Research Article

Statistics in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Asymmetric inner wedge group sequential tests with applications to verifying whether effective drug concentrations are similar in adults and children

Lisa V. Hampson [a] *, Roland Fisch [b], Linh M. Van [c], Thomas Jaki [a]

Extrapolating from information available on one patient group to support conclusions about another is common in clinical research. For example, the findings of clinical trials, often conducted in highly selective patient cohorts, are routinely extrapolated to wider populations by policy makers. Meanwhile, the results of adult trials may be used to support conclusions about the effects of a medicine in children. For example, if the effective concentration of a drug can be assumed to be similar in adults and children, an appropriate paediatric dosing rule may be found by 'bridging', that is, by matching the adult effective concentration. However, this strategy may result in children receiving an ineffective or hazardous dose if, in fact, effective concentrations differ between adults and children. When there is uncertainty about the equality of effective concentrations, some pharmacokinetic-pharmacodynamic data may be needed in children to verify that differences are small. In this paper we derive optimal group sequential tests that can be used to verify this assumption efficiently. Asymmetric inner wedge tests are constructed which permit early stopping to accept or reject an assumption of similar effective drug concentrations in adults and children. Asymmetry arises because the consequences of under- and over-dosing may differ. We show how confidence intervals can be obtained on termination of these tests and illustrate the small sample operating characteristics of designs using simulation.

Keywords: Bayes decision problem; Error spending tests; Extrapolation; Group sequential tests; Optimal tests; Paediatric clinical trials; Pharmacodynamics; Pharmacokinetics; Verifying assumptions.

[a] *Medical and Pharmaceutical Statistics Research Unit, Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, UK* [b] *Biostatistical Science and Pharmacometrics, Novartis Pharma AG, CH-4002 Basel, Switzerland*

[c] *Biostatistical Science and Pharmacometrics, Novartis Pharmaceutical, Cambridge, 02139, MA*

* *Correspondence to: Lisa Hampson, Medical and Pharmaceutical Statistics Research Unit, Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, UK. E-mail: l.v.hampson@lancaster.ac.uk*

Contract/grant sponsor: LVH is grateful to acknowledge funding from the Medical Research Council (grants MR/J014079/1 and MR/M013510/1).This report is independent research, and Dr. Jaki's contribution arises from his Career Development Fellowship (NIHR-CDF-2010-03-32) supported by the National Institute for Health Research. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

1. Introduction

Tests of equivalence have been widely considered in the context of early phase clinical trials for demonstrating the bioequivalence of a test and reference drug [1]. This paper formulates group sequential equivalence tests to verify similarities between the pharmacokinetic-pharmacodynamic (PK-PD) relationships of a drug in adults and children. The need to verify similarities between patient populations may arise when determining which clinical trials are needed to properly evaluate the risks and benefits of a new medicine for children. Medicine development programmes usually begin with early phase trials in adults before a separate sequence of studies is initiated in children. In the US and Europe, details of paediatric studies must be agreed upon ahead of time with the relevant regulatory agency. Smaller or shorter trials in children may suffice if one can extrapolate from existing relevant data to support claims of efficacy in children. Adult data will be relevant for predicting drug effects in children only if disease progression and the PK-PD relationship are consistent across populations for a PD endpoint predictive of clinical efficacy. If these similarities are plausible, the optimal dose for children can be found by conducting PK studies to match the adult effective concentration [2, 3]. Furthermore, given these data, we may then extrapolate from evidence that a drug is efficacious in adults to support a claim of efficacy in children. Such an approach has been referred to as complete extrapolation of efficacy data [4]. However, a bridging approach to dose-finding will result in children receiving an ineffective or hazardous dose if, in fact, there are differences between PK-PD relationships in adults and children. The size and nature of expected differences should be summarised in an ‘extrapolation concept’ [5, 6]: prior uncertainty about potential differences may then prompt investigators to collect data to verify that differences are small.

We propose group sequential tests (GSTs) of equivalence for PK-PD studies in children which allow early stopping either to conclude that there are important differences between PK-PD relationships in adults and children, or to conclude that relationships are similar enough. Throughout we consider frequentist designs since in a Bayesian setting, strong prior opinion on differences may be impossible to shift with limited data. Jennison & Turnbull [7] derive group sequential tests of equivalence from a sequence of repeated confidence intervals. Liu & Li [8] propose exact error spending tests of equivalence based on bivariate non-central t-statistics. Whitehead [9] considers the double triangular test [10] and a design from the power family of tests due to Pampallona & Tsiatis [11], both of which allow early stopping to reject or accept equivalence. Müller & Schäfer [12] find optimal boundaries and interim analysis timings minimising the average expected sample size of GSTs permitting early stopping only to declare equivalence. In this paper, we formulate asymmetric inner wedge group sequential tests (IW GSTs) of equivalence, so-called because the region of the outcome space on which equivalence is accepted forms an inner wedge inside the continuation region. Asymmetry arises because the consequences of under- and over-dosing may differ.

Rather than verify that PK-PD relationships are identical in adults and children, we propose that studies focus on a particular quantile of this relationship. This is because it is differences between effective concentrations that will determine the accuracy of paediatric dosing recommendations when a complete extrapolation of efficacy data is made from adults to children. This approach is commonly taken in bioequivalence studies where traditional criteria are based on one or two summaries of the concentration-time profile. Such studies conclude bioequivalence if the difference between mean pharmacokinetic parameters on the test and reference products is small. The two one-sided tests procedure [13] declares average bioequivalence at the 5% significance level if the 90% confidence interval for the difference in population mean log PK parameters is contained within the interval $(\log(0.8), \log(1.25))$ [14]. Population bioequivalence [15] and the Kullback-Leibler divergence [16] attempt to capture more completely differences between the distributions of PK parameters. Alternative approaches have also been proposed which measure differences between entire concentration-time profiles [17]. Pei and Hughes [18] test whether the probability that a child experiences a PK parameter less than a quantile estimated from adults exceeds a design value by more than an acceptable margin Δ_U.

In Section 2 we formulate flexible asymmetric IW GSTs for verifying similar effective concentrations in adults and children. We derive optimal versions of our designs in Section 3 before using simulation to estimate their small sample operating characteristics in Section 4.

2. Asymmetric error spending IW GSTs

We define a drug's effective concentration as its EC$_\gamma$, that is, the concentration achieving $\gamma\%$ of a drug's maximum efficacy. Assuming a patient's risk of toxicity increases monotonically with dose, γ will reflect a compromise between seeking a dose that is highly effective yet safe. Typically an estimate of the EC$_\gamma$ can be derived from fitting a nonlinear model relating drug exposure to a short-term PD response predictive of the clinical outcome of interest. Let μ_A and μ_C denote the logarithm of the EC$_\gamma$s in adults and children, respectively, and let $\theta = \mu_A - \mu_C$ measure differences between PK-PD relationships in adults and children. If $\theta > 0$, defining the optimal paediatric dose as one that achieves the adult effective concentration will lead to 'over-dosing', as children are exposed to an excessive risk of toxicity. If $\theta < 0$, this dose-selection strategy will result in children receiving a sub-therapeutic dose.

Suppose a bridging approach to dose-finding in children would be acceptable if there was supportive evidence to verify a claim that $\delta_L < \theta < \delta_U$. We propose generating this evidence by comparing existing adult data with data generated by a new PK-PD study in children. These data are then used to test $H_0 : \theta \leq \delta_L$ or $\theta \geq \delta_U$ against $H_1 : \delta_L < \theta < \delta_U$ with power $1 - \beta$ at $\theta = 0$ and type I error rate α at $\theta = \delta_L$ and $\theta = \delta_U$. Note that only children participate in the new PK-PD study; an estimate of μ_A is derived from existing adult data. Since our aim is to verify an assumption rather than test it definitively, tests of H_0 with higher than conventional significance levels may be acceptable. After all, in verifying an assumption we do not wish to find ourselves conducting the trial we had hoped to avoid by extrapolating. Equivalence limits for bioequivalence trials are traditionally taken as $\pm \log(1.25)$ for analyses of log-transformed PK parameters. In this context, however, δ_L and δ_U are likely to be asymmetric about 0 since the effects of over- and under-dosing children may differ in their severity. Section 4 outlines one approach for determining these limits.

We formulate IW GSTs of H_0 which permit early stopping either to declare similar effective concentrations in adults and children or important differences. At the time the PK-PD study in children is designed, the available adult data are summarised by the maximum likelihood estimate (MLE) of μ_A, denoted $\hat{\mu}_A$, obtained from fitting a robust PK-PD model whose goodness-of-fit to the adult data has been established. In most practical applications $\hat{\mu}_A \,\dot{\sim}\, N(\mu_A, \mathcal{I}_A^{-1})$, where $\dot{\sim}$ means approximately distributed and $\mathcal{I}_A$ is the Fisher information for μ_A [19]. Let $\mathcal{D} = \{d_1, \ldots, d_M\}$ be the set of ordered active doses available for the PK-PD study in children. The number and positioning of these doses must be carefully chosen to ensure that candidate PK-PD models can be fitted on termination of the trial. We propose that the PK-PD study in children proceeds group sequentially, randomising each new group of children between doses $d_1, \ldots, d_M$. We note that the proposed study is not a Phase I trial; the drug will already have been well characterised in adults and some PK data may be available in children, so that a fixed randomisation scheme will be acceptable in terms of ethics and safety. Blood samples are taken from each patient at follow-up times $t_1, \ldots, t_B$. From these we can determine a patient's concentration-time profile which we shall assume is summarised by a single measure of exposure (E) predictive of the PD response (Y). If PK sampling is rich, E can be well estimated using non-compartmental methods, otherwise model-based estimates can be derived. Using these estimates of exposure, at interim analysis k all accummulated data from children are used to fit a nonlinear model relating E to Y to obtain a MLE of μ_C, denoted $\hat{\mu}_{Ck}$. It follows from standard sampling theory that at least asymptotically $\hat{\mu}_{Ck} \sim N(\mu_C, \mathcal{I}_{Ck}^{-1})$, where $\mathcal{I}_{Ck}$ is the Fisher information for μ_C at stage k. Define $S_k = \mathcal{I}_k(\hat{\mu}_A - \hat{\mu}_{Ck})$, where $\mathcal{I}_k = \mathcal{I}_{Ck}\mathcal{I}_A/(\mathcal{I}_{Ck} + \mathcal{I}_A)$ is the information for θ at analysis k. We seek IW GSTs of the form:

$$\begin{array}{ll} \multicolumn{2}{l}{\text{At interim analysis } k = 1, \ldots, K\text{:}} \\ \text{If } S_k \geq u_{2,k} & \text{Stop and accept } H_0 \\ \text{If } S_k \leq l_{2,k} & \text{Stop and accept } H_0 \\ \text{If } l_{1,k} \leq S_k \leq u_{1,k} & \text{Stop and reject } H_0 \\ \text{Otherwise} & \text{Continue to stage } (k+1). \end{array} \tag{1}$$

We set $u_{1,K} = u_{2,K}$ and $l_{1,K} = l_{2,K}$ to ensure the test terminates properly at analysis K. Monitoring data group sequentially means we have the flexibility to respond to emerging evidence on the size of differences between populations. If test (1) rejects H_0 and concludes similar PK-PD relationships in adults and children, recruitment would cease and all available PK data in children would be used to derive estimates of the dose(s) achieving expected exposure e^{μ_A}. If, instead, important differences are detected, an estimate of the effective concentration in children could be derived using the data accumulated so far. If the standard error of this estimate is large, PK-PD data from the current study could be used to optimise the decision on which additional data are needed to improve the precision of the current estimate of e^{μ_C}, and could be incorporated into the analysis of these new data as informative Bayesian prior distributions for PK-PD model parameters.

We seek error spending versions [20] of test (1) which are flexible enough to accommodate unpredictable information sequences. ~~The proposed tests are not derived from a sequence of repeated confidence intervals. Instead~~ Critical values for monitoring score statistics are derived ~~directly~~ by extending the approach of Schuirmann [13] to define an error spending test of H_0 as the union of two one-sided error spending tests: a test of $H_{0U} : \theta \geq \delta_U$, which we label Test U, and a test of $H_{0L} : \theta \leq \delta_L$, labelled Test L. Each test is designed with nominal type I error probability α and target information level $\mathcal{I}_{max} > \mathcal{I}_{fix}$, where $\mathcal{I}_{fix}$ is the information required by the fixed sample test of H_0 which rejects H_0 if $l_{fix} \leq S_{fix} \leq u_{fix}$. Boundaries of Tests L and U are found spending error probabilities as a function of the Fisher information for θ. Let $r_k = \mathcal{I}_k / \mathcal{I}_{max}$ and define f and g as monotonic increasing functions satisfying $f(r) = g(r) = 0$ for $r \leq 0$, and $f(r) = \alpha$ and $g(r) = 1 - \alpha$ for $r \geq 1$. Then the stage k boundaries of Test U are found as solutions to

$$\mathbb{P}\{S_1 \in (u_{1,1}, u_{2,1}), \ldots, S_{k-1} \in (u_{1,k-1}, u_{2,k-1}), S_k \leq u_{1,k}; \theta = \delta_U\} = f(r_k) - f(r_{k-1}) \quad \text{and} \tag{2}$$

$$\mathbb{P}\{S_1 \in (u_{1,1}, u_{2,1}), \ldots, S_{k-1} \in (u_{1,k-1}, u_{2,k-1}), S_k \geq u_{2,k}; \theta = \delta_U\} = g(r_k) - g(r_{k-1}), \tag{3}$$

and the stage k boundaries of Test L are found as solutions to

$$\mathbb{P}\{S_1 \in (l_{1,1}, l_{2,1}), \ldots, S_{k-1} \in (l_{1,k-1}, l_{2,k-1}), S_k \geq l_{1,k}; \theta = \delta_L\} = f(r_k) - f(r_{k-1}) \quad \text{and} \tag{4}$$

$$\mathbb{P}\{S_1 \in (l_{1,1}, l_{2,1}), \ldots, S_{k-1} \in (l_{1,k-1}, l_{2,k-1}), S_k \leq l_{2,k}; \theta = \delta_L\} = g(r_k) - g(r_{k-1}). \tag{5}$$

The IW GST in (1) is formed as the superposition of Test U and Test L, with the boundaries of Test U defining the critical values $\{(u_{1,k},\, u_{2,k})\,;\, k = 1, 2, \ldots\}$ in (1) and the boundaries of Test L fixing critical values $\{(l_{2,k},\, l_{1,k})\,;\, k = 1, 2, \ldots\}$. At early interim analyses when information levels are small, we may find that $l_{1,k} > u_{1,k}$, in which case we set $l_{1,k} = u_{1,k}$ to prevent early stopping to reject H_0. Similarly, we set $u_{2,k} = \max\{u_{2,k},\, l_{1,k}\}$ and $l_{2,k} = \min\{l_{2,k},\, u_{1,k}\}$ to ensure final decisions for H_0 are consistent with the conclusions of both one-sided tests. The design of the IW GST is completed by using a numerical search to yield the target information level $\mathcal{I}_{max}$ for which it achieves power $1 - \beta$ at $\theta = 0$ under a plausible assumption for the information timings of interim analyses.

We calculate stopping probabilities in equations (2)-(5) assuming that conditional on $\mathcal{I}_1, \mathcal{I}_2, \ldots$, the sequence of score statistics follows a multivariate normal distribution with the same mean and covariance structure as a Brownian motion with drift θ. Jennison & Turnbull [21] prove that score statistics derived from MLEs of a parameter of a generalised linear model follow this canonical joint distribution at least asymptotically, although no work exists establishing this result for score statistics based on MLEs of a parameter in a non-linear model. The impact of deviations from distributional assumptions on attained type I error rates will be assessed via simulation in Section 4.4. For now we also proceed assuming that information levels $\mathcal{I}_1, \ldots, \mathcal{I}_k$ are known exactly by interim analysis k, although this will not be the case in general since information may depend on unknown variances and/or PK-PD model parameters. In such cases, error spending IW GSTs can proceed calculating stopping probabilities in equations (2)-(5) conditioning on estimated information levels; the impact of such approximations on operating characteristics will be investigated in Section 4.4.

In all evaluations of methods, we will spend cumulative stopping probabilities according to the simple family of $(1-\alpha)-$ and α-spending functions

$$g(r) = (1-\alpha)\min\{1, r^{\rho_1}\} \quad \text{and} \quad f(r) = \alpha\min\{1, r^{\rho_2}\}, \tag{6}$$

where $\rho_1 > 0$ and $\rho_2 > 0$ are constants which must be pre-specified before the GST begins. We adopt this family of spending functions because it is simple and the efficiency of one-sided tests based on the related ρ-family of error spending functions (as defined in Section 7.2 of [22]) has been well characterised and shown to be close to optimal in a range of testing scenarios [23, 24]. Spending stopping probabilities according to functions g and f in (6) also ensures that Test L and Test U terminate properly at the same information level [25].

As an aside, note that rather than finding stopping boundaries as solutions to equations (2)-(5), an alternative approach would be to extend the methods of Kosorok et al. [26] to find critical values as solutions to

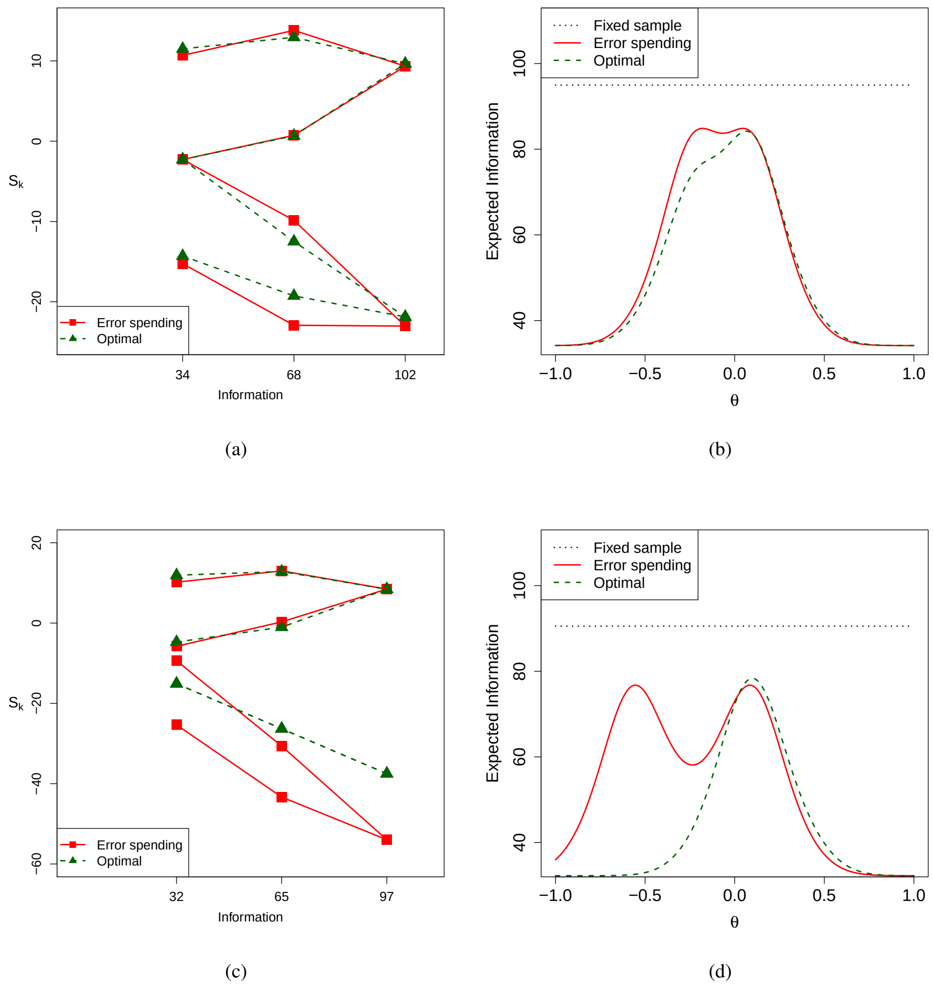
$$\begin{aligned}&\mathbb{P}\{\text{Stop at stage } k \text{ with } S_k \geq u_{2,k};\ \theta = 0\}\\&= \mathbb{P}\{\text{Stop at stage } k \text{ with } S_k \leq l_{2,k};\ \theta = 0\} = (m(r_k) - m(r_{k-1}))/2\end{aligned}$$

and

$$\begin{aligned}&\mathbb{P}\{\text{Stop at Stage } k \text{ with } l_{1,k} \leq S_k \leq u_{1,k};\ \theta = \delta_L\}\\&= \mathbb{P}\{\text{Stop at Stage } k \text{ with } l_{1,k} \leq S_k \leq u_{1,k};\ \theta = \delta_U\} = f(r_k) - f(r_{k-1}),\end{aligned}$$

where m is an error spending function satisfying $m(t) = 0$ for $t \leq 0$, and $m(t) = \beta$ for $t \geq 1$. However, under this approach, it is not guarenteed that sequences of critical values $\{l_{2,1}, l_{2,2}, \ldots\}$ and $\{l_{1,1}, l_{1,2}, \ldots\}$ will cross at the same information level as boundary sequences $\{u_{1,1}, u_{1,2}, \ldots\}$ and $\{u_{2,1}, u_{2,2}, \ldots\}$, in which case it would not be clear how the test should proceed beyond the first observed stage of crossing. Therefore, in what follows, we will restrict attention to error spending IW GSTs found by solving equations (2)-(5).

Appendix A of the Supplementary Materials accompanying this manuscript presents a comprehensive evaluation of the operating characteristics of error spending IW GSTs for various configurations of $(\delta_L,\ \delta_U)$ and the error spending parameters $(\rho_1,\ \rho_2)$. In general, we see that IW GSTs formed as the superposition of Test L and Test U are conservative in the sense that attained type I error rates are less than the nominal level α. Such conservatism arises because at each interim analysis, procedure (1) can continue sampling with values of S_k that would otherwise trigger termination of Test L or Test U. The degree of conservatism decreases as ρ_2 increases so that more type I error is spent at higher information levels. Based on our numerical investigations, we have found setting $\rho_2 = 2$ to be an adequate compromise between achieving a low expected sample size and an attained type I error rate close to α. In all future evaluations of methods, we proceed setting $\rho_1 = 1$ and $\rho_2 = 2$. For example, Figure 1(a) plots the boundaries of a three-stage IW GST designed and conducted



(a)

(b)

(c)

(d)

Figure 1. Stopping rules and expected information on termination of three-stage error spending IW GSTs and optimal versions minimising F when: (a) - (b) $\delta_L = \log(0.7)$ and $\mathcal{I}_{max} = 102.46$; (c) - (d) $\delta_L = \log(0.5)$ and $\mathcal{I}_{max} = 96.802$. All tests are designed and conducted with $\alpha = 0.1$, $\beta = 0.2$, $\delta_U = \log(1.25)$ and for information sequence (7).

with $\rho_1 = 1$, $\rho_2 = 2$, $\alpha = 0.1$, $\beta = 0.2$, $\delta_L = \log(0.7)$ and $\delta_U = \log(1.25)$, scheduling interim analyses at information levels

$$\mathcal{I}_k = k\,\mathcal{I}_{max}/K \quad \text{for } k = 1, \ldots, K. \tag{7}$$

This test attains type I error rate 0.096 under $\theta = \delta_L$ and $\theta = \delta_U$ rather than the nominal level 0.1. The impact of this conservatism on efficiency is evaluated in Section 3.2.

Frequentist group sequential stopping rules can be interpreted using Bayesian predictive probabilities to measure the degree of (in-)consistency between the observed data and an extrapolation assumption needed to trigger termination. Table 1 lists predictive probabilities

$$\mathbb{P}\{S_k \geq s^\star\} = \int_\Theta h(\theta;\xi,\omega^2,\eta)\Phi\left(\frac{\mathcal{I}_k\theta - s^\star}{\sqrt{\mathcal{I}_k}}\right)\mathrm{d}\theta \tag{8}$$

and

$$\mathbb{P}\{s_1^\star \leq S_k \leq s_2^\star\} = \int_\Theta h(\theta;\xi,\omega^2,\eta)\left\{\Phi\left(\frac{u_k - \theta\mathcal{I}_k}{\sqrt{\mathcal{I}_k}}\right) - \Phi\left(\frac{l_k - \theta\mathcal{I}_k}{\sqrt{\mathcal{I}_k}}\right)\right\}\mathrm{d}\theta \tag{9}$$

Table 1. Bayesian predictive probabilities assuming $\theta \sim SN(0.134, 0.063, -1.913)$. Probabilities are evaluated at the boundaries of an error spending IW GST. Figures in parentheses are corresponding probabilities for an optimal IW GST minimising F. Both tests are designed and conducted with $K = 3$, $\alpha = 0.1$, $\beta = 0.2$, $\delta_L = \log(0.7)$, $\delta_U = \log(1.25)$, $\mathcal{I}_{max} = 102.46$ and information sequence (7)

k	$\mathbb{P}\{S_k \leq l_{2,k}\}$	$\mathbb{P}\{S_k \geq u_{2,k}\}$	$\mathbb{P}\{l_{1,k} \leq S_k \leq u_{2,k}\}$
1	0.055 (0.068)	0.069 (0.056)	0 (0)
2	0.091 (0.133)	0.121 (0.135)	0.279 (0.335)
3	0.181 (0.195)	0.260 (0.255)	0.559 (0.550)

for the error spending IW GST in Figure 1(a), with h the prior probability density function (pdf) for θ consistent with an assumption of similar PK-PD relationships in adults and children. Such a prior distribution would be asymmetric, having zero mode and placing small probability mass on the events $\{\theta \geq \delta_U\}$ and $\{\theta \leq \delta_L\}$. We choose to evaluate predictive probabilities setting h as the pdf of a skew normal random variable, written $\theta \sim SN(\xi, \omega^2, \eta)$ [27], although in principle other skew distributions would be acceptable. We choose the parameters of this distribution such that θ has a prior mode of zero and such that there is a prior probability of $\alpha/2$ that $\theta \geq \delta_U$ and probability $\alpha/2$ that $\theta \leq \delta_L$. For this choice of prior

$$h(\theta; \xi, \omega^2, \eta) = \frac{2}{\omega}\phi\left(\frac{\theta - \xi}{\omega}\right)\Phi\left(\frac{\eta(\theta - \xi)}{\omega}\right),$$

where ϕ and Φ are the pdf and cumulative distribution function of a standard normal variate. If predictive probability (8) is close to 0, this implies that observing $S_k = s^\star$ is inconsistent with the extrapolation assumption since we would be unlikely to observe such an event if θ was a sample from a $SN(\xi, \omega^2, \eta)$ distribution. Probabilities (9) close to 1 imply that S_k would lie in the interval $(s_1^\star, s_2^\star)$ with high probability if θ was consistent with the extrapolation assumption. Returning to the error spending IW GST illustrated in Figure 1(a), the skew normal distribution for θ consistent with the extrapolation assumption is defined by $\xi = 0.13$, $\omega^2 = 0.06$ and $\eta = -1.91$. From Table 1, we see that in the early stages of this trial, early stopping to accept H_0 is permitted only when there is strong evidence of discrepancies between the data and the extrapolation assumption. For example, the test stops at the first interim analysis to declare $\theta \leq \delta_L$ ($\theta \geq \delta_U$) if and only if the predictive probability of observing a more extreme test statistic under the extrapolation assumption is less than 0.055 (0.068). It is not possible to reject H_0 at the first interim analysis and so predictive probability (9) is 0.

3. Optimal asymmetric IW GSTs

3.1. Optimising the boundaries of an asymmetric IW GST

To evaluate the efficiency of the proposed error spending IW designs, we seek optimal versions of tests which come to a rapid conclusion when effective concentrations in adults and children are similar. Let T represent the stage at which an IW GST terminates and let $\mathcal{I}_T$ and S_T represent the Fisher information for θ and score statistic on termination. Symmetric IW GSTs minimising the expected sample size on termination have been found by Eales [28] and Chang [29]. We extend this work to find optimal K-stage asymmetric IW GSTs minimising

$$F = \int \mathbb{E}_\theta(\mathcal{I}_T) h(\theta; \xi, \omega^2, \eta) \mathrm{d}\theta, \tag{10}$$

subject to having type I error rate α at $\theta = \delta_L$ and $\theta = \delta_U$, and power $1 - \beta$ at $\theta = 0$. Tests minimising F minimise the average expected information on termination, taking averages across a skew normal distribution for θ chosen to have mode zero and to place probability $\alpha/2$ on events $\{\theta \geq \delta_U\}$ and $\{\theta \leq \delta_L\}$. If $\delta_L = -\delta_U$, the distribution thus defined is equivalent to a $N(0, \delta_U^2[\Phi^{-1}(1-\alpha/2)]^{-2})$ distribution.

Aside from the stated type I and type II error rate constraints, no further restrictions are placed on the form of the boundaries of the optimal test although we will check that they follow the general pattern shown in (1). The optimal test could be found using Lagrangian multipliers [30] or by optimising directly over all possible configurations of the $[4(K-1)+2]$ boundaries. However, as K increases, this latter approach quickly becomes computationally burdonsome. A more feasible approach is to find the optimal frequentist GST as the solution to an unconstrained Bayes problem, a strategy which has been adopted by several authors to find optimal frequentist designs [31, 32, 24]. The novelty of our problem is that optimal tests must control the type I error rate at two values of θ unequally spaced about 0 and minimise the expected information averaging over a skew normal distribution for θ. We define a Bayes decision problem with prior, decision loss and sampling cost functions carefully chosen to ensure the problem has a solution of the form we seek. Specifically, let $L(\mathcal{A}, \theta)$ measure the loss incurred by taking action $\mathcal{A}$ when the true difference between adult and paediatric log effective concentrations is θ. Letting $\mathcal{A}_1$ represent the decision to reject H_0, and $\mathcal{A}_2$ represent the decision to accept H_0, we define $L(\mathcal{A}_1,\, \theta = \delta_L) = d_1$, $L(\mathcal{A}_1,\, \theta = \delta_U) = d_2$, $L(\mathcal{A}_2,\, \theta = 0) = d_3$ and $L(\mathcal{A}, \theta) = 0$ otherwise. We incur a cost of $c(\theta)$ per unit of information sampled, where $c(\delta_L) = c(0) = c(\delta_U) = 0$ and $c(\theta) = 1$ otherwise. Finally, we define a prior distribution for θ which places probability $1/4$ on the cases that $\theta = \delta_L$, $\theta = 0$, $\theta = \delta_U$ and $\theta \sim SN(\xi, \omega^2, \alpha)$. Under this specification, the total expected cost of the IW GST is given by

$$r(\boldsymbol{d}) = 1/4\{d_1\mathbb{P}\{\text{Reject } H_0 \mid \theta = \delta_L\} + d_2\mathbb{P}\{\text{Reject } H_0 \mid \theta = \delta_U\} + d_3\mathbb{P}\{\text{Accept } H_0 \mid \theta = 0\} + F\}, \tag{11}$$

where $\boldsymbol{d} = (d_1, d_2, d_3)$. The Bayes test minimising $r(\boldsymbol{d})$ can be found using backwards induction and the dynamic programming algorithm is listed in Supplementary Appendix B. Let $\alpha_1^\star(\boldsymbol{d})$ and $\alpha_2^\star(\boldsymbol{d})$ denote the attained type I error rates of the Bayes test at $\theta = \delta_L$ and $\theta = \delta_U$, and let $\beta^\star(\boldsymbol{d})$ denote this test's attained type II error rate at $\theta = 0$. Standard arguments imply that the Bayes test minimising $r(\boldsymbol{d})$ also minimises F in the class of tests with frequentist error rates $\alpha_1^\star(\boldsymbol{d})$, $\alpha_2^\star(\boldsymbol{d})$ and $\beta^\star(\boldsymbol{d})$. However, for any particular choice of decision costs, these attained error rates may deviate from their nominal values. We therefore perform an unconstrained three-dimensional search over the vector of log decision costs to find the configuration defining the Bayes problem whose solution minimises

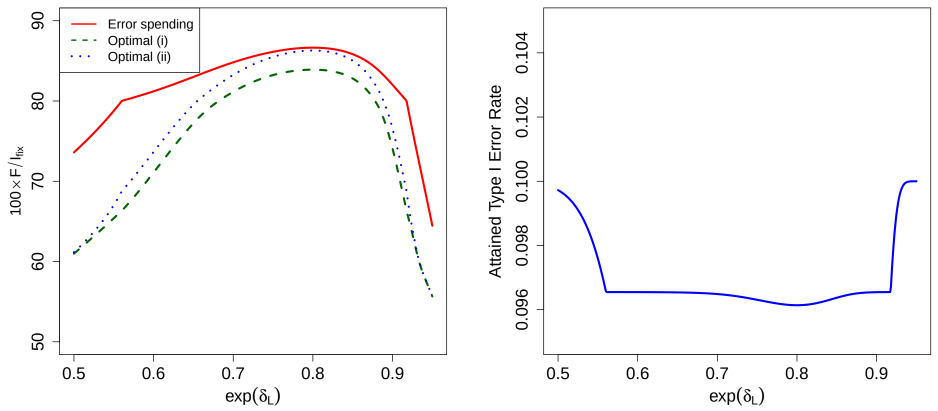
$$\eta(\log(d_1), \log(d_2), \log(d_3)) = \surd\left\{\left[(\alpha_1^\star(\boldsymbol{d}) - \alpha)^2 + (\alpha_2^\star(\boldsymbol{d}) - \alpha)^2 + (\beta^\star(\boldsymbol{d}) - \beta)^2\right]/3\right\}. \tag{12}$$

It follows that the Bayes test minimising (12) with frequentist error rates satisfying $\alpha_1^\star = \alpha_2^\star = \alpha$ and $\beta^\star = \beta$ is the optimal frequentist IW GST we seek.

Figure 1(a)-1(b) plots the boundaries and operating characteristics of the optimal three-stage IW GST minimising F with type I error rate $\alpha = 0.1$ at $\theta = \log(0.7)$ and $\theta = \log(1.25)$, and type II error rate $\beta = 0.2$ at $\theta = 0$, when interim analyses are equally spaced in information up to $\mathcal{I}_{max} = 102.46$. This optimal frequentist test was found as the solution to a Bayes problem of the form described above defined by decision costs $d_1 = 80.1$, $d_2 = 433.2$ and $d_3 = 290.2$.

3.2. *Efficiency of IW error spending tests*

Figure 2(a) plots values of F attained by optimal and error spending IW GSTs as a percentage of $\mathcal{I}_{fix}$. All tests are designed and conducted for information sequence (7) setting $\delta_U = \log(1.25)$. For each pair of equivalence limits, $\mathcal{I}_{max}$ is chosen as the information target needed for the error spending test to achieve power $1-\beta$ at $\theta = 0$. Error spending tests are designed with nominal type I error rate $\alpha = 0.1$ although Figure 2(b) illustrates how attained error rates deviate from this level. Only attained type I error rates at $\theta = \delta_L$ are plotted; lines for type I error rates at $\theta = \delta_U$ and $\theta = \delta_L$ are almost indistinguishable. To assess the impact on efficiency of the conservatism of the error spending test, we consider two versions of the optimal test. Version (i) is derived to attain type I error rate $\alpha = 0.1$. Version (ii) has type I error rate



(a)

(b)

Figure 2. a): Values of F achieved by three-stage error spending and optimal IW GSTs of $H_0 : \theta \leq \delta_L$ or $\theta \geq \delta_U$ expressed as a percentage of $\mathcal{I}_{fix}$; b) Type I error rate at $\theta = \delta_L$ attained by error spending IW GSTs. For each pair (δ_L, δ_U), version (i) of the optimal test has type I error rate α; version (ii) controls the type I error rate at the level attained by the error spending test. All tests are designed and conducted for information sequence (7) setting $\delta_U = \log(1.25)$, $\alpha = 0.1$ and $\beta = 0.2$.

equal to that attained by the error spending test but minimises the average expected sample size across a skew normal distribution for θ placing probability 0.05 on events $\{\theta \geq \delta_L\}$ and $\{\theta \leq \delta_U\}$.

From Figure 2(a) we see that, in general, error spending tests are efficient, attaining values of F within 10% of $\mathcal{I}_{fix}$ of the optimal test in many cases, and offering savings for expected information of more than 13% on the fixed sample test in all cases. When δ_L is close to $\log(0.8)$ and we approach the case of a symmetric hypothesis test, values of F attained by error spending tests tend towards the values attained by optimal versions of IW GSTs with the same attained type I error rates. This implies that in these settings, differences in efficiency between the error spending IW GSTs and version (i) of the optimal tests with type I error rate α are largely due to the conservatism of the former rather than differences between the shapes of the test boundaries. To illustrate this, Figures 1(a)-1(b) compare properties of three-stage error spending and optimal IW GSTs designed with $\delta_L = \log(0.7)$ and $\delta_U = \log(1.25)$. The boundaries of both procedures and the expected information curves are similar, while the power curves are almost indistinguishable. Values of F attained by the optimal and error spending tests are 81.1% and 84.8% of $\mathcal{I}_{fix}$, respectively. However, for more extreme values of δ_L, the error spending test is less efficient. Comparing the performance of versions (i) and (ii) of the optimal test, we deduce that these losses are not due to the conservatism of the error spending test; rather it is the shape of the error spending boundaries that is suboptimal. This is reflected in Figures 1(c)-1(d) comparing properties of tests designed with $\delta_L = \log(0.5)$.

We have investigated the efficiency of the error spending IW GSTs in the somewhat idealised scenario that the target maximum information level is attained exactly and score statistics follow the asymptotic canonical joint distribution defined in [21]. In the absence of similar finite sample distributional results for score statistics based on MLEs of a parameter in a non-linear model, we cannot comment on whether this efficiency is maintained in small samples. However, the flexibility of the error spending designs will likely compensate for small losses in efficiency incurred due to deviations from the canonical joint distribution.

3.3. Optimising interim analysis timings

When differences between adults and children exist, we will measure the efficiency of an IW GST in terms of the accuracy of the final estimate of θ available on termination of the test, where narrow confidence intervals for θ are preferred to better inform subsequent dose-finding trials in children. The analysis schedule of a GST will determine the information available for θ and thus the width of the confidence interval (CI) for θ calculated on termination of the trial. Thus, one may wish to optimise the test analysis schedule to minimise the average expected width of the CI for θ on termination, taking averages across null values of θ. In what follows, let $W(k, s^\star)$ denote the width of the $(1-\alpha)$ CI for θ calculated upon terminating the GST at stage $T = k$ with $S_T = s^\star$. Suppose the first interim analysis is scheduled at information level $\mathcal{I}_1$ and subsequent analyses are timed at

$$\mathcal{I}_k = \mathcal{I}_1 + (k-1)(\mathcal{I}_{max} - \mathcal{I}_1)/(K-1) \quad \text{for } k = 2, \dots, K. \tag{13}$$

Choosing the boundaries of test (1) to minimise F, we seek $\mathcal{I}_1$ minimising

$$G = \frac{\omega_a F}{\mathcal{I}_{fix}} + \frac{\omega_b}{W_{fix}} \left[\frac{\mathbb{E}(W(T, S_T); \theta = \delta_L) + \mathbb{E}(W(T, S_T); \theta = \delta_U)}{2} \right],$$

where ω_a and ω_b are non-negative weights satisfying $\omega_a + \omega_b = 1$, and W_{fix} is the average expected width of the $(1-\alpha)$ CI for θ on termination of the fixed sample test under $\theta = \delta_L$ and $\theta = \delta_U$. The definition of G reflects our wish to come to a rapid conclusion when effective concentrations in adults and children are similar, and to make precise statements about differences when these exist.

We have not yet said how we will calculate the $(1-\alpha)$ CI for θ on termination of test (1). Clearly it is desirable that the CI be consistent with the final decision of the IW GST, being entirely contained within the equivalence interval if and only if the test stops to reject H_0. It would be particularly difficult to claim similar effective concentrations if the CI for θ on termination contained δ_L or δ_U. We therefore seek sharp CIs which are consistent with the final decision of the IW GST with high probability. Hsu et al. [33] consider $(1-\alpha)$ CIs associated with tests of interval null hypotheses. Generalising Theorem 2.5 of that paper to the case of IW GSTs, if $\theta_U(T, S_T)$ and $\theta_L(T, S_T)$ are $(1-\alpha)$ upper and lower confidence limits satisfying

$$\mathbb{P}\{\theta_U(T, S_T) \geq \theta\} = \mathbb{P}\{\theta_L(T, S_T) \leq \theta\} = 1 - \alpha \qquad \text{for all } \theta, \tag{14}$$

the interval

$$\left[\min\left\{ \frac{\delta_L + \delta_U}{2}, \theta_L(T, S_T) \right\}, \max\left\{ \frac{\delta_L + \delta_U}{2}, \theta_U(T, S_T) \right\} \right] \tag{15}$$

has coverage probability 1 if $\theta = (\delta_L + \delta_U)/2$, and $(1-\alpha)$ otherwise. In the case of fixed sample tests, Hsu et al. [33] note that CIs of this form have typically smaller effective length (defined as the supremum of distances between points in the CI and zero) than conventional $(1-\alpha)$ CIs and so will be properly contained in the equivalence interval with higher probability when equivalence holds. We thus prefer to consider CIs of the form (15), despite their drawback that confidence interval limits are allowed to depend on δ_L and δ_U.

We now turn our attention to finding upper and lower confidence limits satisfying equation (14) on termination of the IW GST. For $k = 1, \dots, K$, let $\mathcal{C}_k = (l_{2,k},\, l_{1,k}) \cup (u_{1,k},\, u_{2,k})$ and define $\Omega = \{(k, s_k); s_k \notin \mathcal{C}_k,\, k = 1, \dots, K\}$ as the set of outcomes with which the IW GST can terminate. To find θ_U and θ_L we first note that since the continuation regions are not intervals, the stage-wise ordering [34] cannot be applied to Ω [22]. Instead, we order Ω with respect to the MLE of θ [35]. If the IW GST terminates with $(T, S_T) = (k^\star,\, s^\star)$, this ordering stipulates that an outcome (k', s') lies equal to or above $(k^\star,\, s^\star)$ in the ordering of Ω, that is, $(k', s') \succeq (k^\star,\, s^\star)$, if and only if $s'/\mathcal{I}_{k'} \geq s^\star/\mathcal{I}_{k^\star}$. Clearly the ordering of Ω depends upon information levels $\mathcal{I}_1, \dots, \mathcal{I}_K$, which may be unknown if the test stops early with $\mathcal{I}_T < \mathcal{I}_{max}$ and group sizes are unpredictable. In such cases, we propose calculating the CI assuming that in the absence of early stopping, the test would have evolved with $\mathcal{I}_{T+k} = \mathcal{I}_T + k\,(\mathcal{I}_{max} - \mathcal{I}_T)/(K-T)$, for $k = 1, \dots, K-T$.

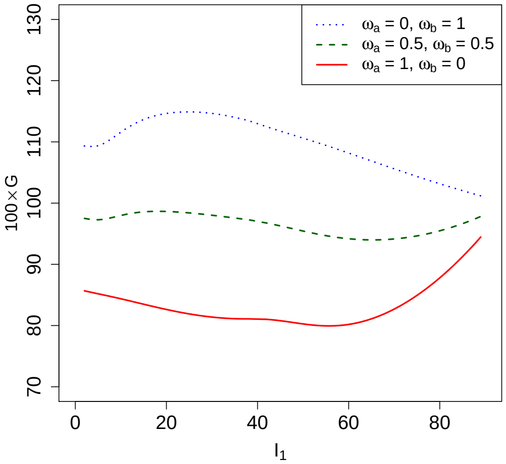


Figure 3. Values of $100 \times G$ attained by optimal three-stage IW GSTs minimising F with $\alpha = 0.1$, $\beta = 0.2$, $\delta_L = \log(0.7)$, $\delta_U = \log(1.25)$, $\mathcal{I}_{max} = 102.46$ and where interim analyses are scheduled according to information sequence (13).

If the distribution of (T, S_T) is stochastically ordered on Ω with respect to θ, the $(1-\alpha)$ confidence limits satisfy

$$\mathbb{P}\{(T, S_T) \succeq (k^\star, s^\star);\ \theta_U(k^\star, s^\star)\} = \mathbb{P}\{(T, S_T) \preceq (k^\star, s^\star);\ \theta_L(k^\star, s^\star)\} = 1 - \alpha.$$

Emerson (Chapter 4 of [36]) has proved that the MLE ordering has this monotonicity property for GSTs with interval continuation regions. It follows from Emerson's proof that this property must also hold for IW GSTs. Given the monotonicity condition, CI (15) will be consistent with the decision of the IW GST if: (a) for $k = 1, \ldots, K-1$, $u_{2,k}/\mathcal{I}_k \geq u_{2,K}/\mathcal{I}_K \geq u_{1,k}/\mathcal{I}_k \geq l_{1,k}/\mathcal{I}_k \geq l_{2,K}/\mathcal{I}_K \geq l_{2,k}/\mathcal{I}_k$; and (b) $\mathbb{P}\{$Accept H_0 and conclude $\theta \geq \delta_U;\ \theta = \delta_U\} = \mathbb{P}\{$Accept H_0 and conclude $\theta \leq \delta_L;\ \theta = \delta_L\} = 1 - \alpha$. Condition (a) stipulates that Ω can be partitioned into three sets, all outcomes for which the GST stops to accept H_0 concluding $\theta \leq \delta_L$ being lower in the ordering than outcomes for which the test stops to reject H_0, which in turn are lower in the ordering than outcomes for which the test stops to accept H_0 concluding $\theta \geq \delta_U$. While this condition is not guarenteed to hold for optimal or error spending IW GSTs, we have found that it does for almost all of the designs we have considered. Similarly, condition (b) is unlikely to hold exactly. Despite this, we have found that the probability of a conflict between the IW GST and the $(1-\alpha)$ CI remains small. For example, boundary condition (a) is satisfied by the three-stage error spending IW GST illustrated in Figure 1(a). For this test, $\mathbb{P}\{$Accept H_0 and conclude $\theta \geq \delta_U;\ \theta = \delta_U\} = 0.903$ and $\mathbb{P}\{$Accept H_0 and conclude $\theta \leq \delta_L;\ \theta = \delta_L\} = 0.903$, thus violating condition (b), although the probability of a conflict between the test and the $(1-\alpha)$ CI is less than 0.008.

Figure 3 compares the efficiencies of three-stage optimal IW GSTs under analysis schedule (13) for a range of values of $\mathcal{I}_1$ and weights (ω_a, ω_b). Setting $\omega_b = 1$ and choosing $\mathcal{I}_1$ to minimise the expected CI width, it is most efficient to time the first interim analysis as late as possible. Setting $\omega_a = 1$, our focus is on minimising F and the optimum analysis schedule sets $\mathcal{I}_1 = 55.7$. If $\omega_a = \omega_b = 0.5$, the trade-off between competing objectives means G is minimised by scheduling $\mathcal{I}_1 = 65.2$.

4. Simulation study

Valsartan is an antihypertensive drug indicated for patients aged 6 years and above. Its antihypertensive effect in adults has been established through seven placebo-controlled trials randomising over 2 000 adults to receive doses between 10 to 320 mg/day, and 800 adults to receive placebo (Section 14.1 of [37]). The clinical benefits of valsartan in hypertensive children have been tested in a dose-ranging trial which randomised 261 children aged 6 to 16 years to a low, medium or high (weight-based) dose between 10 to 160 mg/day. A similar trial was conducted in 90 infants aged 1 to 5 years. For antihypertensives it is plausible, but not obvious, that PK-PD relationships should be similar in adults and children. If relationships were similar for valsartan, we speculate that it may have been sufficient to extrapolate from the substantial adult database to support claims of efficacy in children. In the following, we consider how error spending IW GSTs could have been used to verify whether the EC$_{90}$ of valsartan is similar for adults and children aged 6 to 18 years. For present purposes, we measure valsartan exposure by AUC$(0, \infty)$ and pharmacodynamics by change from baseline in systolic blood pressure at 4 weeks.

4.1. Valsartan population PK and PK-PD models

The population pharmacokinetics of valsartan when administered as a suspension to hypertensive children follow a two compartment model with zero-order absorption, linear elimination and lag time. There is an effect of fat free mass (FFM) on the following parameters: apparent clearance $C\ell$, apparent central volume of distribution V_c, apparent inter-compartmental distributional clearance $C\ell_d$, and apparent peripheral volume of distribution V_p; and an additional effect of age on $C\ell$ [38]. We simulate adult pharmacokinetics using a variation on the paediatric PK model, scaling typical values of population parameters by 1.7 to account for differences in bioavailability as adults receive valsartan as a tablet rather than an oral suspension [39]. Furthermore, we also omit an effect of age on $C\ell$ for adults. Since valsartan does not accummulate markedly in blood plasma with repeated dosing (Section 12.3 of [37]), simulated concentrations depend only on the time since the preceeding dose. Labelling children and adults as populations C and A, let $E_{Ti} = \text{AUC}_{Ti}(0, \infty)$ denote the exposure of patient i in population $T \in \{C, A\}$ who is aged A_{Ti} years and has fat free mass FM$_{Ti}$. We simulate patient-specific apparent clearances as

$$\begin{aligned} C\ell_{Ci} &= \{4.32(\text{FM}_{Ci}/44)^{0.75} + 0.06(A_{Ti} - 7.8)\}e^{b_{Ci}} \\ C\ell_{Ai} &= 1.7\,\{4.32(\text{FM}_{Ai}/44)^{0.75} + 0.06(18 - 7.8)\}e^{b_{Ai}} \end{aligned}$$

where $b_{A1}, b_{A2}, \ldots$ and $b_{C1}, b_{C2}, \ldots$ are independent random effects and each $b_{Ti} \sim N(0, 0.107)$. For a patient receiving dose d, $E_{Ti} = d/C\ell_{Ti}$.

Demographics of hypertensive patients are simulated from log-normal distributions specified based on baseline data from multicentre trials of valsartan [40, 41]. Fat free mass is a function of weight (kg) and BMI (kg/m^2), and the logarithm of these variables for adults are generated as independent realisations of $N(4.52,\, 0.05)$ and $N(3.47,\, 0.03)$ random variables, respectively. Furthermore, adults are female with probability 0.47. The logarithms of the weight and BMI of a hypertensive child are generated as independent realisations of $N(4.07,\, 0.24)$ and $N(3.24,\, 0.12)$ random variables, respectively; children are female with probability 0.39. We simulate 11 PK samples per adult taken at 0.5, 1, 1.5, 2, 2.5, 3, 4, 6, 7, 12, 24 hours post dose (with a sampling window of ± 0.05 hours). Nine PK samples are simulated per child taken at 0.5, 1, 2, 3, 4, 6, 8, 12, 24 hours post dose (± 0.1 hours).

Dose-response curves for many compounds can be adequately approximated by a hyperbolic (i.e., three-parameter) E$_{\text{max}}$ model [42]. Since valsartan exposure is proportional to dose, we simulate the PD response of patient i in population

T as

$$Y_{Ti} = y_{0T} - \frac{0.9\, R_T E_{Ti}}{0.1 e^{\mu_T} + 0.9\, E_{Ti}} + \epsilon_{Ti}, \tag{16}$$

where $\epsilon_{Ti} \sim N(0, \sigma_T^2)$ are independent random errors. In all evaluations we set $y_{0A} = y_{0C} = -5$, $R_A = R_C = 10$, $\sigma_A = 4$, $\mu_A = \log(28)$ and $\mu_C = \mu_A - \theta$. Simulated adult and paediatric data are analysed using a two-stage approach, first approximating a patient's exposure by a non-compartmental estimate of the AUC over their period of observation, and then using these estimates to fit model (16) to obtain $\hat{\mu}_A$ or $\hat{\mu}_C$.

We verify whether PK-PD relationships in adults and children are similar by testing $H_0 : \theta \leq \theta_L$ or $\theta \geq \delta_U$. In practice, the choice of equivalence limits would be informed by clinical considerations. For example, suppose it would be acceptable to dose children targeting the adult EC$_{90}$ if the expected effect of this exposure in children was between 80% and 93% of the maximum expected reduction in systolic blood pressure that can be achieved. These opinions and model (16) would imply that $\delta_L = \log(28/63)$ and $\delta_U = \log(28/19)$. Setting the equivalence limits such that $|\delta_L| < |\delta_U|$ implies the consequences of over-dosing are more severe than those of under-dosing, and vice-versa.

4.2. Simulating a PK-PD study in children to verify similar effective concentrations

We simulated PK-PD studies in children proceeding according to an error spending IW GST with up to $K = 3$ analyses. Here we describe the process used to simulate each study. Locally optimal designs minimising the variance of a quantile of an E$_{\text{max}}$ model are supported by 3 design points [43]. However, in practice, parameters of the exposure-response curve are unlikely to be known exactly at the design stage so that it is good practice to use 4 to 7 active doses in dose-finding studies [44]. We simulate studies randomising equal numbers of children between placebo, 10, 20, 150 and 160 mg/day, where these are the four active doses from the set $\{10, 20, \ldots, 160\}$ maximising the Fisher expected information for μ_C when PK-PD relationships are identical in adults and children. IW GSTs are conducted setting $\alpha = 0.1$, $\rho_1 = 1$ and $\rho_2 = 2$. We choose $\mathcal{I}_{max}$ to ensure the GST has power $1 - \beta$ to reject H_0 when $\theta = 0$ when interim analyses follow pattern (7).

We imagine that before the PK-PD study in children begins, data are available on 2 000 adults of whom equal numbers have received placebo, 20, 30, 40 or 320 mg/day of valsartan. These active doses are the four doses from the set $\{20, 30, \ldots, 320\}$mg maximising the Fisher information for μ_A. Simulated adult data are used to fit model (16) to obtain $\hat{\mu}_A$. MLEs of y_{0A}, R_A and μ_A are found fitting the hyperbolic E_{max} model in R [45] using `nls`; we set $\hat{\sigma}_C$ as the unbiased least squares estimate. We approximate $\mathcal{I}_A$ by an estimate $\hat{\mathcal{I}}_A$ derived substituting estimates for unknown model parameters. From this we can deduce a target information level for μ_C, denoted $\hat{\mathcal{I}}^C_{max}$, and thus a sample size requirement for the paediatric trial. Let $n_{max,0}$ denote our initial estimate of the total number of children needed to generate information $\hat{\mathcal{I}}^C_{max}$ assuming PK-PD parameters in children are equal to the adult estimates. To ensure that target information level $\hat{\mathcal{I}}^C_{max}$ will be met in the absence of early stopping, we adopt an information monitoring approach [46, 47], at each interim analysis using the most recent estimate of σ_C to refine the initial sample size calculation. We denote the stage k estimate by $n_{max,k}$. Further details on simulated trials can be found in Supplementary Appendix C.

Once the PK-PD study in children is underway, interim analysis k, for $k \leq K - 1$, is conducted once data are available on n_k patients. All available paediatric data are then used to fit model (16) to deduce $\hat{\mu}_{Ck}$ and $\hat{\mathcal{I}}_{Ck}$. We do not permit early stopping if MLEs of model parameters cannot be found due to non-convergence of the optimisation routine. Instead we set $n_{max,k} = n_{max,k-1}$ and continue to stage $(k + 1)$ recruiting $(n_{max,k} - n_k)/(K - k)$ additional children. We also prohibit early stopping if MLEs of model parameters can be found but $\hat{\mathcal{I}}_{Ck}\hat{\mathcal{I}}_A/(\hat{\mathcal{I}}_{Ck} + \hat{\mathcal{I}}_A) \leq \hat{\mathcal{I}}_{k-1}$. Otherwise early stopping at stage k is permitted and hypothesis decisions are based on $S_k = (\hat{\mu}_A - \hat{\mu}_C)\hat{\mathcal{I}}_k$. Since this statistic does not follow a known distribution, we approximate and compare S_k with error spending boundaries $(l_{2,k},\, l_{1,k})$ and $(u_{1,k},\, u_{2,k})$ found for monitoring a sequence of score statistics assuming $\mathcal{I}_j = \hat{\mathcal{I}}_j$, for $j = 1, \ldots, k$. If the optimisation routine fitting the E_{max}

model converges and sampling continues to stage $(k+1)$, the current estimate of σ_C^2 is used to refine the target sample size. Decreases in sample size targets are permitted as well as increases. We specify a minimum group size of one child per dose and a maximum group size with the adult sample size per dose.

In the absence of early stopping, if MLEs of the PK-PD model parameters cannot be found at analysis K, we fit a linear model to the paediatric data

$$Y_{Ci} = \gamma_0 + \gamma_1 \log(E_{Ci}+1) + \epsilon_{Ci} \quad \epsilon_{Ci} \sim N(0, \sigma_C^2) \tag{17}$$

and obtain $\hat{\mu}_{CK} = \log(\exp\{(\hat{X} - \hat{\gamma}_0)/\hat{\gamma}_1\} - 1)$, where $\hat{X} = 0.1\,\bar{X}_0 + 0.9\,\bar{X}_4$, and $\bar{X}_0$ and $\bar{X}_4$ are the sample mean PD responses on placebo and the highest valsartan dose. Otherwise, if the E_{max} model can be fitted but $\hat{\mathcal{I}}_{CK}\hat{\mathcal{I}}_A/(\hat{\mathcal{I}}_{CK} + \hat{\mathcal{I}}_A) \leq \hat{\mathcal{I}}_{K-1}$, we recalculate the Fisher information for μ_C using stage $(K-1)$ estimates of model parameters. If the information sequence is still decreasing, we fit model (17). The stage K error spending boundaries are then found to spend the remaining type I error probability.

4.3. *An example simulated trial*

To illustrate how a simulated trial might proceed, suppose that we wish to use existing adult data and data generated from a PK-PD study in children to test H_0 setting $K = 3$, $\alpha = 0.1$, $\beta = 0.3$, $\delta_L = \log(28/108)$ and $\delta_U = \log(28/13.26)$. From the existing adult data, suppose we obtain estimates $\hat{y}_{0A} = -4.98$, $\hat{R}_A = 10.64$, $\hat{\mu}_A = 3.43$, $\hat{\sigma}_A = 4.02$ and thus $\hat{\mathcal{I}}_A = 127.53$. The paediatric trial is planned assuming interim analyses will be equally spaced in information, in which case a three-stage error spending IW GST must be designed to accrue Fisher information $\mathcal{I}_{max} = 6.70$ for θ in the absence of early stopping (corresponding to an information target of 7.07 for μ_C). Assuming PK-PD parameters for children are equal to the adult estimates, we estimate that up to 24 children will be needed per dose level, and the trial begins randomising 8 children to each active dose and placebo.

At the first interim analysis, model (16) is fitted to the first group of responses to obtain $\hat{\mu}_{C,1} = 6.63$ ($\mathcal{I}_{C,1} = 0.34$), $\hat{\sigma}_{C,1} = 3.90$, $\hat{\mathcal{I}}_1 = 0.34$ and $S_1 = -1.08$. Given $\hat{\mathcal{I}}_1 = 0.34$, boundaries of Test L and Test U spending error probabilities $f(r_1) = 2.54 \times 10^{-4}$ and $g(r_1) = 0.05$ are $(-1.44,\ 1.57)$ (Test L) and $(-1.77,\ 1.24)$ (Test U). These are combined so that the stage 1 boundaries for testing H_0 are $l_{2,1} = -1.77$, $l_{1,1} = -0.10$, $u_{1,1} = -0.10$ and $u_{2,1} = 1.56$. Since $l_{2,1} \leq S_1 \leq l_{1,1}$ the trial continues to the next stage. Using the latest estimate of σ_C, we refine our initial sample size calculation to $n_{max,1} = 110$, and the second interim analysis is conducted when data are available on a total of 15 children per dose. At analysis $k = 2$, model (16) is fitted to obtain $\hat{\mu}_{C,2} = 3.78$ ($\hat{\mathcal{I}}_{C,2} = 4.11$), $\hat{\mathcal{I}}_2 = 3.98$ and $S_2 = -1.43$. Second stage boundaries of Test L and Test U spending additional error probabilities $f(r_2) - f(r_1) = 0.04$ and $g(r_2) - g(r_1) = 0.49$ are $(-5.26,\ -1.77)$ (Test L) and $(-0.63,\ 2.85)$ (Test U). These are combined to derive $l_{2,2} = -5.26$, $l_{1,2} = -1.77$, $u_{1,2} = -0.63$ and $u_{2,2} = 2.85$. Since $l_{1,2} \leq S_2 \leq u_{1,2}$ the GST stops to reject H_0. The 90% CI for θ is $(-0.97,\ 0.17)$ assuming that we would have observed $\hat{\mathcal{I}}_3 = \mathcal{I}_{max}$ had the test continued to the final stage. Thus the 90% CI is consistent with the decision of the IW GST since it excludes δ_L and δ_U.

4.4. *Results of the simulation study*

Table 2 lists results for $100\,000$ simulated PK-PD studies. Estimated type I error rates based on $100\,000$ simulations will lie between $(0.098,\ 0.102)$ with probability 0.95 if the true error rate is 0.1. Estimated coverage rates will lie between $(89.8,\ 90.2)$% with probability 0.95 if 90% CIs attain their nominal levels. The proposed designs appear to successfully control error rates at levels reasonably close to their nominal values. Attained type I error rates are consistently higher under $\theta = \delta_L$ than under $\theta = \delta_U$. Discrepancies between attained and nominal error rates are not solely due to deviations from the assumed joint distribution of test statistics used to derive test boundaries. Indeed, in results not presented here, we have simulated three-stage IW GSTs with fixed group sizes basing test statistics on successive MLEs of the log-effective concentration in a hyperbolic E_{max} model, and found the canonical joint distribution described in [21] is accurate for group sizes as small as 75 patients (spread across four active doses and placebo). The simulation study described in Section 4.2 considers a much more complex setting such that several factors may cause deviations from nominal error rates. Allowing mid-trial adaptations to the target paediatric sample size; deriving error spending boundaries conditioning on estimated information levels; and using ad-hoc strategies to proceed when the E_{max} model cannot be fitted will all influence attained error rates.

From Table 2 we see that attained error rates are robust to misspecification of σ_C^2 in the initial sample size calculation for the paediatric trial. Indeed, the procedure is able to adapt to accumulating evidence on σ_C to appropriately refine the sample size requirement. The coverage of 90% CIs remains close to the nominal level in all scenarios, with deviations within $\pm 5\%$. The proposed designs are feasible in the sense that estimation of the hyperbolic E_{max} model rarely failed: the highest failure rate observed in any simulation scenario was 3.1%. The highest observed rate of conflict between the decision of the IW GST and the 90% confidence interval was 1.4%.

Specification of δ_L and δ_U requires careful consideration since these equivalence limits have an important impact on sample size requirements. Large numbers of children are needed to make precise statements about the similarity of effective concentrations in adults and children. On average, setting $1 - \beta = 0.7$, a total of 394 children (interquartile range: 325 - 500) are needed to verify whether the effect of the adult EC$_{90}$ in children is between 70% and 93% of the maximum possible effect when in fact effective concentrations are equal. Relaxing this equivalence criterion slightly so that it would be acceptable to dose children targeting the adult EC$_{90}$ if the expected effect of this exposure was between 70% to 95% of the maximum effect implies a large saving in expected sample size. Under the relaxed criterion, the proposed approach appears feasible, requiring on average a total of 109 children (IQR: 85 - 125), which represents an expected saving of over 9% on the corresponding fixed sample test.

Note that the simulation study described in Section 4.2 was conducted assuming that a hyperbolic E_{max} model would always be an appropriate fit to the adult PK-PD data and the accummulating data in children. In practice, in each population model fit would need to be carefully monitored to ensure the accuracy of the MLEs $\hat{\mu}_A$ and $\hat{\mu}_{Ck}$, for $k = 1, 2, \ldots$, upon which the procedure is based. For instance, the goodness-of-fit of the adult PK-PD model would need to be established in order for it to be sensible to take $\hat{\mu}_A$ forwards to be used as the benchmark for comparison in the paediatric PK-PD study. When this study is underway, the goodness-of-fit of the PK-PD model for children should be assessed at each interim analysis. In the context of the trials simulated in this paper, a lack of fit of the adult PK-PD model to paediatric data would indicate that an alternative to the hyperbolic E_{max} model is needed to derive $\hat{\mu}_{Ck}$. Strictly speaking, this would not trigger termination of the GST in itself since the objective of the procedure is to verify whether a particular quantile of the PK-PD relationship is similar in adults and children, not whether the relationships are identical. However, it is unlikely that qualitatively different PK-PD relationships would be consistent with similar effective concentrations; indeed, different shaped relationships may imply that the drug has a quite different mode of action in children to adults. The decision of whether to continue in such cases should be guided by scientific considerations, and we do not attempt to propose a formal quantitative stopping rule based on goodness-of-fit statistics.

5. Discussion

In this paper we have shown how asymmetric IW GSTs can be used to learn about the similarities between adults and children to support extrapolation decisions. We have derived optimal versions of tests and discussed how confidence intervals can be derived. There are a number of practical and ethical considerations which may influence the conduct

Table 2. Operating characteristics of three-stage error spending IW GSTs of $H_0 : \theta \leq \delta_L$ or $\theta \geq \delta_U$ intended to have power $1 - \beta$ at $\theta = 0$ and type I error rate $\alpha = 0.1$ at $\theta = \delta_L$ and $\theta = \delta_U$. N_T represents the total number of children recruited on termination of the paediatric PK-PD study. IQR represents the interquartile range. Results are based are 100 000 simulations.

Expected effect at adult EC$_{90}$ (% E_{max})		% of times reject H_0 (% times 90% CI contains θ)				
δ_L	δ_U	$\theta = 0$	$\theta = \delta_L$	$\theta = \delta_U$	$\mathbb{E}(N_T; \theta = 0)$ (IQR)	$\mathbb{E}(n_{max,0})$ (IQR)
$1 - \beta = 0.8$, $\sigma_C = 4$						
80	93	77.0 (91.5)	12.3 (87.4)	8.6 (91.2)	602.5 (500, 750)	797.7 (745, 845)
70	93	76.8 (91.4)	12.9 (87.0)	8.6 (91.4)	553.8 (430, 730)	786.1 (735, 835)
80	95	78.0 (91.5)	12.3 (87.2)	10.0 (89.6)	199.5 (165, 235)	235.5 (225, 245)
70	95	77.9 (91.2)	13.6 (86.1)	10.6 (89.2)	141.0 (115, 170)	173.6 (165, 180)
$1 - \beta = 0.7$, $\sigma_C = 4$						
80	93	66.7 (91.2)	12.0 (87.6)	8.6 (91.1)	432.1 (355, 525)	534.5 (500, 565)
70	93	66.9 (91.5)	12.5 (87.4)	8.8 (91.1)	393.7 (325, 500)	516.9 (485, 545)
80	95	68.1 (91.6)	12.1 (87.2)	10.0 (89.3)	165.2 (140, 195)	189.6 (180, 200)
70	95	66.8 (91.4)	13.6 (85.8)	10.4 (89.1)	109.2 (90, 130)	129.4 (125, 135)
$1 - \beta = 0.8$, $\sigma_C = 3$						
80	95	78.5 (91.4)	12.7 (87.0)	10.3 (89.6)	108.7 (85, 125)	235.5 (225, 245)
70	95	77.9 (91.0)	13.9 (85.9)	10.6 (89.3)	78.1 (60, 90)	173.7 (165, 180)
$1 - \beta = 0.7$, $\sigma_C = 5$						
80	95	68.6 (91.8)	12.0 (87.4)	9.7 (89.7)	264.0 (225, 305)	189.6 (180, 200)
70	95	67.2 (91.5)	13.2 (86.2)	10.3 (89.3)	174.7 (140, 205)	129.4 (125, 135)

of the proposed study designs. Specifically, designs assume that the PK-PD study in children will proceed randomising patients between placebo and active doses $d_1, \ldots, d_M$ fixed ahead of time. In practice, specification of these active dose levels will likely be informed by extensive simulations, which in turn would draw on existing relevant adult data and paediatric PK data available, for example, from trials of drugs in the same class as the new medicine of interest. Data from an initial cohort of $\mathcal{N}_1 < (M+1)n_1$ children entered into the PK-PD study could be used to refine the positioning of the active dose levels if discrepancies emerged between the observed PK data and the simulation model used to design the trial. In case of safety concerns, the risk of excessive concentrations could be reduced by staggering the entry of these first $\mathcal{N}_1$ children, so that some PK data are available on lower doses before randomising children to the highest doses. A fixed randomisation scheme may not be ethically acceptable in therapeutic areas such as oncology, where drugs have toxic side-effect profiles. The designs proposed in this paper do not permit response adaptive randomisation, although future work could explore extensions which incorporate PK-PD data into dose assignments while preserving error rate control.

This paper has proposed methods facilitating the extrapolation of efficacy data from adults to children. It should be noted that smaller trials in children will inevitably reduce the size of the safety database on the dose that is eventually recommended. For example, in the second simulation scenario reported in Table 2, between 109 and 432 children spread across four active doses and placebo were, on average, recruited into the paediatric PK-PD trial when the extrapolation assumption was accurate, with the average sample size depending on δ_L and δ_U. The acceptability of such sample sizes for confirming the safety of a drug in children will depend on what is understood about the relationship between dose, PK and safety outcomes. Furthermore, it may not be permissible to augment the safety database in children by extrapolating from adult data if important issues specific to children, such as growth, pubertal or cognitive development, cannot be detected in adults. The question of how to establish the safety of a drug in children is an important and complex issue, but is considered beyond the scope of the current paper.

Throughout, derivations of optimal designs have relied on certain assumptions that may not be appropriate in all cases. For example, for the purposes of optimisation, we have used a skew normal distribution to represent prior beliefs that are consistent with an assumption of similar effective concentrations in adults and children. However, it is clear that the techniques proposed here could be extended to accommodate other choices of prior. In addition, we have formulated designs for the purpose of verifying similarities between adults and children. However, the ideas proposed here are more generally applicable to other scenarios in which we wish to verify similar drug effects in patients from different countries, or patients treated for different indications, in order to support extrapolation decisions.

References

1. Patterson S, Jones B. Bioequivalence and statistics in clinical pharmacology. Chapman & Hall/CRC, Boca Raton, 2006.
2. European Medicines Agency. ICH Topic E11. Clinical investigation of medicinal products in the paediatric population. Available at http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002926.pdf, 2001.
3. Food and Drug Administration. Guidance for Industry. Exposure-response relationships - study design, data analysis, and regulatory applications. Available at http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm072109.pdf, 2003.
4. Dunne J, Rodriguez WJ, Murphy MD, Beasley BN, Burckart GJ, Filie J D et al. Extrapolation of adult data and other data in pediatric drug-development programs. Pediatrics 2011; **128**:e1242–9.
5. European Medicines Agency. Concept paper on extrapolation of efficacy and safety in medicine development. Available at http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2013/04/WC500142358.pdf, 2013.
6. European Medicines Agency. Reflection paper on extrapolation of efficacy and safety in paediatric medicine development. Draft. Available at http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2016/04/WC500204187.pdf, 2016.
7. Jennison C, Turnbull BW. Sequential equivalence testing and repeated confidence intervals, with applications to normal and binary response. Biometrics 1993; **49**:31–43.
8. Liu F, Li Q. Exact sequential test of equivalence hypothesis based on bivariate non-central t-statistics. Computational Statistics and Data Analysis 2014; **77**:14–24.
9. Whitehead J. Sequential designs for equivalence studies. Statistics in Medicine 1996; **15**:2703–2715.
10. Whitehead J, Brunier H. The double triangular test: a sequential test for the two-sided alternative with early stopping under the null hypothesis. Sequential Analysis 1990; **9**:117–136.
11. Pampallona S, Tsiatis, AA. Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favour of the null hypothesis. Journal of Statistical Planning and Inference 1994; **42**:19–35.
12. Müller HH, Schäfer H. Optimization of testing times and critical values in sequential equivalence testing. Statistics in Medicine 1999; **18**:1769–1788.
13. Schuirmann D. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. Journal of Pharmacokinetics and Biopharmaceutics 1987; **15**:657–680.
14. Food and Drug Administration. Guidance for Industry. Bioavailability and bioequivalence studies for orally administered drug products - general considerations. Available at http://www.fda.gov/downloads/drugs/developmentapprovalprocess/howdrugsaredevelopedandapproved/approvalapplications/ abbreviatednewdrugapplicationandagenerics/ucm154838.pdf, 2003.
15. Food and Drug Administration. Guidance for Industry. Statistical approaches to establishing bioequivalence. Available at http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm070244.pdf, 2001.
16. Dragalin V, Fedorov V, Patterson S, Jones B. Kullback-leibler divergence for evaluating bioequivalence. Statistics in Medicine 2003; **22**:913–930.
17. Mauger DT, Chinchilli VM. An alternative index for assessing profile similarity in bioequivalence trials. Statistics in Medicine 2000; **19**:2855–2866.
18. Pei L, Hughes MD. A statistical framework for quantile equivalence clinical trials with application to pharmacokinetic studies that bridge from HIV-infected adults to children. Biometrics 2008; **64**:1117–1125.
19. Davidian M. Introduction to statistical population modeling and analysis for pharmacokinetic data. Available at http://www.epa.gov/ncct/uvpkm, 2006.
20. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika 1983; **70**:659–663.
21. Jennison C, Turnbull BW. Group sequential analysis incorporating covariate information. Journal of the American Statistical Association 1997; **92**:1330–1341.
22. Jennison C, Turnbull BW. Group Sequential Methods with Applications to Clinical Trials. Chapman & Hall/CRC, 2000.
23. Barber S, Jennison C. Optimal asymmetric one-sided group sequential tests. Biometrika 2002; **89**:49–60.
24. Hampson LV, Jennison C. Group sequential tests for delayed responses (with discussion). Journal of the Royal Statistical Society B 2013; **75**:3–54.

25. Stallard N, Facey KM. Comparison of the spending function method and the christmas tree correction for group sequential trials. Journal of Biopharmaceutical Statistics 1996; **6**:361–373.
26. Kosorok MR, Shi Y, DeMets DL. Design and analysis of group sequential clinical trials with multiple primary endpoints. Biometrics 2004; **60**:134–145.
27. Azzalini A, Capitanio A. The Skew-Normal and Related Families. Cambridge, 2014.
28. Eales JD. Optimal group sequential tests. PhD Thesis, University of Bath 1991.
29. Chang M. Optimal designs for group sequential clinical trials. Communications in statistics. Theory and methods 1996; **25**:361–380.
30. Banerjee A, Tsiatis, AA. Adaptive two-stage designs in phase II clinical trials. Statistics in Medicine 2006; **25**:3382–3395.
31. Eales J, Jennison C. Optimal two-sided group sequential tests. Sequential Analysis 1995; **14**:273–286.
32. Öhrn F, Jennison C. Optimal group sequential designs for simultaneous testing of superiority and non-inferiority. Statistics in Medicine 2010; **29**:743–759.
33. Hsu JC, Hwang J, Liu HK, Ruberg S. Confidence intervals associated with tests for bioequivalence. Biometrika 1994; **81**:103–114.
34. Fairbanks K, Madsen R. P-values using a repeated significance test design. Biometrika 1982; **69**:69–74.
35. Emerson SS, Fleming TR. Parameter estimation following group sequential testing. Biometrika 1990; **77**:875–892.
36. Emerson SS. Parameter estimation following group sequential hypothesis testing. PhD Thesis, University of Washington, Seattle 1988.
37. Novartis. Highlights of prescribing information. Diovan (Valsartan). Novartis Pharmaceuticals Corporation, New Jersy, 2014.
38. Habtemariam B, Sallas W, Sunkara G, Kern S, Jarugula V, Pillai G. Population pharmacokinetics of valsartan in pediatrics. Drug Metabolism and Pharmacokinetics 2009; **24**:145–152.
39. Sioufi A, Marfil F, Jaouen A, Cardot J, Godbillon J, Ezzet F, Lloyd P. The effect of age on the pharmacokinetics of valsartan. Biopharmaceutics and Drug disposition 1998; **19**:237–244.
40. Flynn J, Zhang Y, S Solar-Yohay, Shi V. Clinical and demographic characteristics of children with hypertension. Hypertension 2012; **60**:1047–1054.
41. Giles TD, Weber MA, Basile J, Gradman AH, Bharucha DB, Chen W et al. Efficacy and safety of nebivolol and valsartan as fixed-dose combination in hypertension: a randomised, multicentre study. The Lancet 2014; **383**:1889–1898.
42. Thomas N, Sweeney K, Somayaji V. Meta-analysis of clinical dose-response in a large drug development portfolio. Statistics in Biopharmaceutical Research 2014; **6**:302–317.
43. Dette H, Kiss C, Bevanda M, Bretz F. Optimal designs for the emax, log-linear and exponential models. Biometrika 2010; **97**:513–518.
44. European Medicines Agency. Qualification opinion of MCP-Mod as an efficient statistical methodology for model-based design and analysis of Phase II dose finding studies under model uncertainty. Available at http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2014/02/WC500161027.pdf, 2014.
45. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria 2013. URL `http://www.R-project.org/`.
46. Mehta, CR, Tsiatis, A A. Flexible sample size considerations using information based monitoring. Drug Information Journal 2001; **35**:1095–1112.
47. Jennison C, Turnbull, BW. Adaptive seamless designs: Selection and prospective testing of hypotheses. Journal of Biopharmaceutical Statistics 2007; **17**:1135–1161.