JAMA Guide to Statistics and Methods

Bayesian Hierarchical Models

Anna E. McGlothlin, PhD; Kert Viele, PhD

+
Author Audio Interview

Related article page 2344

Treatment effects will differ from one study to another evaluating similar therapies, both because of random variation between individual patients and owing to true differences that exist because of other differences, including inclusion criteria and temporal trends. The sources of variability have many levels; one level involves the random differences between individual patients, and another level involves the systematic differences that exist between studies. This multilevel or hierarchical information occurs in many research settings, such as in cluster-randomized trials and meta-analyses.[1,2] Sources of variation can be better understood and quantified if treatment effect estimates from each individual study are examined in relation to the totality of information available in all the studies.

Bayesian analysis differs from the usual frequentist approach (eg, use of *P* values or confidence intervals). Rather than focusing on the probability of different patterns in outcomes assuming specific treatment effects, Bayesian analysis relies on the use of prior information in combination with data from a study to calculate the probabilities of a treatment effect.[3] Readers may be familiar with Bayesian analysis when used in randomized clinical trials.[4,5] In this type of Bayesian analysis, patients are considered largely equivalent except with respect to the assigned treatment, and the goal is to estimate the probability of an overall treatment effect in the population.

In contrast, a Bayesian hierarchical model (BHM) is a statistical procedure that integrates information across many levels, so multiple quantities are estimated simultaneously, and explicitly separates the observed variability into parts attributable to random differences and true differences.[6] The model has 2 key characteristics. First, there is a hierarchical or multilevel structure. For example, if multiple studies were conducted to evaluate diabetes management strategies, the first-level data may be improvements in hemoglobin A_{1C} values in individual patients, the second-level data may be the mean improvements for patients within each trial, and the third-level data may be the average improvements in trials grouped according to the type of disease management strategy. Second, prior information is used to reflect available information, even if vague, regarding the likely values and variability at each level of the hierarchy (eg, the variability of improvements in patients in a single trial, the variability of average treatment effects between trials using similar disease management strategies, and the variability of treatment effects among groups of trials that use different disease management strategies). Using Bayes theorem, prior information, and the data, the BHM yields estimates of the true effects at each level of the hierarchy.[3,6] Estimates of true treatment effects may be derived for individual patients, patient subgroups, individual trials, or groups of trials. Each of these estimates are informed by the entire data set included in the statistical model.[6]

In this issue of *JAMA*, Stunnenberg and colleagues present the results of a trial that used a BHM to integrate data from a series of N-of-1 crossover trials[7] comparing mexiletine with placebo in the treatment of patients with nondystrophic myotonia.[8] An N-of-1 trial uses a patient as his or her own control by repeatedly exposing the patient to a treatment or placebo and measuring the effect of the intervention. Each N-of-1 trial exposes the patient to between 1 and 4 treatment pairs or sets, with each set randomizing the order of mexiletine and placebo, with a 1-week washout period between therapies. After each treatment set, prespecified rules were used to determine whether the patient should continue to the next treatment set or discontinue, either for evidence of benefit of mexiletine, evidence of no benefit, or for reaching the maximum allowed number of treatment sets. A BHM was used to integrate data from all available N-of-1 trials performed in all the patients to produce estimates of treatment effects for each patient individually and also for 2 genetic subtypes of the disease.

Why Is a BHM Used?

Multilevel data have an underlying hierarchical structure. In the report by Stunnenberg et al, each trial had data from a single patient, and patients were grouped into genetic subtypes. Properly integrating this information required acknowledging the commonalities, eg, data from 2 patients having the same genetic subtype are more likely to be similar than data from 2 patients having different genetic subtypes. Heterogeneity between genetic subtypes and patient-to-patient variability are simultaneously accounted for in the BHM. A pooled analysis, ie, simply combining data from all patients, would not account for systematic patient-to-patient differences. At the other extreme, analyzing each individual patient's trial separately would not account for the information available across all the trials. This could result in underpowered analyses.

By considering the results across all trials, BHMs allow for more accurate estimates of the treatment effects for each individual trial because of a fundamental fact about multilevel data—namely, that regardless of the true systematic differences between the true treatment effects estimated by each trial, random variability is more likely to amplify these differences than diminish them. For example, suppose 4 single-intervention group trials of 100 patients each are conducted to estimate a common rate of a particular patient outcome, which is, hypothetically, 60% for all of the trials. Because of random variability in 100 patients, it is likely that one of the studies will produce an observed rate less than 60%, while another will produce an observed rate greater than 60%. Even though the studies all have exactly equal true underlying rates, numerical simulation demonstrates that the lowest observed value will average 54.9% and the highest will average 65.0%. Even though no true heterogeneity exists, when actual observations are made the results will appear heterogeneous because of random variation. Consider a different scenario in which the 4 trials are truly different, with true underlying rates of 54%, 58%, 62%, and 66%. Although the true rates range from 54% to 66%, the observed rates on average will range from 52.5% to 67.4% because of the additional random variation seen in a trial with a finite sample size. Here again, the observed values tend to be farther apart than the true values. The lowest observed value in a group is likely lower than its true value, and the highest observed value in a group is likely higher than its true value.

Knowing that observed values tend to be farther apart than the true values, the best estimates of the true values are closer together than the observed values. These estimates (which are more accurate than if each estimate were based only on the results from the individual trial) can be obtained using "shrinkage estimation."[6,9] The term "shrinkage" refers to the reduction in the observed differences between the trials. The purpose of the BHM is to determine the proper amount to move the observed treatment differences closer together to obtain the shrinkage estimates. The model estimates the proportion of total variability attributable to random (within-trial) variability and the amount attributable to systematic differences. By eliminating random noise, the resulting estimates are, on average, closer to the underlying truth.[6]

If the observed heterogeneity is consistent entirely with random variation, the resulting estimates for each group will be close to each other. In contrast, if the observed heterogeneity far exceeds what may be explained by random variation, the heterogeneity will be attributed to true differences that exist between the groups, and the treatment effect estimates will not shift much from the observed rates.

Estimates from a BHM typically have reduced variability compared with those from independent analyses, in which each trial is analyzed separately. This results in tighter interval estimates of treatment effects and may result in statistical hypothesis tests with greater power and lower type I error. For these reasons, BHMs are especially promising for studies of rare diseases for which large sample sizes are not feasible.

What Are the Limitations of BHMs?

All statistical models are predicated on assumptions that should be understood before applying the method. Bayesian hierarchical models rely on various assumptions (eg, the number of levels and the prior probability distributions used as the basis for Bayesian estimation of treatment effects) to estimate and separate within- and across-group variability.[6] Additionally, most BHMs assume a certain type of distribution for the across-group variability—for example, a bell-shaped curve. This assumption may fail if there is an outlying group inconsistent with a bell shape, potentially resulting in biased estimates for that outlying group.[10] It is important to consider sensitivity analyses that verify the robustness of the conclusions to changes in the choices of prior distributions.

How Were BHMs Used in This Case?

In the study by Stunnenberg et al,[8] information from 27 N-of-1 trials was integrated to produce estimates of the treatment effect of mexiletine relative to placebo for 2 genetic subgroups and for the overall population. The outcome was a reduction in self-reported muscular stiffness on a 1-to-9 scale using a validated questionnaire. The mean reduction in stiffness was 3.84 (95% CI, 2.52 to 5.16) for the *CLNC1* genotype and 1.94 (95% CI, 0.35 to 3.53) for the *SCN4A* genotype. The mean reduction across all subgroups was 3.06 (95% CI, 1.96 to 4.15). Bayesian hierarchical models were used to successfully and rigorously integrate information with a complex underlying structure: a variable number of treatment sets per patient, with patients grouped into 2 genotype subgroups.

The BHM allows analysis at different levels of aggregation. In the study by Stunnenberg et al,[8] the aggregation occurred at 3 levels. First, data from each patient were aggregated across multiple treatment sets to estimate a single treatment effect for each patient. At the second level of the hierarchy, data were aggregated to estimate the treatment effect within 2 genotype subgroups. The third level described the distribution across subgroups.

How Should BHMs Be Interpreted?

A BHM provides estimates of treatment effects, or other relevant clinical metrics, at each level of the hierarchy, based on all data included in the model. Because of the inclusion of a greater amount of information, these estimates are generally more accurate than if analyses were conducted on subgroups separately, increasing the power of statistical comparisons. For example, in a 3-level model with multiple measurements for each patient, multiple patient subtypes (eg, genetic subtypes), and an overall treatment effect for all patients, the hierarchical model provides estimates for each patient individually, for each patient subtype, and across all subtypes.

ARTICLE INFORMATION

Author Affiliations: Berry Consultants LLC, Austin, Texas.

Corresponding Author: Anna E. McGlothlin, PhD, Berry Consultants LLC, 3345 Bee Caves Rd, Ste 201, Austin, TX 78746 (anna@berryconsultants.com).

Section Editors: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center and David Geffen School of Medicine at UCLA; and Edward H. Livingston, MD, Deputy Editor, *JAMA*.

Conflict of Interest Disclosures: Drs McGlothlin and Viele are employees of Berry Consultants LLC, a private consulting firm specializing in Bayesian adaptive clinical trial design, implementation, and analysis.

REFERENCES

1. Meurer WJ, Lewis RJ. Cluster randomized trials. *JAMA*. 2015;313(20):2068-2069. doi:10.1001/jama.2015.5199

2. Whitehead A. *Meta-analysis of Controlled Clinical Trials*. Sussex, United Kingdom: Wiley West; 2002. doi:10.1002/0470854200

3. Quintana M, Viele K, Lewis RJ. Bayesian analysis: using prior information to interpret the results of clinical trials. *JAMA*. 2017;318(16):1605-1606. doi:10.1001/jama.2017.15574

4. Goligher EC, Tomlinson G, Hajage D, et al. Extracorporeal membrane oxygenation for severe acute respiratory distress syndrome and posterior probability of mortality benefit in a post hoc Bayesian analysis of a randomized clinical trial [published online October 22, 2018]. *JAMA*. doi:10.1001/jama.2018.14276

5. Lewis RJ, Angus DC. Time for clinicians to embrace their inner Bayesian? reanalysis of results of a clinical trial of extracorporeal membrane oxygenation [published online October 22, 2018]. *JAMA*. doi:10.1001/jama.2018.16916

6. Gelman A, Stern HS, Carlin JB, Dunson DB, Vehtari A, Rubin DB. *Bayesian Data Analysis*. 3rd ed. Boca Raton, FL: CRC Press; 2013.

7. Zucker DR, Schmid CH, McIntosh MW, et al. Combining single patient (N-of-1) trials to estimate population treatment effects and to evaluate individual patient responses to treatment. *J Clin Epidemiol*. 1997;50(4):401-410. doi:10.1016/S0895-4356(96)00429-5

8. Stunnenberg BC, Raaphorst J, Groenewoud HM, et al. Effect of mexiletine on muscle stiffness in patients with nondystrophic myotonia evaluated using aggregated N-of-1 trials [published December 11, 2018]. *JAMA*. doi:10.1001/jama.2018.18020

9. Lipsky AM, Gausche-Hill M, Vienna M, Lewis RJ. The importance of "shrinkage" in subgroup analyses. *Ann Emerg Med*. 2010;55(6):544-552. doi:10.1016/j.annemergmed.2010.01.002

10. Neuenschwander B, Wandel S, Roychoudhury S, Bailey S. Robust exchangeability designs for early phase clinical trials with multiple strata. *Pharm Stat*. 2016;15(2):123-134. doi:10.1002/pst.1730