate penalized likelihood estimator for variance parameters in multilevel models, *Psychometrika*, vol. 78, no. 4, pp. 685–709, 2013.

[213] Stan Development Team, Stan modeling language users guide and reference manual, version 2.17.0, 2017.

[214] M. Thomas, B. Bornkamp, and H. Seibold, Subgroup identification in dose-finding trials via model-based recursive partitioning, *Statistics in Medicine*, pp. 1–17, 2017. Available online as Early Access.

[215] M. Rowland and T. N. Tozer, *Clinical Pharmacokinetics and Pharmacodynamics: Concepts and Applications*. Lippincott Williams and Wilkins, 2010.

[216] J. Whitehead, S. Patterson, D. Webber, S. Francis, and Y. Zhou, Easy-to-implement bayesian methods for dose-escalation studies in healthy volunteers, *Biostatistics*, vol. 2, no. 1, pp. 47–61, 2001.

[217] M. D. Luscombe, B. D. Owens, and D. Burke, Weight estimation in paediatrics: a comparison of the APLS formula and the formula Weight= 3 (age)+ 7, *Emergency Medicine Journal*, vol. 28, no. 7, pp. 590–593, 2011.

Appendices

Appendix A

Supplementary material to accompany Chapter 3.

This section contains the following three appendices:

A.1: Systematic review search strategy;

A.2: Data extraction form used to record relevant information from articles identified by the systematic review;

A.3: 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 D can be found at: `http://www.research.lancs.ac.uk/portal/en/publications/-(8911844e-2638-4dec-a844-8b842f034168).html`

A.1 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))
```

A.2 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? ☐

A.3 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.

Appendix B

Online supplementary material to accompany Chapter 6.

Below is the online supplementary material provided for the manuscript "A proposal for a quantitative framework to inform extrapolation decisions in children" by Wadsworth I, Hampson LV, Jaki T, Sills GJ, Marson AG and Appleton R

This document contains the following appendices:

B.1: Screen captures of the Shiny app developed for prior elicitation of expert opinion;

B.2: Additional plots of the probability of the extrapolation assumption holding for scenarios 2, 5 and 6;

B.3: Supplementary tables ST1 - ST13.

B.1 Prior elicitation scheme screenshots

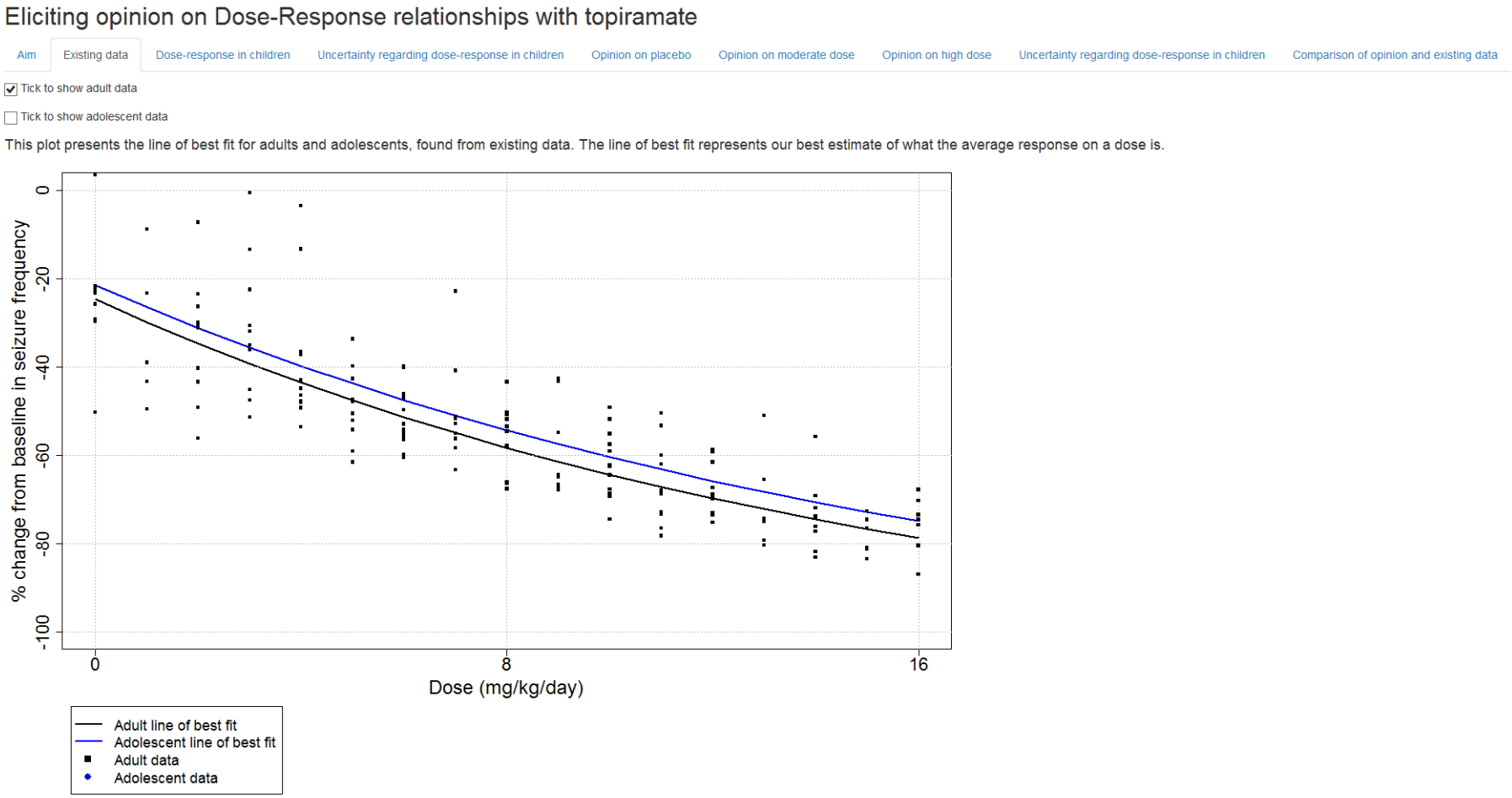


Figure B.1: App screen capture illustrating step 2 of the elicitation scheme.

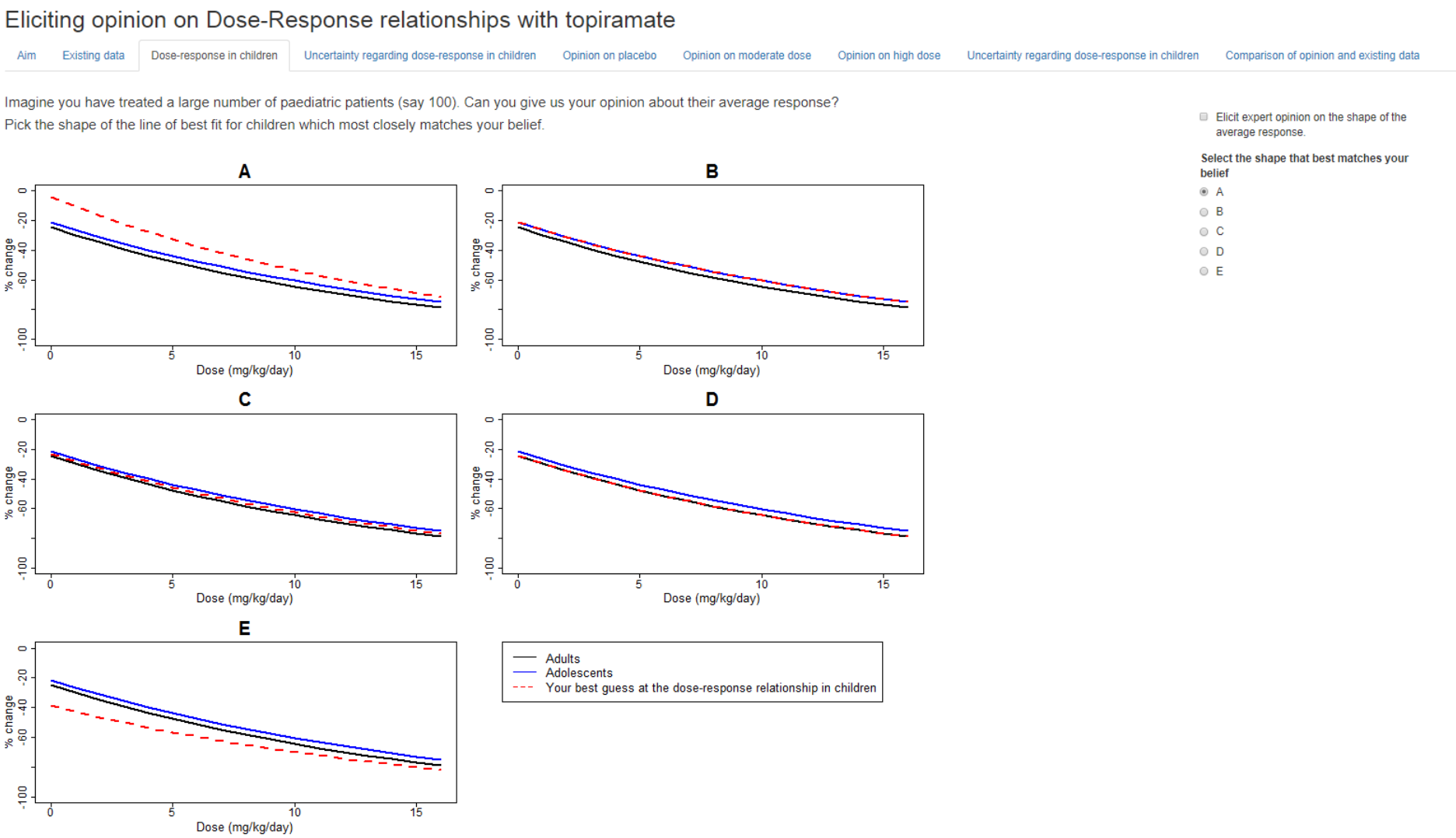


Figure B.2: App screen capture illustrating step 3 of the elicitation scheme.

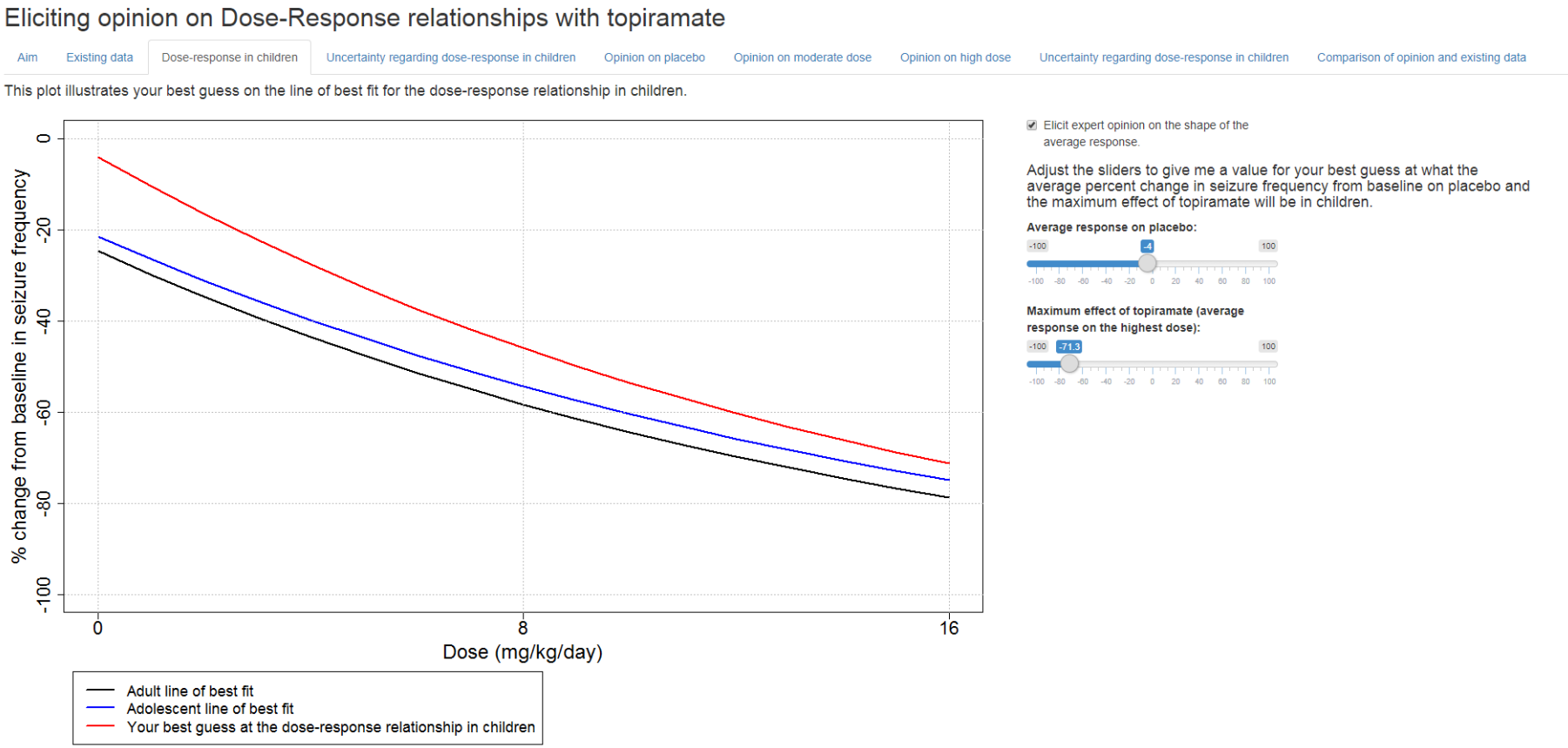


Figure B.3: App screen capture illustrating step 3 of the elicitation scheme.

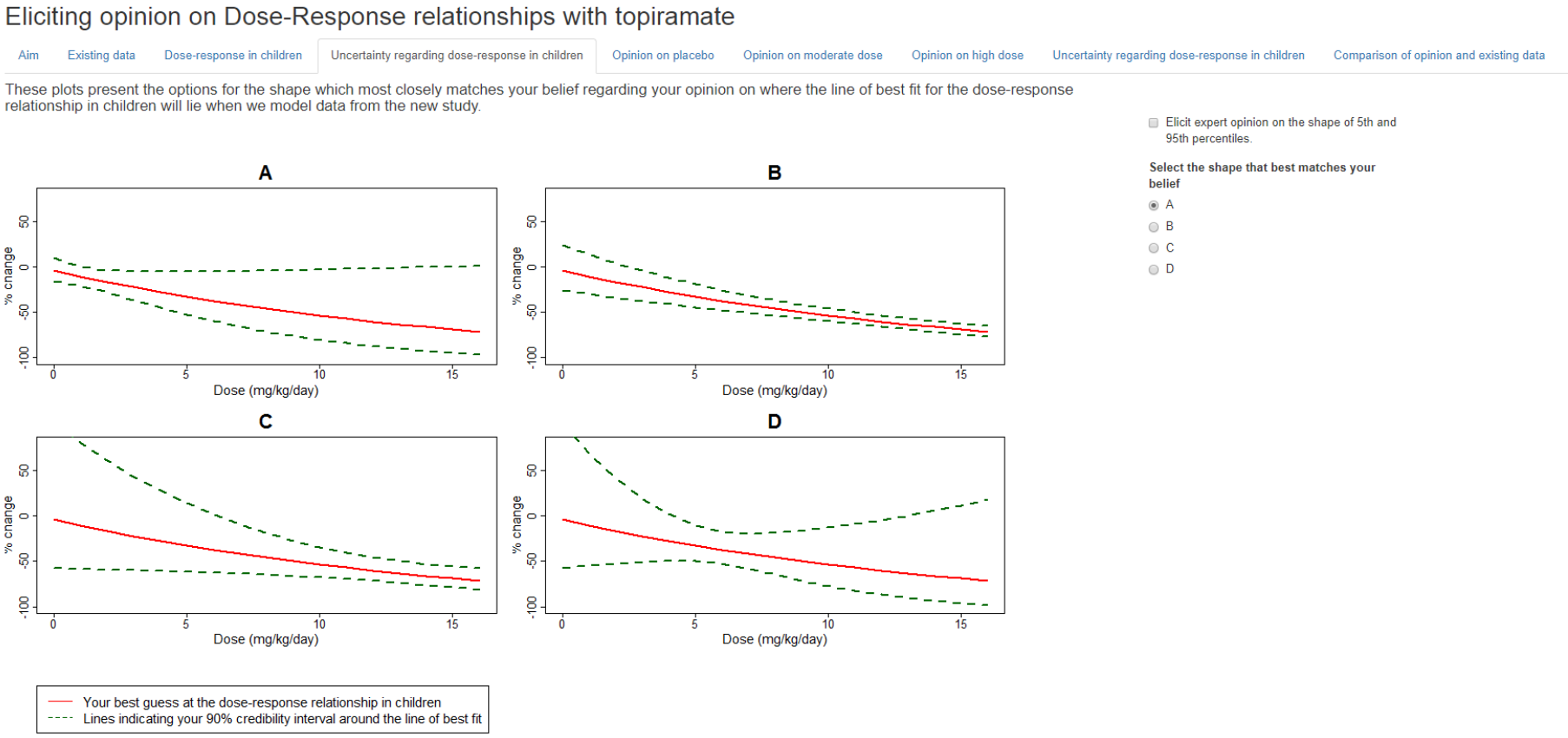


Figure B.4: App screen capture illustrating step 4 of the elicitation scheme.

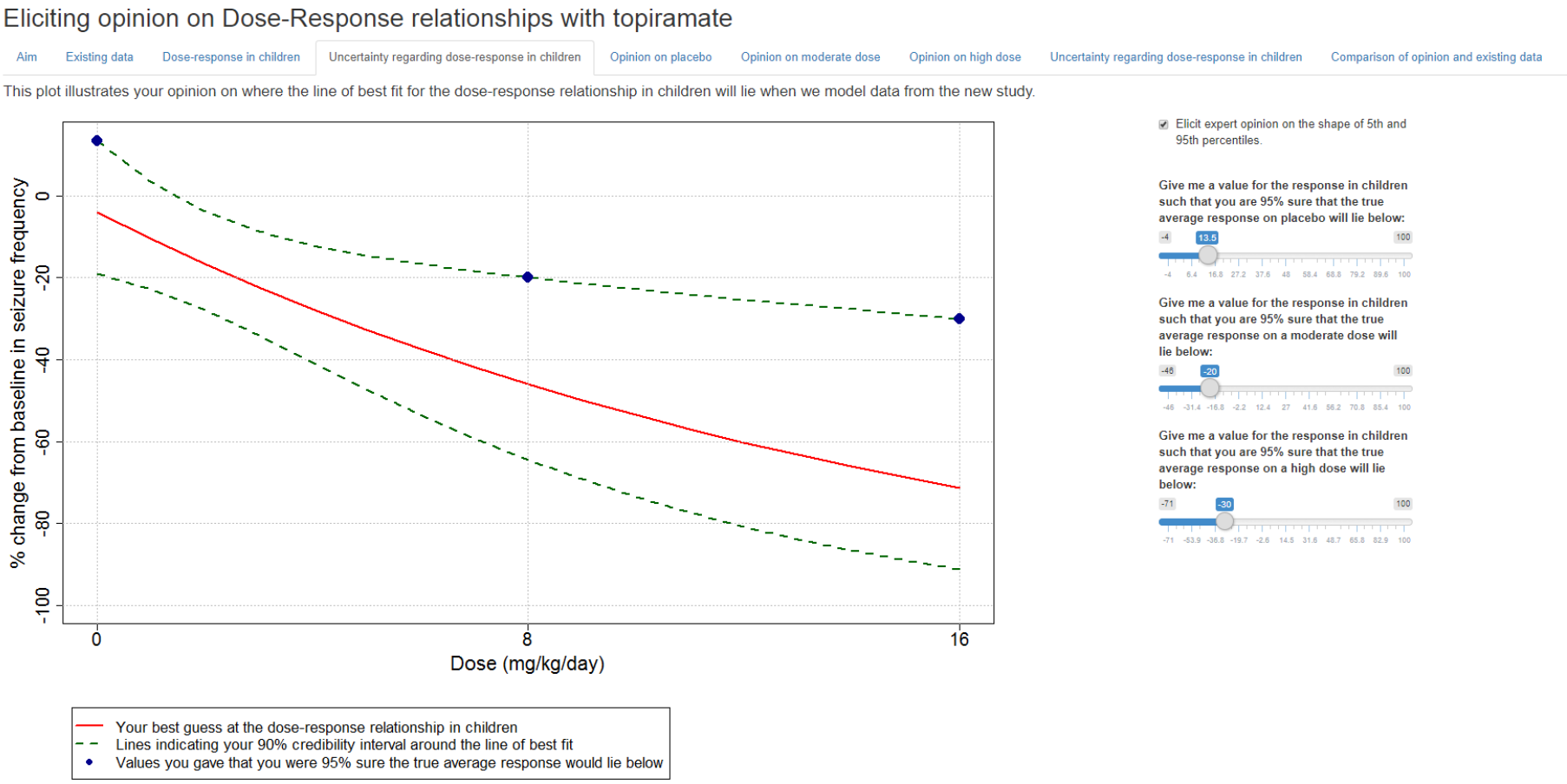


Figure B.5: App screen capture illustrating step 4 of the elicitation scheme.

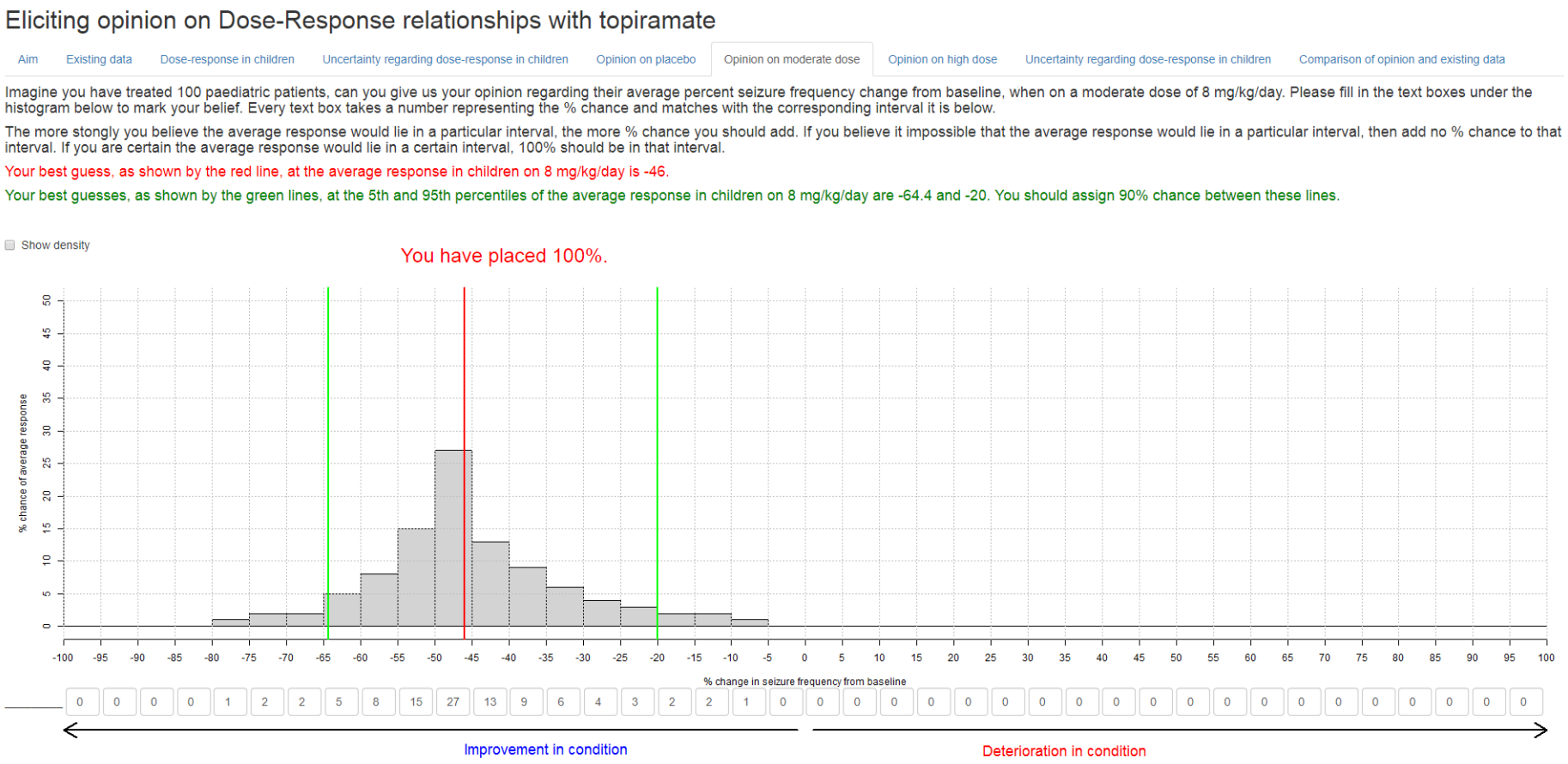


Figure B.6: App screen capture illustrating step 5 of the elicitation scheme.

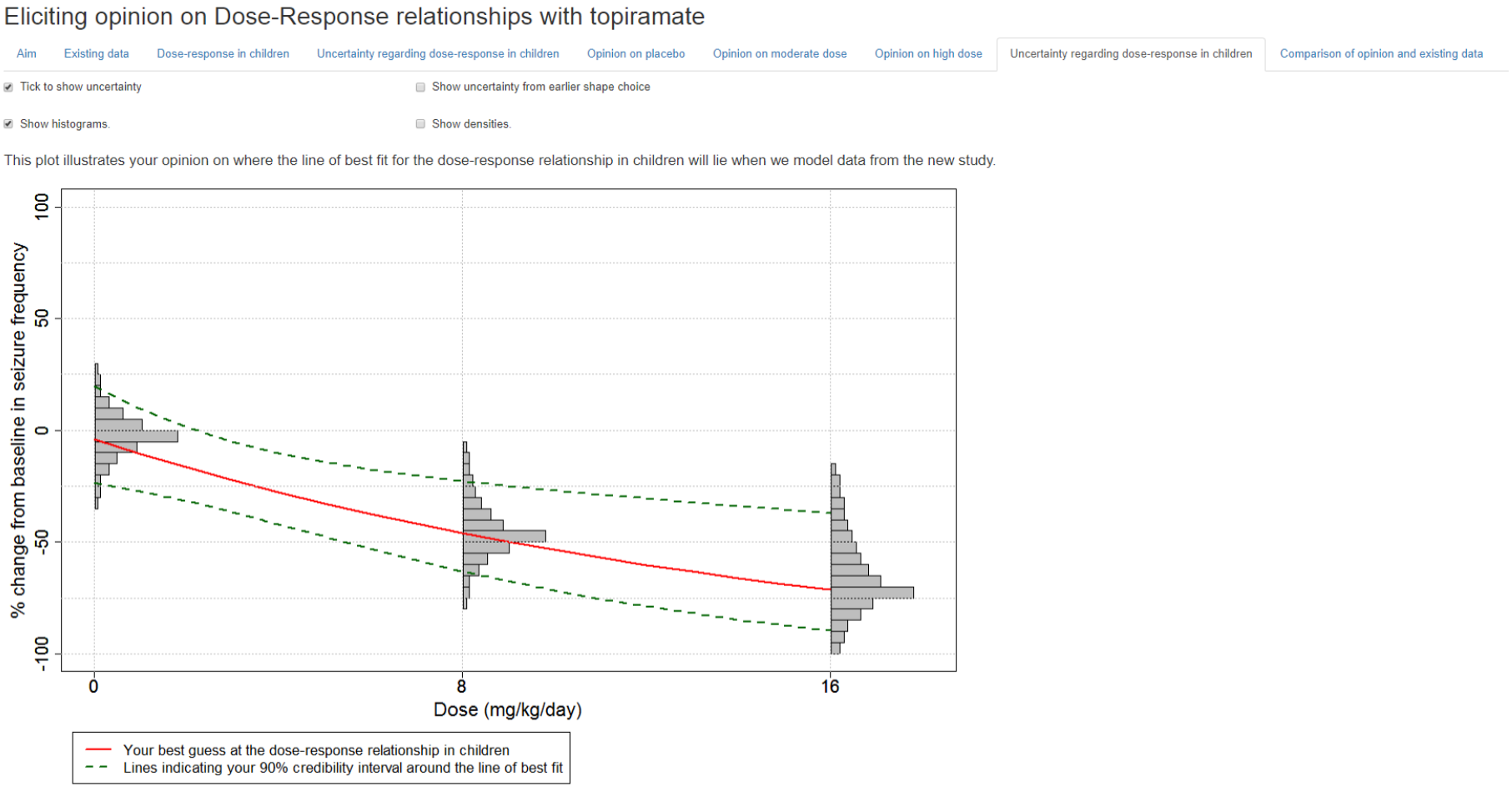


Figure B.7: App screen capture illustrating step 6 of the elicitation scheme.

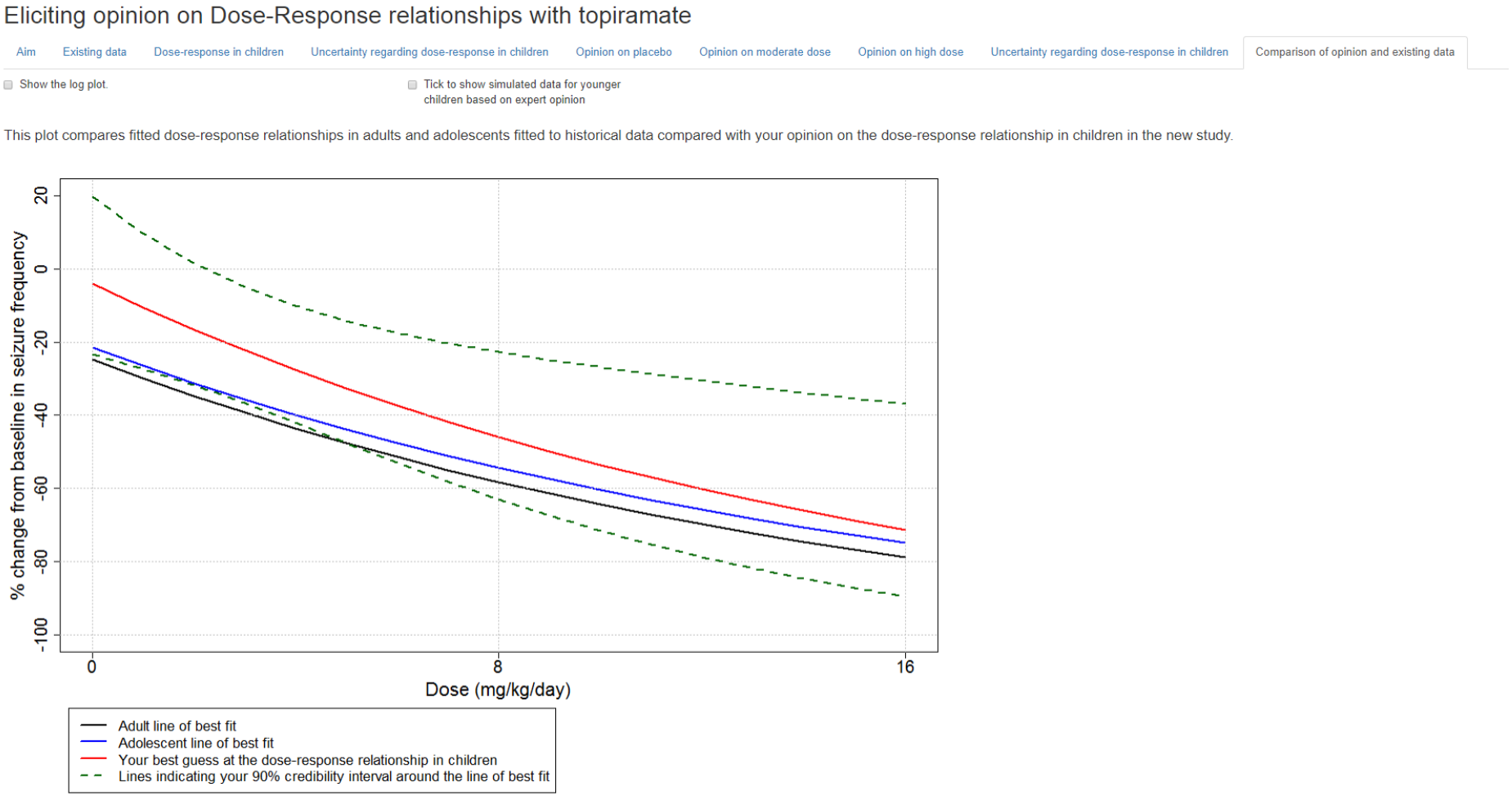


Figure B.8: App screen capture illustrating step 6 of the elicitation scheme.

B.2 Additional plots of the probability of the extrapolation assumption holding

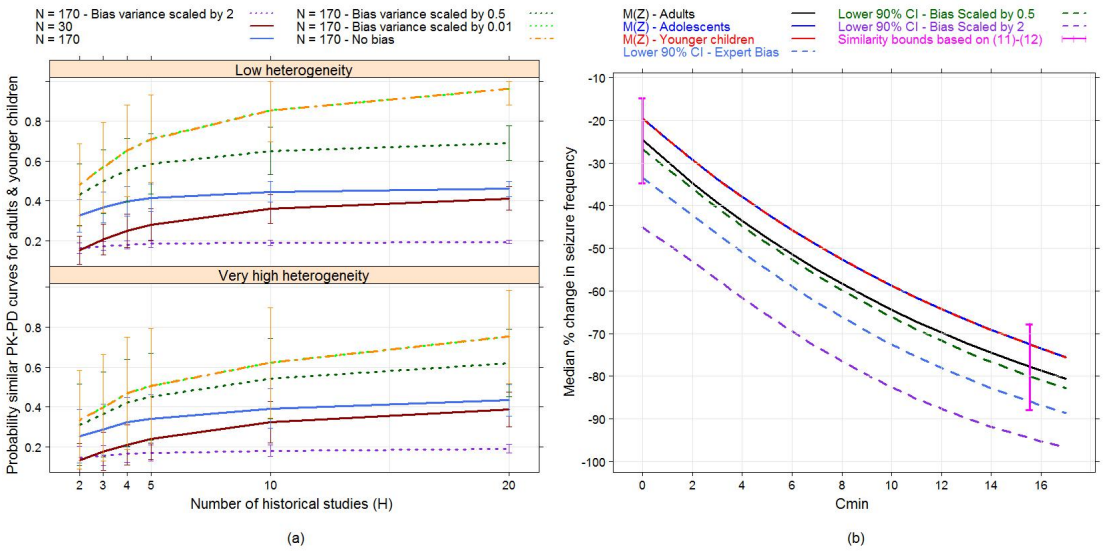


Figure B.9: (a) Average probability of the extrapolation assumption under bias prior E1, with bars of $\pm$ 1 standard deviations of the mean, when the true PK-PD relationships in adults and adolescents follows model S2. (b) Median PK-PD curves for adults, adolescents and younger children following model S2, along with lower bounds of 90% empirical credibility intervals for the median response in younger children resulting from the expert bias prior E1 when its variance matrix is unscaled; or scaled by a factor of 0.5 or 2. Credibility intervals are calculated conditioning on the true values of the adult and adolescent PK-PD parameters. Similarity bounds at placebo and the EC_{90} given by criteria (11) and (12) are also shown.

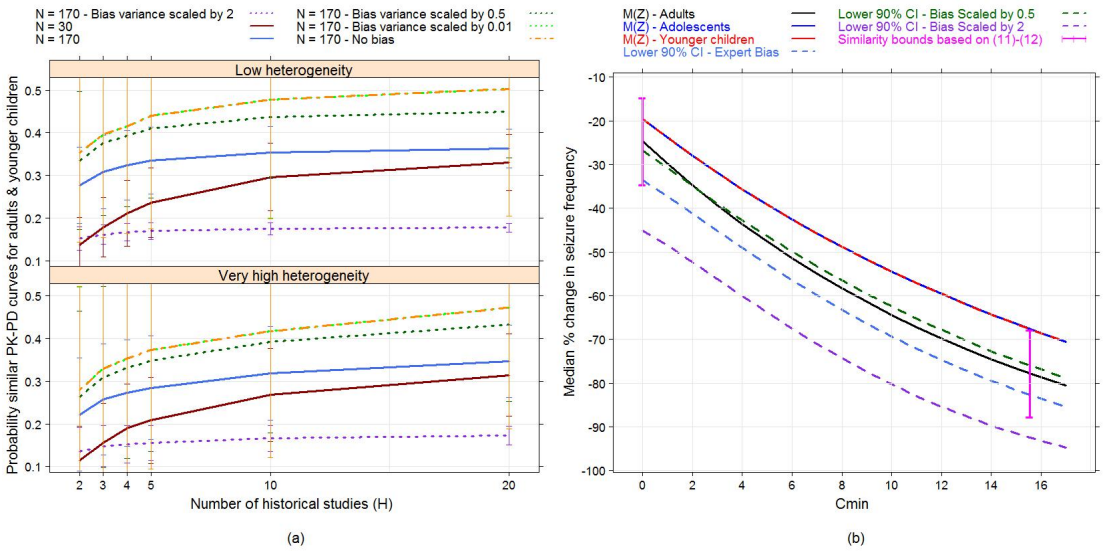


Figure B.10: (a) Average probability of the extrapolation assumption under bias prior E1, with bars of $\pm$ 1 standard deviations of the mean, when the true PK-PD relationships in adults and adolescents follows model S5. (b) Median PK-PD curves for adults, adolescents and younger children following model S5, along with lower bounds of 90% empirical credibility intervals for the median response in younger children resulting from the expert bias prior E1 when its variance matrix is unscaled; or scaled by a factor of 0.5 or 2. Credibility intervals are calculated conditioning on the true values of the adult and adolescent PK-PD parameters. Similarity bounds at placebo and the EC_{90} given by criteria (11) and (12) are also shown.

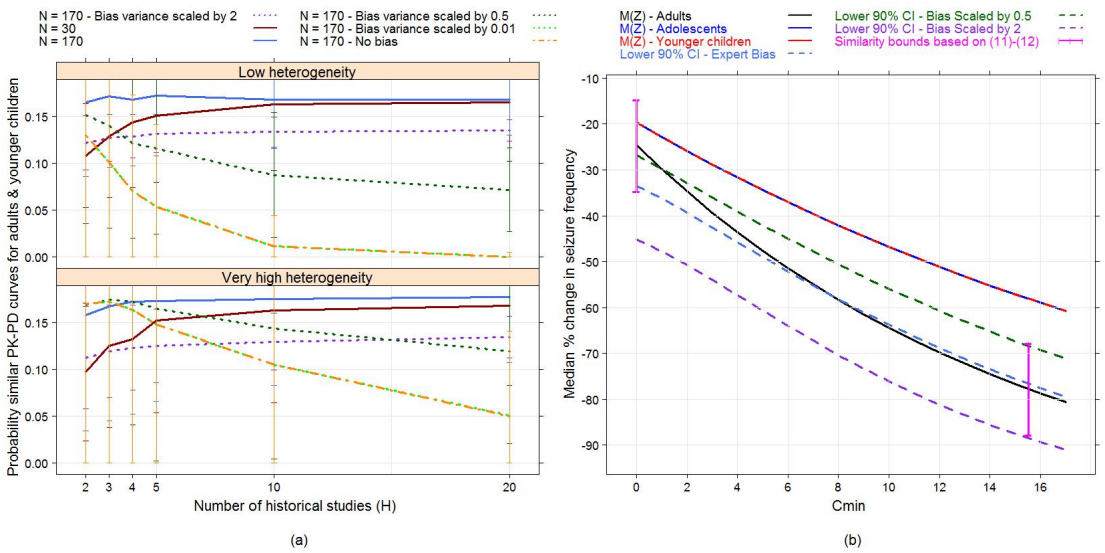


Figure B.11: (a) Average probability of the extrapolation assumption under bias prior E1, with bars of $\pm$ 1 standard deviations of the mean, when the true PK-PD relationships in adults and adolescents follows model S6. (b) Median PK-PD curves for adults, adolescents and younger children following model S6, along with lower bounds of 90% empirical credibility intervals for the median response in younger children resulting from the expert bias prior E1 when its variance matrix is unscaled; or scaled by a factor of 0.5 or 2. Credibility intervals are calculated conditioning on the true values of the adult and adolescent PK-PD parameters. Similarity bounds at placebo and the EC_{90} given by criteria (11) and (12) are also shown.

B.3 Supplementary tables

Table B.1: Means, empirical standard deviations and mean squared error over the 1000 replications of each scenario's meta-analysis results for the effects of age and the interaction between age and exposure - Low between-trial heterogeneity 1 of 2

Low between-trial heterogeneity								
		γ_A				γ_I		
		mean	SD	MSE		mean	SD	MSE
N = 30	Truth:	0			Truth:	0		
H = 2		0.0013	0.0481	0.0227		-0.0018	0.1065	0.0056
H = 3		0.0022	0.0126	0.0083		0	0.0010	0.0005
H = 4		0.0021	0.0114	0.0064		0.0010	0.0231	0.0042
H = 5		-0.0002	0.0060	0.0047		0.0010	0.0046	0.0007
H = 10		-0.0018	0.0026	0.0020		0.0002	0.0013	0.0002
H = 20		-0.0007	0.0011	0.0009		-0.0006	0.0031	0.0002
N = 170	Truth:	0			Truth:	0		
H = 2		0.0002	0.0054	0.0031		0.0002	0.0002	0.0001
H = 3		0.0011	0.0030	0.0020		0.0001	0.0001	0.0001
H = 4		-0.0003	0.0021	0.0017		0.0004	0.0001	0.0001
H = 5		0.0017	0.0015	0.0013		0.0001	0.0001	0
H = 10		-0.0002	0.0007	0.0007		0.0001	0	0
H = 20		0.0007	0.0003	0.0003		-0.0001	0	0
N = 30	Truth:	0.0569			Truth:	0.0056		
H = 2		0.0637	0.0320	0.0160		0.0030	0.0054	0.0012
H = 3		0.0560	0.0139	0.0091		0.0045	0.0017	0.0007
H = 4		0.0551	0.0081	0.0057		0.0058	0.0006	0.0004
H = 5		0.0615	0.0061	0.0042		0.0049	0.0004	0.0002
H = 10		0.0551	0.0026	0.0022		0.0064	0.0010	0.0002
H = 20		0.0576	0.0011	0.0010		0.0058	0.0002	0.0001
N = 170	Truth:	0.0569			Truth:	0.0056		
H = 2		0.0578	0.0055	0.0029		0.0058	0.0002	0.0001
H = 3		0.0576	0.0030	0.0020		0.0060	0.0001	0.0001
H = 4		0.0584	0.0021	0.0015		0.0053	0.0001	0
H = 5		0.0561	0.0015	0.0013		0.0058	0.0001	0
H = 10		0.0577	0.0007	0.0006		0.0056	0	0
H = 20		0.0563	0.0003	0.0003		0.0058	0	0
N = 30	Truth:	0.1108			Truth:	0.0103		
H = 2		0.1057	0.0269	0.0128		0.0114	0.0026	0.0010
H = 3		0.1093	0.0128	0.0075		0.0093	0.0010	0.0005
H = 4		0.1141	0.0082	0.0055		0.0097	0.0006	0.0003
H = 5		0.1098	0.0060	0.0043		0.0106	0.0004	0.0002
H = 10		0.1123	0.0025	0.0020		0.0100	0.0001	0.0001
H = 20		0.1096	0.0011	0.0010		0.0103	0.0001	0.0001
N = 170	Truth:	0.1108			Truth:	0.0103		
H = 2		0.1106	0.0054	0.0029		0.0103	0.0002	0.0001
H = 3		0.1108	0.0031	0.0022		0.0102	0.0001	0.0001
H = 4		0.1105	0.0020	0.0015		0.0105	0.0001	0
H = 5		0.1107	0.0015	0.0012		0.0105	0.0001	0
H = 10		0.1129	0.0007	0.0006		0.0102	0	0
H = 20		0.1113	0.0003	0.0003		0.0104	0	0

Table B.2: Means, empirical standard deviations and mean squared error over the 1000 replications of each scenario's meta-analysis results for the effects of age and the interaction between age and exposure - Low between-trial heterogeneity 2 of 2

Low between-trial heterogeneity								
		γ_A				γ_I		
		mean	SD	MSE		mean	SD	MSE
N = 30	Truth:	0.2105			Truth:	0.0175		
H = 2		0.2185	0.0726	0.0195		0.0141	0.0055	0.0014
H = 3		0.2102	0.0126	0.0073		0.0180	0.0010	0.0005
H = 4		0.2093	0.0081	0.0055		0.0172	0.0006	0.0003
H = 5		0.2155	0.0059	0.0042		0.0163	0.0004	0.0002
H = 10		0.2121	0.0026	0.0019		0.0171	0.0009	0.0002
H = 20		0.2101	0.0014	0.0011		0.0175	0.0011	0.0001
N = 170	Truth:	0.2105			Truth:	0.0175		
H = 2		0.2084	0.0054	0.0030		0.0178	0.0002	0.0001
H = 3		0.2115	0.0031	0.0019		0.0175	0.0001	0.0001
H = 4		0.2107	0.0021	0.0016		0.0176	0.0001	0
H = 5		0.2085	0.0016	0.0013		0.0175	0.0001	0
H = 10		0.2108	0.0007	0.0006		0.0175	0	0
H = 20		0.2116	0.0003	0.0003		0.0175	0	0
N = 30	Truth:	0.0569			Truth:	0.0137		
H = 2		0.0601	0.0457	0.0140		0.0083	0.1246	0.0244
H = 3		0.0605	0.0127	0.0072		0.0132	0.0010	0.0005
H = 4		0.0564	0.0084	0.0055		0.0141	0.0009	0.0003
H = 5		0.0553	0.0061	0.0045		0.0137	0.0025	0.0003
H = 10		0.0564	0.0024	0.0021		0.0136	0.0001	0.0001
H = 20		0.0566	0.0013	0.0010		0.0142	0.0014	0.0001
N = 170	Truth:	0.0569			Truth:	0.0137		
H = 2		0.0558	0.0055	0.0034		0.0137	0.0002	0.0001
H = 3		0.0545	0.0030	0.0022		0.0138	0.0001	0.0001
H = 4		0.0589	0.0021	0.0015		0.0134	0.0001	0.0001
H = 5		0.0572	0.0015	0.0012		0.0135	0.0001	0
H = 10		0.0560	0.0007	0.0006		0.0137	0	0
H = 20		0.0577	0.0003	0.0003		0.0136	0	0
N = 30	Truth:	0.0569			Truth:	0.0274		
H = 2		0.0506	0.0412	0.0187		0.0281	0.0049	0.0018
H = 3		0.0607	0.0125	0.0073		0.0268	0.0010	0.0005
H = 4		0.0597	0.0082	0.0054		0.0268	0.0006	0.0003
H = 5		0.0618	0.0102	0.0055		0.0282	0.0064	0.0024
H = 10		0.0558	0.0025	0.0021		0.0278	0.0007	0.0001
H = 20		0.0563	0.0011	0.0011		0.0276	0.0001	0.0001
N = 170	Truth:	0.0569			Truth:	0.0274		
H = 2		0.0558	0.0053	0.0031		0.0276	0.0002	0.0001
H = 3		0.0573	0.0030	0.0021		0.0272	0.0001	0.0001
H = 4		0.0574	0.0020	0.0016		0.0276	0.0001	0
H = 5		0.0565	0.0015	0.0011		0.0273	0.0001	0
H = 10		0.0567	0.0007	0.0006		0.0275	0	0
H = 20		0.0574	0.0003	0.0003		0.0275	0	0

Table B.3: Means, empirical standard deviations and mean squared error over the 1000 replications of each scenario's meta-analysis results for the effects of age and the interaction between age and exposure - Moderate between-trial heterogeneity 1 of 2

Moderate between-trial heterogeneity								
		γ_A				γ_I		
		mean	SD	MSE		mean	SD	MSE
N = 30	Truth:	0			Truth:	0		
H = 2		0.0039	0.0269	0.0156		0.0003	0.0027	0.0012
H = 3		-0.0019	0.0201	0.0119		0.0005	0.0166	0.0030
H = 4		0.0007	0.0090	0.0073		-0.0005	0.0019	0.0005
H = 5		0.0025	0.0062	0.0056		0.0001	0.0004	0.0003
H = 10		-0.0028	0.0027	0.0028		0	0.0024	0.0003
H = 20		0.0016	0.0012	0.0013		-0.0003	0.0001	0.0001
N = 170	Truth:	0			Truth:	0		
H = 2		0.0018	0.0063	0.0052		0.0006	0.0002	0.0002
H = 3		0	0.0039	0.0033		0.0001	0.0001	0.0001
H = 4		0.0019	0.0029	0.0027		-0.0002	0.0001	0.0001
H = 5		0.0002	0.0023	0.0022		0	0.0001	0.0001
H = 10		0.0007	0.0011	0.0012		-0.0001	0	0
H = 20		-0.0005	0.0006	0.0005		0.0002	0	0
N = 30	Truth:	0.0569			Truth:	0.0056		
H = 2		0.0525	0.0307	0.0175		0.0059	0.0040	0.0015
H = 3		0.0523	0.0134	0.0108		0.0072	0.0011	0.0006
H = 4		0.0539	0.0090	0.0067		0.0066	0.0023	0.0005
H = 5		0.0583	0.0063	0.0054		0.0057	0.0011	0.0004
H = 10		0.0564	0.0028	0.0026		0.0061	0.0023	0.0003
H = 20		0.0562	0.0014	0.0013		0.0050	0.0040	0.0007
N = 170	Truth:	0.0569			Truth:	0.0056		
H = 2		0.0552	0.0063	0.0053		0.0059	0.0002	0.0002
H = 3		0.0587	0.0040	0.0037		0.0064	0.0001	0.0001
H = 4		0.0559	0.0029	0.0026		0.0058	0.0001	0.0001
H = 5		0.0584	0.0023	0.0021		0.0058	0.0001	0.0001
H = 10		0.0556	0.0011	0.0011		0.0056	0	0
H = 20		0.0572	0.0006	0.0005		0.0057	0	0
N = 30	Truth:	0.1108			Truth:	0.0103		
H = 2		0.1154	0.0322	0.0190		0.0093	0.0041	0.0024
H = 3		0.1089	0.0132	0.0107		0.0095	0.0011	0.0006
H = 4		0.1067	0.0087	0.0074		0.0105	0.0006	0.0004
H = 5		0.1112	0.0063	0.0052		0.0108	0.0006	0.0003
H = 10		0.1115	0.0027	0.0026		0.0098	0.0002	0.0001
H = 20		0.1103	0.0013	0.0013		0.0101	0.0002	0.0001
N = 170	Truth:	0.1108			Truth:	0.0103		
H = 2		0.1110	0.0063	0.0058		0.0105	0.0002	0.0002
H = 3		0.1124	0.0040	0.0038		0.0105	0.0001	0.0001
H = 4		0.1126	0.0030	0.0027		0.0100	0.0001	0.0001
H = 5		0.1129	0.0023	0.0022		0.0099	0.0001	0.0001
H = 10		0.1114	0.0011	0.0011		0.0099	0	0
H = 20		0.1105	0.0006	0.0006		0.0103	0	0

Table B.4: Means, empirical standard deviations and mean squared error over the 1000 replications of each scenario's meta-analysis results for the effects of age and the interaction between age and exposure - Moderate between-trial heterogeneity 2 of 2

Moderate between-trial heterogeneity								
		γ_A				γ_I		
		mean	SD	MSE		mean	SD	MSE
N = 30	Truth:	0.2105			Truth:	0.0175		
H = 2		0.2099	0.0256	0.0155		0.0179	0.0025	0.0014
H = 3		0.2110	0.0135	0.0087		0.0174	0.0011	0.0006
H = 4		0.2086	0.0085	0.0077		0.0180	0.0006	0.0004
H = 5		0.2101	0.0066	0.0053		0.0167	0.0068	0.0011
H = 10		0.2105	0.0028	0.0025		0.0177	0.0004	0.0001
H = 20		0.2092	0.0013	0.0013		0.0176	0.0001	0.0001
N = 170	Truth:	0.2105			Truth:	0.0175		
H = 2		0.2126	0.0063	0.0057		0.0172	0.0002	0.0002
H = 3		0.2098	0.0039	0.0036		0.0173	0.0001	0.0001
H = 4		0.2110	0.0030	0.0028		0.0175	0.0001	0.0001
H = 5		0.2128	0.0023	0.0024		0.0179	0.0001	0.0001
H = 10		0.2091	0.0011	0.0011		0.0174	0	0
H = 20		0.2103	0.0006	0.0006		0.0174	0	0
N = 30	Truth:	0.0569			Truth:	0.0137		
H = 2		0.0500	0.0269	0.0161		0.0149	0.1033	0.0012
H = 3		0.0560	0.0132	0.0100		0.0142	0.0011	0.0006
H = 4		0.0564	0.0095	0.0069		0.0142	0.0040	0.0007
H = 5		0.0567	0.0068	0.0056		0.0138	0.0011	0.0003
H = 10		0.0548	0.0027	0.0026		0.0140	0.0002	0.0001
H = 20		0.0576	0.0022	0.0016		0.0138	0.0012	0.0001
N = 170	Truth:	0.0569			Truth:	0.0137		
H = 2		0.0574	0.0064	0.0056		0.0142	0.0002	0.0002
H = 3		0.0593	0.0040	0.0038		0.0138	0.0001	0.0001
H = 4		0.0576	0.0029	0.0028		0.0140	0.0001	0.0001
H = 5		0.0551	0.0023	0.0022		0.0135	0.0001	0.0001
H = 10		0.0580	0.0011	0.0010		0.0136	0	0
H = 20		0.0572	0.0006	0.0005		0.0134	0	0
N = 30	Truth:	0.0569			Truth:	0.0274		
H = 2		0.0621	0.0304	0.0174		0.0297	0.0243	0.0067
H = 3		0.0496	0.0192	0.0123		0.0297	0.0196	0.0059
H = 4		0.0564	0.0089	0.0070		0.0279	0.0012	0.0004
H = 5		0.0594	0.0072	0.0053		0.0277	0.0011	0.0003
H = 10		0.0587	0.0029	0.0027		0.0269	0.0028	0.0003
H = 20		0.0558	0.0012	0.0012		0.0276	0.0001	0.0001
N = 170	Truth:	0.0569			Truth:	0.0274		
H = 2		0.0598	0.0063	0.0052		0.0275	0.0002	0.0002
H = 3		0.0569	0.0039	0.0039		0.0276	0.0001	0.0001
H = 4		0.0541	0.0029	0.0029		0.0274	0.0001	0.0001
H = 5		0.0563	0.0023	0.0021		0.0276	0.0001	0.0001
H = 10		0.0583	0.0011	0.0011		0.0275	0	0
H = 20		0.0571	0.0006	0.0006		0.0272	0	0

Table B.5: Means, empirical standard deviations and mean squared error over the 1000 replications of each scenario's meta-analysis results for the effects of age and the interaction between age and exposure - High between-trial heterogeneity 1 of 2

High between-trial heterogeneity								
		γ_A				γ_I		
		mean	SD	MSE		mean	SD	MSE
N = 30	Truth:	0			Truth:	0		
H = 2		-0.0081	0.1871	0.0474		0.0018	0.0098	0.0022
H = 3		0.0028	0.0148	0.0112		-0.0008	0.0127	0.0015
H = 4		0.0004	0.0089	0.0084		0.0007	0.0006	0.0005
H = 5		0.0019	0.0075	0.0060		-0.0014	0.0099	0.0018
H = 10		0.0044	0.0029	0.0032		-0.0002	0.0002	0.0002
H = 20		-0.0007	0.0014	0.0016		0.0003	0.0001	0.0001
N = 170	Truth:	0			Truth:	0		
H = 2		-0.0050	0.0070	0.0084		-0.0005	0.0002	0.0002
H = 3		-0.0016	0.0046	0.0055		-0.0003	0.0001	0.0002
H = 4		-0.0030	0.0036	0.0039		0.0005	0.0001	0.0001
H = 5		0.0015	0.0030	0.0030		0	0.0001	0.0001
H = 10		0.0009	0.0015	0.0015		0.0002	0	0
H = 20		0.0001	0.0008	0.0007		-0.0001	0	0
N = 30	Truth:	0.0569			Truth:	0.0056		
H = 2		0.0590	0.0481	0.0199		0.0037	0.0123	0.0016
H = 3		0.0602	0.0138	0.0113		0.0046	0.0011	0.0006
H = 4		0.0528	0.0088	0.0081		0.0060	0.0008	0.0004
H = 5		0.0560	0.0071	0.0065		0.0058	0.0015	0.0004
H = 10		0.0543	0.0029	0.0033		0.0060	0.0002	0.0002
H = 20		0.0566	0.0017	0.0018		0.0024	0.0221	0.0108
N = 170	Truth:	0.0569			Truth:	0.0056		
H = 2		0.0542	0.0070	0.0075		0.0062	0.0002	0.0002
H = 3		0.0537	0.0055	0.0051		0.0060	0.0056	0.0003
H = 4		0.0595	0.0037	0.0039		0.0057	0.0001	0.0001
H = 5		0.0560	0.0030	0.0030		0.0052	0.0001	0.0001
H = 10		0.0564	0.0016	0.0016		0.0058	0	0
H = 20		0.0566	0.0008	0.0008		0.0057	0	0
N = 30	Truth:	0.1108			Truth:	0.0103		
H = 2		0.1194	0.0549	0.0204		0.0078	0.0060	0.0015
H = 3		0.1095	0.0134	0.0115		0.0116	0.0011	0.0006
H = 4		0.1171	0.0090	0.0090		0.0093	0.0009	0.0004
H = 5		0.1096	0.0067	0.0060		0.0105	0.0005	0.0003
H = 10		0.1093	0.0030	0.0030		0.0108	0.0005	0.0002
H = 20		0.1107	0.0014	0.0015		0.0102	0.0001	0.0001
N = 170	Truth:	0.1108			Truth:	0.0103		
H = 2		0.1065	0.0072	0.0070		0.0108	0.0002	0.0002
H = 3		0.1070	0.0048	0.0052		0.0096	0.0001	0.0001
H = 4		0.1101	0.0038	0.0039		0.0102	0.0001	0.0001
H = 5		0.1109	0.0030	0.0030		0.0104	0.0001	0.0001
H = 10		0.1135	0.0016	0.0015		0.0102	0	0
H = 20		0.1123	0.0008	0.0008		0.0104	0	0

Table B.6: Means, empirical standard deviations and mean squared error over the 1000 replications of each scenario's meta-analysis results for the effects of age and the interaction between age and exposure - High between-trial heterogeneity 2 of 2

High between-trial heterogeneity								
		γ_A				γ_I		
		mean	SD	MSE		mean	SD	MSE
N = 30	Truth:	0.2105			Truth:	0.0175		
H = 2		0.2055	0.1025	0.0333		0.0192	0.0085	0.0024
H = 3		0.2168	0.0134	0.0111		0.0160	0.0015	0.0007
H = 4		0.2127	0.0101	0.0079		0.0157	0.0010	0.0004
H = 5		0.2088	0.0066	0.0066		0.0183	0.0016	0.0004
H = 10		0.2143	0.0029	0.0033		0.0177	0.0002	0.0001
H = 20		0.2096	0.0014	0.0014		0.0173	0.0001	0.0001
N = 170	Truth:	0.2105			Truth:	0.0175		
H = 2		0.2045	0.0070	0.0083		0.0180	0.0002	0.0002
H = 3		0.2121	0.0049	0.0053		0.0179	0.0002	0.0002
H = 4		0.2117	0.0036	0.0041		0.0171	0.0001	0.0001
H = 5		0.2133	0.0030	0.0029		0.0169	0.0001	0.0001
H = 10		0.2065	0.0016	0.0017		0.0176	0	0
H = 20		0.2112	0.0008	0.0008		0.0173	0	0
N = 30	Truth:	0.0569			Truth:	0.0137		
H = 2		0.0615	0.0432	0.0236		0.0090	0.0775	0.0184
H = 3		0.0566	0.0132	0.0100		0.0144	0.0011	0.0006
H = 4		0.0495	0.0087	0.0081		0.0143	0.0007	0.0004
H = 5		0.0583	0.0067	0.0069		0.0123	0.0065	0.0010
H = 10		0.0550	0.0031	0.0034		0.0134	0.0022	0.0004
H = 20		0.0592	0.0014	0.0017		0.0136	0.0002	0.0001
N = 170	Truth:	0.0569			Truth:	0.0137		
H = 2		0.0575	0.0071	0.0078		0.0142	0.0002	0.0002
H = 3		0.0565	0.0049	0.0056		0.0137	0.0002	0.0002
H = 4		0.0603	0.0037	0.0038		0.0139	0.0001	0.0001
H = 5		0.0565	0.0030	0.0032		0.0137	0.0001	0.0001
H = 10		0.0569	0.0016	0.0015		0.0138	0	0
H = 20		0.0581	0.0008	0.0008		0.0136	0	0
N = 30	Truth:	0.0569			Truth:	0.0274		
H = 2		0.0573	0.1333	0.1001		0.0350	0.2044	0.0318
H = 3		0.0585	0.0138	0.0120		0.0269	0.0011	0.0007
H = 4		0.0517	0.0141	0.0096		0.0304	0.0324	0.0056
H = 5		0.0563	0.0067	0.0072		0.0277	0.0005	0.0004
H = 10		0.0555	0.0029	0.0031		0.0277	0.0007	0.0002
H = 20		0.0561	0.0019	0.0016		0.0275	0.0006	0.0001
N = 170	Truth:	0.0569			Truth:	0.0274		
H = 2		0.0529	0.0074	0.0074		0.0267	0.0002	0.0002
H = 3		0.0572	0.0048	0.0055		0.0272	0.0001	0.0001
H = 4		0.0593	0.0037	0.0040		0.0269	0.0001	0.0001
H = 5		0.0570	0.0030	0.0030		0.0274	0.0001	0.0001
H = 10		0.0583	0.0015	0.0017		0.0275	0	0
H = 20		0.0553	0.0008	0.0007		0.0275	0	0

Table B.7: Means, empirical standard deviations and mean squared error over the 1000 replications of each scenario's meta-analysis results for the effects of age and the interaction between age and exposure - Very high between-trial heterogeneity 1 of 2

Very high between-trial heterogeneity								
		γ_A				γ_I		
		mean	SD	MSE		mean	SD	MSE
N = 30	Truth:	0			Truth:	0		
H = 2		0.0047	0.0636	0.0339		0.0029	0.1632	0.0143
H = 3		0.0013	0.0221	0.0162		-0.0024	0.0313	0.0130
H = 4		-0.0016	0.0093	0.0103		0.0004	0.0009	0.0005
H = 5		0.0047	0.0070	0.0078		-0.0010	0.0011	0.0004
H = 10		0.0013	0.0032	0.0043		0.0004	0.0002	0.0002
H = 20		0.0026	0.0017	0.0022		0.0003	0.0008	0.0001
N = 170	Truth:	0			Truth:	0		
H = 2		0	0.0082	0.0106		-0.0001	0.0003	0.0003
H = 3		0.0020	0.0057	0.0071		0.0004	0.0002	0.0002
H = 4		0.0021	0.0046	0.0059		-0.0005	0.0001	0.0002
H = 5		-0.0032	0.0040	0.0045		-0.0004	0.0001	0.0001
H = 10		0.0005	0.0021	0.0023		-0.0003	0.0001	0.0001
H = 20		0.0009	0.0011	0.0011		0.0002	0	0
N = 30	Truth:	0.0569			Truth:	0.0056		
H = 2		0.0649	0.0903	0.0268		0.0051	0.0126	0.0015
H = 3		0.0567	0.0141	0.0146		0.0053	0.0016	0.0007
H = 4		0.0538	0.0114	0.0107		0.0090	0.0213	0.0054
H = 5		0.0548	0.0072	0.0080		0.0061	0.0008	0.0004
H = 10		0.0550	0.0042	0.0043		0.0066	0.0067	0.0017
H = 20		0.0583	0.0017	0.0020		0.0051	0.0001	0.0001
N = 170	Truth:	0.0569			Truth:	0.0056		
H = 2		0.0500	0.0083	0.0112		0.0064	0.0002	0.0003
H = 3		0.0568	0.0060	0.0073		0.0052	0.0002	0.0002
H = 4		0.0584	0.0048	0.0054		0.0054	0.0001	0.0002
H = 5		0.0542	0.0038	0.0042		0.0061	0.0001	0.0001
H = 10		0.0555	0.0022	0.0023		0.0057	0.0001	0.0001
H = 20		0.0566	0.0011	0.0012		0.0055	0	0
N = 30	Truth:	0.1108			Truth:	0.0103		
H = 2		0.1113	0.0290	0.0250		0.0113	0.1037	0.0017
H = 3		0.1115	0.0145	0.0135		0.0077	0.0102	0.0024
H = 4		0.1145	0.0093	0.0102		0.0093	0.0007	0.0005
H = 5		0.1089	0.0069	0.0079		0.0107	0.0006	0.0004
H = 10		0.1122	0.0033	0.0041		0.0098	0.0002	0.0002
H = 20		0.1140	0.0017	0.0019		0.0104	0.0015	0.0002
N = 170	Truth:	0.1108			Truth:	0.0103		
H = 2		0.1135	0.0080	0.0111		0.0101	0.0002	0.0003
H = 3		0.1107	0.0060	0.0075		0.0110	0.0002	0.0002
H = 4		0.1084	0.0046	0.0053		0.0096	0.0001	0.0002
H = 5		0.1120	0.0039	0.0043		0.0112	0.0001	0.0001
H = 10		0.1119	0.0021	0.0022		0.0102	0.0001	0.0001
H = 20		0.1116	0.0011	0.0011		0.0104	0	0

Table B.8: Means, empirical standard deviations and mean squared error over the 1000 replications of each scenario's meta-analysis results for the effects of age and the interaction between age and exposure - Very high between-trial heterogeneity 2 of 2

Very high between-trial heterogeneity								
		γ_A				γ_I		
		mean	SD	MSE		mean	SD	MSE
N = 30	Truth:	0.2105			Truth:	0.0175		
H = 2		0.2122	0.0300	0.0258		0.0168	0.0030	0.0015
H = 3		0.2074	0.0137	0.0140		0.0164	0.0011	0.0008
H = 4		0.2080	0.0108	0.0103		0.0195	0.0116	0.0015
H = 5		0.2106	0.0071	0.0082		0.0162	0.0080	0.0011
H = 10		0.2144	0.0032	0.0035		0.0170	0.0002	0.0002
H = 20		0.2107	0.0017	0.0019		0.0176	0.0001	0.0001
N = 170	Truth:	0.2105			Truth:	0.0175		
H = 2		0.2137	0.0080	0.0108		0.0172	0.0002	0.0003
H = 3		0.2106	0.0060	0.0077		0.0173	0.0002	0.0002
H = 4		0.2163	0.0048	0.0054		0.0177	0.0001	0.0002
H = 5		0.2098	0.0039	0.0046		0.0179	0.0001	0.0001
H = 10		0.2113	0.0021	0.0024		0.0171	0.0001	0.0001
H = 20		0.2087	0.0011	0.0011		0.0177	0	0
N = 30	Truth:	0.0569			Truth:	0.0137		
H = 2		0.0601	0.0582	0.0283		0.0145	0.0052	0.0020
H = 3		0.0575	0.0141	0.0144		0.0140	0.0011	0.0008
H = 4		0.0568	0.0093	0.0095		0.0140	0.0007	0.0005
H = 5		0.0504	0.0073	0.0080		0.0159	0.0069	0.0021
H = 10		0.0562	0.0033	0.0039		0.0145	0.0002	0.0002
H = 20		0.0595	0.0017	0.0020		0.0133	0.0001	0.0001
N = 170	Truth:	0.0569			Truth:	0.0137		
H = 2		0.0546	0.0082	0.0109		0.0144	0.0002	0.0003
H = 3		0.0566	0.0061	0.0071		0.0136	0.0002	0.0002
H = 4		0.0570	0.0048	0.0053		0.0134	0.0001	0.0002
H = 5		0.0559	0.0040	0.0045		0.0140	0.0001	0.0001
H = 10		0.0581	0.0021	0.0022		0.0139	0.0001	0.0001
H = 20		0.0569	0.0011	0.0011		0.0136	0	0
N = 30	Truth:	0.0569			Truth:	0.0274		
H = 2		0.0513	0.0428	0.0275		0.0287	0.0044	0.0019
H = 3		0.0606	0.0136	0.0138		0.0258	0.0012	0.0008
H = 4		0.0596	0.0093	0.0107		0.0273	0.0007	0.0005
H = 5		0.0576	0.0072	0.0082		0.0284	0.0196	0.0045
H = 10		0.0593	0.0034	0.0040		0.0275	0.0002	0.0002
H = 20		0.0588	0.0017	0.0021		0.0274	0.0001	0.0001
N = 170	Truth:	0.0569			Truth:	0.0274		
H = 2		0.0619	0.0082	0.0111		0.0268	0.0002	0.0003
H = 3		0.0558	0.0059	0.0077		0.0273	0.0002	0.0002
H = 4		0.0559	0.0047	0.0055		0.0276	0.0001	0.0002
H = 5		0.0584	0.0039	0.0046		0.0272	0.0001	0.0001
H = 10		0.0569	0.0021	0.0021		0.0275	0.0001	0.0001
H = 20		0.0562	0.0011	0.0011		0.0272	0	0

Table B.9: Means and empirical standard deviations over the 1000 replications of each scenario's probability of the extrapolation assumption holding - Low between-trial heterogeneity

Low between-trial heterogeneity						
		mean	SD		mean	SD
Scenario 1	N = 30			N = 170		
H = 2		0.170	0.073		0.387	0.068
H = 3		0.236	0.079		0.451	0.065
H = 4		0.286	0.079		0.479	0.057
H = 5		0.330	0.082		0.506	0.048
H = 10		0.437	0.069		0.548	0.032
H = 20		0.512	0.045		0.572	0.021
Scenario 2	N = 30			N = 170		
H = 2		0.152	0.070		0.327	0.083
H = 3		0.207	0.078		0.368	0.078
H = 4		0.250	0.085		0.397	0.074
H = 5		0.279	0.081		0.415	0.069
H = 10		0.360	0.073		0.445	0.053
H = 20		0.412	0.059		0.461	0.039
Scenario 3	N = 30			N = 170		
H = 2		0.122	0.064		0.236	0.085
H = 3		0.156	0.069		0.254	0.085
H = 4		0.174	0.073		0.267	0.076
H = 5		0.192	0.074		0.273	0.073
H = 10		0.236	0.069		0.284	0.057
H = 20		0.268	0.055		0.292	0.042
Scenario 4	N = 30			N = 170		
H = 2		0.056	0.040		0.083	0.052
H = 3		0.058	0.041		0.077	0.043
H = 4		0.061	0.039		0.076	0.042
H = 5		0.061	0.040		0.076	0.037
H = 10		0.064	0.033		0.070	0.025
H = 20		0.064	0.025		0.068	0.018
Scenario 5	N = 30			N = 170		
H = 2		0.137	0.064		0.277	0.089
H = 3		0.178	0.070		0.308	0.086
H = 4		0.211	0.077		0.323	0.081
H = 5		0.236	0.082		0.335	0.079
H = 10		0.296	0.079		0.354	0.061
H = 20		0.330	0.066		0.363	0.045
Scenario 6	N = 30			N = 170		
H = 2		0.108	0.055		0.165	0.079
H = 3		0.129	0.065		0.171	0.075
H = 4		0.144	0.069		0.168	0.071
H = 5		0.151	0.071		0.172	0.064
H = 10		0.163	0.070		0.168	0.052
H = 20		0.165	0.063		0.168	0.038

Table B.10: Means and empirical standard deviations over the 1000 replications of each scenario's probability of the extrapolation assumption holding - Moderate between-trial heterogeneity

Moderate between-trial heterogeneity						
		mean	SD		mean	SD
Scenario 1	N = 30			N = 170		
H = 2		0.163	0.077		0.347	0.096
H = 3		0.227	0.085		0.409	0.087
H = 4		0.269	0.089		0.436	0.080
H = 5		0.313	0.088		0.466	0.072
H = 10		0.416	0.080		0.523	0.047
H = 20		0.493	0.053		0.555	0.032
Scenario 2	N = 30			N = 170		
H = 2		0.144	0.073		0.303	0.106
H = 3		0.188	0.082		0.341	0.109
H = 4		0.238	0.093		0.373	0.099
H = 5		0.267	0.094		0.385	0.093
H = 10		0.346	0.088		0.432	0.074
H = 20		0.404	0.070		0.451	0.053
Scenario 3	N = 30			N = 170		
H = 2		0.115	0.068		0.219	0.104
H = 3		0.152	0.081		0.241	0.105
H = 4		0.172	0.084		0.257	0.100
H = 5		0.187	0.085		0.267	0.098
H = 10		0.235	0.085		0.285	0.080
H = 20		0.264	0.070		0.292	0.061
Scenario 4	N = 30			N = 170		
H = 2		0.057	0.048		0.088	0.075
H = 3		0.060	0.046		0.086	0.065
H = 4		0.064	0.052		0.083	0.059
H = 5		0.065	0.048		0.077	0.052
H = 10		0.066	0.042		0.077	0.039
H = 20		0.068	0.031		0.072	0.026
Scenario 5	N = 30			N = 170		
H = 2		0.132	0.072		0.252	0.112
H = 3		0.175	0.085		0.283	0.109
H = 4		0.209	0.089		0.301	0.102
H = 5		0.227	0.091		0.321	0.100
H = 10		0.287	0.091		0.344	0.080
H = 20		0.324	0.079		0.362	0.062
Scenario 6	N = 30			N = 170		
H = 2		0.102	0.064		0.158	0.096
H = 3		0.129	0.075		0.168	0.097
H = 4		0.139	0.080		0.173	0.091
H = 5		0.148	0.082		0.169	0.081
H = 10		0.164	0.082		0.169	0.065
H = 20		0.168	0.071		0.172	0.050

Table B.11: Means and empirical standard deviations over the 1000 replications of each scenario's probability of the extrapolation assumption holding - High between-trial heterogeneity

High between-trial heterogeneity						
		mean	SD		mean	SD
Scenario 1	N = 30			N = 170		
H = 2		0.154	0.081		0.316	0.113
H = 3		0.212	0.087		0.373	0.104
H = 4		0.255	0.093		0.410	0.098
H = 5		0.295	0.093		0.434	0.084
H = 10		0.398	0.085		0.501	0.060
H = 20		0.482	0.061		0.543	0.040
Scenario 2	N = 30			N = 170		
H = 2		0.142	0.080		0.280	0.118
H = 3		0.189	0.087		0.319	0.115
H = 4		0.227	0.092		0.345	0.113
H = 5		0.257	0.099		0.374	0.105
H = 10		0.337	0.095		0.413	0.084
H = 20		0.399	0.076		0.444	0.067
Scenario 3	N = 30			N = 170		
H = 2		0.112	0.075		0.217	0.117
H = 3		0.142	0.086		0.247	0.121
H = 4		0.166	0.091		0.252	0.116
H = 5		0.187	0.094		0.256	0.106
H = 10		0.231	0.089		0.274	0.091
H = 20		0.261	0.077		0.285	0.069
Scenario 4	N = 30			N = 170		
H = 2		0.058	0.052		0.098	0.088
H = 3		0.064	0.057		0.089	0.078
H = 4		0.067	0.056		0.091	0.075
H = 5		0.067	0.055		0.085	0.064
H = 10		0.066	0.050		0.084	0.051
H = 20		0.070	0.035		0.074	0.033
Scenario 5	N = 30			N = 170		
H = 2		0.128	0.076		0.236	0.120
H = 3		0.172	0.089		0.273	0.120
H = 4		0.205	0.096		0.287	0.114
H = 5		0.222	0.100		0.306	0.110
H = 10		0.279	0.101		0.333	0.091
H = 20		0.319	0.088		0.352	0.070
Scenario 6	N = 30			N = 170		
H = 2		0.096	0.065		0.165	0.112
H = 3		0.121	0.077		0.172	0.112
H = 4		0.138	0.085		0.174	0.108
H = 5		0.142	0.085		0.172	0.097
H = 10		0.165	0.092		0.171	0.080
H = 20		0.171	0.078		0.173	0.060

Table B.12: Means and empirical standard deviations over the 1000 replications of each scenario's probability of the extrapolation assumption holding - Very high between-trial heterogeneity

Very high between-trial heterogeneity						
		mean	SD		mean	SD
Scenario 1	N = 30			N = 170		
H = 2		0.141	0.085		0.273	0.128
H = 3		0.199	0.098		0.332	0.125
H = 4		0.245	0.103		0.367	0.111
H = 5		0.280	0.101		0.396	0.102
H = 10		0.379	0.095		0.466	0.074
H = 20		0.464	0.072		0.520	0.053
Scenario 2	N = 30			N = 170		
H = 2		0.131	0.086		0.253	0.133
H = 3		0.177	0.095		0.287	0.128
H = 4		0.208	0.102		0.322	0.124
H = 5		0.239	0.105		0.340	0.120
H = 10		0.325	0.104		0.392	0.100
H = 20		0.386	0.088		0.433	0.079
Scenario 3	N = 30			N = 170		
H = 2		0.108	0.077		0.201	0.132
H = 3		0.140	0.090		0.222	0.131
H = 4		0.159	0.094		0.247	0.129
H = 5		0.182	0.108		0.240	0.120
H = 10		0.224	0.101		0.269	0.101
H = 20		0.252	0.089		0.282	0.084
Scenario 4	N = 30			N = 170		
H = 2		0.061	0.062		0.097	0.098
H = 3		0.071	0.072		0.099	0.094
H = 4		0.069	0.068		0.090	0.081
H = 5		0.072	0.068		0.094	0.080
H = 10		0.070	0.058		0.085	0.057
H = 20		0.069	0.041		0.078	0.044
Scenario 5	N = 30			N = 170		
H = 2		0.114	0.081		0.222	0.133
H = 3		0.156	0.091		0.257	0.130
H = 4		0.190	0.104		0.273	0.125
H = 5		0.209	0.101		0.284	0.121
H = 10		0.268	0.109		0.319	0.110
H = 20		0.315	0.097		0.347	0.085
Scenario 6	N = 30			N = 170		
H = 2		0.097	0.074		0.157	0.123
H = 3		0.125	0.087		0.167	0.122
H = 4		0.132	0.091		0.172	0.120
H = 5		0.152	0.098		0.173	0.107
H = 10		0.163	0.098		0.175	0.091
H = 20		0.168	0.085		0.177	0.070

Table B.13: Means and empirical standard deviations of the average ESS over the 1000 replications - Model 1

		mean	SD		mean	SD
Low between-trial heterogeneity	N = 30			N = 170		
H = 2		5.425	2.312		13.504	1.043
H = 3		7.648	2.601		17.093	1.266
H = 4		9.465	2.588		19.122	1.278
H = 5		11.394	2.503		20.478	1.084
H = 10		16.482	1.899		22.821	0.621
H = 20		20.305	1.127		23.828	0.444
Moderate between-trial heterogeneity	N = 30			N = 170		
H = 2		5.226	2.294		13.102	1.140
H = 3		7.549	2.577		16.305	1.663
H = 4		9.146	2.581		18.199	1.712
H = 5		11.038	2.394		19.487	1.557
H = 10		15.986	1.866		21.924	1.089
H = 20		19.807	0.997		23.292	0.576
High between-trial heterogeneity	N = 30			N = 170		
H = 2		5.206	2.345		12.932	1.151
H = 3		7.361	2.568		15.815	1.834
H = 4		9.176	2.567		17.544	2.138
H = 5		10.534	2.574		18.644	2.146
H = 10		15.555	1.856		21.252	1.360
H = 20		19.493	1.095		22.749	0.718
Very high between-trial heterogeneity	N = 30			N = 170		
H = 2		5.260	2.294		12.596	1.121
H = 3		7.232	2.546		15.197	2.229
H = 4		8.894	2.545		16.571	2.562
H = 5		10.466	2.489		17.534	2.588
H = 10		15.148	1.808		20.023	1.817
H = 20		19.039	1.28		21.936	0.902

Appendix C

Online supplementary material to accompany Chapter 7.

Below is the supplementary material for Chapter 7: "Exposure-response modelling approaches for determining optimal dosing rules in children" by Wadsworth I, Hampson LV, Jaki T, Bornkamp B

This document contains the following appendices:

C.1: Worked example of methods;

C.2: Simulation scenarios in detail;

C.3: Inclusion of additional covariate;

C.4: Supplementary tables;

C.5: Supplementary plot.

C.1 Worked example of methods

This appendix aims to give an illustration of the output that would be seen from fitting each of the methods in Section 7.5 to a single set of simulated data. For each of the approaches we estimate the relationship between intercept or slope and age. For $i = 1, \ldots, 100$ subjects, we simulate the response as:

$$Y_i = \begin{cases} 5.1 - 0.010C_i + \epsilon_i, & \text{for } A_i \in (0, 4] \\ 4.8 - 0.035C_i + \epsilon_i, & \text{for } A_i \in (4, 10] \\ 4.4 - 0.075C_i + \epsilon_i, & \text{for } A_i \in (10, 14] \\ 3.9 - 0.125C_i + \epsilon_i, & \text{for } A_i \in (14, 18] \end{cases}$$

where the C_i exposure values are simulated as in Section 7.7 following Wadsworth et al.[209] and the ϵ_i are random errors simulated from a normal distribution with mean 0 and variance 0.02. The A_i age values are simulated from four Uniform distributions such that there are 25 subjects in each of four age groups: 0 to 4 years; 4 to 10 years; 10 to 14 years; and 14 to 18 years.

First, the linear model with categorical covariates as shown in Section 7.5.1 is fitted to the simulated example data. Using the true age groups to define $A_{1i}, \ldots, A_{Hi}$ (the age groups used for the categorical covariates), the following intercepts and slopes are estimated in turn for each of the four age groups: intercept estimates are 5.13, 4.80,

4.42 and 3.84; and slope estimates are -0.010, -0.031, -0.080 and -0.120.

Now, we fit a single PALM tree model as described in Section 7.5.2. Figure C.1 shows the results of a PALM tree fitted to simulated data; this is standard output from the 'partykit' package.[2,3] Four nodes (here, age groups) have been found in the following age groupings: 0 to 3.89 years; 3.89 to 9.94 years; 9.94 to 14.00 years; and 14.00 to 18 years. Other than 0 and 18 (fixed based on the paediatric population), the age group bounds are observed age values from the data; were there an age data point less than 4, but closer to 4 than 3.89, this age boundary could be even closer to the truth. Regardless, these age groups are very close to the true age groups and estimate the underlying PK-PD parameters well also. For each age group in turn, the intercepts are 5.13, 4.80, 4.42 and 3.84 and the slopes are -0.010, -0.031, -0.080 and -0.120. For this data, this model gives identical estimates to the linear model with categorical covariates, to six decimal places.

We then extend to the bootstrapped PALM trees also described in Section 7.5.2. Figure C.2 presents plots of the bootstrapped PALM fits of intercept and slope parameters over age constructed by following the approach given in Section 7.5.2. We plot the relationship

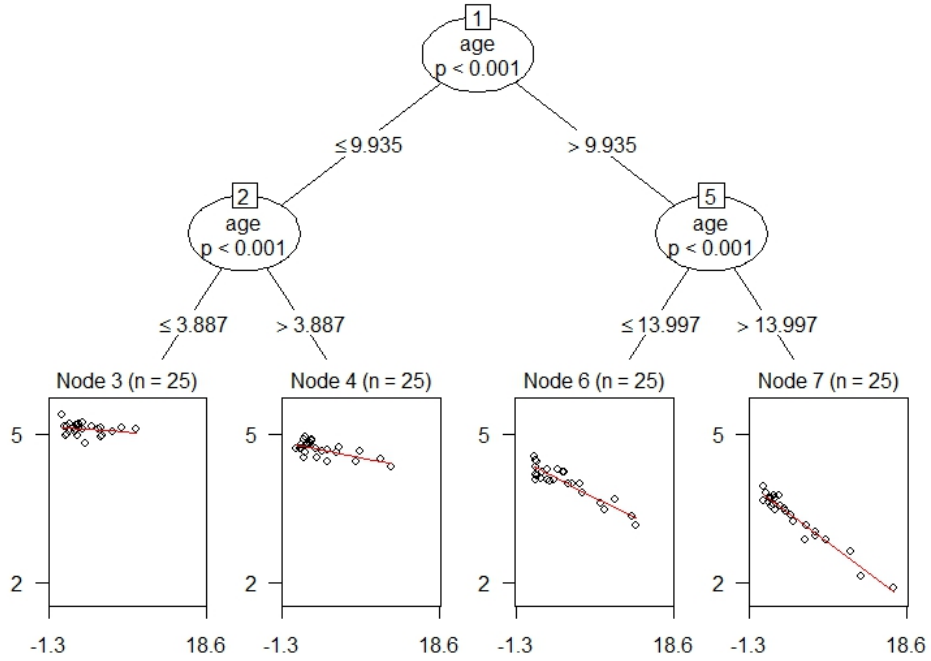
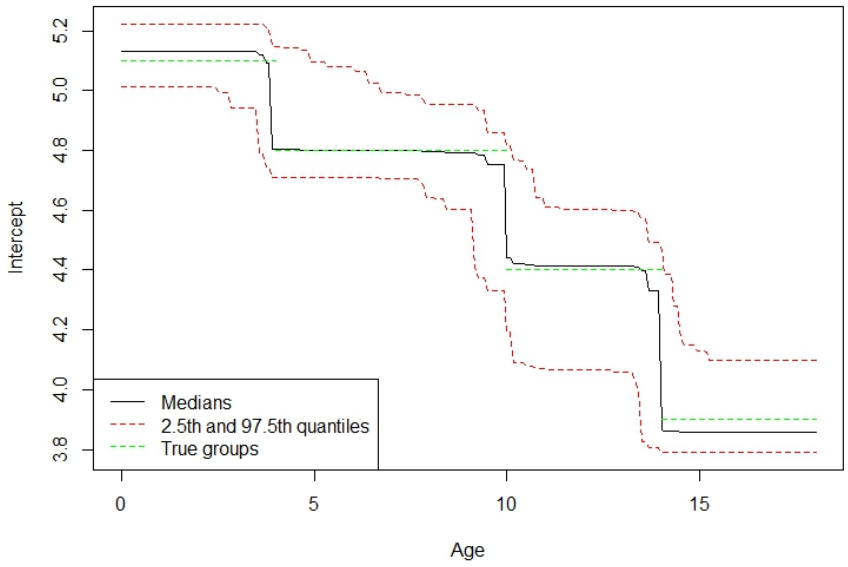


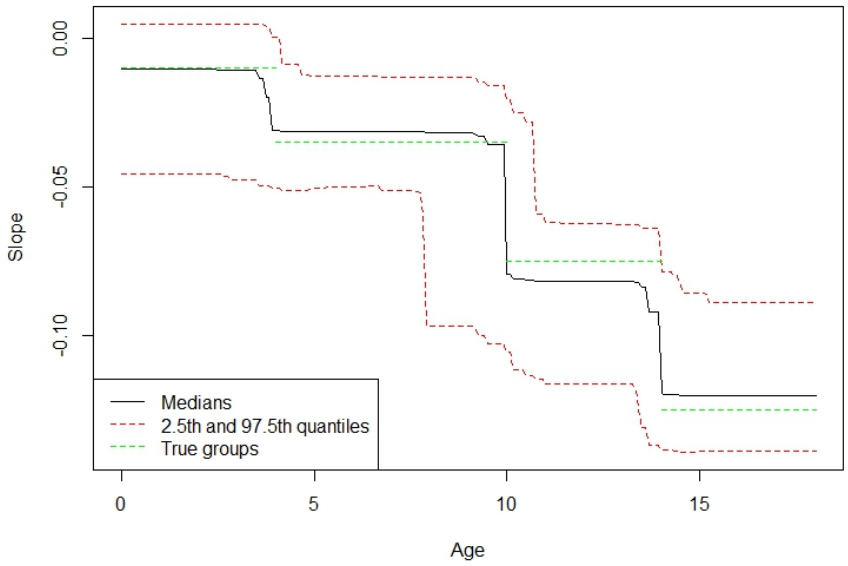
Figure C.1: Example to demonstrate the structure of a single PALM tree fitted to simulated data, produced from the 'partykit' package.[2,3]

between intercept or slope against age using the bootstrap averaged median, 2.5th and 97.5th quantiles at a continuum of ages from 0 to 18 years, also highlighting the true underlying intercept/slope values by green dashed lines. The 2.5th and 97.5th quantile lines are asymmetric as the distribution (over the bootstrap samples) of the intercept / slope values is asymmetric at many age values from 0 to 18.

Next, we apply the B-splines approach described in Section 7.5.3. Figure C.3 presents plots of the median, 2.5th and 97.5th quantiles of the posterior distributions of the intercept/slope parameters at each A_i from the MCMC output of the B-spline model fit, also highlighting



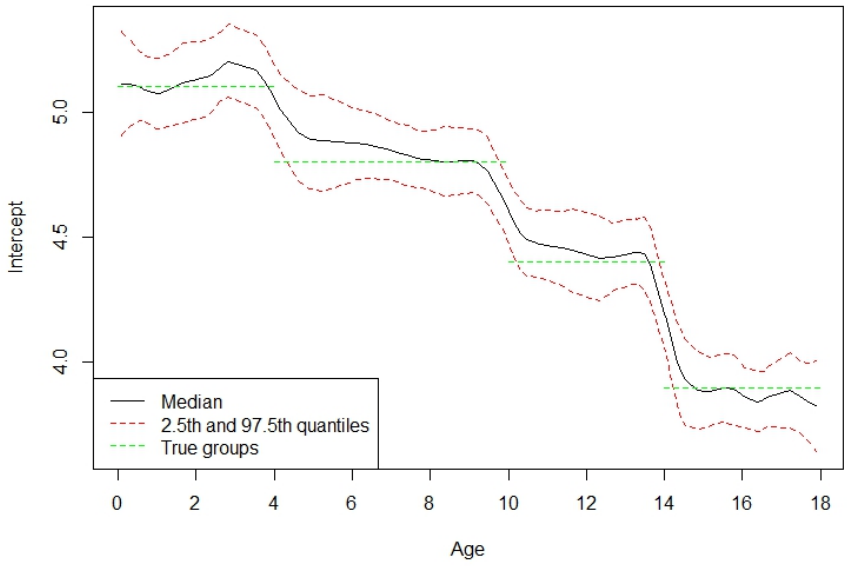
(a)



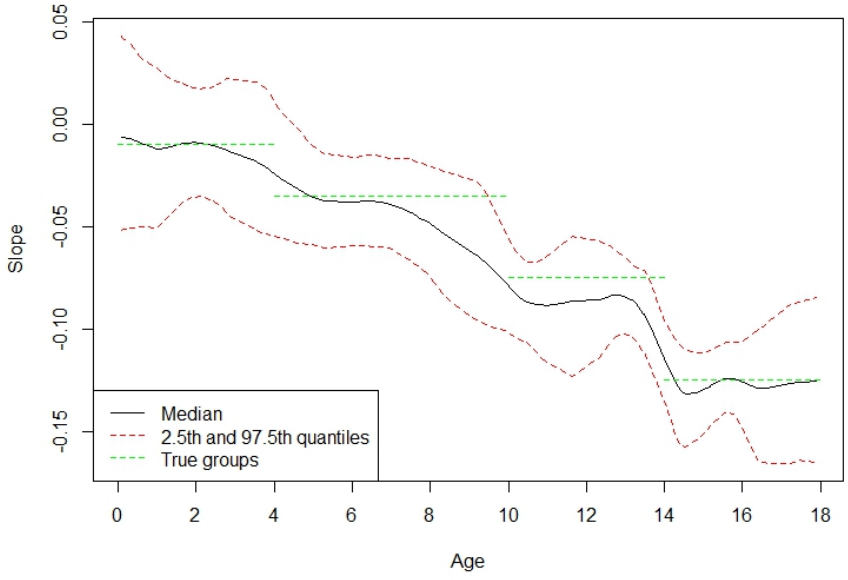
(b)

Figure C.2: (a) Plot of the the intercept parameter over age from the bootstrapped PALM fit to the simulated example data, showing the median intercept, with 2.5th and 97.5th quantiles, over the 1000 simulated bootstrap samples and true parameter values given by the green dotted lines. (b) Plot of the the slope parameter over age from the bootstrapped PALM fit to the simulated example data, showing the median slope, with 2.5th and 97.5th quantiles, over the 1000 simulated bootstrap samples and true parameter values given by the green dotted lines.

the true underlying intercept/slope values by green dashed lines.



(a)



(b)

Figure C.3: (a) Plot of the the intercept parameter over age from the B-spline fit to the simulated example data, showing the median intercept with 2.5th and 97.5th quantiles and true parameter values given by the green dotted lines. (b) Plot of the the slope parameter over age from the B-spline fit to the simulated example data, showing the median slope with 2.5th and 97.5th quantiles and true parameter values given by the green dotted lines.

C.2 Simulation scenarios in detail

In terms of how the E-R model parameters change over age, we consider 11 scenarios for data generation:

1. Step function relates how E-R model parameters change over age, following ICH guidance document age groupings with substantial differences between E-R model parameter values in each age group;

2. E-R model parameters have a less steep linear transition between age groups, with parameter values and age groups the same as scenario 1;

3. E-R model parameter values have same change over age as in scenario 2, though now age groups parameter values are more similar between age groups to more closely resemble what might be observed in reality;

4. Step function relates how E-R model parameters change over age, as in scenario 1, but now there is a deviation from the ICH age groups so that there are more distinct age groups amongst younger children, following NICHD groupings;

5. E-R model parameters constant over all age groups;

6. E-R model parameters have a constant linear decrease over the whole age range;

7. Intercept term as in scenario 3, slope term constant over age as in scenario 6;

8. Intercept term constant over age as in scenario 6, slope term as in scenario 3;

9. Same as scenario 3, though now the true age groups will be 0 to 6 months, 6 months to 3 years, 3 years to 11 years, 11 years to 18 years;

10. Same as scenario 3, though now the true age groups will be 0 to 2 year, 2 year to 6 years, 6 years to 14 years, 14 years to 18 years;

11. Same as scenario 3, though now the true age groups will be 0 to 8 years, 8 years to 12 years, 12 years to 16 years, 16 years to 18 years.

C.3 Inclusion of additional covariate

For all scenarios in Chapter 7, the response has been modelled as in equation (7.1) without additional covariates $x_{1i}, \ldots, x_{Pi}$. In this appendix, the data are generated such that there is a relationship between response and an additional covariate.

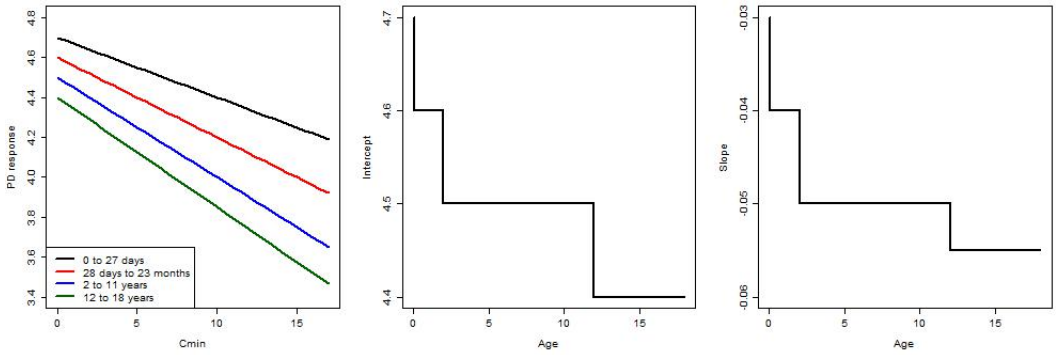


Figure C.4: Plots for the supplementary scenario showing (1) the true underlying E-R relationship; (2) how the intercept of the E-R model changes with age; (3) how the slope of the E-R model changes with age.

Assume we have data on body weight, x_w, which is modelled as a linear function of age; the linear relationship we use, $x_w = 3A + 7$, is based on the use of weight estimation in paediatrics (1 to 13 years, inclusive) suggested by Luscombe et al.,[217] though other suggested weight/age relationships exist. We assume that, like age, this covariate has an effect on response. However, as body weight here is largely explained by age it feels more natural to regress body weight against age and to consider the effect of the fitted residuals, r_w, in the model, essentially, what effect of body weight on response remains after adjusting for age. We simulate the response in this example by having the body weight residuals, r_w (which have a standard normal distribution), included in the simulation model as follows:

$$Y_i = \gamma_0(A_i) + 0.4r_w + \gamma_C(A_i)C_i + \epsilon_i. \tag{C.1}$$

When fitting the model, we consider two approaches: one approach where body weight has been observed and the residuals are included

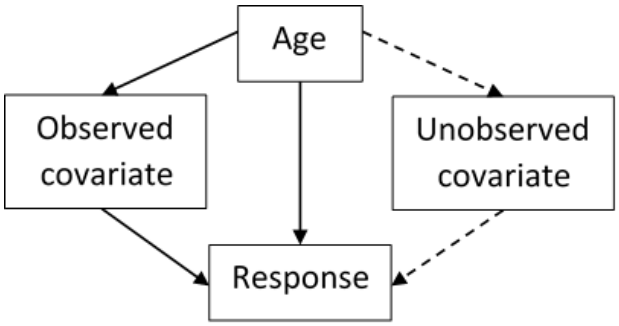


Figure C.5: Directed acyclic graph illustrating the causal relationships between age, observed/unobserved covariates and response.

in the model; and a variation where the effect of the body weight residuals still exists, though body weight is now an unobserved covariate and not included in the model. The directed acyclic graph shown in Figure C.5 aims to visualise this causal relationship. For this second modelling approach, we seek to identify how the methods cope when there is an effect that we are unable to observe and control for. We will therefore fit the models in Section 7.5 to this scenario, modelling the PD response in two ways.

Comparing the panels in Figures C.6 representing the supplementary scenario with and without r_w, it is clear that when the additional covariate is included in the simulation model, but not included in the analysis model, all approaches do not perform as well at estimating the underlying relationship between age and the exposure-response model slope or intercept parameters. However, the B-splines approach again seems to perform better than the other approaches in terms of accuracy and precision.

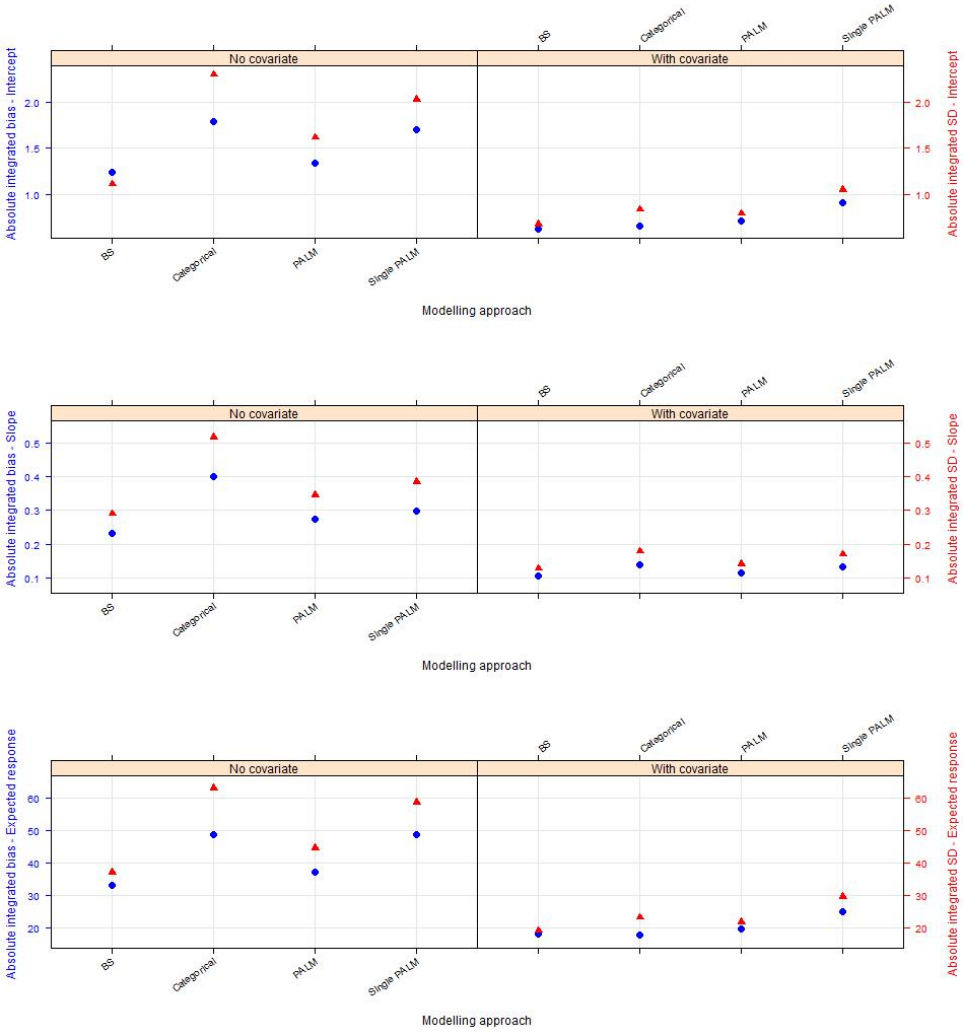


Figure C.6: Integrated absolute bias (blue circles) and integrated empirical SD (red triangles) for (1) E-R model intercept; (2) E-R model slope; (3) expected response. On the horizontal axis, 'BS' refers to the Bayesian penalised B-splines approach, 'Categorical' the linear model adjusted for a categorical age covariate, and 'PALM' and 'singlePALM' label the bootstrapped PALM tree approach and single PALM tree, respectively. Panels display the supplementary scenario with and without the additional covariate in the analysis model.

Figure C.7 shows that when the additional covariate is not included in the analysis model, the accuracy of the K-group optimal dosing rule is lower and the true expected response (derived from the simulation model when children are dosed according to the estimated optimal dosing rule) is further from the target response, for both the bootstrapped PALM trees and Bayesian penalised B-splines ap-

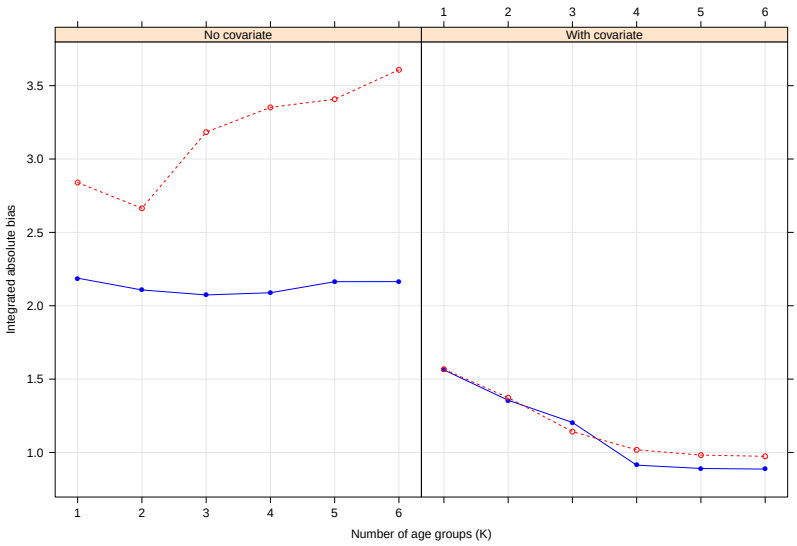


Figure C.7: Integrated absolute difference between the target response and true expected response when children are dosed according to the K group optimal dosing rule, for the supplementary scenario. Results are shown for dosing rules obtained modelling the E-R relationship using Bayesian penalised B-splines (solid blue line) and bootstrapped PALM trees (dashed red line).

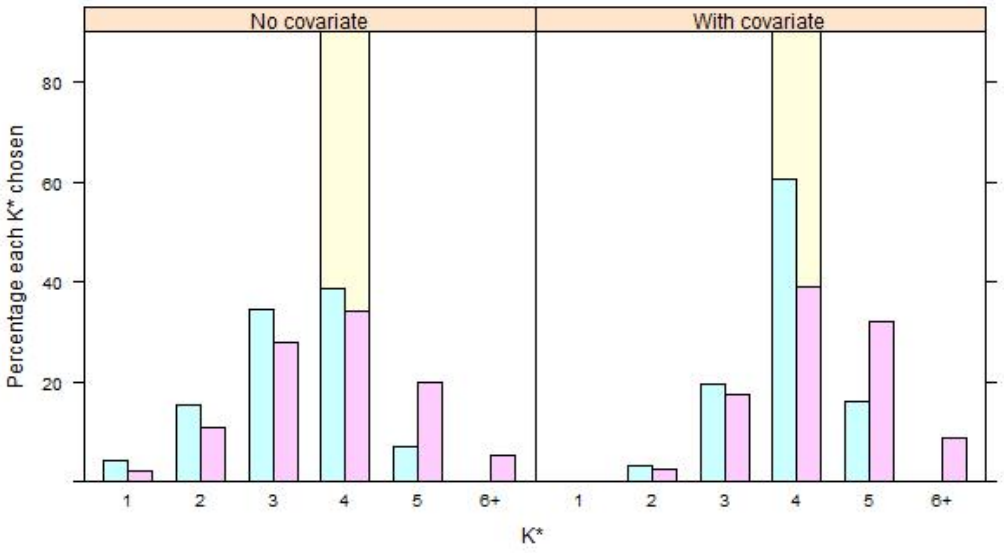


Figure C.8: Percentage of 1000 simulations in which K^*, the optimal number of age groups in the dosing rule, takes each value shown. K^* is selected according to the algorithm described in Section 7.6.2 for Bayesian penalised B-spline (blue) and bootstrapped PALM tree (pink) approaches. The values of K^* chosen by applying the algorithm in Section 7.6.2 to the true underlying E-R relationships in the supplementary scenario are shown by the yellow bars.

proaches. However, the Bayesian penalised B-splines approach does seem to provide better accuracy than the bootstrapped PALM trees. Figure C.8 shows that when the additional covariate is included in the analysis model, the majority of simulated datasets would lead to the investigator selecting a global optimum dosing rule with $K^* = 4$, especially for the Bayesian penalised B-splines approach.

C.4 Supplementary tables

Table C.1: True E-R model parameter values for age groupings in scenarios 1 to 5

		Simulation scenario			
Parameter	Age group	1 & 2	3	4	5
γ_0	0 to 28 days	4.85	4.60	4.60	4.50
	28 days to 1 year	4.65	4.55	4.52	4.50
	1 to 2 years	4.65	4.55	4.44	4.50
	2 to 6 years	4.45	4.50	4.35	4.50
	6 to 12 years	4.45	4.50	4.27	4.50
	12 to 18 years	4.25	4.45	4.20	4.50
γ_E	0 to 28 days	-0.045	-0.040	-0.035	-0.050
	28 days to 1 year	-0.055	-0.045	-0.045	-0.050
	1 to 2 years	-0.055	-0.045	-0.055	-0.050
	2 to 6 years	-0.065	-0.050	-0.065	-0.050
	6 to 12 years	-0.065	-0.050	-0.075	-0.050
	12 to 18 years	-0.075	-0.055	-0.085	-0.050

Table C.2: True E-R model parameter values for age groupings in scenarios 6 to 8.

		Simulation scenario		
Parameter	Age group	6	7	8
γ_0	0 to 28 days	4.700 – 4.698	4.60	4.50
	28 days to 1 year	4.698 – 4.680	4.55	4.50
	1 to 2 years	4.680 – 4.660	4.55	4.50
	2 to 6 years	4.660 – 4.580	4.50	4.50
	6 to 12 years	4.580 – 4.460	4.50	4.50
	12 to 18 years	4.460 – 4.340	4.45	4.50
γ_E	0 to 28 days	-0.050 – -0.050	-0.050	-0.040
	28 days to 1 year	-0.050 – -0.052	-0.050	-0.045
	1 to 2 years	-0.052 – -0.053	-0.050	-0.045
	2 to 6 years	-0.053 – -0.059	-0.050	-0.050
	6 to 12 years	-0.059 – -0.068	-0.050	-0.050
	12 to 18 years	-0.068 – -0.077	-0.050	-0.055

Table C.3: Average absolute bias, empirical standard deviation (ESD) and empirical mean squared error (EMSE) for the estimated B-spline intercept/slope over age curve, integrated over age for each scenario considered.

	Intercept			Slope			Expected response		
Scenario	Bias	ESD	EMSE	Bias	ESD	EMSE	Bias	ESD	EMSE
1	0.831	0.811	0.063	0.110	0.132	0.001	21.994	21.214	2.672
2	0.720	0.790	0.045	0.103	0.125	0.001	18.879	20.195	1.927
3	0.496	0.557	0.022	0.091	0.113	0.001	14.058	16.574	1.152
4	0.644	0.726	0.036	0.122	0.145	0.001	20.405	21.440	2.355
5	0.339	0.421	0.010	0.074	0.093	0.000	10.381	13.254	0.675
6	0.571	0.706	0.030	0.100	0.126	0.001	15.568	18.994	1.439
7	0.526	0.544	0.024	0.087	0.103	0.001	12.501	15.018	0.934
8	0.346	0.434	0.011	0.083	0.100	0.001	11.940	14.354	0.877
9	0.526	0.580	0.024	0.090	0.113	0.001	14.175	16.622	1.159
10	0.557	0.611	0.027	0.097	0.123	0.001	14.946	17.706	1.305
11	0.570	0.578	0.029	0.093	0.117	0.001	14.950	17.128	1.280

Table C.4: Average absolute bias, empirical standard deviation (ESD) and empirical mean squared error (EMSE) for the estimated bootstrapped PALM intercept/slope over age curve, integrated over age for each scenario considered.

	Intercept			Slope			Expected response		
Scenario	Bias	ESD	EMSE	Bias	ESD	EMSE	Bias	ESD	EMSE
1	0.811	0.875	0.063	0.129	0.162	0.002	21.690	23.930	2.905
2	0.763	0.860	0.052	0.125	0.157	0.001	20.588	23.762	2.473
3	0.553	0.668	0.027	0.103	0.130	0.001	16.089	19.138	1.514
4	0.702	0.810	0.043	0.130	0.156	0.002	22.774	24.781	2.987
5	0.392	0.508	0.014	0.076	0.098	0.001	9.690	12.429	0.596
6	0.759	0.845	0.051	0.116	0.145	0.001	21.741	23.790	2.662
7	0.573	0.694	0.029	0.092	0.118	0.001	13.359	16.536	1.041
8	0.437	0.548	0.017	0.093	0.111	0.001	12.668	14.691	0.997
9	0.564	0.682	0.028	0.103	0.130	0.001	16.379	19.371	1.551
10	0.583	0.715	0.030	0.104	0.131	0.001	16.786	20.444	1.634
11	0.593	0.711	0.032	0.105	0.132	0.001	16.923	19.635	1.733

Table C.5: Average absolute bias, empirical standard deviation (ESD) and empirical mean squared error (EMSE) for the estimated single PALM tree intercept/slope over age step function curve, integrated over age for each scenario considered.

	Intercept			Slope			Expected response		
Scenario	Bias	ESD	EMSE	Bias	ESD	EMSE	Bias	ESD	EMSE
1	0.821	1.051	0.076	0.142	0.183	0.002	21.921	28.674	3.548
2	0.856	1.086	0.073	0.143	0.181	0.002	23.066	29.816	3.434
3	0.689	0.825	0.041	0.116	0.148	0.001	21.454	25.290	2.514
4	0.942	1.117	0.078	0.166	0.205	0.003	32.076	37.016	5.766
5	0.340	0.446	0.011	0.069	0.088	0.0004	8.686	11.151	0.480
6	1.070	1.174	0.097	0.141	0.179	0.002	31.382	34.460	5.209
7	0.760	0.905	0.048	0.097	0.129	0.001	16.499	20.326	1.490
8	0.391	0.512	0.015	0.106	0.111	0.001	15.696	15.713	1.472
9	0.711	0.848	0.043	0.118	0.150	0.001	22.087	25.688	2.645
10	0.742	0.896	0.048	0.119	0.149	0.001	22.596	26.905	2.858
11	0.689	0.837	0.044	0.116	0.149	0.001	20.588	24.893	2.607

Table C.6: Average absolute bias, empirical standard deviation (ESD) and empirical mean squared error (EMSE) for the estimated intercept/slope over age step function curve based on the categorical covariates model approach, integrated over age for each scenario considered.

	Intercept			Slope			Expected response		
Scenario	Bias	ESD	EMSE	Bias	ESD	EMSE	Bias	ESD	EMSE
1	0.645	0.819	0.037	0.133	0.172	0.002	16.881	21.832	1.868
2	0.729	0.837	0.048	0.138	0.177	0.002	19.589	22.973	2.419
3	0.648	0.807	0.037	0.134	0.170	0.002	17.299	21.917	1.910
4	0.837	0.889	0.062	0.165	0.200	0.003	26.975	25.637	4.257
5	0.657	0.823	0.038	0.135	0.172	0.002	17.434	22.263	1.929
6	0.996	0.904	0.086	0.162	0.198	0.002	29.114	25.421	4.722
7	0.664	0.827	0.039	0.137	0.179	0.002	17.761	23.049	2.094
8	0.651	0.821	0.037	0.137	0.176	0.002	17.628	22.626	2.020
9	0.708	0.845	0.044	0.141	0.180	0.002	19.509	23.098	2.398
10	0.776	0.851	0.054	0.150	0.186	0.002	23.047	23.907	3.179
11	0.765	0.867	0.051	0.152	0.190	0.002	22.264	24.177	2.971

C.5 Supplementary plot

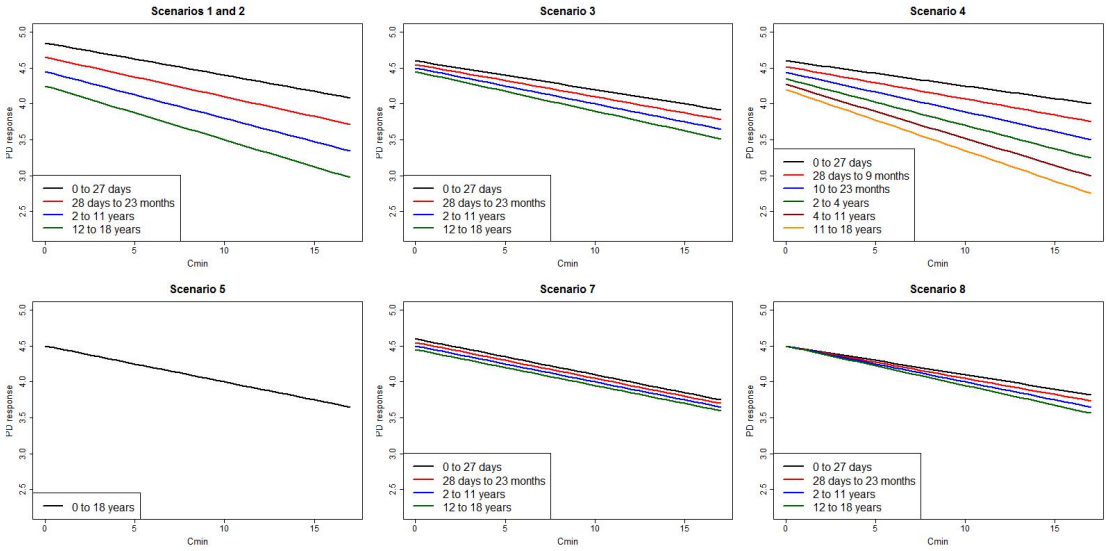


Figure C.9: True underlying E-R relationships for scenarios 1 to 5, 7 and 8, as described in Appendix C.2.