Submitted to the Annals of Applied Statistics

BAYESIAN AGGREGATION OF AVERAGE DATA

By Sebastian Weber[‡] Andrew Gelman[§,*] Daniel Lee[§,*]
Michael Betancourt[§,*]
Aki Vehtari[¶,†] and Amy Racine-Poon[‡]

Novartis Pharma AG[‡], Columbia University[§], Aalto University[¶]

How can we combine datasets collected under different conditions? This is the well-known problem of *meta-analysis*, for which Bayesian methods have long been used to achieve partial pooling. Here we consider the challenge when one dataset is given as raw data while the second dataset 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.

The need for meta-analytic methods applied to complex models arises frequently in clinical drug development. Throughout the different phases of a drug development program, randomized trials are used to establish in stages 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. Moreover, realistic models suitable for the desired extrapolation are often complex (longitudinal and nonlinear). 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 demonstrate its application to the problem in drug development that motivated this research, a hierarchical nonlinear model in pharmacometrics, implementing the computation in Stan.

1. Introduction. Modern drug development proceeds in stages to establish the tolerability, safety, and efficacy of a candidate drug (Sheiner, 1997). At each stage it is essential to plan the next steps, using all relevant information. The collected raw data are measurements of individual patients over time. Pharmacometric models of such raw data are commonly

[*]Institute for Education Sciences R305D140059-16, Office of Naval Research N00014-15-1-2541, Sloan Foundation G-2015-13987

[†]Academy of Finland (grant 298742)

Keywords and phrases: Meta-Analysis, Hierarchical modeling, Bayesian computation, Pharmacometrics, Stan

nonlinear longitudinal differential equation models 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). However, available meta-analytic methods can be applied either to summary or individual patient data, but not in the important mixed data case when models are complex.

This leads to the awkward situation that the nonlinear model is likely relevant for extrapolating from the raw data to the summary data as it can (one hopes) adequately account for confounding factors, but the summary data cannot be included directly, as the likelihood function for the aggregated data in general has no closed-form expression. The standard EM or Bayesian approach in this case is to consider the unavailable individual data points as missing data, but such an approach can be computationally prohibitive as it can vastly increase the dimensionality of the problem space.

This paper describes 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 datasets.

In the remainder of this section we describe the algorithm; section 2 demonstrates a simulated linear example for which our approach is compared to an exact analytic reference; and section 3 applies the method to the problem that motivated this work, a nonlinear differential equation model in pharmacometrics.

1.1. *General formulation.* We shall work in a hierarchical Bayesian framework. Suppose we have data $y = (y_{jt}; j = 1, \ldots, J; t = 1, \ldots, T)$ 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 dataset*, $y' = (y'_{jt}; j = 1, \ldots, J'; t = 1, \ldots, T')$, 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, for example, 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 to 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 datasets. 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 concep-

tually simplest way to proceed would be to consider the individual measurements y'_{jt} as missing data, and to perform Bayesian inference jointly on all unknowns, obtaining draws from the posterior distribution, $p(\phi, \delta, \alpha'|y, \bar{y}')$. 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 datasets such that the dimensionality of δ may be smaller than ϕ. This imposes that some components of δ are exactly 0.

We resolve the second difficulty via a normal approximation, taking 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.

1.2. *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 whose parameters depend 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 from 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 within commonly used MCMC software, including Stan. To overcome these computational issues we instead use simulated plug-in estimates. We calculate in each iteration s of the MCMC run 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 $\tilde{y}_{jt}, j = 1, \ldots, \tilde{J}, t = 1, \ldots, T'$, for some number of hypothetical new individuals $\tilde{J}$, drawn from the

distribution $p(y'|\phi'_s)$, 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 dataset; rather, we simulate them only as a device for approximating the likelihood of average data, $\bar{y}$, under these conditions. The choice of 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.

1.3. *Computational issues: tuning and convergence.* By construction of Hamiltonian Monte Carlo (HMC), as used in Stan, no random numbers can be drawn during sampling as part of the model. However, the 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 random numbers initially. 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 $\tilde{J}$ was chosen too small, standard diagnostics like $\widehat{R}$ (Gelman et al., 2013) 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 4 parallel MCMC chains as this is free on a four-processor laptop or desktop computer). The $\widehat{R}$ calculation 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.

2. Hierarchical linear regression. The first example, hierarchical linear regression, is simple enough such that we can easily 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 model motivating this work which is presented in Section 3.

For simplicity, we consider a linear regression with a continuous covariate x (corresponding to time) in which the intercept varies by individual and the slope varies by group. That is, for the main dataset y, the model is $y_{jt} \sim \mathrm{N}(\alpha_{j1} + \alpha_{j2}\, x_t + \beta\, x_t^2, \sigma_y^2)$, 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 1.1, the vector of shared parameters α is $\phi = (\mu_{\alpha1}, \mu_{\alpha2}, \beta, \sigma_{\alpha1}, \sigma_{\alpha2}, \sigma_y)$; we assume the number of individuals J is large enough that we can assign a noninformative prior to ϕ.

For the external dataset y', the model is $y_{jt} \sim \mathrm{N}(\alpha'_{j1} + \alpha'_{j2}\, x_t + \beta\, x_t^2, \sigma_y^2)$, 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 $\phi' = (\mu_{\alpha1} + \delta_1, \mu_{\alpha2} + \delta_2, \beta, \sigma_{\alpha1}, \sigma_{\alpha2}, \sigma_y)$.

Assumed parameter values. We create simulations assuming the following conditions, which we set to roughly correspond to the features of the pharmacometrics example in Section 3:

- $J = 100$ individuals in the original dataset, each measured $T = 13$ times (corresponding to measurements once per month for a year), $x_t = 0, \frac{1}{12}, \ldots, 1$.
- $J' = 100$ individuals in the external dataset, also measured at these 13 time points.
- $(\mu_{\alpha1}, \sigma_{\alpha1}) = (0.5, 0.1)$, 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 5 points on the original scale of the data.

Finally, we set δ to $(0.1, 0.1)$, which represents a large difference between groups 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 μ_1, μ_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 1 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 4 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 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 left panel of Figure 1. 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

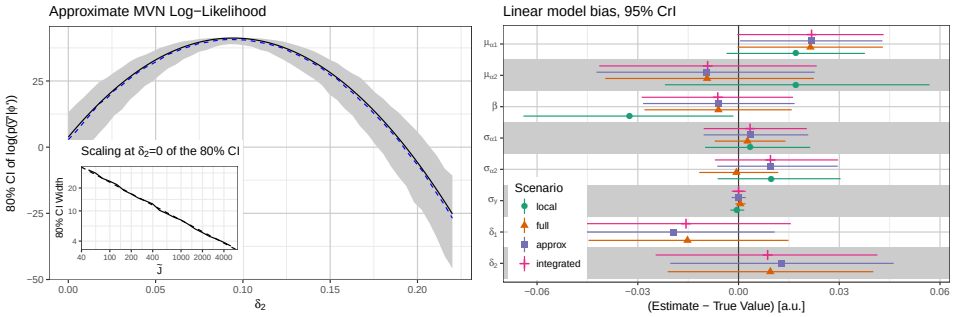

FIG 1. *Hierarchical linear model example. (Left) Comparison of the analytical expression for* $\log p(\bar{y}'|\phi')$*, shown as a solid black line, to the simulation based multivariate normal approximation* $N(\bar{y}'|\tilde{M}_s, \frac{1}{J'}\tilde{\Sigma}_s)$*. The simulation includes* $\tilde{J} = 100$ *hypothetical individuals, and* 1000 *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. (Right) 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}*.*

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.

approx: 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 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 right panel of Figure 1 shows the results of the parameter estimates as bias. We are using informative priors and so we neither desire or 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 $p(\phi, \delta|y, y')$ which is based on the individual raw data y and y' instead of y and mean data $\bar{y}'$. The appropriate comparison is wrt 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 *delta*.

3. Application: A differential equation model from pharmacometrics. This work derived from a drug development program to investi-

gate new treatment options for patients with wet age-related macular degeneration (wetAMD). This disease is the leading cause of severe vision loss in the elderly population. 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 anti-VEGF agent first introduced was Ranibizumab while another anti-VEGF agent, Aflibercept, was 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. A key requirement for a 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 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 time-points to a given patient.

The clinical trial simulations are performed using nonlinear population drug-disease models. These describe the response of an individual patient to a treatment over time. A key assessment of visual acuity in clinical studies is the number of letters a patient can read from an ETDRS (Early Treatment Diabetic Retinopathy Study) chart, expressed as best-corrected visual acuity (BCVA) score, i.e. 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 her/his individual dosing history. This enables realistic extrapolation to future designs with alternative dosing regimens.

A drug-disease model 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, EXCITE and ANCHOR (Rosenfeld et al., 2006; Brown et al., 2006; Schmidt-Erfurth et al., 2011). We chose to model the nominal visual acuity value as recorded instead of the commonly reported baseline change. The visual acuity value is limited to the range of 0–100, which corresponds to the number of letters read on the ETDRS chart. We decided to model the measured letter value, y_{jt}, of a patient j at time-point t on a logit transformed scale, $R_j(t) = \text{logit}(y_{jt}/100)$. The drug-disease model used was derived from the frequently used 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 this 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 $t_{1/2}$ from the vitreous to 9 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. The modified turnover model used is defined by the ordinary differential equation (ODE),

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

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 Eq. (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 Eq. (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, $S_j(t) = s_j = 0$, is present and 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 drug-related parameters ($t_{1/2}$, $E_{\max}$ and $EC50$) from non-drug related parameters which are the remaining parameters.

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 since the three drug-related parameters ($t_{1/2}$, $E_{\max}$ and $EC50$) are expected to differ from the particular values for Ranibizumab, which led us to this work. In the following we will first assess the feasibility of our proposed approach in a simulation study, i.e. if we can successfully learn differences in drug related parameters from average data only, given that individual raw data is available for only some but not all drug treatments. Then we will apply the approach to the real data and perform a model qualification by demonstrating that the final model can do realistic out-of-sample predictions for Aflibercept.

3.1. *Simulation study.* The function $R_j(t)$ in Eq. (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 $R_j(t) = R_j^{\text{ss}} + \left(R_{0j} - R_j^{\text{ss}}\right) \exp\left(-k_j^{\text{out}} t\right)$. In the following we consider 3 different cohorts of patients (placebo, treatment 1 and 2) observed at times x_t. Data for treatment 2 will be considered as the external dataset and given as average data only to evaluate our approach. Measurements y_{jt} of a patient j are assumed to be iid normal, $y_{jt}/100 \sim \text{N}(\text{logit}^{-1}(R_j(x_t)), \sigma_y^2)$. 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 $x_t = 0, \frac{1}{12}, \ldots, 1$.
- $J' = 100$ patients in the external dataset, measured at the same $T' = 13$ time points.
- $R_{0j} \sim \text{N}(L\alpha_0, \sigma^2_{L\alpha_0})$ 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 $L\alpha_0 \sim \text{N}(0, 2^2)$ and $\sigma_{L\alpha_0} \sim \text{N}^+(0, 1^2)$.
- $k_j^{\text{in}}/k_j^{\text{out}} = L\alpha_s$ 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 $L\alpha_s = \text{logit}(35/100)$ and the prior to $L\alpha_s \sim \text{N}(-1, 2^2)$.
- $\log(1/k_j^{\text{out}}) \sim \text{N}(l\kappa, \sigma^2_{l\kappa})$ 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 $l\kappa \sim \text{N}(\log(1/4), \log(2)^2)$ and $\sigma_{l\kappa} \sim \text{N}^+(0, 1^2)$.
- $\log(E_{\max j})$ 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 $\log(E_{\max j}) = lE_{\max j} = \log(\text{logit}(60/100) - \text{logit}(35/100))$ which represents a gain of 25 points in comparison to placebo. Pa-

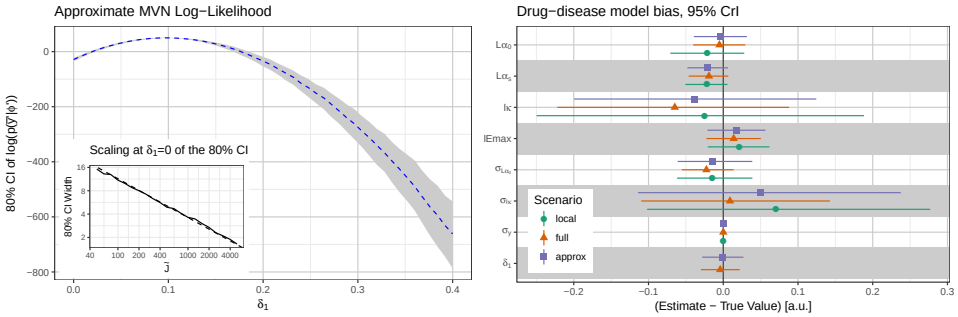

FIG 2. *Drug-disease model example: (Left) Assessment of the distribution of the multivariate normal approximation to* $\log p(\bar{y}'|\phi')$ *at a simulation size of* $\tilde{J} = 10^2$ *hypothetical fake patients using* 10^3 *replicates for varying* δ_1*. The gray area marks the 80% confidence interval, 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 a linear model. (Right) 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* $2 \cdot 10^{-3}$*.*

tients 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 2. In the left panel of Fig. 2 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 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 4 chains and choose to set $\tilde{J}$ to $5 \cdot 10^2$ as before. The model estimates are shown as bias in the right panel of Figure 2. The precision of the estimates from the local fit (*green line with a dot*) increases

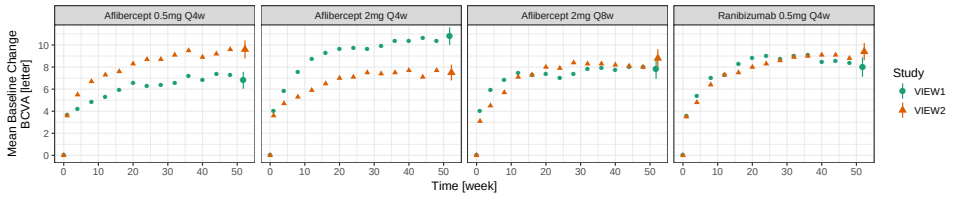

FIG 3. *Published average data of the VIEW1+2 studies (Heier et al. (2012)). Shown is the reported mean baseline change best-corrected visual acuity (BCVA) over a time period of one year. The vertical line at the last time point marks one standard error of the reported mean.*

when adding the external data. While population mean parameters gain in precision in the full (*orange line with a triangle*) and approx (*blue line with a square*) scenario, 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 approx scenario are almost identical to the full scenario, including the important parameter δ_1.

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.

3.2. *Application in wetAMD.* Now we apply our approach to the actual dataset which motivated this work. The drug-disease model developed had to account for a number of additional complexities compared to the simulated example, e.g. a fast onset of the drug effect, a truncated distribution of the baseline BCVA value due to enrollment restrictions and per study varying mean baseline BCVA values. We provide the Stan model used in the supplementary material and concentrate here on the steps which were needed to include the average data of interest. The only drug related parameters of the 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 compromise a total of $N = 1342$ patients from the studies MARINA, EXCITE and ANCHOR, the data from

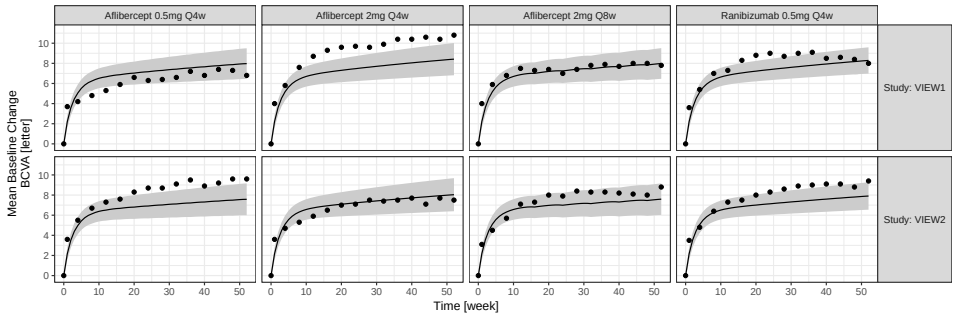

FIG 4. *Application in wetAMD: Shown is the predicted mean baseline change BCVA as solid line. The gray area marks one standard error for the predicted mean assuming a sample size as reported per arm (about 300 each). The dots mark the reported mean baseline change BCVA and are shown as reference. The two left columns for Q4w Aflibercept are out-of-sample predictions while the two right columns are within-sample predictions, since only the data of the two right columns were included in the model fit.*

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 $\mathrm{N}(0, 1^2)$) for each of the drug related parameters for patients on Aflibercept treatment. 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. For one we consider it reasonable to assume that the natural disease progression does not change and for two it is practically impossible to infer differences in the natural disease progression as compared to our dataset with the VIEW1+2 data since no placebo patients were included in either study for ethical reasons.

Figure 3 shows the published mean baseline change BCVA data of the VIEW1+2 studies. We choose to include only the mean BCVA data of the dosing regimens Q8w Aflibercept and Ranibizumab into our model as these are the registered dosing regimens and are hence of greatest interest to describe these as accurately as possible. The included treatment arms from VIEW1+2 compromise a total of 1202 patients. 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 decided to use the remaining Q4w Aflibercept regimens for out-of-sample model qualification.

The final results are shown in Figure 4. The posterior predictive of the mean baseline change BCVA response of the two included regimens (Q8w

Aflibercept and Ranibizumab) are shown in the two right columns. The model predictions are in good agreement with the reported data. The out-of-sample predictions of Aflibercept are in good agreement with the observed data for the dosing regimens 0.5mg Q4w of both studies. The 2.0mg Q4w of VIEW2 is well predicted by the model, while the 2.0mg Q4w held-out regimen of 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.

In summary, our approach enables reliable predictions of the longitudinal BCVA response for individual patients for the available registered wetAMD treatments under varying dosing regimens. The reported mean data of Aflibercept is sufficient to estimate drug specific parameters in the context of a nonlinear drug-disease model. This is of great benefit for robust decisions for the development of new treatments in wetAMD.

4. Discussion. We constructed this method in response to three different issues that arose with the integration of external data into a statistical analysis:

1. Our new data were in aggregated average form; the raw data y'_{jt} 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 2 (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 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 due to 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 enough 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 whenever a Monte Carlo integration is needed during a MCMC run.

Considering our presented idea in a more general sense, we have effectively reversed the common Bayesian approach in which external data is 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 made explicitly 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 arises oftentimes in the domain of biostatistics which uses meta-analytic approaches in various stages of drug development. However, the ideas presented are general and should find application in other domains.

References.

Betancourt, M. (2016). Diagnosing Suboptimal Cotangent Disintegrations in Hamiltonian Monte Carlo. *arXiv:1604.00695 [stat]*. arXiv: 1604.00695.

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. *New England Journal of Medicine* **355** 1432–1444.

Dominici, F., Parmigiani, G., Wolpert, R. L. and Hasselblad, V. (1999). Meta-analysis of migraine headache teatments: Combining information from heterogeneous designs. *Journal of the American Statistical Association* **94** 16–28.

Gelman, A. (2004). Parameterization and Bayesian modeling. *Journal of the American Statistical Association* **99** 537–545.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013). *Bayesian Data Analysis*, third ed. CRC Press.

Hart, W. M., ed. (1992). *Adler's Physiology Of The Eye: Clinical Application, 9e*, 9th edition ed. Mosby, St. Louis.

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. *Statistics in Medicine* **15** 2733–2749.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15** 1593–1623.

Jusko, W. J. and Ko, H. C. (1994). Physiologic indirect response models characterize diverse types of pharmacodynamic effects. *Clinical Pharmacology and Therapeutics* **56** 406–419.

Pocock, S. J. (1976). The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases* **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. *New England Journal of Medicine* **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. *Clinical Pharmacology & Therapeutics* **61** 275–291.

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.

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 ApproachRanibizumab Pharmacokinetics in AMD. *Investigative Ophthalmology & Visual Science* **54** 1616–1624.

ACKNOWLEDGEMENTS

We thank Bob Carpenter for fruitful discussions on the manuscript.

Novartis Pharma AG
Basel, Switzerland

Department of Statistics
Columbia University
New York, USA

Helsinki Institute for
Information Technology HIIT
Department of Computer Science
Aalto University, Finland