The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1583–1604
https://doi.org/10.1214/17-AOAS1122

BAYESIAN AGGREGATION OF AVERAGE DATA: AN APPLICATION IN DRUG DEVELOPMENT

BY SEBASTIAN WEBER*, ANDREW GELMAN†,[1], DANIEL LEE‡, MICHAEL BETANCOURT†,[1], AKI VEHTARI§,[2] AND AMY RACINE-POON*

Novartis Pharma AG, Columbia University†, Generable‡ and Aalto University§*

Throughout the different phases of a drug development program, randomized trials are used to establish the tolerability, safety and efficacy of a candidate drug. At each stage one aims to optimize the design of future studies by extrapolation from the available evidence at the time. This includes collected trial data and relevant external data. However, relevant external data are typically available as averages only, for example, from trials on alternative treatments reported in the literature. Here we report on such an example from a drug development for wet age-related macular degeneration. This disease is the leading cause of severe vision loss in the elderly. While current treatment options are efficacious, they are also a substantial burden for the patient. Hence, new treatments are under development which need to be compared against existing treatments.

The general statistical problem this leads to is *meta-analysis*, which addresses the question of how we can combine data sets collected under different conditions. Bayesian methods have long been used to achieve partial pooling. Here we consider the challenge when the model of interest is complex (hierarchical and nonlinear) and one data set is given as raw data while the second data set is given as averages only. In such a situation, common meta-analytic methods can only be applied when the model is sufficiently simple for analytic approaches. When the model is too complex, for example, nonlinear, an analytic approach is not possible. We provide a Bayesian solution by using simulation to approximately reconstruct the likelihood of the external summary and allowing the parameters in the model to vary under the different conditions. We first evaluate our approach using fake data simulations and then report results for the drug development program that motivated this research.

1. Introduction. Modern drug development proceeds in stages to establish the tolerability, safety and efficacy of a candidate drug [Sheiner (1997)]. At each stage and using all relevant information, it is essential to plan the next steps. The

Received July 2017; revised October 2017.

[1]Supported by Institute for Education Sciences R305D140059-16, Office of Naval Research N00014-15-1-2541 & N00014-16-P-2039, Sloan Foundation G-2015-13987, National Science Foundation CNS-1205516, Defense Advanced Research Projects Agency DARPA BAA-16-32.

[2]Supported by Academy of Finland Grant 298742.

Key words and phrases. Meta-analysis, hierarchical modeling, Bayesian computation, pharmacometrics, Stan.

collected raw data are measurements of individual patients over time. Pharmacometric models of such raw data commonly use nonlinear longitudinal differential equations with hierarchical structure (also known as population models), which can, for example, describe the response of patients over time under different treatments. Such models typically come with assumptions of model structure and variance components that offer considerable flexibility and allow for meaningful extrapolation to new trial designs. While these models can be fit to raw data, we often wish to consider additional data which may be available only as averages or aggregates. For example, published summary data of alternative treatments are critical for planning comparative trials. Such external data would allow for indirect comparisons as described in the Cochrane Handbook [Higgins and Green (2011)].

Methods for the mixed case of individual patient data and aggregate data are recognized as important but are limited in their scope so far. For example, in the field of pharmaco-economics, treatments need to be assessed which have never been compared in a head-to-head trial. Methods such as matching-adjusted indirect comparisons (MAIC) [Signorovitch et al. (2010)] and simulated treatment comparisons (STC) [Caro and Ishak (2010), Ishak, Proskorovsky and Benedict (2015)] have been proposed to address the problem of mixed data in this domain. The focus of these methods is a retrospective comparison of treatments while we seek a prospective comparison under varying designs. That is, in the MAIC approach the individual patient data is matched to the reported aggregate data using baseline covariates. While simple in its application, its utility is limited for a prospective planning of new trials which vary in design properties. The STC approach offers additional flexibility as it is based on the simulation of an index trial to which other trials are matched using predictive equations. However, the approach requires calibration for which individual patient data is recommended. Hence, the effort of an STC approach is considerable, and its flexibility is still limited, since the simulated quantities are densities of the endpoints. In contrast, longitudinal nonlinear hierarchical pharmacometric models have the ability to simulate the individual patient response over time and, hence, give the greatest flexibility for prospective clinical trial simulation, which provides valuable input to strategic decisions for a drug development program.

Here we report on an example of a drug development program to investigate new treatment options for wet age-related macular degeneration (wetAMD); see Ambati and Fowler (2012), Buschini et al. (2011), Khandhadia et al. (2012), Kinnunen et al. (2012). This disease is the leading cause of severe vision loss in the elderly [Augood et al. (2006)]. Available drugs include anti-vascular endothelial growth factor (anti-VEGF) agents, which are repeatedly administered as direct injections into the vitreous of the eye. The first anti-VEGF agent was Ranibizumab [Brown et al. (2006), Rosenfeld et al. (2006)], with another, Aflibercept [Heier et al. (2012)], introduced several years later. Initially, anti-VEGF intravitreal injections were given monthly, and more flexible schemes with longer breaks between dosings evolved over recent years to reduce the burden for patients and their

caregivers. In addition, a reduced dosing frequency also increases compliance to treatment, which ensures sustained long-term efficacy.

A key requirement for any new anti-VEGF agent is an optimized dosing scheme to compare favorably to existing treatment options. For a prospective evaluation of new trials, we simulate clinical trials using nonlinear hierarchical pharmacometric models in which a new anti-VEGF agent is compared to available treatments with various design options. Important design options include the patient population characteristics and the dosing regimen, which specifies what dose amount is to be administered at which timepoints to a given patient.

In clinical studies, visual acuity is assessed by the number of letters a patient can read from an Early Treatment Diabetic Retinopathy Study (ETDRS) chart, expressed as best corrected visual acuity (BCVA) score, where the patient is allowed to use glasses for the assessment. A nonlinear pharmacometric drug-disease model is able to longitudinally regress the efficacy response as a function of the patients' characteristics and individual dosing history. This flexibility reduces confounding (through covariates and accounting for noncompliance) during inference and enables realistic extrapolation to future designs with alternative dosing regimens. However, these models do require certain raw data that are commonly not reported in the literature. In our example, raw patient data from Ranibizumab trials were available to us, but we only had aggregate data available for Aflibercept. This creates the awkward situation that the reported aggregate data on Aflibercept cannot be used to obtain accurate model predictions despite our understanding that the nonlinear model is appropriate for the same patient population and we are moreover only interested in population predictions, that is, the interest lies in population parameters and not in patient specific parameters. The problem is that the likelihood function for the aggregated data in general has no closed form expression. The standard expectation–maximization or Bayesian approach in this case is to consider the unavailable individual data points as missing data, but this can be computationally prohibitive as it will vastly increase the dimensionality of the problem space in an experiment with hundreds of patients and multiple measurements per patient.

This paper describes how we enabled accurate clinical trial simulations to inform the design of future studies in wetAMD, which aim at improving the dosing regimens of anti-VEGF agents. This led us to develop a novel statistical computational approach for integrating averaged data from an external source into a linear or nonlinear hierarchical Bayesian analysis. The key point is that we use an approximate likelihood of the external average data instead of using an approximate prior derived from the external data. Doing so enables coherent joint Bayesian inference of raw and summary data. The approach takes account of possible differences in the model in the two data sets.

In Section 2, we describe the data and model for our study, and Section 3 lays out our novel approach for including aggregate data into the pharmacometric model. Section 4 demonstrates our approach using simulation studies of a linear and a nonlinear example. In the linear example we compare our approach to

an exact analytic reference, the nonlinear case is constructed to be similar in its properties to the actual pharmacometric model. We present results for our main problem in Section 5 and conclude with a discussion in Section 6. Source code of R and Stan programs of simulation studies and drug-disease model can be found in the Supplementary Material [Weber et al. (2018)].

2. Data and pharmacometric model.

2.1. *Study data*. We included in the analysis data set the raw data from the studies MARINA, ANCHOR and EXCITE [Rosenfeld et al. (2006), Brown et al. (2006), Schmidt-Erfurth et al. (2011)]. In MARINA and ANCHOR, a monthly (q4w) treatment with Ranibizumab was compared to placebo and active control, respectively. In MARINA, a high and a low dose regimen treatment arm with Ranibizumab were included in the trial. The EXCITE study tested the feasibility of an alternative dosing regimen with longer, three months (q12w), treatment intervals after an initial three month loading phase of monthly treatments with Ranibizumab. We restricted our analysis to the efficacy data only for up to one year which is the follow-up time for the primary endpoints of these studies. We consider the reported BCVA measure of the number of letters read from the ETDRS chart which contains 0–100 letters.

For Aflibercept no raw data from patients are available in the public domain; only literature data of reported mean responses are available from the VIEW1 and VIEW2 studies [Heier et al. (2012)]. These studies assessed noninferiority of a low/high dose q4w and an eight week (q8w) dosing regimen with Aflibercept in comparison to 0.5 mg q4w Ranibizumab treatment, which was also included in these studies as reference arm. Figure 1 shows the reported mean BCVA data of VIEW1+2. In Table 1, we list the baseline characteristics for all the included study arms in the analysis.

2.2. *Pharmacometric model*. We use a drug-disease model, which is informed on the basis of raw measurements of individual patients over time. Such a model [Weber et al. (2014)] was developed on the available raw data for Ranibizumab using the studies MARINA, ANCHOR and EXCITE. The visual acuity measure (BCVA) is limited to the range of 0–100 (letters read from the ETDRS chart), so, we modeled it on a logit transformed scale, $R_j(t) = \text{logit}(y_{jk}/100)$, where y_{jk} is the measurement for patient j at time $t = x_k$. The drug-disease model used was derived from the semimechanistic turnover model [Jusko and Ko (1994)], which links a drug concentration, $C_j(t)$, with a pharmacodynamic response, $R_j(t)$. The drug concentration, $C_j(t)$, is determined by the dose amount and dosing frequency as defined by the regimen. In our case the drug concentration, $C_j(t)$, is latent, since no measurements of $C_j(t)$ in the eye of a patient is possible for ethical and practical reasons. Therefore, we used a simple mono exponential elimination model and fixed the vitreous volume to 4mL [Hart (1992)] and the elimination half-life $t_{1/2}$

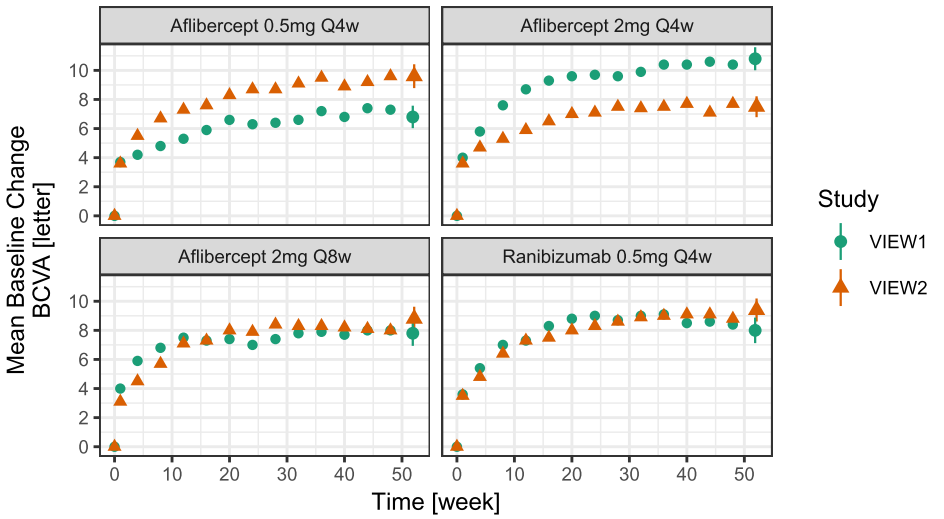

FIG. 1. *Published average data of the VIEW1+2 studies* [*Heier et al.* (2012)]. *Shown is the reported mean baseline change best-corrected visual acuity* (*BCVA*) *over a time period of one year. The vertical line at the last time point marks one standard error of the reported mean.*

TABLE 1

Baseline data of trials included in the analysis. The reported baseline BCVA and age are the respective mean values and their standard deviations

Study	**Data**	**Compound**	***N***	**Freq.**	**Dose [mg]**	**BCVA (SD) [letter]**	**Age (SD) [*y*]**
MARINA	patient	Ranibizumab	238	Q4w	0.3	53.1 (12.9)	77.4 (7.6)
MARINA	patient	Ranibizumab	239	Q4w	0.5	53.7 (12.8)	76.8 (7.6)
MARINA	patient	Placebo	236	Q4w	sham	53.9 (13.7)	77.1 (6.6)
ANCHOR	patient	Ranibizumab	137	Q4w	0.3	47.1 (12.8)	77.3 (7.3)
ANCHOR	patient	Ranibizumab	139	Q4w	0.5	47.1 (13.2)	75.9 (8.5)
EXCITE	patient	Ranibizumab	120	Q12w	0.3	55.8 (11.8)	75.1 (7.5)
EXCITE	patient	Ranibizumab	118	Q12w	0.5	57.7 (13.1)	75.8 (7.0)
EXCITE	patient	Ranibizumab	115	Q4w	0.3	56.5 (12.2)	75.0 (8.3)
VIEW1	average	Aflibercept	301	Q4w	0.5	55.6 (13.1)	78.4 (8.1)
VIEW1	average	Aflibercept	304	Q4w	2.0	55.2 (13.2)	77.7 (7.9)
VIEW1	average	Aflibercept	301	Q8w	2.0	55.7 (12.8)	77.9 (8.4)
VIEW1	average	Ranibizumab	304	Q4w	0.5	54.0 (13.4)	78.2 (7.6)
VIEW2	average	Aflibercept	296	Q4w	0.5	51.6 (14.2)	74.6 (8.6)
VIEW2	average	Aflibercept	309	Q4w	2.0	52.8 (13.9)	74.1 (8.5)
VIEW2	average	Aflibercept	306	Q8w	2.0	51.6 (13.9)	73.8 (8.6)
VIEW2	average	Ranibizumab	291	Q4w	0.5	53.8 (13.5)	73.0 (9.0)

from the vitreous to nine days [Xu et al. (2013)]. The standard turnover model assumes that the response $R_j(t)$ can only take positive values, which is not given on the logit transformed scale. A modified turnover model is therefore used, which is defined by the ordinary differential equation (ODE)

$$\frac{dR_j(t)}{dt} = k_j^{\text{in}} - k_j^{\text{out}}\big[R_j(t) - E_{\max j} S_j\big(C_j(t)\big)\big]. \tag{1}$$

The drug effect enters this equation via the function S_j, which is typically chosen to be a Hill function of the concentration $C_j(t)$. The Hill function is a logistic function of the log drug concentration, $\text{logit}^{-1}(\log EC50 - \log C_j(t))$. At baseline, $R_j(t=0) = R_{0j}$ defines the initial condition for the ODE. The model in equation (1) has an important limit whenever a time constant stimulation, $S_j(t) = s_j$, is applied. Then, the ODE system drives $R_j(t)$ towards its stable steady-state, which is derived from equation (1) by setting the left-hand side to 0, $R_j^{\text{ss}} = (k_j^{\text{in}}/k_j^{\text{out}}) + E_{\max j} s_j$. In absence of a drug treatment, no stimulation is present; that is, $S_j(t) = s_j = 0$, hence, the ratio $k_j^{\text{in}}/k_j^{\text{out}}$ is of particular importance, as for placebo patients it holds that $\lim_{t\to\infty} R_j(t) = k_j^{\text{in}}/k_j^{\text{out}}$. The drug-disease model describes treated patients in relation to placebo patients and separates the drug-related parameters ($t_{1/2}$, $E_{\max}$ and $EC50$) from the remaining nondrug-related parameters.

3. Bayesian aggregation of average data.

3.1. *General formulation.* We shall work in a hierarchical Bayesian framework. Suppose we have data $y = (y_{jk}; j = 1, \ldots, J; k = 1, \ldots, T)$ on J individuals at T time points, where each $y_j = (y_{j1}, \ldots, y_{jT})$ is a vector of data with model $p(y_j|\alpha_j, \phi)$. Here, each α_j is a vector of parameters for individual j, and ϕ is a vector of shared parameters and hyperparameters so that the joint prior is $p(\alpha, \phi) = p(\phi)\prod_{j=1}^{J} p(\alpha_j|\phi)$, and the primary goal of the analysis is inference for the parameter vector ϕ.

We assume that we can use an existing computer program such as Stan [Stan Development Team (2017)] to draw simulations from the posterior distribution, $p(\alpha, \phi|y) \propto p(\phi)\prod_{j=1}^{J} p(\alpha_j|\phi)\prod_{j=1}^{J} p(y_j|\alpha_j, \phi)$.

We then want to update our inference using an *external data set*, $y' = (y'_{jk}; j = 1, \ldots, J'; k = 1, \ldots, T')$, on J' individuals at T' time points, assumed to be generated under the model, $p(y'_j|\alpha'_j, \phi')$. There are two complications:

- The external data, y', are modeled using a process with parameters ϕ' that are similar to but not identical to those of the original data. We shall express our model in terms of the difference between the two parameter vectors, $\delta = \phi' - \phi$. We assume the prior distribution factorizes as $p(\phi, \delta) = p(\phi)p(\delta)$.

  We assume that all the differences between the two studies, and the populations which they represent, are captured in δ. One could think of ϕ and ϕ' as two

instances from a population of studies. If we were to combine data from several external trials, it would make sense to include between trial variation using an additional set of hyperparameters in the hierarchical model.

- We do not measure y' directly; instead, we observe the time series of averages, $\bar{y}' = (\bar{y}'_1, \ldots, \bar{y}'_T)$. And, because of nonlinearity in the data model, we cannot simply write the model for the external average data, $p(\bar{y}'|\alpha', \phi')$, in closed form.

This is a problem of meta-analysis, for which there is a longstanding concern when the different pieces of information to be combined come from different sources or are reported in different ways [see, e.g., Higgins and Whitehead (1996), Dominici et al. (1999)].

The two data issues listed above lead to corresponding statistical difficulties:

- If the parameters ϕ' of the external data were completely unrelated to the parameters of interest, ϕ—that is, if we had a noninformative prior distribution on their difference, δ—then there would be no gain from including the external data into the model, assuming the goal is to learn about ϕ.

  Conversely, if the two parameter vectors were identical, so that $\delta \equiv 0$, then we could just pool the two data sets. The difficulty arises because the information is partially shared, to an extent governed by the prior distribution on δ.
- Given that we see only averages of the external data, the conceptually simplest way to proceed would be to consider the individual measurements y'_{jk} as missing data, and to perform Bayesian inference jointly on all unknowns, obtaining draws from the posterior distribution, $p(\phi, \delta, \alpha, \alpha'|y, \bar{y}')$. The difficulty here is computational. Every missing data point adds to the dimensionality of the joint posterior distribution, and the missing data can be poorly identified from the model and the average data. Weak data in a nonlinear model can lead to a poorly regularized posterior distribution that is hard to sample from.

As noted, we resolve the first difficulty using an informative prior distribution on δ. Specifically, we consider in the following that not all components of ϕ, but only a few components, differ between the data sets, such that the dimensionality of δ may be smaller than that of ϕ. This imposes that some components of δ are exactly 0.

We resolve the second difficulty via a normal approximation and take advantage of the fact that our observed data summaries are averages. That is, as we cannot construct the patient specific likelihood contribution for the external data set, $\prod_{j=1}^{J'} p(y'_j|\alpha'_j, \phi')$, directly, instead we approximate this term by a multivariate normal, $\mathrm{N}(\bar{y}'|\tilde{M}_s, \frac{1}{J'}\tilde{\Sigma}_s)$ to be introduced below.

3.2. *Inclusion of summary data into the likelihood.* Our basic idea is to approximate the probability model for the external average data, $p(\bar{y}'|\phi')$, by a multivariate normal with parameters depending on $\bar{y}'$. For a linear model this is the

analytically exact representation of the average data in the likelihood. For nonlinear models the approximation is justified by the central limit theorem if the summary is an average over many data points. This corresponds in essence to a Laplace approximation to the marginalization integral over the unobserved (latent) individuals in the external data set y' as $p(\bar{y}'|\phi') = \int p(\bar{y}'|\alpha', \phi')\, d\alpha'$.

The existing model on y is augmented by including a suitably chosen prior on the parameter vector δ and the log-likelihood contribution implied by the external average data $\bar{y}'$. As such, the marginalization integral must be evaluated in each iteration s of the MCMC run. Evaluating the Laplace approximation requires the mode and the Hessian at the mode of the integrand. Both are unavailable in commonly used MCMC software, including Stan. To overcome these computational issues, we instead use simulated plug-in estimates. In each iteration s of the MCMC run we calculate the Laplace approximation of the marginalization integral as follows:

1. Compute $\phi'_s = \phi_s + \delta_s$.

2. Simulate parameters $\tilde{\alpha}_j$ and then data $\tilde{y}_{jk}$, $j = 1, \ldots, \tilde{J}$, $k = 1, \ldots, T'$, for some number $\tilde{J}$ of hypothetical new individuals, drawn from the distribution $p(y'|\phi'_s)$ and corresponding to the conditions under which the external data were collected (hence, the use of the same number of time points T'). The $\tilde{J}$ individuals do *not* correspond to the J' individuals in the external data set; rather, we simulate them only for the purpose of approximating the likelihood of the external average data, $\bar{y}'$, under these conditions. The choice of $\tilde{J}$ must be sufficiently large, as is discussed below.

3. Compute the mean vector and the $T' \times T'$ covariance matrix of the simulated data $\tilde{y}$. Call these $\tilde{M}_s$ and $\tilde{\Sigma}_s$.

4. Divide the covariance matrix $\tilde{\Sigma}_s$ by J' to get the simulation estimated covariance matrix for $\bar{y}'$, which is an average over J' individuals whose data are modeled as independent conditional on the parameter vector ϕ'.

5. Approximate the marginalization integral over the individuals in the external y' data set with the probability density of the observed mean vector of the T' external data points using the multivariate normal distribution with mean $\tilde{M}_s$ and covariance matrix $\frac{1}{J'}\tilde{\Sigma}_s$, which are the plug-in estimates for the mode and the Hessian at the mode of the Laplace approximation. The density $\mathrm{N}(\bar{y}'|\tilde{M}_s, \frac{1}{J'}\tilde{\Sigma}_s)$ then represents the information from the external mean data.

3.3. *Computational issues—Tuning and convergence.* For the simulation of the $\tilde{J}$ hypothetical new individuals we do need random numbers which are independent of the model. However, as Bayesian inference results in a joint probability density, we cannot simply declare an extra set of parameters in our model during an MCMC run. That is, we can only control for the prior density of these extra parameters but not so for the posterior density, which is generated by the sampler. This is an issue, as by construction of Hamiltonian Monte Carlo (HMC), as

used in Stan, no random numbers can be drawn independently from the model during sampling. However, our algorithm does not require that the random numbers change from iteration to iteration. Hence, we can simply draw a sufficient amount of random numbers per chain and include these as data for a given chain. As consequence, different chains may converge to different distributions due to different initial sets of random numbers. However, with increasing simulation size $\tilde{J}$, the simulations have a decreasing variability in their estimates, as the standard error scales with $\tilde{J}^{-1/2}$. Therefore, the tuning parameter $\tilde{J}$ must be chosen sufficiently large to ensure convergence of all chains to the same result. This occurs once the standard error is decreased below the overall MC error. Whenever $\tilde{J}$ is chosen too small, standard diagnostics like $\widehat{R}$ [Gelman et al. (2014)] will indicate nonconvergence. We assess this by running each odd chain with $\tilde{J}$ and each even chain with $2\tilde{J}$ hypothetical new individuals (typically we run four parallel MCMC chains as this is free on a four processor laptop or desktop computer). The calculation of $\widehat{R}$ then considers chains with different $\tilde{J}$, and, so, a too low $\tilde{J}$ will immediately be detected, in which case the user can increase $\tilde{J}$.

For models with a Gaussian residual error model, Step 2 above can be simplified. Instead of simulating observed fake data $\tilde{y}$, it suffices to simulate the averages of the hypothetical new individuals $\tilde{J}$ at the T' time points. The residual error term can be added to the variance–covariance matrix $\tilde{\Sigma}_s$ as diagonal matrix. Should the sampling model not be normal, then normal approximations should be considered to use. The benefit is a much reduced simulation cost in each iteration of the MCMC run.

4. Simulation studies.

4.1. *Hierarchical linear regression*. We begin with a fake data hierarchical linear regression example, which is simple enough that we can compare our approximate inferences to a closed form analytic solution to the problem as the unobserved raw data can be marginalized over in a full analytic approach. We set up this example to correspond in its properties to the longitudinal nonlinear drug-disease model.

We consider a linear regression using a continuous covariate x (corresponding to time) with an intercept, a linear, and a quadratic slope term. The intercept and linear slope term vary in two ways which is by individual and data set. The quadratic term does not vary by individual or data set. This allows us to check two aspects: (a) if we can learn differences between data sets (intercept and slope) and (b) if the precision on fully shared parameters (quadratic term) increases when combining data sets. That is, for the main data set y, the model is $y_{jk} \sim \mathrm{N}(\alpha_{j1} + \alpha_{j2}x_k + \beta x_k^2, \sigma_y^2)$, with prior distribution $\alpha_j \sim \mathrm{N}(\mu_\alpha, \Sigma_\alpha)$ for which we set the correlations $\rho_{\alpha_{j1}\alpha_{j2}}$ (the off-diagonal elements of Σ_α) to 0. Using the notation from Section 3.1, the vector of shared parameters ϕ is $\phi =$

$(\mu_{\alpha 1}, \mu_{\alpha 2}, \beta, \sigma_{\alpha 1}, \sigma_{\alpha 2}, \sigma_y)$. We assume that the number of individuals J is large enough such that we can assign a noninformative prior to ϕ.

For the external data set y', the model is $y'_{jk} \sim \mathrm{N}(\alpha'_{j1} + \alpha'_{j2} x_k + \beta x_k^2, \sigma_y^2)$, with the prior distribution $\alpha'_j \sim \mathrm{N}(\mu'_\alpha, \Sigma_\alpha)$. In this simple example, we assign a noninformative prior distribution to $\delta = \mu'_\alpha - \mu_\alpha$ while we assign a δ of exactly 0 to all other components in ϕ such that $\phi' = (\mu_{\alpha 1} + \delta_1, \mu_{\alpha 2} + \delta_2, \beta, \sigma_{\alpha 1}, \sigma_{\alpha 2}, \sigma_y)$.

Assumed parameter values. We create simulations assuming the following conditions, which we set to roughly correspond to the features of the drug-disease model:

- $J = 100$ individuals in the original data set, each measured $T = 13$ times (corresponding to measurements once per month for a year), $x_k = 0, \frac{1}{12}, \ldots, 1$.
- $J' = 100$ individuals in the external data set, also measured at these 13 time points.
- $(\mu_{\alpha 1}, \sigma_{\alpha 1}) = (0.5, 0.1)$, corresponding to intercepts that are mostly between 0.4 and 0.6. The data from our actual experiment roughly fell on a 100-point scale, which we are rescaling to 0–1 following the general principle in Bayesian analysis to put data and parameters on a unit scale [Gelman (2004)].
- $(\mu_{\alpha 2}, \sigma_{\alpha 2}) = (-0.2, 0.1)$, corresponding to an expected loss of between 10 and 30 points on the 100-point scale for most people during the year of the trial.
- $\rho_{\alpha_{j1}\alpha_{j2}} = 0$, no correlation assumed between individual slopes and intercepts.
- $\beta = -0.1$, corresponding to an accelerating decline representing an additional drop of 10 points over the one-year period.
- $\sigma_y = 0.05$, indicating a measurement and modeling error on any observation of about five points on the original scale of the data.

Finally, we set δ to $(0.1, 0.1)$, which represents a large difference between the two data set in the context of this problem and allows us to test how well the method works when the shift in parameters needs to be discovered from data.

In our inferences, we assign independent unit normal priors for all the parameters $\mu_{\alpha 1}$, $\mu_{\alpha 2}$, β, δ_1, and δ_2; and independent half unit normal priors to the variance components $\sigma_{\alpha 1}$, $\sigma_{\alpha 2}$, and σ_y. Given the scale of the problem (so that parameters should typically be less than one in absolute value, although this is not a hard constraint), the unit normals represent weak prior information which just serves to keep the inferences within generally reasonable bounds.

Conditions of the simulations. We run four chains using the default sampler in Stan, the HMC variant No-U-Turn Sampler (NUTS) [Hoffman and Gelman (2014), Betancourt (2016)], and set $\tilde{J}$ to 500, so that every odd chain will simulate 500 and every even 1000 hypothetical individuals, thus allowing us to easily check if the number of internal simulations is enough for stable inference. If there were notable differences between the inferences from even and odd chains, this would suggest that $\tilde{J} = 500$ is not enough and should be increased.

Computation and results. We simulate data y and y'. For simplicity we do our computations just once in order to focus on our method only. If we wanted to evaluate the statistical properties of the examples, we could nest all this in a larger simulation study.

We first evaluate the simulation based approximation of the log-likelihood contribution of the mean data $\bar{y}'$. This is shown in the top panel of Figure 2. The plot shows $\log p(\bar{y}'|\phi')$ evaluated at the true value of ϕ' for varying values of δ_2. The gray band marks the 80% confidence interval of 10^3 replicates when simulating per replicate a randomly chosen set of $\tilde{J} = 10^2$ patients. The dotted blue line is the median of these simulations and the black solid line is the analytically computed expression for $\log p(\bar{y}'|\phi')$, which we can compute for this simple model directly. Both lines match respectively, which suggests that the simulation approximation is consistent with the analytical result. The width of the gray band is determined by the number of hypothetical fake patients $\tilde{J}$. The inset plot shows at a fixed value of $\delta_2 = 0$ the width of the 80% confidence interval as a function of $\tilde{J}$ in a log-log plot. The solid black line marks the simulation results while the dashed line has a fixed slope of $-1/2$ and a least-squares estimated intercept. As both lines match each other, we can conclude that the scaling of the confidence interval width is consistent with $\propto \tilde{J}^{-1/2}$.

We run the algorithm as described below and reach approximate convergence in that the diagnostic $\widehat{R}$ is near 1 for all the parameters in the model. We then compare the inferences for the different scenarios:

local: The posterior estimates for the shared parameters ϕ using just the model fit to the local data y.

full: The estimates for all the parameters ϕ, δ using the complete data y, y', which would not in general be available—from the statement of the problem we see only the averages for the new data y'—but we can do so here as we have simulated data.

approximate: The estimates when using the approximation scheme for all the parameters ϕ, δ using the actual available data y, $\bar{y}'$.

integrated: The estimates when using an analytical likelihood for all of the parameters ϕ, δ using the actual available data y, $\bar{y}'$. In general, it would not be possible to compute this inference directly, as we need the probability density for the averaged data, but in this linear model this distribution has a closed-form expression which we can calculate.

The bottom panel of Figure 2 shows the results of the parameter estimates as bias. We are using informative priors and so we neither desire nor expect expect a bias of exactly 0. Rather we would like to see for each parameter a match of the approximate estimate (*blue line with a square*) with the estimate of the full scenario (*orange line with a triangle*), which corresponds to the correct Bayes estimate. However, we cannot expect that the full scenario matches the approximate estimate, since the correct Bayes estimate for the full scenario is given by $p(\phi, \delta|y, y')$,

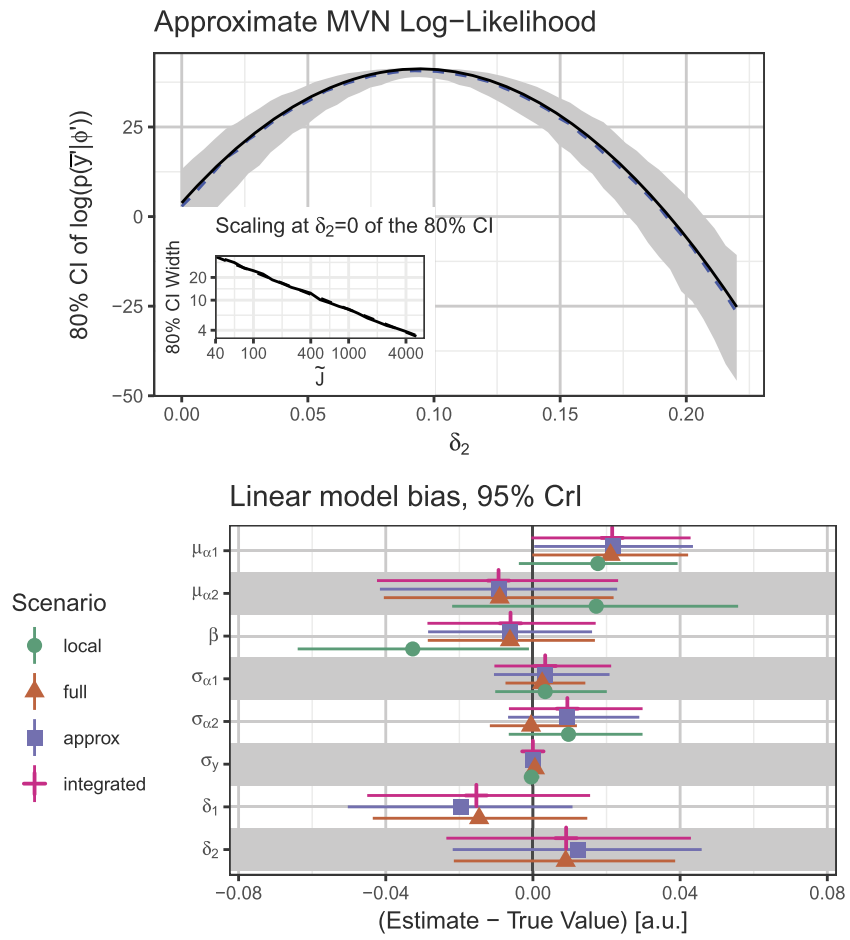

FIG. 2. *Hierarchical linear model example.* (*Top*) *Comparison of the analytical expression for* $\log p(\bar{y}'|\phi')$, *shown as a solid black line, to the simulation based multivariate normal approximation* $\mathrm{N}(\bar{y}'|\tilde{M}_s, \frac{1}{J}\tilde{\Sigma}_s)$. *The simulation includes* $\tilde{J} = 10^2$ *hypothetical individuals, and* 10^3 *replicates were performed to assess its distribution. The gray area marks the* 80% *confidence interval and the dotted blue line is the median of the simulations. The inset shows the width of the* 80% *confidence interval at* $\delta_2 = 0$ *as a function of the simulation size* $\tilde{J}$ *on a* log-log *scale. The dotted line has a fixed slope of* $-1/2$ *and the intercept was estimated using least squares.* (*Bottom*) *The model estimates are shown as bias for the four different scenarios as discussed in the text. Lines show the* 95% *credible intervals of the bias and the center point marks the median bias. The MCMC standard error of the mean is for all quantities below* 10^{-3}.

which is based on the individual raw data y and y' instead of y and mean data $\bar{y}'$. The appropriate comparison is with reference to the integrated scenario (*red line with a cross*), which is the correct Bayes estimate of $p(\phi, \delta|y, \bar{y}')$. The integrated and the approximate scenarios do match closely for all parameters.

When comparing the full scenario with the approximate and integrated result, one can observe that the variance components $\sigma_{\alpha 1}$ and $\sigma_{\alpha 2}$ are estimated with higher precision in the full scenario. This is a direct consequence of using the reported means only for the external data.

Including the averaged data $\bar{y}'$ into the model does not inform the variance components $\sigma_{\alpha 1}$ and $\sigma_{\alpha 2}$, but it does increase the precision of all other parameters in ϕ. This can be observed by considering the reduced width of the credible intervals when comparing the local scenario (*green line with a dot*) to the others, in particular for $\mu_{\alpha 2}$ and β. The estimates of δ_1 and δ_2 are similar across all cases—whenever these can be estimated. This suggests that the external averaged data $\bar{y}'$ are just as informative for the δ vector as the individual raw data y' themselves. The main reason as to why the precision of the δ estimate is a little higher for the full scenario is related to the estimates of the variance components $\sigma_{\alpha 1}$ and $\sigma_{\alpha 2}$. These variance components are estimated from the complete individual raw data (y and y') to be smaller in comparison to the other scenarios. As a result the overall weight of each patient to the log-likelihood is larger. This leads to a higher precision of the population parameters which can be observed in particular for the parameters $\mu_{\alpha 1}$ and δ.

4.2. *Hierarchical nonlinear pharmacometric model.* Next, we perform a fake data study that is closely adapted to our application of interest. The function $R_j(t)$ in equation (1) is only implicitly defined; no closed form solution is available for the general case. For the simulation study we consider the special case of constant maximal drug effect at all times; that is, $S_j(t) = s_j = 1$ for a patient j who receives treatment or $S_j(t) = s_j = 0$ for placebo patients otherwise. The advantage of this choice is that the ODE can then be solved analytically as $R_j(t) = R_j^{\mathrm{ss}} + (R_{0j} - R_j^{\mathrm{ss}}) \exp(-k_j^{\mathrm{out}} t)$. In the following we consider three different cohorts of patients (placebo, treatment 1 and treatment 2) observed at times $t = x_k$. Data for treatment 2 will be considered as the external data set and given as average data only to evaluate our approach. Measurements y_{jk} of a patient j are assumed to be i.i.d. normal, $y_{jk}/100 \sim \mathrm{N}(\mathrm{logit}^{-1}(R_j(x_k)), \sigma_y^2)$. We assume that the number of patients is large enough such that weakly informative priors, which identify the scale of the parameters, are sufficient. The above quantities are parametrized and assigned the simulated true values and priors for inference as:

- $J = 100$ patients in the data set with raw measurements per individual patient. The first $j = 1, \ldots, 50$ patients are assigned a placebo treatment ($E_{\max j} = 0$) and the remaining $j = 51, \ldots, 100$ patients are assigned a treatment with nonzero drug effect ($E_{\max j} > 0$). All patients are measured at $T = 13$ time points corresponding to one measurement per month over a year. We rescale time accordingly to $x_k = 0, \frac{1}{12}, \ldots, 1$.
- $J' = 100$ patients in the external data set, measured at the same $T' = 13$ time points.

- $R_{0j} \sim \mathrm{N}(L\alpha_0, \sigma^2_{L\alpha_0})$ is the unobserved baseline value of each patient j on the logit scale which we set to $L\alpha_0 = 0$ corresponding to 50 on the original scale and $\sigma_{L\alpha_0} = 0.2$. We set the weakly informative prior to $L\alpha_0 \sim \mathrm{N}(0, 2^2)$ and $\sigma_{L\alpha_0} \sim \mathrm{N}^+(0, 1^2)$.
- $k_j^{\mathrm{in}}/k_j^{\mathrm{out}} = L\alpha_s$ is the placebo steady state, the asymptotic value patients reach if not on treatment (or treatment is stopped). In the example, lower values of the response correspond to worse outcome. We set the simulated values to $L\alpha_s = \mathrm{logit}(35/100)$ and the prior to $L\alpha_s \sim \mathrm{N}(-1, 2^2)$.
- $\log(1/k_j^{\mathrm{out}}) \sim \mathrm{N}(l\kappa, \sigma^2_{l\kappa})$ determines the patient-specific time scale of the exponential changes (k_j^{out} is a rate of change). We assume that changes in the response happen within 10/52 time units, which led us to set $l\kappa = \log(10/52)$ and we defined as a prior $l\kappa \sim \mathrm{N}(\log(1/4), \log(2)^2)$ and $\sigma_{l\kappa} \sim \mathrm{N}^+(0, 1^2)$.
- $\log(E_{\max j})$ is the drug effect for patient j. If patient j is in the placebo group, then $E_{\max j} = 0$. For patients receiving the treatment 1 drug we assumed $\log(E_{\max j}) = lE_{\max j} = \log(\mathrm{logit}(60/100) - \mathrm{logit}(35/100))$, which represents a gain of 25 points in comparison to placebo. Patients in the external data set y' are assumed to have received the treatment 2 drug and are assigned a different $lE'_{\max}$. We consider $\delta = lE'_{\max} - lE_{\max} = 0.1$, which corresponds to a moderate to large difference [$\exp(0.1) \approx 1.1$]. As priors we use $lE_{\max} \sim \mathrm{N}(\log(0.5), \log(2)^2)$ and $\delta \sim \mathrm{N}(0, 1^2)$.
- $\sigma_y = 0.05$ is the residual measurement error on the original letter scale divided by 100. The prior is assumed to be $\sigma_y \sim \mathrm{N}^+(0, 1^2)$.

All simulation results are shown in Figure 3. In the top panel of Figure 3 an assessment of the sampling distribution of our approximation is shown for a simulation size of $\tilde{J} = 10^2$ hypothetical fake patients and 10^3 replicates. Since for this nonlinear example we cannot integrate out analytically the missing data in the external data set such that there is no black reference line as before. However, we can conclude that the qualitative behavior of a maximum around the simulated true value is like that in the linear case. Moreover, the inset confirms that the scaling of the precision of the approximation with increasing simulation size $\tilde{J}$ of hypothetical fake patients scales as a power law consistent with $\propto \tilde{J}^{-1/2}$.

For the model we run four chains and set $\tilde{J}$ to 500 as before. The model estimates are shown as bias in the bottom panel of Figure 3. The precision of the estimates from the local fit (*green line with a dot*) increases when adding the external data. While population mean parameters gain in precision in the full (*orange line with a triangle*) and approximate (*blue line with a square*) scenarios, the precision of variance component parameters like $\sigma_{L\alpha_0}$ and $\sigma_{l\kappa}$ only increase in the full scenario. This is expected as the mean data $\bar{y}'$ does not convey information on between-subject variation. However, it is remarkable that the population mean parameter estimates for the approximate scenario are almost identical to the full scenario, including the important parameter δ_1.

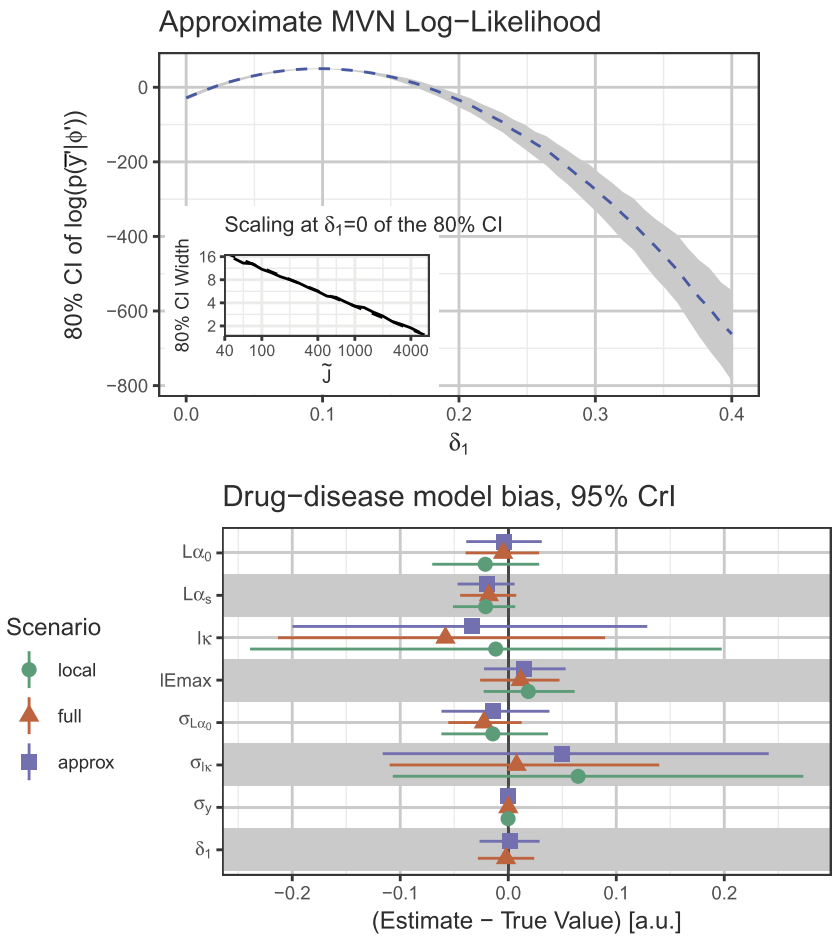

FIG. 3. *Drug-disease model example*: (*Top*) *Assessment of the distribution of the multi-variate normal approximation to* $\log p(\bar{y}'|\phi')$ *at a simulation size of* $\tilde{J} = 10^2$ *hypothetical fake patients using* 10^3 *replicates for varying* δ_1. *The gray area marks the* 80% *confidence interval, the blue dotted line is the median of the simulations. The inset shows the width of the* 80% *confidence interval at* $\delta_1 = 0$ *as a function of the simulation size* $\tilde{J}$ *on a* log-log *scale. The dotted line has a fixed slope of* $-1/2$ *and the intercept was estimated using least squares.* (*Bottom*) *The model estimates are shown as bias for the three different scenarios as discussed in the text. Lines show the* 95% *credible intervals of the bias and the center point marks the median bias. The MCMC standard error of the mean is for all quantities below* 10^{-3}.

We can conclude that possible differences in a drug-related parameter, δ_1, can equally be identified from individual raw data as from the external mean data only. The mean estimate for δ_1 and its 95% credible interval in the full scenario (y, y') and the approximate scenario $(y, \bar{y}')$ do match one another closely.

5. Results for the drug development application. We now turn to the application of our approach for the development of a new drug for wetAMD. For Aflibercept no raw data from patients is available in the public domain; only literature data of reported mean responses are available [Heier et al. (2012)]. Hence, extrapolation for Aflibercept treatments on the basis of the developed drug-disease model was not possible. The drug-related parameters of the drug-disease model are the elimination half-life $t_{1/2}$, the maximal drug effect, $lEmax$ and the concentration at which 50% of the drug effect is reached, $lEC50$ (both parameters are estimated on the log scale). The elimination half-life is fixed with a drug specific value in our model from values reported in the literature for each drug. We can inform the latter two parameters for Ranibizumab from our raw data, which comprise a total of $N = 1342$ patients from the studies MARINA, EXCITE and ANCHOR; the data from the VIEW1+2 studies [Heier et al. (2012), $N = 1210 + 1202$] enables us to estimate these parameters for Aflibercept. Following our approach, we modified the existing model on Ranibizumab to include a δ parameter [with a weakly informative prior of N(0, 1)] for each of the drug-related parameters for patients on Aflibercept treatment. In addition, we also allowed the baseline BCVA of VIEW1+2 to differ as compared to the chosen reference study MARINA. We did not include a δ parameter for any other parameter in the model, since the remaining parameters characterize the natural disease progression in absence of any drug. We consider it reasonable to assume that the natural disease progression does not change under the two conditions, and in any case it is impossible to infer differences in the natural disease progression as compared to our data set with the VIEW1+2 data since no placebo patients were included in either study for ethical reasons.

It is important to note that the VIEW1+2 studies included each a 0.5 mg q4w treatment arm with Ranibizumab. For these arms only the mean data is reported as well, and we include these into our model as a reference—assuming that the drug specific parameters are exactly the same for all data sets.

Figure 1 shows the published mean baseline change BCVA data of the VIEW1+2 studies. From the VIEW1+2 studies we choose to include only the mean BCVA data of the dosing regimens 2 mg q8w Aflibercept and 0.5 mg q4w Ranibizumab into our model, as these are used in clinical practice and are hence of greatest interest to describe these as accurately as possible. The total data set then included raw data from $N = 1342$ patients from MARINA, ANCHOR and EXCITE (different Ranibizumab regimens and a placebo arm) and $N = 1202$ patients from the reported mean data in VIEW1+2 (2 mg q8w Aflibercept and 0.5 mg q4w Ranibizumab). Since our model is formulated on the scale of the nominally observed BCVA measurements, we shifted the reported baseline change BCVA values by the per study mean baseline BCVA value. We used the remaining data from the 2 mg q4w and 0.5 mg q4w Aflibercept regimens for an out-of-sample model qualification.

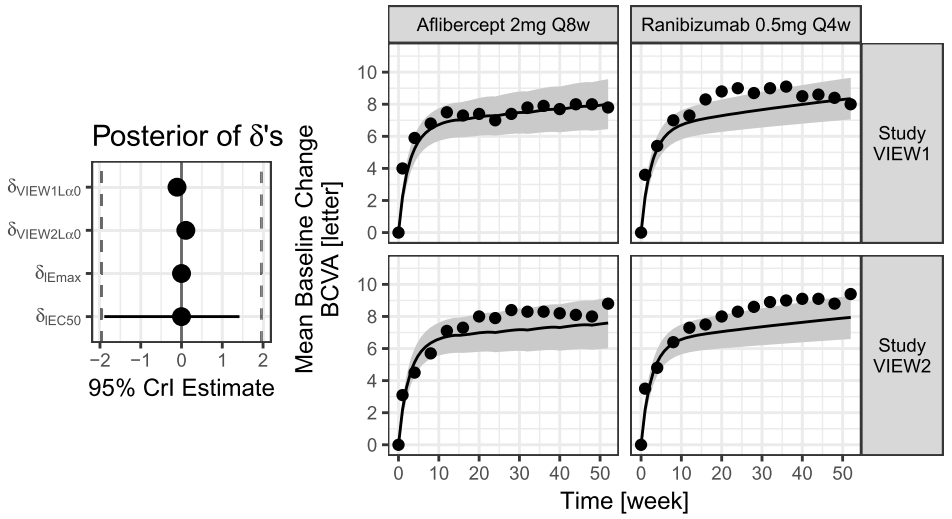

FIG. 4. *Main analysis results*: (*Left*) *Shows the posterior* 95% *credible intervals of the estimated* δ *parameters. The dotted lines mark the* 95% *credible interval of the prior.* (*Right*) *Shows the predicted mean baseline change BCVA as solid line for the study arms included in the model fit. The gray area marks one standard error for the predicted mean, assuming a sample size as reported per arm* (*about* 300 *each, see Table* 1). *The dots mark the reported mean baseline change BCVA and are shown as reference.*

The final result of the fitted model, which uses our internal patient-level data, and the VIEW1+2 summary data of the 2 mg q8w Aflibercept and 0.5 mg q4w Ranibizumab arms, are shown in Figure 4. Presented are the posteriors of the δ parameters (left) and the posterior predictive of the mean baseline change BCVA response of the two included regimens of VIEW1+2 (right).

The posterior predictive distribution of the mean baseline change BCVA is in excellent agreement with the reported data for the 2 mg q8w Aflibercept arms of VIEW1+2. The posterior predictive distribution of the 0.5 mg q4w Ranibizumab mean data in VIEW1+2 suggests a slight underprediction from the model. However, the prediction is for one standard error corresponding to a 68% credible interval, and, hence, the observed data is well in the usual 95% credible interval.

When comparing the posteriors of the δ parameters to their standard normal priors (corresponding to a prior 95% credible interval from -1.96 to $+1.96$), we observe that the information implied by the aggregate data of VIEW1+2 for each parameter varies substantially. While the δ_{lEmax} parameter is estimated with great precision to be close to 0, the precision of the δ_{lEC50} posterior is only increased slightly from a prior standard deviation of 1 to a posterior standard deviation of 0.8. This is a consequence of the dosing regimens in VIEW1+2, which keep patients at drug concentrations well above the $lEC50$ in order to ensure maximal drug effect at all times. In fact, the only trial in our Ranibizumab database where concentrations vary around the range of the $lEC50$ is the EXCITE study. This study

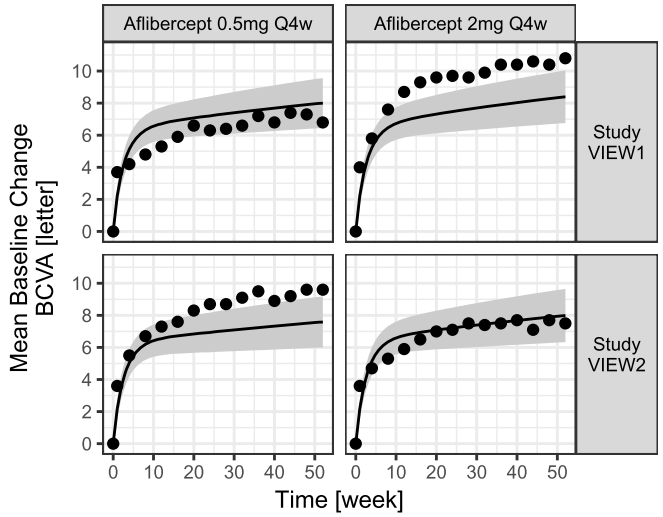

FIG. 5. *Out-of-sample model qualification*: *Shown is the predicted mean baseline change BCVA as solid line for the study arms of VIEW*1+2 *which were not included in the model fitting. The gray area marks one standard error for the predicted mean assuming a sample size as reported per arm* (*about* 300 *each*, *see Table* 1). *The dots mark the reported mean baseline change BCVA and are shown as reference*.

included two q12w Ranibizumab arms which showed a decrease of the BCVA after the loading phase such that drug concentrations have apparently fallen below the $lEC50$ which makes its estimation possible; see Schmidt-Erfurth et al. (2011).

The out-of-sample model qualifications are shown in Figure 5. The 2 mg q4w Aflibercept of VIEW2 arm is well predicted by the model, while the respective regimen in VIEW1 is predicted less successfully. This arm was reported to have an unusually high mean baseline change BCVA outcome for reasons which are still not well understood such that we did not investigate further. Moreover, the regimen 0.5 mg q4w Aflibercept appears to be under predicted in VIEW2 and slightly over predicted in VIEW1. However, when considering that VIEW1+2 are exactly replicated trials, the observed differences in this arm (see Figure 4) are not expected (also note that the ordering for each regimen reversed when comparing these in VIEW1 and VIEW2). If we were to compare our model predictions against an averaged result from VIEW1+2, these comparisons would look more favorable as the study differences would average out. We conclude that the average outcomes are well captured while the per arm variations are within limits which are known and still unexplained.

In summary, our final model is able to predict accurately the 2 mg q8w Aflibercept regimen which is our main focus when including the VIEW1+2 data into our analysis. The 2 mg q8w Aflibercept regimen is one of the treatments for wetAMD applied in clinical practice.

6. Discussion. Model-based drug development hinges on the amount of information which we can include into models. While hierarchical patient-level nonlinear models offer the greatest flexibility, they make raw patient-level data a requirement. This can limit the utility of such models considerably, as relevant information may only be available to the analyst in aggregate form from the literature. For our wetAMD drug development program the presented approach enabled patient-level clinical trial simulations for most wetAMD treatments used in the clinic. Our approach was used to plan confirmatory trials which test a new treatment regimen with less frequent dosing patterns against currently established regimens. In particular, these results were used to plan the confirmatory studies HARRIER and HAWK, which evaluate Brolucizumab in comparison to Aflibercept. These trials test a new and never observed dosing regimen aiming at a reduced dosing frequency while maintaining maximal efficacy. Within this regimen patients are assessed for their individual treatment needs during a q12w-learning cycle. Depending on this assessment, patients are allocated to a q12w or a q8w schedule. A key outcome of the trials is the proportion of patients allocated to the q12w regimen. Through the use of our approach it was possible to include highly relevant information from the literature into a predictive model which supported strategic decision making for the drug development program in wetAMD.

The critical step in our analysis was to model jointly our study data and external aggregate data. We constructed a novel Bayesian aggregation of average data which had to overcome three different issues:

1. Our new data were in aggregated average form; the raw data y' were not available, and we could not directly write or compute the likelihood for the observed average data $\bar{y}'$.
2. The new data were conducted under different experimental conditions. This is a standard problem in statistics and can be handled using hierarchical modeling, but here the number of "groups" is only two (the old data and the new data), so it would not be possible to simply fit a hierarchical model estimating group-level variation from data.
3. It was already possible to fit the model to the original data y, hence, it made sense to construct a computational procedure that made use of this existing fit.

We handled the first issue using the central limit theorem (CLT), which was justified by the large sample size of the external data. This allowed us to approximate the sampling distribution of the average data by a multivariate normal and using simulation to compute the mean and covariance of this distribution, for any specified values of the model parameters.

We handled the second issue by introducing a parameter δ governing the difference between the two experimental conditions. In some settings it would make sense to assign a weak prior on δ and essentially allow the data to estimate the parameters separately for the two experiments. In other cases a strong prior on δ would express the assumption that the underlying parameters do not differ much

between groups. Seen from a different perspective, the new experimental condition is considered as a biased observation of an already observed experimental condition, which goes back to Pocock (1976).

Finally, we formulated our approach by extending an existing model. That is, we added a term to the log-likelihood of the original model. This term represents the information from the external means. We used a nested simulation scheme which we ran during the MCMC fit. The key step to perform the nested simulation scheme was to generate a sufficiently large sample of random numbers prior to the MCMC run and to then use this sample for each iteration of the running MCMC to perform effectively a Monte Carlo integration. We expect this nested integration approach to be useful in general, since its applicability is not restricted to the presented application of marginalizing the likelihood over a latent variable space, but can be applied in general during a MCMC run.

Our proposed approach is an approximate solution with respect to the alternative approach, which is to represent the patient-level data of the external data set as latent. As our simulation studies have revealed, we are still able to obtain accurate estimates of the δ parameter vector, which is our main objective here. The reason is the large sample size of the external data, which ensures that the assumption of the CLT holds well. The use of our approximate procedure does lead to a reduction of computational resources needed to integrate the external average data. Thus, we can then use these freed-up computational resources to model more accurately the patient-level data and obtain in return better predictions. As external data sets of interest are usually of considerable sample size, we expect this to be an advantageous choice to spend our finite computational resources in these applications.

Considering our idea more generally, we have effectively reversed the common Bayesian approach in which external data are commonly used to elicit a prior, which is then updated with experimental data through the model likelihood. In our approach, this paradigm is conceptually reversed. The external data is explicitly made part of the model likelihood, which then informs our parameters of interest. In this light, we expect that our ideas will allow for future developments of general interest, such as the formulation of implicit priors or the definition of an effective sample size for complex models using a normal approximation.

In this work we have expanded the applicability of Bayesian meta-analysis to the broad class of nonlinear hierarchical models for the case whenever we wish to learn from aggregated average data, which renders data from individuals latent and only indirectly reported via means. This situation often times arises in the domain of biostatistics which uses meta-analytic approaches in various stages of drug development. However, the ideas presented are general and should also find application in other domains. For our specific case this work enabled accurate clinical trial simulations which supported the design of large phase III trials aiming to establish better treatments in wetAMD.

SUPPLEMENTARY MATERIAL

Supplement: Program sources (DOI: 10.1214/17-AOAS1122SUPP; .zip). Source code of R and Stan programs of simulation studies and drug-disease model.

REFERENCES

AMBATI, J. and FOWLER, B. J. (2012). Mechanisms of age-related macular degeneration. *Neuron* **75** 26–39.

AUGOOD, C. A., VINGERLING, J. R., DE JONG, P. T., CHAKRAVARTHY, U., SELAND, J., SOUBRANE, G., TOMAZZOLI, L., TOPOUZIS, F., BENTHAM, G., RAHU, M., VIOQUE, J., YOUNG, I. S. and FLETCHER, A. E. (2006). Prevalence of age-related maculopathy in older Europeans. *Arch. Ophthalmol.* **124** 529–535.

BETANCOURT, M. (2016). Diagnosing suboptimal cotangent disintegrations in Hamiltonian Monte Carlo. Preprint. Available at arXiv:1604.00695 [stat].

BROWN, D. M., KAISER, P. K., MICHELS, M., SOUBRANE, G., HEIER, J. S., KIM, R. Y., SY, J. P. and SCHNEIDER, S. (2006). Ranibizumab versus Verteporfin for Neovascular age-related macular degeneration. *N. Engl. J. Med.* **355** 1432–1444.

BUSCHINI, E., PIRAS, A., NUZZI, R. and VERCELLI, A. (2011). Age related macular degeneration and drusen: Neuroinflammation in the retina. *Prog. Neurobiol.* **95** 14–25.

CARO, J. J. and ISHAK, K. J. (2010). No head-to-head trial? Simulate the missing arms. *PharmacoEcon.* **28** 957–967.

DOMINICI, F., PARMIGIANI, G., WOLPERT, R. L. and HASSELBLAD, V. (1999). Meta-analysis of migraine headache teatments: Combining information from heterogeneous designs. *J. Amer. Statist. Assoc.* **94** 16–28.

GELMAN, A. (2004). Parameterization and Bayesian modeling. *J. Amer. Statist. Assoc.* **99** 537–545. MR2109315

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. CRC Press, Boca Raton, FL. MR3235677

HARRIER. Efficacy and Safety of RTH258 Versus Aflibercept - Study 2 - ClinicalTrials.gov. Available at https://clinicaltrials.gov/ct2/show/NCT02434328.

HART, W. M., ed. (1992). *Adler's Physiology of the Eye*: *Clinical Application*, 9th ed. Mosby, St. Louis.

HAWK. Efficacy and Safety of RTH258 Versus Aflibercept - ClinicalTrials.gov. Available at https://clinicaltrials.gov/ct2/show/NCT02307682.

HEIER, J. S., BROWN, D. M., CHONG, V., KOROBELNIK, J.-F., KAISER, P. K., NGUYEN, Q. D., KIRCHHOF, B., HO, A., OGURA, Y., YANCOPOULOS, G. D., STAHL, N., VITTI, R., BERLINER, A. J., SOO, Y., ANDERESI, M., GROETZBACH, G., SOMMERAUER, B., SANDBRINK, R., SIMADER, C. and SCHMIDT-ERFURTH, U. (2012). Intravitreal Aflibercept (VEGF trap-eye) in wet age-related macular degeneration. *Ophthalmology* **119** 2537–2548.

HIGGINS, J. P. T. and GREEN, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions*, Version 5.1.0 ed. The Cochrane Collaboration.

HIGGINS, J. P. T. and WHITEHEAD, A. (1996). Borrowing strength from external trials in a meta-analysis. *Stat. Med.* **15** 2733–2749.

HOFFMAN, M. D. and GELMAN, A. (2014). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15** 1593–1623. MR3214779

ISHAK, K. J., PROSKOROVSKY, I. and BENEDICT, A. (2015). Simulation and matching-based approaches for indirect comparison of treatments. *PharmacoEcon.* **33** 537–549.

JUSKO, W. J. and KO, H. C. (1994). Physiologic indirect response models characterize diverse types of pharmacodynamic effects. *Clin. Pharmacol. Ther.* **56** 406–419.

KHANDHADIA, S., CIPRIANI, V., YATES, J. R. W. and LOTERY, A. J. (2012). Age-related macular degeneration and the complement system. *Immunobiology* **217** 127–146.

KINNUNEN, K., PETROVSKI, G., MOE, M. C., BERTA, A. and KAARNIRANTA, K. (2012). Molecular mechanisms of retinal pigment epithelium damage and development of age-related macular degeneration. *Acta Ophthalmol.* **90** 299–309.

POCOCK, S. J. (1976). The combination of randomized and historical controls in clinical trials. *J. Chronic Dis.* **29** 175–188.

ROSENFELD, P. J., BROWN, D. M., HEIER, J. S., BOYER, D. S., KAISER, P. K., CHUNG, C. Y. and KIM, R. Y. (2006). Ranibizumab for neovascular age-related macular degeneration. *N. Engl. J. Med.* **355** 1419–1431.

SCHMIDT-ERFURTH, U., ELDEM, B., GUYMER, R., KOROBELNIK, J.-F., SCHLINGEMANN, R. O., AXER-SIEGEL, R., WIEDEMANN, P., SIMADER, C., GEKKIEVA, M. and WEICHSELBERGER, A. (2011). Efficacy and safety of monthly versus quarterly Ranibizumab treatment in neovascular age-related macular degeneration: The EXCITE study. *Ophthalmology* **118** 831–839.

SHEINER, L. B. (1997). Learning versus confirming in clinical drug development. *Clin. Pharmacol. Ther.* **61** 275–291.

SIGNOROVITCH, J. E., WU, E. Q., YU, A. P., GERRITS, C. M., KANTOR, E., BAO, Y., GUPTA, S. R. and MULANI, P. M. (2010). Comparative effectiveness without head-to-head trials. *PharmacoEcon.* **28** 935–945.

STAN DEVELOPMENT TEAM (2017). Stan: A C++ library for probability and sampling.

WEBER, S., CARPENTER, B., LEE, D., BOIS, F. Y., GELMAN, A. and RACINE, A. (2014). Bayesian drug disease model with Stan: Using published longitudinal data summaries in population models, Population Approach Group Europe Meeting 2014, Alicante, Spain. Available at http://page-meeting.org/?abstract=3200.

WEBER, S., GELMAN, A., LEE, D., BETANCOURT, M., VEHTARI, A. and RACINE-POON, A. (2018). Supplement to "Bayesian aggregation of average data: An application in drug development." DOI:10.1214/17-AOAS1122SUPP.

XU, L., LU, T., TUOMI, L., JUMBE, N., LU, J., EPPLER, S., KUEBLER, P., DAMICO-BEYER, L. A. and JOSHI, A. (2013). Pharmacokinetics of Ranibizumab in patients with neovascular age-related macular degeneration: A population approach. *Investig. Ophthalmol. Vis. Sci.* **54** 1616–1624.

S. WEBER
A. RACINE-POON
NOVARTIS PHARMA AG
BASEL, 4002
SWITZERLAND
E-MAIL: sebastian.weber@novartis.com
amy.racine@novartis.com

A. GELMAN
M. BETANCOURT
DEPARTMENT OF STATISTICS
COLUMBIA UNIVERSITY
NEW YORK, NEW YORK 10027
USA
E-MAIL: gelman@stat.columbia.edu
betanalpha@gmail.com

D. LEE
GENERABLE
BROOKLYN, NEW YORK 11205
USA
E-MAIL: daniel@generable.com

A. VEHTARI
HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY HIIT
DEPARTMENT OF COMPUTER SCIENCE
AALTO UNIVERSITY
AALTO, FI-00076
FINLAND
E-MAIL: Aki.Vehtari@aalto.fi