MAIN PAPER

Pharmaceutical Statistics

(wileyonlinelibrary.com) DOI: 10.1002/pst.1730

Published online 18 December 2015 in Wiley Online Library

Robust exchangeability designs for early phase clinical trials with multiple strata

Beat Neuenschwander,[a]* Simon Wandel,[a] Satrajit Roychoudhury,[b] and Stuart Bailey[c]

Clinical trials with multiple strata are increasingly used in drug development. They may sometimes be the only option to study a new treatment, for example in small populations and rare diseases. In early phase trials, where data are often sparse, good statistical inference and subsequent decision-making can be challenging. Inferences from simple pooling or stratification are known to be inferior to hierarchical modeling methods, which build on exchangeable strata parameters and allow borrowing information across strata. However, the standard exchangeability (EX) assumption bears the risk of too much shrinkage and excessive borrowing for extreme strata. We propose the exchangeability–nonexchangeability (EXNEX) approach as a robust mixture extension of the standard EX approach. It allows each stratum-specific parameter to be exchangeable with other similar strata parameters or nonexchangeable with any of them. While EXNEX computations can be performed easily with standard Bayesian software, model specifications and prior distributions are more demanding and require a good understanding of the context. Two case studies from phases I and II (with three and four strata) show promising results for EXNEX. Data scenarios reveal tempered degrees of borrowing for extreme strata, and frequentist operating characteristics perform well for estimation (bias, mean-squared error) and testing (less type-I error inflation).

Keywords: Bayesian; between-strata heterogeneity; exchangeability–nonexchangeability; hierarchical model; pooling; shrinkage; stratification

1. INTRODUCTION

In recent years, clinical trials with multiple strata (indications, regions, subgroups) have become more popular, in particular, in the early phases of drug development. Various reasons explain this trend, including the aim to identify interesting strata earlier or to combine information for small populations and rare diseases in a single trial.

Statistical discussions for subgroup analyses in clinical trials have a long history and can be controversial [1–3]. There are issues in the exploratory as well as confirmatory setting, and irrespective of whether one adopts a stratified or pooling perspective. Under stratification, where subgroups are analyzed separately, there is the danger of over-interpreting results for extreme subgroups, and results may be unreliable because of small sample sizes. On the other hand, complete pooling may lead to very precise but inaccurate conclusions.

Here, we consider early phase trials where the number of strata is small, strata are of small to moderate size, and therefore, borrowing information across similar strata is of interest. The ability to effectively use information may often be a crucial factor when deciding on the feasibility of the study. For example, complete pooling across regions in a trial is usually performed if regional results look sufficiently similar; otherwise, pooling is not advised, for example if a drug works well in a subset of strata only.

The idea to exploit the similarity and borrow information across strata has a long tradition (see, e.g., [4–6]), which originates in Stein's ground-breaking work [7,8]. The methods are known to improve estimation accuracy over estimates obtained from stratification or complete pooling. Standard hierarchical models assume full exchangeability of parameters, which is usually expressed via a random effects (exchangeability) model. The respective exchangeability distributions range from parsimonious models, often used if the number of strata is small, to high-dimensional or nonparametric models, for example for large data problems. For clinical trials, methods that borrow information across strata based on different exchangeability assumptions have been considered [9–14] and successfully applied in numerous trials, the most prominent being the *Imatinib* trial [15,16] and the *BATTLE* and *I-SPY 2* trials [17–19]. While these trials highlight the potential for more efficient trial designs, their implementation proved to be complex and required careful considerations with respect to statistical and operational challenges (such as randomized allocations and other trial adaptations).

For few strata of small to moderate size, typical for early phase trials, we propose models that allow each stratum-specific parameter to be exchangeable with parameters from other similar strata, or nonexchangeable with any of the other strata parameters. These exchangeability–nonexchangeability (EXNEX) models are robust versions of the often used full exchangeability (EX)

[a] *Novartis Pharma AG, Basel, Switzerland*

[b] *Novartis Pharmaceuticals, East Hanover, NJ, USA*

[c] *Novartis Pharmaceuticals, Cambridge, MA, USA*

**Correspondence to: Oncology, Novartis Pharma AG, Basel, Switzerland. E-mail: beat.neuenschwander@novartis.com*

model. They allow borrowing information across similar strata while avoiding too optimistic borrowing for extreme strata. EXNEX models are parametric, have easily interpretable parameters, allow for a few special partial exchangeability patterns, and can be implemented with standard Bayesian software (e.g. `WinBUGS` [20,21]). The EXNEX approach is inspired by ideas discussed in [22,23], which robustify the full exchangeability assumption with nonparametric models.

Section 2 provides an overview of hierarchical models and details of the EXNEX methodology and uses the phase II sarcoma trial by Chugh *et al.* [15,16] for illustration. In Section 3, we discuss two applications: a phase IIa trial with four strata and a phase I dose-escalation trial with three strata. Section 4 concludes the paper with a discussion. Technical details for prior distributions and `WinBUGS` code for EXNEX models with binary data are given in the on-line Appendix.

2. METHODOLOGY

This section provides an overview of hierarchical models and a more detailed discussion of the EXNEX approach. Throughout the section, we use the phase II trial by Chugh *et al.* [16] for illustration. Chugh *et al.* assessed the effect of Imatinib (Novartis) in 10 histological subtypes of sarcoma, after the compound had shown promising effects for gastrointestinal tumors [24,25]. The data are shown in Table I (left panel): 179 patients were available for analysis; sample sizes ranged from 2 to 29; and response rates for clinical benefit response (CBR) varied from 0% for strata 2 and 9 to 24% for stratum 5. The trial design was based on a standard hierarchical model (Section 2.1) with fully exchangeable strata parameters [15], which has recently been used to illustrate more flexible exchangeability structures [22,23].

2.1. Hierarchical models

Hierarchical models are widely used when data arise from different strata, such as trials, regions, and subgroups. They have two main components: a data model and a parameter model. Data Y_j from stratum $j = 1, \ldots, J$ follow a distribution F parameterized by a strata-specific parameter θ_j

$$Y_j|\theta_j \sim F(\theta_j), \tag{1}$$

and the strata parameters θ_j follow a distribution G

$$\theta_j|\eta \sim G(\eta), \tag{2}$$

where additional (nuisance) parameters have been omitted from notation. The parameter η determines the similarity (borrowing) across strata. Inference for strata parameters can be performed in a classical or Bayesian way. The simplest hierarchical model assumes (approximately) normal data. In many applications, the Y_j are sufficient statistics. For this case,

$$Y_j|\theta_j \sim N\left(\theta_j, s_j^2\right) \tag{3}$$

and

$$\theta_j|\mu, \tau \sim N(\mu, \tau^2). \tag{4}$$

For fixed s_j, a flat prior on μ, and fixed τ, classical and Bayesian results for μ and strata parameters θ_j are equivalent. For inverse-variance weights w_j (precisions) and shrinkage parameters B_j

$$w_j = 1/\left(s_j^2 + \tau^2\right), \quad B_j = s_j^2/\left(s_j^2 + \tau^2\right), \tag{5}$$

the posterior distribution of μ is

$$\mu|Y_1, \ldots, Y_J \sim N\left(\sum w_j Y_j/w_+, 1/w_+\right), \tag{6}$$

where $w_+ = \sum w_j$. Posterior distributions of the strata parameters θ_j are

$$\theta_j|Y_1, \ldots, Y_J \sim N\left(B_j E_\mu + (1 - B_j)Y_j, B_j\left(\tau^2 + B_j V_\mu\right)\right), \tag{7}$$

where E_μ and V_μ are the mean and variance in (6). Classical equivalents to the posterior means and standard deviations

Table I. Sarcoma trial by Chugh *et al.* (JCO 2009).

	Original data				Nugget scenario			
Subtype	r/n %	EX	EXNEX	STRAT	r/n %	EX	EXNEX	STRAT
1. Angiosarcoma	2/15 13	15	15	12	**7/15 47**	27	40	45
2. Ewing	0/13 0.0	13	2.7	1.3	0/13 0.0	13	2.4	1.3
3. Fibrosarcoma	1/12 8.3	14	13	7.3	1/12 8.3	15	13	7.3
4. Leiomyosarcoma	6/28 21	16	18	21	6/28 21	19	19	21
5. Liposarcoma	7/29 24	17	19	23	7/29 24	20	21	23
6. MFH	3/29 10	14	13	9.7	3/29 10	14	13	9.7
7. Osteosarcoma	5/26 19	16	17	18	5/26 19	18	18	18
8. MPNST	1/5 20	15	16	17	1/5 20	18	17	16
9. Rhabdomysarcoma	0/2 0.0	15	12	4.0	0/2 0.0	16	13	4.0
10. Synovial	3/20 15	15	15	14	3/20 15	17	16	14
DIC		37.0	38.4	42.1		45.8	42.1	43.2

Number with clinical benefit response (r) and number of patients (n) for original data and nugget scenario; posterior medians (%) for response rate are given for exchangeability (EX), exchangeability–nonexchangeability (EXNEX), and stratified (STRAT) analyses. DIC, deviance information criterion.
MFH, Malignant fibrous histiocytoma; MPNST, Malignant peripheral nerve sheath tumor.

are maximum likelihood estimates and their standard errors. The special cases of complete pooling and stratification arise for $\tau = 0$ and $\tau = \infty$, respectively.

Hierarchical models have the desirable properties one would expect from an approach that aims to improve inference by taking advantage of ('borrowing from') data from similar strata:

- Compared with the stratified estimate Y_j, the hierarchical model estimate is shrunk toward the population mean estimate, which is a safeguard against over-interpreting extreme strata estimates. Shrinkage is determined by two factors: stratum size (for large strata, i.e., small s_j, shrinkage is small) and between-strata heterogeneity (notable shrinkage is only possible if τ is of small to moderate size).
- The hierarchical model leads to precision gains: because

$$B_j\left(\tau^2 + B_j V_\mu\right) = s_j^2 - s_j^2 B_j(1 - w_j/w_+), \tag{8}$$

the variance s_j^2 under stratification is always larger than the variance (8).

Although these properties are attractive, there are caveats. First, borrowing and precision gains are of course only advantageous if the exchangeability assumption (4) is reasonable. This may not be the case when there are outlying strata, or when strata cluster around more than one mean parameter μ. Second, unknown between-strata heterogeneity poses a major problem when the number of strata is small, because τ, which drives the amount of borrowing, can then not be well inferred from the data.

The inferential challenge for τ applies to both the classical and Bayesian setting. For the former, two problems arise: first, many estimates have been proposed [26], and they may vary considerably for small J; second, estimation uncertainty for τ is large for small J and should therefore be accounted for in the estimation of the strata parameters θ_j. In the Bayesian setting, on the other hand, the prior for τ matters. In situations where there is no reliable *a priori* information for between-strata heterogeneity, we will account for the uncertainty about τ by prior distributions that put most of their probability mass to values representing small to large heterogeneity; half-normal, half-Cauchy, and half-t distributions have been suggested in this context [27–29].

For instance, for binary data and the between-strata standard deviation τ on the log-odds scale, Table II shows four typical values representing *small*, *moderate*, *substantial*, and *large* heterogeneity. For example, for $\tau = 0.25$, the 97.5%-*quantile-to-mean odds ratio* for strata parameters is 1.63, which implies considerable concentration of strata parameters around the mean μ. Alternatively, for two arbitrary strata parameters, there is a 95% probability that the *two-parameter odds ratio* is between 0.25 and 4. For $\tau = 0.5$ and 1, the respective odds ratios show considerably larger heterogeneity. The lower part of Table II summarizes three priors that cover the range of small to large between-strata heterogeneity: two half-normal priors (with scale parameters 0.5 and 1) and a uniform (0,2) prior. In Section 3.1, we will use half-normal priors as proposed in [27].

For the example in Table I, the analysis of Chugh *et al.* was based on a hierarchical model [15] with a binomial sampling model

$$r_j|\pi_j \sim \text{Binomial}(\pi_j, n_j), \quad j = 1, \ldots, 10, \tag{9}$$

where n_j and r_j are the number of patients and the corresponding number of patients with CBR for each stratum. Exchangeable log-odds parameters $\theta_j = \log(\pi_j/(1-\pi_j))$ were specified as

$$\theta_j|\mu, \tau \sim N(\mu, \tau^2), \quad j = 1, \ldots, 10. \tag{10}$$

The degree of borrowing across strata was small, because the prior for τ, obtained after elicitation from experts, was concentrated in a range representing large to very large between-strata heterogeneity: $\tau^{-2} \sim \text{Gamma}(2, 20)$, with prior 95% interval (1.89, 9.09) for τ. Despite the fact that the prior excluded the possibility of substantial pooling, the posterior probabilities that CBR rates exceed the clinically relevant threshold 30% were small for all strata (between 0 and 0.134, table 4 in [16]), which was sufficient to conclude that 'imatinib is not an active agent in advanced sarcoma in these subtypes' [16].

2.2. Exchangeability–nonexchangeability models for binary data

We now extend the full exchangeability model (10) by allowing for nonexchangeability (NEX). In view of the application of Section 3.1, we confine the discussion to binary data. The EXNEX approach allows each log-odds parameter $\theta_j = \log(\pi_j/(1-\pi_j))$ to be either exchangeable with some of the other strata parameters, or nonexchangeable with any of them. Thus, for

Table II. Between-strata heterogeneity (small, moderate, substantial, and large) as a function of the between-strata standard deviation τ for log-odds parameters θ: (1) odds ratio (97.5%-quantile to mean), (2) 95% interval for odds ratio of two strata parameters, and (3) three priors for τ with median (95% interval).

	Small $\tau = 0.125$	Moderate $\tau = 0.25$	Substantial $\tau = 0.5$	Large $\tau = 1$
(1) $\exp(\theta_{97.5\%} - \mu)$	1.28	1.63	2.66	7.10
(2) $\exp(\theta_1 - \theta_2)_{95\%}$	(1/2,2)	(1/4,4)	(1/16,16)	(1/256,256)
(3) Prior distributions: median (95% interval)				
$\tau \sim$ half-normal(scale = 0.5)		0.34 (0.016,1.12)		
$\tau \sim$ half-normal(scale = 1)		0.67 (0.031,2.24)		
$\tau \sim$ uniform(0,2)		1.00 (0.050,1.95)		

each stratum two possibilities arise, with respective fixed *a priori* weights p_j and $1 - p_j$:

(i) EX: with probability p_j, θ_j follows a normal distribution with exchangeability parameters μ and τ:

$$\theta_j|\mu, \tau \sim N(\mu, \tau^2). \quad (11)$$

In the sequel, we refer to (11) as the *exchangeability distribution*.

(ii) NEX: with probability $1-p_j$, θ_j is nonexchangeable with any of the other parameters. For this case, strata-specific priors will be used

$$\theta_j \sim N(m_j, v_j). \quad (12)$$

A method for setting weakly informative priors in the aforementioned model is given in the on-line Appendix. Alternatively, if it is *a priori* suspected that a stratum behaves systematically different than the other strata, this can be represented by an increased nonexchangeability weight p_j and an informative stratum-specific NEX prior in (12). For example, this could be the case in a situation with children and adults, where children are suspected to be more prone to toxicity. The example shows that the proposed model is flexible in the way it handles borrowing, which is driven by two inputs: the global exchangeability parameter τ and the local parameters p_j.

The EXNEX model (11) and (12) with *a priori* weights p_j and $(1 - p_j)$ covers the special cases of fully exchangeable or stratified Bayesian analyses by setting $p_j = 1$ or $p_j = 0$ for all strata parameters. The model can be extended, for example:

(i) A single exchangeability distribution (11) may be too restrictive. The extension to more than one exchangeability distribution is technically straightforward. However, the choice of mixture weights and prior distributions will require more thought. An example with two EX distributions is discussed in application 3.1.

(ii) Another extension refers to the *a priori* weights. Even though in the aforementioned formulation they were assumed fixed (a practice commonly applied in settings with Bayes factors), standard Bayesian calculus for mixture models [27,30] shows that the posterior distribution of θ_j is again a mixture, with component-wise posterior distributions and updated mixture weights. The latter depend on the *a priori* weights and on how likely the data are under the mixture components EX and NEX. Yet, instead of assuming fixed *a priori* weights, they could be represented with an additional hierarchical layer via Dirichlet distributions. While in our applications, where so far the data have been sparse, we have not seen substantial differences between fixed and uncertain weights, allowing for uncertain weights could be a better approach in other applications.

Finally, it should be noted that, strictly speaking, if mixture weights p_j and prior parameters m_j and v_j in (12) are the same for all strata, the parameters θ_j are *a priori* exchangeable, because the joint distribution of strata parameters is permutation invariant. Thus, the terminology 'exchangeable–nonexchangeable' in (11) and (12) is not precise. It is motivated by the fact that borrowing (no-borrowing) is often used interchangeably with exchangeability (nonexchangeability). Of course, unequal mixture weights across strata imply nonexchangeable parameters, even though equal weights for all strata may be a reasonable choice in many situations. We recommend to tailor the weights to the actual application. For the case of little *a priori* knowledge, we suggest non-extreme weights, such that exchangeability and nonexchangeability are *a priori* likely.

We now further investigate the example in Table I with two data sets: the *original trial data* (left panel) and a hypothetical *nugget scenario* (right panel), for which stratum 1 shows very promising efficacy: seven responders in 15 patients. The *nugget scenario* will highlight differences between the results for the three models, with respective EXNEX mixture weights (1,0), (0.5,0.5), and (0,1) for each stratum, respectively:

(i) EX: the standard *exchangeability* model (11), with a weakly informative prior for μ and a half-normal (scale 1) prior for τ. The half-normal prior covers very small to very large heterogeneity (95% interval (0.031,2.24)), Table II. For μ, the normal $N(-1.73, 2.616^2)$ prior is centered at logit(0.15) and results in a marginal variance for θ_j ((22) in the on-line Appendix) that is approximately worth one observation.

(ii) EXNEX: the EXNEX model, which allows each stratum parameter θ_j to be either exchangeable with other strata parameters (11) or nonexchangeable with any of them. For EX, the specifications (i) were used. For NEX, the prior was chosen as weakly informative, with mean logit(0.15) and variance approximately worth one observation: $\theta_j \sim N(-1.734, 2.801^2)$. The resulting prior 95% interval for each stratum response rate under NEX is (0.00, 0.98).

(iii) *stratified* analyses, with independent weakly informative priors $N(-1.734, 2.801^2)$ for all θ_j.

For the original data (left panels in Table I and Figure 1), the EX analysis shows the strongest borrowing. Under EXNEX borrowing is less because of the the additional NEX component. Compared with the stratified analyses (Bayesian or frequentist), EX and EXNEX deliver considerable precision gains. This is useful for the very small strata 8 and 9, for which stratified results would have been inconclusive. According to the deviance information criterion [31], EX and EXNEX perform better than the stratified analysis.

Of interest are cases where the exchangeability assumption is in doubt. For the *nugget scenario*, the EX analysis pulls back the parameter estimate of the promising stratum 1 quite strongly (right panels of Table I and Figure 1): from 47% (=7/15) to 27%, compared with 40% for EXNEX. A similar pattern emerges for the other extreme stratum 2, with an observed CBR rate of 0%, and EX and EXNEX estimates of 13% and 2.4%. The EXNEX analysis is more robust for extreme strata, while borrowing across the remaining strata is still guaranteed. The deviance information criterion indicates the benefit of adding a robust NEX component for EXNEX, which performs better than EX as well as stratification. Also, the stratified analysis would be preferred over EX, which for the case of outlying strata, shows the limitations of the full exchangeability model (10).

How EX and EXNEX capture heterogeneity can be seen from the posterior summaries of the global heterogeneity parameter τ and the local mixture weights p_j (Table III). For the *original data*, despite considerable borrowing under EX, uncertainty for the between-strata standard deviation τ is still quite large. The posterior median and 95% interval are 0.28 (0.01,1.01), which still represent a wide range of heterogeneity, although considerably less than under the half-normal prior 0.67 (0.03,2.24). For EXNEX, the posterior summaries are 0.29 (0.01,1.22).

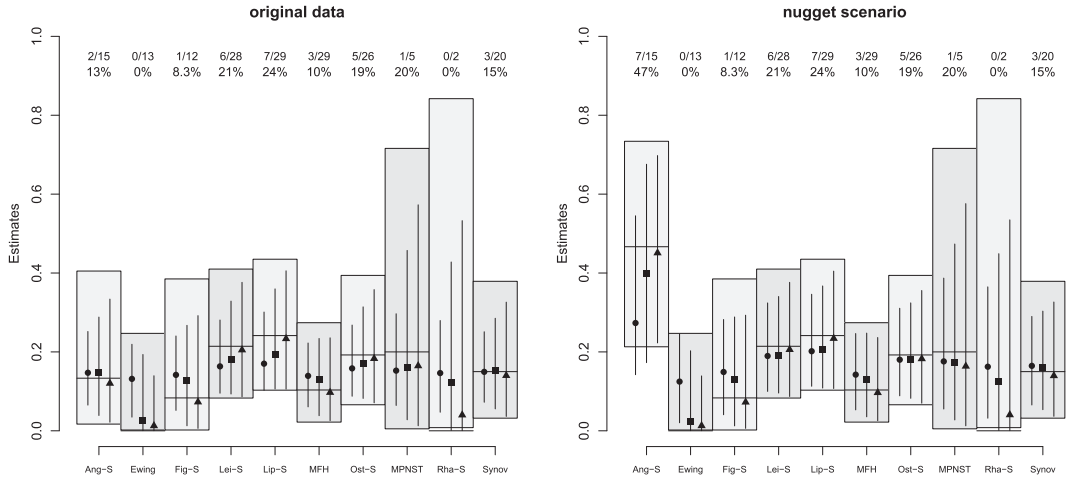

Figure 1. Chugh *et al.*: Exchangeability (circle), exchangeability–nonexchangeability (square), and stratified (triangle) posterior medians and 95% intervals for original data and nugget scenario, observed rates (horizontal lines), and exact frequentist 95% intervals (rectangles).

Table III. EX and EXNEX analyses for sarcoma trial by by Chugh *et al.* (JCO 2009): posterior summaries for heterogeneity parameters τ (median and 95%) and strata mixture weights for exchangeability; prior mixture weights for EXNEX = 0.5 for the EX and NEX mixture components.

	Original data		Nugget scenario	
	EX	EXNEX	EX	EXNEX
	τ: median (95% interval)			
	0.28 (0.01,1.01)	0.29 (0.01,1.22)	0.51 (0.04,1.36)	0.36 (0.02,1.40)
	Mixture weights p_j			
1. Angiosarcoma	–	0.74	–	0.36
2. Ewing	–	0.29	–	0.26
3. Fibrosarcoma	–	0.66	–	0.62
4. Leiomyosarcoma	–	0.76	–	0.76
5. Liposarcoma	–	0.72	–	0.74
6. MFH	–	0.72	–	0.67
7. Osteosarcoma	–	0.77	–	0.77
8. MPNST	–	0.69	–	0.68
9. Rhabdomysarcoma	–	0.54	–	0.53
10. Synovial	–	0.77	–	0.75

EX, exchangeability; EXNEX, exchangeability–nonexchangeability.

The similarity of strata for the *original data* is supported by the increasing mixture weights for the EX component for all strata except stratum 2. For the former, weights increase from their prior values (0.5) to 0.54 (for stratum 9) to 0.77 (strata 7 and 10). For stratum 2, the mixture weight decreases to 0.29, showing that this stratum (with 0 responders in 10 patients) is a potential outlier.

Similar results for the heterogeneity parameters can be seen for the *nugget scenario*, except for stratum 1 (seven responders in 15 patients), for which the mixture weight for the EX component drops from 0.5 to 0.36. For τ, the posterior summaries show slightly larger uncertainty compared with the ones for the *original data*.

In summary, extending the full exchangeability model to mixtures that allow for exchangeability and nonexchangeability looks promising. However, design specifications for EXNEX are more involved, which requires careful considerations for the specification of prior distributions and mixture weights. Also, properties of the design should be thoroughly investigated by hypothetical data scenarios and frequentist operating characteristics for various simulation scenarios (Section 3).

2.3. Exchangeability–nonexchangeability models for binary dose-toxicity data

In the following, we consider binary dose-toxicity data for a phase I dose-escalation trial with several strata, which will be discussed in application 3.2. For each stratum $j = 1, \ldots, J$, assume that at dose d there are n_{jd} patients, of which r_{jd} experience a dose-limiting toxicity (DLT). The statistical model is binomial,

$$r_{jd}|\pi_{jd} \sim \text{Binomial}(n_{jd}, \pi_{jd}). \quad (13)$$

For each stratum, the dose-toxicity model is logistic

$$\text{logit}(\pi_{jd}) = \log(\alpha_j) + \beta_j \log(d/d^\star), \tag{14}$$

where $\alpha_j, \beta_j > 0$ and $d^\star$ is a fixed scaling dose, implying that α_j is the odds of a DLT at $d^\star$ [32–35]. Inferential results are used to guide dose escalations during the trial and to eventually determine suitable doses for investigation in later trials. We confine the model to one exchangeability distribution and nonexchangeability, with fixed *a priori* weights p_j and $1 - p_j$:

(i) EX: with probability p_j, the logistic parameter vector

$$\theta_j = (\log(\alpha_j), \log(\beta_j))$$

is exchangeable with at least one of the other strata parameters and follows a bivariate normal distribution with mean vector μ and covariance matrix Σ:

$$\theta_j | \mu, \Sigma \sim N_2(\mu, \Sigma) \tag{15}$$

with

$$\mu = (\mu_1, \mu_2), \quad \Sigma = \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{pmatrix}. \tag{16}$$

(ii) NEX: with probability $1 - p_j$, θ_j is nonexchangeable with any of the other strata parameters and follows a prior distribution with mean vector m_w and covariance matrix S_w:

$$\theta_j \sim N_2(m_w, S_w). \tag{17}$$

3. APPLICATIONS

This section presents the design of two trials that were based on the EXNEX approach. The properties of the designs were investigated with data scenarios and frequentist operating characteristics for various simulation scenarios assuming true DLT rates. Data scenarios help clinical teams and review boards to understand decisions derived from the model-based inferential summaries, in particular how and why they may differ from alternative designs. Here, we will compare EXNEX results with those of standard exchangeability and stratified analyses.

3.1. Application 1: phase II trial with four indications

We consider the design of a phase II trial in four indications. The goal of the trial was to investigate potential activity for two previously tested indications as well as two new indications. It was decided to perform a small non-randomized phase II trial, comparing treatment response rates for each indication with 10%, the approximate standard of care rate seen previously in the four indications.

Because for all indications the treatment inhibits the same target for tumor growth, similarity of rates was considered possible, although some uncertainty remained. Given the limitation for total sample size (50–60 patients), the clinical team encouraged a design that takes advantage of potentially similar response rates.

For each indication, two criteria were required for trial success:

(1) the estimated response rate (posterior mean) is at least 20%;
(2) and the posterior probability that the rate exceeds 0.1 is at least 90% for indications 1 and 2, and at least 80% for indications 3 and 4.

Data scenarios and frequentist operating characteristics for simulation scenarios were used to assess the design (see the succeeding texts). Sample sizes were set to 20 for indications 1 and 2, and 10 for indications 3 and 4. The smaller sample sizes and lower success requirements for indications 3 and 4 were chosen because they were considered less important.

The model specifications were as follows: due to the possibility of activity for indications 1 and 2 and no activity for the other indications (or other constellations for which a single exchangeability distribution could be too restrictive), an EXNEX model with two exchangeability distributions was chosen. The weights for the two EX distributions and NEX were set to

$$p_j = (0.25, 0.25, 0.5)$$

for all indications. This reflects the uncertainty about exchangeability and assigns non-negligible probabilities to the various combinations of EX (one or two distributions) and NEX. For comparisons with alternative designs (EX and stratified), the mixture weights p_j=(1,0,0) and p_j=(0,0,1) were used. Finally, weakly informative prior distributions were chosen for all model parameters; see on-line Appendix for details.

Table IV shows data scenarios (observed response rates) for which decisions (+/− for success/failure) are given for EXNEX, EX,

Table IV. Data scenarios for phase II trial (application 1): observed rates for four indications, and the three decisions (+/− = success/failure) for EXNEX [1], EX [2], and stratified [3] analyses; see text.

Data scenario	Data (%) r_1/n_1	r_2/n_2	r_3/n_3	r_4/n_4	Indication 1 $n_1 = 20$	Indication 2 $n_2 = 20$	Indication 3 $n_3 = 10$	Indication 4 $n_4 = 10$
1	20	20	30	20	$+^1$ $+^2$ $-^3$	+ + −	+ + +	+ + −
2	20	20	30	30	+ + −	+ + −	+ + +	+ + +
3	20	25	20	20	+ + −	+ + +	+ + −	+ + −
4	20	25	30	20	+ + −	+ + +	+ + +	+ + −
5	20	25	30	30	+ + −	+ + +	+ + +	+ + +
6	25	25	20	20	+ + +	+ + +	+ + −	+ + −
7	25	25	30	20	+ + +	+ + +	+ + +	+ + −
8	10	10	10	50	− − −	− − −	− − −	+ + +
9	10	50	50	50	− − −	+ + +	+ + +	+ + +
10	10	30	30	50	− − −	+ + +	+ + +	+ + +
11	10	30	50	70	− − −	+ + +	+ + +	+ + +

EX, exchangeability; EXNEX, exchangeability–nonexchangeability.

and stratified analyses. Scenarios 1–7 represent fairly homogenous data, with observed rates of 20% to 30%. Under stratification, success is achieved if the observed rate is at least 25% (5/20) for indications 1 and 2, and at least 30% (3/10) for indications 3 and 4. Therefore, scenarios 1–7 represent cases of borderline activity, for which different decisions may arise under stratification (no borrowing) and EXNEX or EX (with borrowing). Scenarios 8–11 represent substantially different observed rates for the four indications.

For example, under stratification, indication 1 fails for scenarios 1–5 (observed 20% instead of the required 25%), which differs from the EXNEX and EX analyses; both borrow from the better looking indications, which suffices to declare success for the borderline observed rate of 20%.

Interestingly, conclusions for EXNEX and EX are the same for all data scenarios in Table IV. This can be explained by the rather conservative prior for the borrowing parameter (between-indication standard deviation) τ, a half-normal distribution with scale parameter 1. Of course, this will not always be the case: For example, if a more optimistic half-normal prior with scale 0.5 had been used, indication 1 would have been declared successful for scenario 9 under EX, whereas under EXNEX, it would still have failed.

In addition to assure sensible inferences for data scenarios, it is important to assess the operating characteristics of the design. Table V shows frequentist metrics for various simulation scenarios. Because success is only declared if there is reasonable evidence that the response rate exceeds 10%, this threshold would serve as the null hypothesis in the classical (frequentist) setting. For a true rate of 10%, and the required 90% (indications 1 and 2) and 80% (indications 3 and 4) posterior probabilities that the response rates exceed 10%, one would therefore expect to approximately control the type-I error (falsely declare success) at 10% and 20%, respectively [36,37].

In fact, for stratification and EXNEX, type-I errors are fairly well controlled. However, due to more borrowing under EX compared with EXNEX, type-I error inflation can be substantial if response rates differ: for scenarios 5–7, the type-I errors for indication 1 are 0.32, 0.26, and 0.25, respectively. The respective type-I errors under EXNEX, however, are much lower (0.13). For true rates of 0.3, 0.5, and 0.7, gain in power for EXNEX and EX compared with stratification can be substantial if response rates are similar.

Estimation metrics (bias and mean squared error (MSE)) show a similar behavior. Results for EX become problematic for the heterogenous scenarios 5–7: biases for extreme indications can be as high as 12.6% (scenario 4 and indication 4), considerably larger than the ones for EXNEX. On the other hand, for MSE, EX is best if response rates are equal (scenarios 1 and 2) but gets increasingly worse compared with the other analyses if rates are dissimilar. Overall, MSE for EXNEX look good and are always better than under stratification.

In summary, EXNEX performs well for data scenarios and operating characteristics for various assumed parameters. As expected, due to the robust component, borrowing is performed

Table V. Operating characteristics for phase II trial (application 1): probability of trial success, bias, and mean-squared error for EXNEX [1], EX [2] and stratified [3] analyses; see text.

						% success				100 × bias				100 × MSE			
Scenario						Indication											
	π_1	π_2	π_3	π_4	(%)	1	2	3	4	1	2	3	4	1	2	3	4
						5[1]	5	8	8	0.8	0.8	1.9	1.7	0.33	0.34	0.59	0.56
1:	10	10	10	10		2[2]	2	4	3	0.3	0.3	0.7	0.6	0.19	0.19	0.23	0.22
						4[3]	4	18	17	0.9	0.9	2.4	2.1	0.37	0.37	0.71	0.66
						88	88	85	85	-0.1	-0.2	0.1	-0.3	0.83	0.84	1.43	1.37
2:	30	30	30	30		93	92	92	91	-0.1	-0.1	0.2	0	0.49	0.51	0.61	0.59
						75	76	75	74	-0.3	-0.4	0	-0.4	0.95	0.96	1.74	1.66
						10	9	69	70	1.3	1.3	-0.9	-1.4	0.37	0.38	1.63	1.57
3:	10	10	30	30		10	10	61	60	3.3	3.3	-5.9	-6.2	0.35	0.36	1.35	1.34
						4	4	75	74	0.9	0.9	0	-0.5	0.37	0.37	1.74	1.66
						7	7	15	96	1.1	1.1	2.4	-3.1	0.36	0.37	0.67	2.26
4:	10	10	10	50		8	8	16	88	2.5	2.5	3.9	-12.6	0.34	0.34	0.53	3.66
						4	4	18	97	0.9	0.9	2.4	-1.9	0.37	0.37	0.71	2.1
						13	100	99	99	1.6	-1.1	-1.7	-1.9	0.44	1.08	1.94	1.87
5:	10	50	50	50		32	100	99	99	7.4	-3.3	-4.5	-4.6	1.05	1.08	1.68	1.64
						4	99	97	97	0.9	-1.1	-1.7	-1.9	0.37	1.15	2.19	2.1
						13	86	84	98	1.7	-0.2	0.1	-3.2	0.42	0.91	1.57	2.06
6:	10	30	30	50		26	88	84	98	6.4	-1.1	-0.9	-9.7	0.78	0.67	0.95	2.48
						4	76	75	97	0.9	-0.4	0	-1.9	0.37	0.96	1.74	2.1
						13	84	99	100	1.6	0.2	-1.9	-4.7	0.42	0.97	2.1	1.92
7:	10	30	50	70		25	90	99	100	6.1	1	-4.3	-10.8	0.81	0.78	1.8	2.87
						4	76	97	100	0.9	-0.4	-1.7	-3.6	0.37	0.96	2.19	1.82

EX, exchangeability; EXNEX, exchangeability–nonexchangeability.

more cautiously for dissimilar rates, resulting in better results for type-I error, bias, and MSE. Judgment about the likelihood of dissimilar rates helps to specify the mixture weights p_j. If similar rates are expected, nonexchangeability weights (0.5 in the application) should be decreased, which would then lead to similar results for EXNEX and EX.

3.2. Application 2: phase I trial with three indications

This application considers a phase I study to determine the maximum tolerable dose (MTD) and recommended phase II dose of a new compound. Dose escalations in the study happened concurrently in three indications (strata), for which dose-toxicity profiles were expected to be similar. However, disregarding indications (complete pooling) was considered inappropriate. Instead, it was decided to use an EXNEX approach.

In phase I trials, cohorts of three patients are usually enrolled to a new dose. After all patients in the cohort have successfully finished their minimum observation period or experienced a DLT, the dose for the next cohort is selected. The decision can be to (i) escalate to a higher dose, (ii) remain at the same dose, (iii) reduce the dose, or (iv) declare the MTD. A cohort consisted of patients from three indications. For each cohort, patients from the same indication were given the same dose, but doses between indications could vary. Doses for this trial were 2, 4, 6, 12, 24, 40, and 56 mg.

Due to expected similarities of dose-toxicity profiles, it was decided to start with two patients per indication, providing a total of six patients per cohort. Some flexibility with regard to sample size and dose selection was allowed for patient cohorts, depending on accumulating safety, pharmacokinetic, and efficacy data.

The most important and therefore binding decision rule refers to patient safety: *escalation with overdose control* (EWOC [32]) must always be guaranteed. That is, for each indication (j = 1,2,3), only doses d can be administered to patients if the probability of DLT π_{jd} fulfills the following overdose criterion:

$$\Pr(\pi_{jd} \geqslant 0.33|\text{data}) < 0.25. \tag{18}$$

Here, 0.33 is the upper bound of the target interval (0.16-0.33) for π_{jd}.

For the logistic EXNEX model of Section 2.2, weakly informative priors were obtained using similar ideas as described in Section 3.1; for details of the prior distributions see on-line Appendix. For the three indications, the mixture weights were chosen as $p_j = (0.9, 0.1)$, reflecting the *a priori* confidence in the similarity of indications. For the comparisons with alternative designs (EX and stratified), the mixture weights $p_j = (1, 0)$ and (0,1) were used.

Table VI shows data scenarios and indication-specific dose recommendations derived from EXNEX, EX, and stratified analyses. Two points should be noted when interpreting the results. First, for simple designs (with only one indication), if cohorts are of size 3 and *a priori* information for DLT rates is weak, model-based (EWOC compliant) dose recommendations for the first cohorts

Table VI. Data scenarios for phase I trial (application 2): for EXNEX[1], EX[2], and stratified[3] analyses, escalations refer to dose recommendations under EWOC, that is, the requirement to de-escalate (↓), and the possibilities to escalate (↑) or stay at the same dose (→); see text.

Data scenario	Cohort	Indication 1 dose: r_1/n_1	Indication 2 dose: r_2/n_2	Indication 3 dose: r_3/n_3
1	1	6mg: 1/3	6mg: 0/2	6mg: 0/2
		↑[1] ↑[2] →[3]	↑ ↑ ↑	↑ ↑ ↑
2	1	6mg: 1/3	6mg: 1/3	6mg: 0/2
		→ → →	→ → →	→ → ↑
3	1	6mg: 1/3	6mg: 1/3	6mg: 1/3
		↓ ↓ →	↓ ↓ →	↓ ↓ →
4	1	6mg: 0/2	6mg: 0/2	6mg: 0/2
	2	12mg: 1/3	12mg: 1/3	12mg: 0/2
		→ → →	→ → →	→ → ↑
5	1	6mg: 0/2	6mg: 0/2	6mg: 0/2
	2	12mg: 1/3	12mg: 0/2	12mg: 0/2
		↑ ↑ →	↑ ↑ ↑	↑ ↑ ↑
6	1	6mg: 0/2	6mg: 0/2	6mg: 0/2
	2	12mg: 0/2	12mg: 0/2	12mg: 0/2
	3	24mg: 0/2	24mg: 1/3	24mg: 2/3
		→ ↓ ↑	↓ ↓ →	↓ ↓ ↓
7	1	6mg: 0/2	6mg: 0/2	6mg: 0/2
	2	12mg: 0/2	12mg: 0/2	12mg: 0/2
	3	24mg: 0/2	24mg: 0/2	24mg: 2/3
		→ → ↑	→ → ↑	↓ → ↓

EX, exchangeability; EXNEX, exchangeability–nonexchangeability; EWOC, escalation with overdose control.

are usually similar to recommendations from the simple 3 + 3 design [38]: 0/3 (1/3) leads to dose escalation (staying at the same dose). Second, if more data are available, either later in the trial or from more than one indication, model-based dose recommendations will usually deviate from 3 + 3 rules. The reason is that model-based designs account for all the data, which explains why they perform better than simplistic algorithms like the 3 + 3.

For cohort 1 (data scenarios 1–3), EXNEX and EX conclusions are identical. This is not surprising, because the weight for EX is 0.9 and the amount of data is small. More interesting are the

Table VII. Operating characteristics for EXNEX and stratified analyses (application 2): probability to declare doses with less than 10%, 16%, 25%, or more than 25% DLT rate, or to stop the trial without declaring an MTD (all doses too toxic).

	True DLT rate by dose (%)							Probability (%) of MTD with DLT rate				
Scenario	π_2	π_4	π_6	π_{12}	π_{24}	π_{40}	π_{56}	$< 10\%$	16%	25%	$> 25\%$	Stop
						EXNEX						
	1	2	3	5	16	25	40	10	54	28	9	0
1	1	2	3	5	16	25	40	9	54	29	9	0
	1	2	3	5	16	25	40	9	54	28	9	0
	2	3	5	16	25	40	60	4	52	40	4	0
2	2	3	5	16	25	40	60	4	52	41	4	0
	2	3	5	16	25	40	60	4	52	41	3	0
	3	5	16	25	40	60	80	1	35	55	6	4
3	3	5	16	25	40	60	80	1	37	54	5	4
	3	5	16	25	40	60	80	1	36	55	4	4
	3	5	16	25	40	60	80	0	21	60	15	4
4	3	5	16	25	40	60	80	1	24	59	13	3
	1	2	3	5	16	25	40	47	40	8	4	1
	1	2	3	5	16	25	40	24	53	16	6	0
5	1	2	3	5	16	25	40	26	52	15	7	0
	3	5	16	25	40	60	80	1	14	52	30	3
	1	2	3	5	16	25	40	38	48	10	4	0
6	2	3	5	16	25	40	60	4	52	40	4	0
	3	5	16	25	40	60	80	0	15	62	21	2
						Stratified						
	1	2	3	5	16	25	40	15	46	25	13	0
1	1	2	3	5	16	25	40	15	48	23	14	0
	1	2	3	5	16	25	40	14	47	25	14	0
	2	3	5	16	25	40	60	5	45	43	6	1
2	2	3	5	16	25	40	60	5	48	40	7	1
	2	3	5	16	25	40	60	4	47	42	6	1
	3	5	16	25	40	60	80	1	29	51	10	9
3	3	5	16	25	40	60	80	1	29	50	10	10
	3	5	16	25	40	60	80	1	27	50	11	10
	3	5	16	25	40	60	80	1	27	53	12	8
4	3	5	16	25	40	60	80	1	25	53	12	9
	1	2	3	5	16	25	40	21	47	20	11	0
	1	2	3	5	16	25	40	15	45	24	15	1
5	1	2	3	5	16	25	40	15	47	24	14	0
	3	5	16	25	40	60	80	1	28	51	10	10
	1	2	3	5	16	25	40	13	47	25	14	0
6	2	3	5	16	25	40	60	4	47	41	6	1
	3	5	16	25	40	60	80	1	28	50	11	10

For each scenario, the three rows correspond to the three indications; see text. EXNEX, exchangeability–nonexchangeability; DLT, dose-limiting toxicity; MTD, maximum tolerable dose.

divergent conclusions for the stratified analysis: for scenario 1, which appears fairly safe overall (one DLT in seven patients, no DLT in indications 2 and 3), one DLT in three patients in indication 1 does not allow escalation under stratification. For EXNEX and EX, however, escalation is feasible because of borrowing across indications. On the other hand, for scenario 3, which suggests similar toxicity for all indications, recommendations go in the other direction: 1/3 allows to stay at the same dose under stratification, while under EXNEX and EX, de-escalation is required. Of note, the latter appears more sensible, because under pooling 3/9 appears too toxic, so de-escalation would be advised.

For scenarios 4–5 and 6–7, with two and three cohorts per indication, similar explanations (borrowing versus no borrowing) apply if stratified and EXNEX or EX recommendations differ. Recommendations for indication 1 in scenario 6 are particularly interesting, with three different outcomes. After no DLTs had been seen in cohorts 1 and 2 (doses 6 and 12), observed DLT rates are quite different for the third cohort (dose 24), with (in total) three out of eight patients experiencing a DLT. Do these data suggest similar or dissimilar DLT rates? Given the sparse data, there is no clear answer, and one would expect a good analysis to avoid extreme conclusions. Even though the two patients in indication 1 had no DLT at dose 24, under EX, this dose is considered too toxic for indication 1 (too much borrowing?), whereas under stratification, escalation would be advised (not enough borrowing?). EXNEX strikes a balance and allows re-dosing at 24.

In addition to data scenarios, operating characteristics for various scenarios were assessed by simulation (Table VII). Scenarios 1–3 and 4–6 assumed homogeneity and heterogeneity, respectively. For the latter, the MTD differed more than three-fold across indications (e.g., MTD at 12, 24, and 40 mg), which represent extreme and rather unlikely situations, because the context of the study suggested similarity across indications to be likely.

The metrics investigated in the simulations were the probabilities of declaring an MTD with DLT rate of 16% or 25%, declaring lower or higher doses, or to stop the trial (all doses too toxic). These metrics were assessed for EXNEX with mixture weights $p_j = (0.9, 0.1)$ and for the stratified design ($p_j = (0, 1)$). For the simulations, we decided to include a dose with DLT rate 16% (at the boundary of the target interval) in order to assess the operating characteristics for cases where two or only one dose lies in the target interval.

In the simulations, upon declaration of an MTD for a specific indication, further enrollment to this indication was stopped, and the trial was stopped after an MTD was declared for each indication. Other details of the simulations were as follows:

- number of simulated trials per scenario: 2000;
- cohort size per indication: 2 (increased to 3 upon observation of at least one DLT in this indication);
- next dose selection: highest dose fulfilling EWOC and maximal 100% increase from current dose;
- maximum sample size per indication: 40;
- MTD declaration by indication: next dose selected identical to the current dose; at least six patients at this dose enrolled; at least 12 (EXNEX) and 21 (stratified) patients enrolled or $P(\pi_{jd} \in [0.16, 0.33)|\text{data}) \geqslant 0.5$.

For all scenarios, probabilities to declare an MTD with a DLT rate of 16% or 25% are generally high for the EXNEX and stratified design. Not surprisingly, EXNEX performs better than stratification for scenarios 1–3 by taking advantage of homogeneity across indications but can perform worse for some indications for the heterogeneity scenarios 4–6. For these scenarios, the probabilities to declare a too toxic dose (with DLT rate 40% or higher) for EXNEX can be increased. For example, for scenario 5, these risks are 6%, 7%, and 30% for EXNEX, and 15%, 14%, and 10% for stratification. If this scenario had been considered likely and its risks of overdosing unacceptable, a more robust EXNEX model with mixture weights $p_j = (0.5, 0.5)$ could have been used. For these mixture weights, the respective probabilities to select a too toxic dose would have been 12%, 13%, and 16%, which are similar to the ones under stratification. If only one MTD were considered (dose with 16% DLT rate excluded), a less likely scenario given the width of the target interval, the targeting probabilities would be considerably lower. This is because the lower dose with DLT rate 16% is generally favored, which reflects the cautious, safety-centric EWOC approach and has been reported in other simulation studies [39–42].

Finally, the decision to select an EXNEX rather than a stratified design should be made carefully. The former is only advised if homogeneity across some of the indications is judged likely. Otherwise, either EXNEX with greatly increased mixture weights for nonexchangeability or a stratified design is recommended.

4. DISCUSSION

Since the seminal work on multi-parameter problems by James and Stein [7,8], it is well known that parameter estimates can often be improved by shrinkage (borrowing). While 'shrinkage wins' under fairly general conditions, shrinkage estimators are particularly attractive for multi-strata problems, where strata parameters are often similar.

Yet, simplistic and extreme approaches (complete pooling or stratification) prevail in clinical trials. While they play a role in confirmatory trials, they are less useful in earlier stages of drug development. Too much pooling across strata bears the danger of overlooking interesting strata, whereas full stratification usually comes with considerable uncertainty due to small amounts of data. Hence, for early phase trials, which increasingly investigate small and potentially non-homogeneous strata, the need for approaches that allow for tailored borrowing across strata is obvious.

Here, we have discussed EXNEX models, which allow for a small number of partial exchangeability structures as well as nonexchangeability. The weakly informative NEX mixture component makes the prior for each stratum parameter heavy tailed, and therefore robust [43,44]; for similar proposals to robustify historical data priors using different exchangeability structures, see [45,46].

For the example of Section 2 and the two applications of Section 3, EXNEX results look promising. Data scenarios show that, even for sparse data, the degree of borrowing adapts well to the different data constellations, and frequentist operating characteristics look good in comparison with alternative approaches.

Limitations of the proposed and other borrowing methods should be kept in mind. First, operating characteristics, if mainly reduced to type-I error control, are ambiguous [18,47]. Due to the many possible parameter constellations and the complexity of the models, exact type-I error control is not feasible. We think that for early phase trials, this concern is minor and clearly outweighed by the improved quality of shrinkage estimators. Second, there are technical considerations. While for binary data the EXNEX analysis is straightforward with `WinBUGS` (on-line

Appendix), implementation can be harder and requires specialized expertise for more complex data and model structures. In addition, because the analyses are MCMC-based, trial simulations for operating characteristics can be time-consuming. Third, and most importantly, a good understanding of the context is needed. While we think this applies to any inferential approach, for EXNEX it requires carefully selected mixture weights and prior distributions. Furthermore, if relevant *a priori* information exists for a specific stratum, this should be represented in its respective nonexchangeability distribution and mixture weight.

We have confined the discussion of the EXNEX approach to simple data structures. Examples for extensions include the following: other data types (e.g., time-to-event or longitudinal data); inclusion of covariates allowing for potentially relevant predictors as in [22]; random rather than fixed mixture weights (Section 2.3); and exchangeability in more than one dimension, like in the *BATTLE* trial, which used exchangeability assumptions over treatments and biomarker groups [17,18].

Irrespective of the details of the selected approach, designs that aim to borrow information across strata require careful considerations with regard to statistical properties, which include investigations of data scenarios and frequentist operating characteristics. Even though our experience with EXNEX designs in phases I and II is limited so far, we find the approach useful and see potential for applications in other areas, which will help to further improve and extend the methodology.

Acknowledgements

The authors like to thank all Novartis personnel involved in the trials of Section 3. For critical reviews of the manuscript, our special thanks go to Peter Müller (University of Texas Austin), Soumi Lahiri, Heinz Schmidli, Matt Whiley (all Novartis), and two anonymous reviewers.

REFERENCES

[1] Berry DA. Subgroup analyses. *Biometrics* 1990; **46**:1227–1230.
[2] Grouin JM, Coste M, Lewis J. Subgroup analyses in randomized clinical trials: statistical and regulatory issues. *Journal of Biopharmaceutical Statistics* 2005; **15**:869–882.
[3] Freidlin B, Korn EL. Borrowing information across subgroups in phase II trials: is it useful? *Clinical Cancer Research* 2013; **19**(6): 1326–1334.
[4] Efron B, Morris C. Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association* 1975; **70**:311–319.
[5] Louis TA. Estimating a population of parameter values using Bayes and empirical Bayes methods. *Journal of the Amerian Statistical Association* 1984; **79**:393–398.
[6] Carlin B, Louis TA. *Bayes and Empirical Bayes Methods for Data Analysis*. Taylor and Francis, Chapman & Hall/CRC: Boca Raton, FL, 2000.
[7] Stein C. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, Berkeley and Los Angeles, 1956; pp. 197-206.
[8] James W, Stein C. Estimation with quadratic loss. In *Breakthroughs in Statistics*, Johnson NL, Kotz S (eds), Vol. I. Springer: New York; pp. 443–460.
[9] Davis CE, Leffingwell DP. Empirical Bayes estimates of subgroup effects in clinical trials. *Controlled Clinical Trials* 1990; **11**:37–42.
[10] Dixon DO, Simon R. Bayesian subset analysis. *Biometrics* 1991; **47**:871–881.
[11] Dixon DO, Simon R. Bayesian subset analysis in a colorectal cancer clinical trial. *Statistics in Medicine* 1992; **11**:13–22.
[12] Simon R. Bayesian subset analysis: application to studying treatment-by-gender interactions. *Statistics in Medicine* 2002; **21**: 2909–2916.
[13] Jones HE, Ohlssen DI, Neuenschwander B, Racine A, Branson M. Bayesian models for subgroup analysis in clinical trials. *Clinical Trials* 2011; **8**:129–143.
[14] Berry S, Broglio KR, Susan Groshen S, Berry DA. Bayesian hierarchical modeling of patient subpopulations: efficient designs of phase II oncology clinical trials. *Clinical Trials* 2013; **10**:720–734.
[15] Thall PF, Wathen JK, Bekele BN, Champlin RE, Baker LH, Benjamin RS. Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes. *Statistics in Medicine* 2003; **22**:763–780.
[16] Chugh R, Wathen K, Maki RG, Benjamin RS, Patel RS, Myers PA, Priebat DA, Reinke DK, Thomas DG, Keohan ML, Samuels BL, Baker LH. Phase II multicenter trial of imatinib in 10 histologic subtypes of sarcoma using a Bayesian hierarchical statistical model. *Journal of Clinical Oncology* 2009; **27**(19):3148–3153.
[17] Zhou X, Liu S, Kim ES, Herbst RS, Lee JJ. Bayesian adaptive design for targeted therapy development in lung cancer – a step toward personalized medicine. *Clinical Trials* 2008; **5**:181–193.
[18] Berry SM, Carlin BP, Lee JJ, Müller P. *Bayesian Adaptive Methods for Clinical Trials*. Chapman and Hall: Boca Raton, 2010.
[19] I-SPY 2 trial. Available at: http:/ispy.org (accessed 1.12.2015).
[20] Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 2000; **10**:325–337.
[21] Lunn DJ, Jackson C, Best N, Thomas A, Spiegelhalter DJ. *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Chapman & Hall/CRC: Boca Raton, FL, 2012.
[22] Leon-Novelo LG, Bekele BN, Müller P, Quintana F, Wathen K. Borrowing strength with nonexchangeable priors over subpopulations. *Biometrics* 2012; **68**:550–558.
[23] Müller P, Mitra R. Bayesian nonparametric inference: why and how. *Bayesian Analysis* 2013; **8**(2):269–302.
[24] Demetri GD, von Mehren M, Blanke CD, Van den Abbeele AD, Eisenberg B, Roberts PJ, Heinrich MC, Tuveson DA, Singer S, Janicek M, Fletcher JA, Silverman SG, Silberman SL, Capdeville R, Kiese B, Peng B, Dimitrijevic S, Druker B, Corless C, Fletcher CDM, Joensuu H. Efficacy and safety of imatinib mesylate in advanced gastrointestinal stromal tumours. *New England Journal of Medicine* 2002; **347**:472–480.
[25] Verweij J, Casali PG, Zalcberg J, LeCesne A, Reichardt P, Blay JY, Issels R, van Oosterom A, Hogendoorn PC, Van Glabbeke M, Bertulli R, Judson I. Progression free-survival in gastrointestinal stromal tumours with high dose of imatinib: randomizsed trial. *Lancet* 2004; **364**:1127–1134.
[26] DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. *Contemporary Clinical Trials* 2007; **28**:105–114.
[27] Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley: New York, 2004.
[28] Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 2006; **1**(3):515–533.
[29] Polson NG, Scott JG. On the Half-Cauchy prior for a global scale parameter. *Bayesian Analysis* 2012; **7**:887–902.
[30] O'Hagan A, Forster J. *Bayesian Inference*, Kendall's Advanced Theory of Statistics, Volume 2B. Wiley: Chichester, 2004.
[31] Spiegelhalter D, Best N, Carlin B, van der Linde A. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society Series B* 2002; **64**:1–34.
[32] Babb J, Rogatko A, Zacks S. Cancer phase I clinical trials: efficient dose escalation with overdose control. *Statistics in Medicine* 1998; **17**:1103–20.
[33] Neuenschwander B, Branson M, Gsponer G. Critical aspects of the Bayesian approach to phase I cancer trials. *Statistics in Medicine* 2008; **27**:2420–39.
[34] Bailey S, Neuenschwander B, Laird G, Branson M. A Bayesian case-study in oncology phase I combination dose-finding using logistic regression with covariates. *Journal of Biopharmaceutical Statistics* 2009; **19**:469–484.
[35] Neuenschwander B, Matano A, Tang Z, Roychoudhury S, Wandel S, Bailey S. A Bayesian industry approach to phase I combination trials in oncology. In *Statistical Methods in Drug Combination Studies*, Zhao W, Yang H (eds). Chapman & Hall/CRC Press: Boca Raton, FL; 2015.

[36] Casella G, Berger RL. Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association* 1987; **82**:106–111.
[37] Berger J, Sellke T. Testing a point null hypothesis: the irreconcilability of P-values and evidence. *Journal of the American Statistical Association* 1987; **82**:112–139.
[38] Storer B. Design and analysis of phase I clinical trials. *Biometrics* 1989; **45**:925–937.
[39] Chu PL, Lin Y, Shih WJ. Unifying CRM and EWOC designs for phase I cancer clinical trials. *Journal of Statistical Planning and Inference* 2009; **139**:1146–1163.
[40] Maughen A, Le Deley MC, Zohar S. Dose-finding approach for dose escalation with overdose control considering incomplete observations. *Statistics in Medicine* 2011; **30**:1584–1594.
[41] Azriel D. Optimal sequential designs in phase I studies. *Computational Statistics and Data Analysis* 2014; **71**:288–297.
[42] Huang B, Kuan PF. Time-to-event continual reassessment methods incorporating treatment cycle information with application to an oncology phase I trial. *Biometrical Journal* 2014; **6**:933–946.
[43] O'Hagan A. On outlier rejection phenomena in Bayes inference. *Journal of the Royal Statistical Society Series B* 1979; **41**:358–367.
[44] O'Hagan A, Pericchi L. Bayesian heavy-tailed models and conflict resolution: a review. *Brazilian Journal of Probability and Statistics* 2012; **26**:372–401.
[45] Hobbs BP, Sargent DJ, Carlin BP. Commensurate priors for incorporating historical information in clinical trials using generaland generalized linear models. *Bayesian Analysis* 2012; **7**(3):639–674.
[46] Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter DJ, Neuenschwander B. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* 2014; **70**:1023–1032.
[47] Viele K. Hierarchical borrowing in personalized medicine. *Berry Consultants* 2012, Tessella Webinar Available at: https://www.youtube.com/watch?v=H7C1lPvybOk (accessed 1.12.2015).

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web-site.