DIA

Therapeutic Innovation
& Regulatory Science
2018, Vol. 52(5) 546-559

DOI: 10.1177/2168479018778282
tirs.sagepub.com

Minimizing Patient Burden Through the Use of Historical Subject-Level Data in Innovative Confirmatory Clinical Trials: Review of Methods and Opportunities

Jessica Lim, MA[1], Rosalind Walley, MA[2], Jiacheng Yuan, PhD[3], Jeen Liu, PhD[3], Abhishek Dabral, MS[3], Nicky Best, PhD[4], Andrew Grieve, PhD[2], Lisa Hampson, PhD[5], Josephine Wolfram, MSc[6], Phil Woodward, MPhil[7], Florence Yong, PhD[8], Xiang Zhang, PhD[9], and Ed Bowen, MSc[10]

Abstract

The goal of clinical trial research is to deliver safe and efficacious new treatments to patients in need in a timely and cost-effective manner. There is precedent in using historical control data to reduce the number of concurrent control subjects required in developing medicines for rare diseases and other areas of unmet need. The purpose of this paper is to provide a review for a regulatory and industry audience of the current state of relevant statistical methods, and of the uptake of these approaches and the opportunities for broader use of historical data in confirmatory clinical trials. General principles to consider when incorporating historical control data in a new trial are presented. Bayesian and frequentist approaches are outlined including how the operating characteristics for such a trial can be obtained. Finally, examples of approved new treatments that incorporated historical controls in their confirmatory trials are presented.

Keywords

TransCelerate, historical controls, propensity score, Bayesian, informative prior

Introduction

In 2004, the United States (US) Food and Drug Administration (FDA) launched the Critical Path Initiative, which sought to determine the root causes for the latency between laboratory discoveries and their translation into clinical therapies delivered to patients.[1] This initiative acknowledged that a significant opportunity to speed the delivery of new therapies exists and that "[drug] developers have no choice but to use the tools and concepts of the last century to assess this century's candidates."[1] Meanwhile, in this century, the availability of standardized data and the evolution of statistical methods gives us the ability to assess safety and efficacy more efficiently and with reduced patient burden.

Regulators have demonstrated willingness to accept the use of historical controls in rare disease, where subjects are scarce. However, significant challenges also exist in more common disease areas, for example, Alzheimer disease (AD). Recruiting subjects for AD clinical trials is increasingly difficult because of logistical and patient burden issues, resulting in increased clinical trial timelines. Hurdles to getting potential subjects to participate include the request to provide cerebrospinal fluid (CSF) via an uncomfortable spinal tap, injections of tracking agents to support imaging, and a need for a study partner to provide assessments regarding daily functioning.[2]

There is an opportunity to leverage the considerable investments made in high-quality, curated, and trusted clinical data

[1] Clinical Statistics, GlaxoSmithKline, Collegeville, PA, USA
[2] Centre for Excellence in Statistical Innovation, UCB, UK
[3] Statistical Science and Programming, Allergan, Irvine, CA, USA
[4] Advanced Biostatistics and Data Analytics Centre of Excellence, GlaxoSmithKline, Uxbridge, Middlesex, UK
[5] Statistical Methodology & Consulting, Novartis Pharma AG, Basel
[6] Data Science Biostatistics, Astellas, Leiden, the Netherlands
[7] Independent Consultant, Newmarket, UK
[8] Biostatistics, Worldwide Research & Development, Pfizer, Cambridge, MA, USA
[9] Global Statistical Science, Eli Lilly and Company, Indianapolis, IN, USA
[10] R&D Data Centre of Excellence, GlaxoSmithKline, Collegeville, PA, USA

Submitted 18-Dec-2017; accepted 30-Apr-2018

Corresponding Author:
Jessica Lim, MA, GlaxoSmithKline, 1250 S. Collegeville Road, Collegeville, PA 19426-0989, USA.
Email: jessica.w.lim@gsk.com

Step 1: Assess whether the new trial might be an appropriate candidate for complete or partial replacement of concurrent controls with historical controls.

Step 2: Prospectively establish a systematic search plan for selecting a set of historical trials,[4] considering consistency of trial design & conduct, as well as trends over time in standard of care (SOC) and placebo response. Pocock describes conditions that a historical trial must meet before it can be used in this manner.[5] These criteria are stringent; for areas of high unmet need, consideration should be given to whether they can be relaxed.

Step 3: If subject-level historical control data are available, consider refinement of the historical control set based on pertinent criteria (e.g., intended inclusion/exclusion criteria or geographical location).

Step 4: Pre-specify analysis method to incorporate the historical controls, which may include down-weighting the historical controls versus concurrent controls in case of discordance, or selecting a subset of controls (e.g., using propensity scores) to match the population eventually recruited in the ongoing trial (see Review of Bayesian Approaches and Review of Frequentist Propensity Score Approaches sections for further information). Give consideration to the handling of missing data if subject-level historical control data are available. Also, it is self-evident that historical control data from completed randomized clinical trials (RCTs) are no longer considered to be RCT data, but are considered non-randomized, high-quality observational data once they are removed from their original context and used in the concurrent trial.

Step 5: Consider inclusion of an interim analysis, which may allow possible adaptation of the ongoing trial (e.g., changing sample size or randomization allocation, depending on the consistency of historical and concurrent controls; see Review of Bayesian Approaches section).

Step 6: Assess operating characteristics of the current trial design including impact of historical information. Simulations will probably be required (see Review of Methods section).

Figure 1. High-level steps for designing a study incorporating historical control data.

collected over the last 20 years to speed the delivery of medicines in areas of unmet medical need. Statistical methods—discussed in detail in this paper—provide us with confidence that a sound scientific basis exists for the use of historical data to inform clinical research. The broad application of these methods could have a profound impact in reducing patient burden and accelerating clinical research timelines. For example, between 2000 and 2015, there were 267 AD clinical trials that, in aggregate, enrolled over 150,000 subjects.[3] If we assume that one-third of those subjects were in control arms, and that by using historical control data we could conservatively reduce that number by 25%, we might have saved over 12,000 subjects from painful procedures that reproduced information that already existed while also accelerating clinical trial decision timelines.

Therefore, it is proposed that, with the proper understanding and matching of study design and demographic parameters, historical data can be used in a supplementary manner to reduce the number of concurrent control subjects required during late-phase clinical development. The use of historical data in confirmatory trials is a large and complex topic, and the criteria for deciding whether historical controls are appropriate for a particular trial and for selecting the historical controls to ensure they are comparable with those in the current trial is a topic that needs further exploration. Use of appropriate statistical methods for implementing historical control trial designs can help to reduce the risk of selection bias. The specific focus of this paper is a review of statistical methods for historical borrowing in drug development, with the aim of highlighting how the various approaches can help to balance the potential benefits and risks of including historical data in confirmatory trial.

Principles for the Use of Historical Controls in a New Clinical Trial

This section describes general principles to be considered when incorporating historical control data in a new trial. Suggested steps are outlined in Figure 1.

One can consider a continuum of approaches for incorporating historical controls into a clinical trial based on the severity of need and/or rarity of available subjects, ranging from single-arm studies where the only controls are obtained from historical trials for disease areas of very high unmet need through supplementation (but not complete replacement) of concurrent

Table 1. Key Regulatory References.

Reference Document	Key Message(s)
ICH E4: Dose Response Information to Support Drug Registration[9]	"Agencies should also be open to the use of various statistical and pharmacometric techniques such as Bayesian and population methods, modeling, and pharmacokinetic-pharmacodynamic approaches."
ICH E9: Statistical Principles for Clinical Trials[10]	"The use of Bayesian and other approaches may be considered when the reasons for their use are clear and when the resulting conclusions are sufficiently robust."
ICH E10: Choice of Control Group and Related Issues in Clinical Trials[11]	Guideline expresses a major concern about only using historical controls (ie, the inability to control bias), but also describes the usefulness of such controls under certain scenarios. Guideline describes situations where appropriately and carefully chosen historical controls are more persuasive and potentially less biased.
FDA: The Use of Bayesian Statistics in Medical Device Clinical Trials[12]	Guidance states that in some circumstances "the prior information for a device may be a justification for a smaller-sized or shorter-duration pivotal trial."
EMA: Concept Paper on Extrapolation of Efficacy and Safety in Medicine Development[13]	Paper proposes a framework to establish a systematic approach for extrapolation of efficacy and safety data from a source population to a target population.
EMA: Guideline on Clinical Trials in Small Populations[14]	A Bayesian methodology with an informative prior built on historical data may be suitable. Recommends sensitivity analyses for the choice of prior. "Under exceptional circumstances," historical controls with no concurrent control may be acceptable.
European Commission: Ethical Considerations for Clinical Trials on Medicinal Products Conducted with the Paediatric Population[15]	"Adaptive, Bayesian, or other designs may be used to minimize the size of the trial."

controls in disease areas where controls are easier to recruit and an acceptable standard of care (SOC) is available. Eichler et al proposed using historical control data to set a threshold as a benchmark for the primary analysis of a single-arm trial.[6] If the success threshold is crossed, the treatment is determined to be effective. If the futility threshold is crossed, the treatment is determined to be ineffective. If there is an intermediate outcome, then a further trial is performed; this would either be a randomized control trial (RCT) if practical or another single-arm trial if not. This approach could be applied broadly, particularly for disease areas where control arms are not feasible. When concurrent controls are not precluded by ethical or practical concerns, supplementation of concurrent controls with historical controls is preferable to complete replacement of concurrent controls because it allows assessment of the comparability of the historical data to the concurrent controls.

Key considerations when using historical data for confirmatory (eg, phase 3) trials:

- Inflation of type I error rate: A nonrandomized comparison may introduce bias and as a result potentially increase the type I error rate. Inflation of type I error rate in earlier-phase studies that are not used for registration is a risk to both the sponsor and study subjects. For example, the sponsor may continue a development program that should be terminated and expose study subjects to a non-effective treatment in future trials. Of greater concern, inflation of type I error rate in confirmatory studies could lead to increased risk of approving medicines that aren't effective. (Conversely, the bias may lead to a loss of power and a deflation of type I error rate, which could result in a failed trial.) Since the amount of bias can't be known, simulations should be conducted to examine "long-run" estimates of type I error rates in plausible scenarios (see Review of Methods section).
- Interaction with regulatory agencies: Discussions with regulatory agencies about the acceptability of this approach and the selection of historical controls should occur as early as possible and well in advance of a confirmatory trial, with a review of the proposed application of the historical data within the product development plan.

Table 1 lists key regulatory references relevant to the use of historical data in clinical trials. A survey carried out by the Drug Information Association (DIA) Bayesian Scientific Working Group of industry statisticians in 2012 identified "a lack of clarity of the regulatory position and/or lack of guidance" as one of the 4 main barriers to the implementation of Bayesian methodology.[7] In 2016, representatives from FDA's Centers for Drug Evaluation and Research (CDER), Biologic Evaluation and Research (CBER), and Devices and Radiological Health (CDRH) participated in a workshop "Substantial Evidence in 21st Century Regulatory Science: Borrowing Strength from Accumulating Data" that focused on methods to incorporate historical information.[8] This

Table 2. Options for Specifying a Prior Based on Historical Data.

Assumed Relationship Between p_H & p	Details
1. Equal	Assumes differences between observed historical & concurrent control response rates are solely attributable to sampling variation. Equivalent to pooling historical & concurrent controls.
2. Functional Dependence	Assumes differences between historical and concurrent control response rates can be explained by known covariates. Requires estimate of covariate-response relationship (eg, from historical data) which can be used to construct a predicted prior distribution for p based on historical control response rate and observed covariates for historical and concurrent controls.
3. Equal but Discounted	Assumes $p_H = p$ as in option 1, but discounts the historical information by inflating the variance of the historical prior (eg, using power prior).[21] Can also be thought of as reducing the effective sample size of the historical controls on which the prior is based. Amount of discounting is subjective and has no operational interpretation.
4. Biased	Assumes concurrent control response is a biased (shifted or rescaled) version of the historical response (ie, $p = p_H + \delta$ or $p = p_H \times \delta$). Similar to functional dependence (option 2), except that the precise form of dependence is unknown. Therefore, the dependence is captured by a generic bias term δ. Prior for p is constructed by combining the historical prior for p_H with a prior for δ which is typically chosen to reflect judgments about the relevance and quality of the historical study to the current setting.[22]
5. Exchangeable	Assumes p_H and p are "similar" (ie, assumes a distribution across studies with parameter σ^2 that reflects heterogeneity between historical and concurrent control response rates). Equivalent to a random effects meta-analysis of the historical and current trials.

indicates a growing acceptance by regulators of using historical controls for late-stage drug development.

Review of Bayesian Approaches in Confirmatory Clinical Trials With Historical Controls

Introduction to Bayesian Approaches

The Bayesian approach to evaluating new medicines is "the explicit quantitative use of external evidence in the design, monitoring, analysis, interpretation and reporting of a healthcare evaluation."[16] Prior distributions can be used to summarize information available from completed RCTs, registries, real-world evidence (see Bayesian Methodology section), and expert opinion. Even seemingly "objective" data-based priors will involve some element of subjectivity reflecting choices that were made about how to design the systematic review or otherwise specify the selection criteria which identified the historical data and choices about how to weigh them.[17]

Since Bayesian methods are now widely used in the early stages of drug development, the case for using a Bayesian modeling approach to draw strength from historical controls when both designing and analyzing any clinical trial is compelling. The Bayesian paradigm of formally quantifying current knowledge and then updating that knowledge in the light of new data fits perfectly with the idea that there is useful information contained within historical data available prior to a clinical trial. Lee and Chu identified 121 publications reporting a Bayesian analysis of a clinical trial; 54 of these publications described use of an informative prior,[18] leading us to speculate that many of these used historical data.

Guidelines exist for the reporting of a Bayesian analysis. For example, Spiegelhalter et al and Sung et al both provide helpful guidelines with a great deal of overlap.[19,20] When using priors that incorporate historical data, details of the historical data sources, the method used to identify these sources, and the weight assigned to each historical dataset need to be reported to ensure the reproducibility of the Bayesian analysis.

One of the advantages of working with subject-level historical control data (as opposed to aggregate-level data) is that they enable the analyst to characterize the prognostic effects of baseline covariates. This information can be used to formulate priors for prognostic effects in models used to analyze the proposed trial. Alternatively (as discussed in the Bayesian Methodology section), quantifying differences between the distributions of baseline covariates for historical and concurrent controls and estimating the covariate-response relationship can be used to inform our understanding of how these 2 subject groups may differ in terms of their underlying response rate, for example. This applies when covariates can be assumed to completely explain between-group differences (see Method 2 of Table 2), and, more generally, when they cannot, this information may still be used to inform the specification of a prior for the bias inherent in the existing data (see Method 4 in Table 2).

One may want to be assured that the results of a Bayesian analysis incorporating historical controls are relatively insensitive to the range of assumptions that are credible, or at least that any sensitivity to assumptions is well understood so that

decisions can be made with this knowledge. The Bayesian approach to incorporating historical controls, via an explicitly defined model that quantifies the relationship between these data and the data from the new clinical trial, is a transparent way to show the effects of the assumptions.

One may be cautious of Bayesian approaches for incorporating historical controls into confirmatory trials. In this case, reassurance could be provided by examples of compounds that have demonstrated efficacy and safety in phase 2 based on a Bayesian incorporation of historical controls into those clinical trials and subsequently go on to confirm these results in conventional phase 3 trials and receive marketing approval. That is, it may be necessary to show that the failure rate in a conventional phase 3 setting following a predominantly Bayesian phase 2 development plan is, at a minimum, no worse than when following a conventional development plan throughout. To generate this evidence, we need to encourage even greater use of Bayesian methods utilizing informative priors in phase 2. Over the long run, this could show that the Bayesian approach does not lead to unacceptable bias or inflation of frequentist error rates. Those sponsors who take the plunge should reap the reward of faster and cheaper routes to phase 3, with the secondary aim of generating the evidence that will allow wider exploitation of these savings in future phase 3 trials.

Bayesian Methodology

To illustrate the Bayesian paradigm for the inclusion of historical information, consider a clinical trial with a primary endpoint defined in terms of a dichotomous outcome (eg, responder vs non-responder), and focus on estimation of the true control response rate p. A Bayesian analysis requires specification of a prior probability distribution for p reflecting that which is currently known about the plausible values of p. The prior distribution for p is combined with information about the control response rate observed in the current trial to form an updated (posterior) distribution for p. This posterior is a weighted average of the information in the prior and the information in the current data, weighted by their relative precisions or sample sizes. Thus, a natural way of including historical controls in the analysis of a current trial is by using historical control data to construct a prior distribution for p. When a single historical study is available, the most direct way to do this is to use the sampling distribution of the response rate in the historical trial as the prior for p in the current trial. This turns out to be equivalent to pooling the historical and current trial data. The approach can be extended to multiple historical studies by pooling the historical studies and treating them as a single large historical trial. However, pooling historical and concurrent controls only seems justifiable under very specific and tightly controlled situations where it is reasonable to assume that the true underlying control rate in the population is the same in both historical and current settings.

Several other options are also available for specifying a prior based on historical data. These options reflect a range of different, and less stringent, assumptions about the relevance of the historical data and the relationship between the true control response rates, p_H and p, in the populations represented by the historical and current trials, respectively. These are summarized in Table 2, based on the structure proposed by Spiegelhalter et al.[16]

Options 2 through 5 in Table 2 are ways of discounting historical prior information. They are all mathematically related, but differ conceptually and in terms of the quantities that require subjective specification.[23] It is important to note that some degree of expert judgment is necessary for all these options. Indeed, it can also be argued that ignoring historical control information completely represents a very strong subjective judgment that the historical data are irrelevant and provide no useful information about the current setting. By requiring explicit specification of the assumed relationship between the historical and concurrent controls, the above approaches provide a valuable mechanism for formalizing the assumptions being made and provide a useful framework for sensitivity analyses to assess the impact of varying these assumptions.

For confirmatory trials, the choice of discounting method and the specification of subjective "tuning" parameters (eg, bias parameter for option 4 or the down-weighting factor for option 3) may also need to be guided by examination of frequentist operating characteristics (see later).

In an effort to introduce greater objectivity into the amount of discounting applied to the historical data, various "dynamic borrowing" methods have been proposed for constructing a prior based on historical data. Such methods allow the amount of historical information borrowed to depend on the agreement between the concurrent and historical control data. Of course, it is entirely possible that an observed divergence between historical control response rates and a single contemporary trial's control response rate is caused by simple random variability alone – and a "pure Bayesian" would say that it is not an issue at all. Nevertheless, the clinical trial practitioner operates in a world where the inclusion of historical information is likely to be viewed as suspect, and a world, moreover, that gives greatest weight to the most recent events. Therefore, some method of down-weighting the impact of historical data on the basis of observed divergence can be desirable. Table 3 summarizes the main approaches for dynamic borrowing within a Bayesian framework.

An attractive design option when using any of the dynamic borrowing methods is to plan for adaptive adjustment of the concurrent control sample size (Figure 1, step 5).[24,29,32] The target control sample size, N_c, is prespecified, and an interim analysis is carried out after n_c ($<N_c$) concurrent controls have been recruited. An interim posterior distribution for the control response rate is calculated by combining the historical and concurrent controls using one of the dynamic borrowing methods discussed above, and the amount of information contributed by the historical controls is then quantified in terms of an

Table 3. Dynamic Borrowing Approaches Within a Bayesian Framework.

Dynamic Borrowing Method	Details
Hierarchical meta analytic models[16,23,24]	Option 5 in Table 2; amount of borrowing depends on between-trial heterogeneity σ^2, which may be estimated from the data.
	Large differences between the concurrent and historical controls $\Rightarrow$ large $\sigma^2 \Rightarrow$ little borrowing of historical information, and vice versa.
	When only a few historical studies are available, σ^2 can be difficult to estimate. In this case, a weakly informative prior distribution for σ^2 is recommended (see Rhodes et al and Turner et al for derivation of evidence-based priors for σ^2 or Friede et al for priors reflecting judgments about degree of similarity of historical and concurrent controls).[22,25,26]
	The model can be defined in 2 stages: (1) define prior for p at design stage based on meta-analysis of historical controls (MAP prior); (2) at end of current trial, combine MAP prior with concurrent control data in standard Bayesian analysis.
Commensurate priors[27]	Alternative type of hierarchical model that assumes historical response rate p_H is a nonsystematically biased version of the current response rate p, rather than assuming that p_H and p are exchangeable (ie, drawn from the same distribution).
	Prior for p has mean $= p_H$ and variance $= \tau^2$, which is estimated from the data.
	Small $\tau^2 \Rightarrow$ little bias $\Rightarrow$ high "commensurability" between historical and concurrent control response rates $\Rightarrow$ strong borrowing from historical data and vice versa.
Power priors with estimated power parameter[28]	Extension of conditional power prior (Option 3 in Table 2), where power parameter is assigned a prior distribution rather than fixed in advance.
	Aim is to learn about amount of downweighting from the observed difference between historical and concurrent controls, but the method tends to heavily discount historical data unless a very informative prior is used for the power parameter.
	Some recent alternative methods to estimate power parameter allow greater borrowing when historical and concurrent controls are similar.[29,30]
Robust MAP priors[24]	Reflects hybrid of assumptions in Options 3 and 5 in Table 2. Historical control response rates are assumed to be exchangeable with concurrent control response rates, but are discounted.
	Prior for p is a weighted mixture of a historical MAP prior and a comparatively vague prior.
	Weights on each mixture component are updated based on relative likelihood of the concurrent control data under the historical MAP prior vs vague prior. Large divergence between concurrent & historical controls $\Rightarrow$reduced weight on historical MAP component and little influence on posterior distribution for p, and vice versa.
	"Discounting" (weight on vague component) has operational interpretation as the probability that the historical data is not relevant (Cromwell's rule).[31]

effective historical sample size, ESS_h. As divergence between the historical and concurrent interim controls increases, the contribution of the former to the interim posterior will be down-weighted and hence ESS_h will be reduced; in contrast, close agreement between historical and concurrent interim controls will result in more historical borrowing and hence larger ESS_h. ESS_h is then used to determine how many more controls will be randomized post interim analysis so that the final effective control sample size (historical + concurrent controls) is at least N_c. Because such an interim analysis would require unblinded data, appropriate steps would be needed (eg, analysis done by an Independent Data Monitoring Committee) to maintain the integrity of the trial.

Viele et al review several (mainly Bayesian) methods for incorporating historical controls into a current trial, including a dynamic borrowing method, and provide a comparison of their frequentist operating characteristics via a simulation study.[33] Their example study provides a useful illustration of the potential advantages and disadvantages of these methods and clearly demonstrates the trade-off between precision and type I error that is a feature of all historical borrowing designs. They divide the range of possible values for the true concurrent control response rate, p, into 3 regions: a "sweet spot" corresponding to values for p that are similar to the observed historical control response rate, and regions on either side of this where p is either somewhat smaller or somewhat larger than the observed historical control response rate. When p falls into the sweet spot, historical borrowing leads to increased power and precision with negligible inflation of the type I error rate and bias compared to an analysis of the current trial data alone. The regions outside of the sweet spot correspond to reduced power (when the true concurrent control response rate p corresponds to "worse" outcomes than the observed historical response rate) and type I error rate inflation (when p corresponds to "better" outcomes than historical).

Similar comparative simulation studies have been reported by several other authors,[29,34] while Wadsworth et al provides a comprehensive review of Bayesian and frequentist methods used to include historical information in pediatric trials.[35] While the latter review focuses on extrapolation of historical efficacy data

(usually from adult studies) rather than historical control information, many of the methods used are similar. No single method has yet emerged as consistently superior. For a confirmatory trial, it may be necessary to conduct a simulation tailored to the specific trial and available historical data to optimize the choice of method and specification of fixed values or priors for any "tuning" parameters (eg, heterogeneity variance, power parameter, mixture weights) in terms of the size of the sweet spot region and the trade-off between type I error rate/bias inflation and gain in power/precision, taking into consideration the likelihood of the true control response rate actually falling into each of the 3 regions (sweet spot, reduced power, or type I error rate inflation). Wadsworth et al[35] also refers to points that should be agreed between the regulator and the sponsor.

Review of Frequentist Propensity Score Approaches in Confirmatory Clinical Trials With Historical Controls

This section reviews frequentist methods for utilizing patient-level historical control data in combination with concurrent control data in clinical trials. It covers the situation when there are no concurrent controls as a special case. A potential risk of using data from nonrandomized sources such as historical control data is the influence of subject characteristics, both observed and unobserved. The propensity score method[36-38] plays a very important role in eliminating or reducing the potential bias in estimated effects obtained from nonrandomized comparative studies. Two examples from medical device studies illustrate the use of propensity scores in incorporating historical controls.[39,40] In both examples, because of lack of randomization, important differences between comparison groups at baseline were observed and the propensity score-matching method was implemented to control such bias. The propensity score is defined as the conditional probability of assignment to a particular treatment given a vector of observed covariates, and adjustment for the scalar propensity score is sufficient to remove bias due to all observed covariates. Traditional propensity score analysis compares one treatment group to one control group where the two are not from an RCT. The procedure starts with propensity score estimation. Then, the estimated propensity scores are used to balance the 2 treatment groups. Stuart and Rubin extend the propensity score method to compare a treatment group to multiple control groups in broader contexts than clinical trials.[41] A particularly important implementation of this extension is an RCT with a reduced concurrent control plus a historical control. This section discusses the traditional situation with historical controls only and the extended situation with both historical and concurrent controls.

Traditional Propensity Score Method (Historical Controls Only)

Logistic regression or probit models are often used to estimate propensity scores, where the propensity to treatment, a dichotomous criterion variable, is estimated from a set of baseline covariate measures. Depending on the sample size, the researcher may include all available baseline covariates in the analysis or start with either a forward or backward stepwise regression when sample size might not support the inclusion of all variables. Overfitting is not a concern in estimating a propensity score, and a large number of covariates do not introduce bias in treatment comparison.[42] Once the propensity score (ie, the estimated probability of being in the treated group) is obtained, it is important to assess balance in propensity score analyses. Methods for assessing the comparability of treated and control subjects in a propensity score–matched sample have been discussed elsewhere.[43,44]

The derived propensity score can be used for treatment comparison and related inference; some common methods include matching, stratification, inverse probability of treatment weights (IPTW), and covariate adjustment on propensity score (CAPS).[44-47] Pair matching requires the number of control subjects to be larger than the number of treated subjects; hence, it will not perform well when the treatment group is larger than or about the same in size as the control group. However, except for this limitation, matching seems to be preferable to the other 3 methods to achieve balance in the baseline covariates.[44] Given the context of this paper—that rich, high-quality historical control data exists—matching is therefore the method of focus.

Matching aims to form matched sets of treated and control subjects who share similar propensity scores to allow for comparability. The treatment effect on outcomes is then estimated in the matched sample consisting of all matched sets. The most common implementation of propensity score matching is one-to-one or pair matching without replacement. An important concept in matching is caliper width, which is a maximum allowable difference between propensity scores if 2 subjects are allowed to be matched. Choice of the caliper is a trade-off between variance and bias. Narrower calipers lead to better-matched subjects that correspond to less bias; however, it also leads to fewer matched subjects, resulting in higher variance. A caliper width equal to 0.2 of the standard deviation of the logit of the propensity score when estimating the difference in means (for continuous outcomes) and risk difference (for binary outcomes) has been suggested.[48]

Extended Propensity Score Method (Historical and Concurrent Controls)

Stuart and Rubin developed a method to compare a treatment group to multiple control groups while accounting for the difference between control groups in addition to adjusting for differences between treatment and control groups in the observed covariates.[41] Although the method was developed mainly outside of a clinical setting, the paper mentioned supplementing an RCT with historical control data as a potential application. The method may discard some concurrent control subjects, which can be difficult to justify in the context of a

confirmatory clinical trial. To adapt this method for clinical trials, we propose finding matches from the treatment arm for all the concurrent control subjects. This would be the case of infinite caliper.[41] Bias is not a concern because the treatment arm and the concurrent control arm are both from the same RCT. To estimate the propensity score, the concurrent control and historical control groups are pooled together as if from a large control group.

Denote the i th subject's outcome with $Y_i(1)$ had he/she been on treatment, and $Y_i(0)$ had he or she been on control. The goal is estimating the average treatment effect in the full treatment group. That is, we want to estimate the following quantity:

$$\tau = \frac{1}{N_t}\sum_{i\in \mathrm{T}}[Y_i(1) - Y_i(0)]$$

where T is the set of all subjects in the treated group.

Of course, we never observe $Y_i(0)$ in the treated group, so it will be imputed by observation from the control group subject matched to the treated subject on all covariates (ie, the subjects have similar propensity scores).

Assume in the confirmatory trial that there are N_t treated subjects with outcome Y_{ti} with $i = 1, \cdots, N_t$ and N_c concurrent control subjects with outcome Y_{cj} with $j = 1, \cdots, N_c$, and $N_t > N_c$.

Stuart and Rubin provided an algorithm to compare 1 treatment group and 2 control groups assuming the outcome variable is normally distributed.[41] The adaptation in the confirmatory clinical trial setting follows all steps of that algorithm but with revisions in the first step so that every primary control unit (ie, concurrent control subject) is matched with a treated unit (ie, active treatment subject) in contrast to the original method in which some primary control units may be discarded (Figure 2).

Discussion of Propensity Scores

Analyses utilizing propensity score methods can only account for measured baseline covariates and are still susceptible to bias due to unmeasured confounding covariates. To address this, sensitivity analyses allow one to assess how strongly an unmeasured confounder would have to be associated with treatment selection in order for a previously statistically significant treatment effect to become statistically nonsignificant if the unmeasured confounder had been considered.[36]

The propensity score method, both traditional and extended, applies to studies with one treatment arm. In practice, some confirmatory studies have more than one treatment arm (eg, 2 dose levels). For these situations especially, control of the Type 1 error rate needs to be considered. The propensity score estimation can be performed by either combining treatment arms as a single treatment arm or the procedure being repeated for each treatment arm. The latter may be preferred when matching is used because of its theoretical requirement of unlimited size for the historical control group.

There are some regulatory issues regarding application of the propensity score method in clinical trials.[52-55] The bias and confounding common in nonrandomized comparison trials include subject selection bias coming from physician judgment or subject preference, temporal bias caused by evolution of medical practice or technology, heterogeneity in subject population, differences in definition and adjudication of clinical outcomes, confounding by important baseline covariates, different lengths of follow-up, etc.

The propensity score method has seen increased use in recent years. However, the focus of the research work so far has generally been data analysis; there is a need for more research in study design work, as has been done for Bayesian approaches.

Review of Methods for Determining the Sample Size and Operating Characteristics of Comparative Clinical Trials That Include Historical Controls

The determination of sample size and operating characteristics (OC) is an important element in the planning of any clinical trial, especially any trial including historical controls. There are multiple approaches based on developing analytical approximations and using simulation. For example, there are approximate methods for determining the sample size for trials with a historical control group, assuming that the observed response rate of the historical control group is the true control response rate.[56,57] A uniform power method can be used to control the expected power, accounting for uncertainty in the historical control response rate,[58] and similar solutions can be used for continuous outcomes.[59] Other approaches are based on exact unconditional estimation,[60] a maximization approach,[61] and a flexible sample size formula for survival outcomes that controls arbitrary percentiles of the conditional power and type I error rate, conditional on the historical control response rate.[62] The approach with the greatest flexibility for determining the sample size is based on simulation.[63] In many cases (particularly many Bayesian applications), it is the only approach.

FDA guidelines on the use of Bayesian methods and adaptive designs in clinical trials of medical devices, drugs, and biologics have stressed the importance of investigating and reporting the OC of the chosen design or analysis.[12,64,65] Each guideline emphasizes the importance of demonstrating control of type I error rate with subtle differences. For example, the CDRH Bayesian guideline states that they "strive for reasonable control of the type I error,"[12] but the CDER/CBER adaptive design guideline states that the "primary statistical concern of an adequate and well-controlled trial study is to control the overall studywide type I error rate for all involved hypotheses."[64] The distinction is between approaches which are "well-calibrated" and those that are "perfectly calibrated."[66]

The context in which Bayesian methods are used to incorporate historical data is important. The motivation to use historical data is strongest in small orphan or rare disease

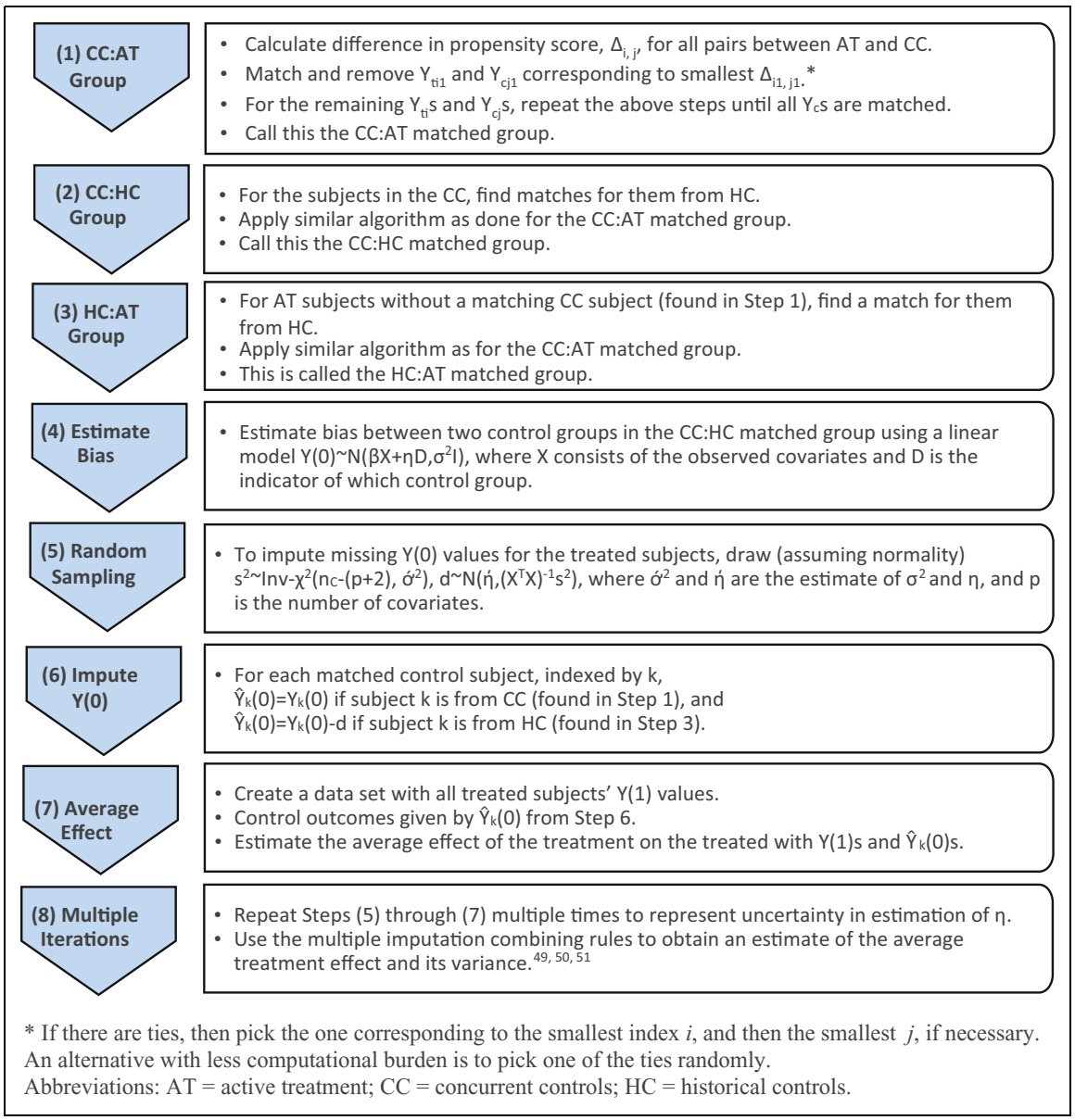

Figure 2. Algorithm to compare active treatment vs concurrent controls and historical controls.

populations, pediatric populations, and cases of high unmet medical need. In these cases, there is often a willingness of regulators to implicitly relax the level of type I error rate control. We can formally account for the relative consequences of type I and type II error rates,[67] potentially leading to larger type I error rates, or accept well-calibrated as sufficient.

It is important in planning clinical trial simulations to adhere to good practices.[68-71] Most of these practices stress the importance of a simulation plan and recommend its content include recording the randomization seed to facilitate replication, the choice of which may be controversial.[66,72] The simulations should cover a wide range of scenarios, and it may be advisable to consider formal experimental design methods in planning and analyzing the simulations.[66] The number of simulation replications should be adequate for the purpose of the investigation and should be justified.

Examples of Successful Use of Historical Controls Resulting in Regulatory Approval of New Treatments

There are several successful examples of drug approvals using historical control data in rare diseases, oncology, and other life-threatening and debilitating diseases (Table 4). These examples

Table 4. Historical Control Groups in Confirmatory Clinical Trials Leading to Regulatory Approval of New Treatments.

Disease Description and/or Indication Subject Population	1. Historical Control Group 2. Primary Efficacy Endpoint 3. Method	Drug (Approval Year) Clinical Impact of Drug
Acute hyperammonemia and associated encephalopathy Subjects with severe urea cycle disorders (UCDs) due to enzyme deficiencies	1. Untreated UCD subjects followed between 1975 and 1995 2. Overall survival 3. Survival rates were compared*	AMMONUL (sodium phenylacetate and sodium benzoate) (2005)[73-75] Adjunctive therapy
Pompe Disease, also known as acid alpha-glucosidase deficiency Subjects with infantile-onset disease	1. Untreated subjects diagnosed by age 6 mo, born between 1982 and 2002, were identified by a retrospective review of medical charts. 2. Ventilator-free survival at 18 mo of age 3. Proportions (95% confidence intervals [95% CIs]) of active-treated subjects who died or needed invasive ventilator support were compared with the mortality experience of the historical control group with similar age and disease severity.*	MYOZYME (alglucosidase alfa) (2006)[76-79] First treatment approved for any of >40 neuromuscular diseases covered by the Muscular Dystrophy Association
Toxic plasma methotrexate concentrations caused by delayed methotrexate clearance due to impaired renal function Mainly subjects with cancers such as osteosarcoma, leukemia, or lymphoma	1. Extensive data (>40 y of clinical trials) with well-characterized methotrexate excretion curves 2. Rapid ($\leq$15 min) and sustained ($\geq$8 d) clinically important reduction (CIR) in plasma methotrexate concentration 3. Point estimate and 95% CIs for the CIR rate (number [%] of subjects with a CIR)*	VORAXAZE (glucarpidase) (2012)[80-82] Reduction of toxic plasma methotrexate (a chemotherapeutic drug) concentrations
Relapsed or refractory acute lymphoblastic leukemia (ALL) Subjects with Philadelphia chromosome-negative relapsed or refractory B-cell precursor ALL	1. Historical data pooled from European national study groups and large individual sites across Europe and the United States 2. Complete remission (CR) rate, proportion of subjects with minimal residual disease (MRD) with partial hematological recovery, duration of response 3. Two analytical approaches: a. A weighted analysis with outcomes from the historical data set were weighted according to the frequency distribution of predetermined prognostic baseline factors in the blinatumomab clinical trial population b. Propensity score analysis balancing the populations with respect to important baseline factors and enabling quantification of differences in outcomes between the two groups. This was performed in response to regulatory feedback in 2015.	BLINCYTO (blinatumomab) (2014)[83-85] A second-line treatment
Hypophosphatasia (HPP), a rare genetic progressive metabolic disorder Subjects with perinatal/infantile- and juvenile-onset disease	1. Untreated subjects with similar clinical characteristics as the STRENSIQ-treated subjects 2a. Survival and ventilator-free survival in perinatal/ infantile-onset HPP 2b. Growth and bone health in juvenile-onset HPP subjects 3a. Proportions, hazard ratio (95% CIs), and Kaplan-Meier estimates 3b. Height and weight z-scores; proportion of responders*	STRENSIQ™ (asfotase alfa) (2015)[86-88] First approved treatment for HPP

(continued)

Table 4. (continued)

Disease Description and/or Indication Subject Population	1. Historical Control Group 2. Primary Efficacy Endpoint 3. Method	Drug (Approval Year) Clinical Impact of Drug
Lysosomal Acid Lipase (LAL) deficiency, a rare genetic metabolic disorder known as Wolman disease Subjects with rapidly progressive LAL deficiency presenting within the first 6 mo of life	1. Untreated subjects with a similar age at disease presentation and clinical characteristics as the KANUMA-treated subjects 2. Survival at 12 mo of age 3. Proportions, median age of survival*	KANUMA (sebelipase alfa) (2015)[89,90] First approved treatment for LAL deficiency
Neuronal ceroid lipofuscinosis type 2 (CLN2 disease) Subjects $\geq$3 y of age with symptomatic CLN2 disease	1. Independent historical control group with similar but not identical baseline characteristics as the active-treated subjects in the single-arm, open-label clinical trial 2. Clinician-reported outcome (ClinRo), the CLN2 rating scale (motor domain): Motor function (walking or crawling ability) was assessed using the motor domain of the CLN2 clinical rating scale, which could range from a score of 3 (normal) to a score of zero (profoundly impaired). 3. Efficacy conclusions were based on multiple analyses of the best-matched subjects in the two cohorts; the analyses accounted for several confounding factors (age, genotype, screening motor score).	BRINEURA (cerliponase alfa) (2017)[91] First approved treatment for CLN2 disease First enzyme replacement therapy (ERT) to use intracerebroventricular administration

*These examples, for the most part, did not use the conventional approach of hypothesis testing to compare the results between historical controls and the treatment arm in the current trial.

for the most part did not use conventional hypothesis testing to compare the results between historical controls and treatment arms in the current trial. The use of historical controls is beginning to gain momentum, both for supplementing as well as replacing control arms in confirmatory trials. However, there is some way to go in gaining broader acceptance of historical data in confirmatory trials, as so far the examples are only in less common disease areas. Furthermore, only one of the examples in Table 4 used a propensity score method to adjust for selection bias in the historical controls, and none used any of the Bayesian approaches reviewed in this paper. There are, however, several illustrative examples in the recent literature of how the types of historical borrowing methodologies discussed here *could* be applied in a late phase development for a broader range of disease areas. For example, Wandel and Roychoudhary discuss how Bayesian meta-analytic priors could be used to reduce the sample size requirements for a new phase 3 study in schizophrenia using historical data from two phase 2 studies and one previous phase 3 study[92]; Dejardin et al compare several of the dynamic borrowing methods summarized in Table 3 for including historical controls in the design of a phase 3 non-inferiority trial of a novel antibacterial agent.[93] In the near future, we expect and hope to see more examples of successful regulatory approvals in larger disease areas and/or larger trials, based on the approaches discussed in the Review of Bayesian Approaches and Review of Frequentist Propensity Score Approaches sections of this paper.

Conclusion

> All the forces in the world are not so powerful as an idea whose time has come.
>
> —Victor Hugo[94]

There continues to be a sense of urgency in developing medicines for patients in need. Patients, academics, drug development companies, and regulators are all incentivized to accelerate our ability to test new interventions for efficacy and safety while minimizing subject exposure. Regulators have a record of accepting historical control data for interventions for medical devices and/or indications with small populations.

The methods covered in this paper give us the tools to use fewer subjects in late-phase confirmatory clinical trials. It is our opinion that this is an idea whose time has come. The industry and regulatory science has matured to the point where high-quality data exists to support these approaches; the statistical methods have evolved to provide a robust understanding of risk; and our evolution to a patient-centric model demands that we leverage these methods more broadly. We encourage regulators, industry, and academia to develop a

framework for implementing these approaches more broadly in clinical research.

Acknowledgments

The authors gratefully acknowledge the support of TransCelerate BioPharma Inc, a nonprofit organization dedicated to improving the health of people around the world by accelerating and simplifying the research and development (R&D) of innovative new therapies. The organization's mission is to collaborate across the global biopharmaceutical R&D community to identify, prioritize, design, and facilitate implementation of solutions designed to drive the efficient, effective, and high-quality delivery of new medicines.

The authors also gratefully acknowledge the support of the following TransCelerate working team members who contributed to the development of the manuscript: Graeme Archer, GlaxoSmithKline; Matt Austin, Amgen; Julian Desmond, Amgen; Ryan Feld, Accenture; Daniel Jia, Allergan; Hassan Lakkis, Allergan; Jingyi Liu, Eli Lily; Frances Pu, Renaissance Writing Services; Jihao Zhou, Allergan; and Ray Zhu, Allergan.

Declaration of Conflicting Interests

No potential conflicts were declared.

Funding

Lisa Hampson's contribution was partly funded by the UK Medical Research Council (grant MR/M013510/1). Frances Pu's contribution was funded by TransCelerate BioPharma Inc.

References

1. Food and Drug Administration (FDA). *Innovation or Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products*. Rockville, MD: FDA; 2004.
2. Watson JL, Ryan L, Silverberg N, Cahan V, Bernard MA. Obstacles and opportunities in Alzheimer's clinical trial recruitment. *Health Affairs (Millwood)*. 2014;33:574-579.
3. ClinicalTrials.gov Web site. https://clinicaltrials.gov/ Accessed July 24, 2017.
4. Higgins JPT, Altman DG, Gøtzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*. 2011;343: d5928.
5. Pocock SJ. The combination of randomized and historical controls in clinical trials. *J Chronic Dis*. 1976;29:175-188.
6. Eichler HG, Bloechl-Daum B, Bauer P, et al. Threshold-crossing: a useful way to establish the counterfactual in clinical trials? *Clin Pharmacol Ther*. 2016;100:699-712.
7. Natanegara F, Neuenschwander B, Seaman JW, et al. The current state of Bayesian methods in medical product development: survey results and recommendations from the DIA Bayesian Scientific Working Group. *Pharm Stat*. 2014;13:3-12.
8. Substantial evidence in 21st century regulatory science: borrowing strength from accumulating data. Workshop sponsored by the American Course on Drug Development and Regulatory Sciences (ACDRS); April 21, 2016; Washington, DC. Available at: http://pharmacy.ucsf.edu/events/2016/04/evidence. Accessed July 19, 2017.
9. International Council on Harmonisation (ICH). *Dose Response Information to Support Drug Registration E4*. Geneva, Switzerland: ICH; 1994.
10. International Council on Harmonisation (ICH). *Statistical Principles for Clinical Trials E9*. Geneva, Switzerland: ICH; 1998.
11. International Council on Harmonisation (ICH). *Choice of Control Group and Related Issues in Clinical Trials E10*. Geneva, Switzerland: ICH; 2000.
12. Food and Drug Administration Center for Devices and Radiological Health (FDA CDRH). *Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials*. Rockville, MD: FDA CDRH; 2010.
13. European Medicines Agency (EMA). *Concept Paper on Extrapolation of Efficacy and Safety in Medicine Development—EMA/129698/2012*. London, UK: EMA; 2013.
14. European Medicines Agency (EMA). *Guideline on Clinical Trials in Small Populations*. London, UK: EMA; 2006.
15. Ad Hoc Group for the Development of Implementing Guidelines for Directive 2001/20/EC. Ethical Considerations for Clinical Trials on Medicinal Products Conducted with the Paediatric Population. 2008. http://ec.europa.eu/health//sites/health/files/files/eudralex/vol-10/ethical_considerations_en.pdf. Accessed July 19, 2017.
16. Spiegelhalter D, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health Care Evaluation*. Chichester, England: John Wiley & Sons, Ltd; 2004.
17. Lunn D, Jackson C, Best N, Thomas A, Spiegelhalter D. *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Boca Raton, FL: CRC Press, Taylor & Francis Group, LLC; 2013.
18. Lee JJ, Chu CT. Bayesian clinical trials in action. *Stat Med*. 2012;31:2955-2972.
19. Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. Bayesian methods in health technology assessment: a review. *Health Technol Assess*. 2000;4:1-130.
20. Sung L, Hayden J, Greenberg ML, Koren G, Feldman BM, Tomlinson GA. Seven items were identified for inclusion when reporting a Bayesian analysis of a clinical study. *J Clin Epidemiol*. 2005;58:261-268.
21. Ibrahim JG, Chen MH. Power prior distributions for regression models. *Statist Sci*. 2000;15:46-60.
22. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JP. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. *Int J Epidemiol*. 2012;41:818-827.
23. Neuenschwander B, Capkun-Niggli G, Branson MS, Spiegelhalter D. Summarizing historical information on controls in clinical trials. *Clin Trials*. 2010;7:5-18.
24. Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*. 2014;70:1023-1032.
25. Rhodes KM, Turner RM, Higgins JP. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *J Clin Epidemiol*. 2015;68:52-60.

26. Friede T, Röver C, Wandel S, Neuenschwander B. Meta-analysis of two studies in the presence of heterogeneity with applications in rare diseases. *Biom J*. 2017;59:658-671.
27. Hobbs BP, Sargent DJ, Carlin BP. Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Anal*. 2012;7:639-674.
28. Neelon B, O'Malley J. Bayesian analysis using power priors with application to pediatric quality of care. *J Biometrics Biostats*. 2010;1:1-9.
29. Bennett M, White SR, Mander AP. A novel equivalence probability weighting of historical data in an adaptive clinical trial design: A comparison to standard methods. MRC Biostatistics Unit, Cambridge, UK; In press.
30. Gravestock I, Held L; COMBACTE-Net consortium. Adaptive power priors with empirical Bayes for clinical trials. *Pharm Stat*. 2017;16:349-360.
31. Lindley DV. Cromwell's Rule. *Encyclopedia of Statistical Sciences*. 2006. doi:10.1002/0471667196.ess0622.pub2.
32. Hobbs BP, Carlin BP, Sargent DJ. Adaptive adjustment of the randomization ratio using historical control data. *Clin Trials*. 2013;10:430-440.
33. Viele K, Berry S, Neuenschwander B, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharm Stat*. 2014;13:41-54.
34. van Rosmalen J, Dejardin D, van Norden Y, Löwenberg B, Lesaffre E. Including historical data in the analysis of clinical trials: is it worth the effort? *Stat Method Med Res*. 2017 Jan 1: 962280217694506. In press.
35. Wadsworth I, Hampson LV, Jaki T. Extrapolation of efficacy and other data to support the development of new medicines for children: a systematic review of methods. *Stat Methods Med Res*. 2018;27:398-413.
36. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41-55.
37. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc*. 1984;79:516-524.
38. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat*. 1985;39:33-38.
39. Wachter R, Halbach M, Bakris GL, et al. An exploratory propensity score matched comparison of second-generation and first-generation baroreflex activation therapy systems. *J Am Soc Hypertens*. 2017;11:81-91.
40. Tarricone R, Boscolo PR, Armeni P. What type of clinical evidence is needed to assess medical devices? *Eur Respir Rev*. 2016; 25:259-265.
41. Stuart EA, Rubin DB. Matching with multiple control groups and adjusting for group differences. *J Educ Behav Stat*. 2008;33: 279-306.
42. Rubin DB. Estimating casual effects from large data sets using propensity scores. *Ann Intern Med*. 1997;127:757-763.
43. Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Med Decis Making*. 2009;29: 661-677.
44. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46:399-424.
45. Braitman L, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores. *Ann Intern Med*. 2002; 137:693-696.
46. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163:1149-1156.
47. Shadish WR, Steiner PM. A primer on propensity score analysis. *Newborn Infant Nurs Rev*. 2010;10:19-26.
48. Austin PC. Optimal caliper widths for propensity score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat*. 2011;10:150-161.
49. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc; 2002.
50. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley & Sons, Inc; 1987.
51. Rubin DB. On principles for modeling propensity scores in medical research. *Pharmacoepidemiol Drug Saf*. 2004;13:855-857.
52. Yue LQ. Statistical and regulatory issues with the application of propensity score analysis to nonrandomized medical device clinical studies. *J Biopharm Stat*. 2007;17:1-13.
53. Wright D, Day S. Discussion of: Statistical and regulatory issues with the application of propensity score analysis to nonrandomized medical device clinical studies. *J Biopharm Stat*. 2007; 17:15-17.
54. Yue LQ. Regulatory considerations in the design of comparative observational studies using propensity scores. *J Biopharm Stat*. 2012;22:1272-1279.
55. Levenson MS, Yue LQ. Regulatory issues of propensity score methodology application to drug and device safety studies. *J Biopharm Stat*. 2013;23:110-121.
56. Makuch RW, Simon RM. Sample size considerations for nonrandomized comparative studies. *J Chronic Dis*. 1980;33: 175-181.
57. Dixon DO, Simon R. Sample size considerations for studies comparing survival curves using historical controls. *J Clin Epidemiol*. 1988;41:1209-1213.
58. Lee JJ, Tseng C. Uniform power method for sample size calculation in historical control studies with binary response. *Control Clin Trials*. 2001;22:390-400.
59. Zhang S, Cao J, Ahn C. Calculating sample size in trials using historical controls. *Clin Trials*. 2010;7:343-353.
60. Shan G, Moonie S, Shen J. Sample size calculation based on efficient unconditional tests for clinical trials with historical controls. *J Biopharm Stat*. 2016;26:240-249.
61. Lloyd CJ. A new exact and more powerful unconditional test of no treatment effect from binary matched pairs. *Biometrics*. 2008; 64:716-723.
62. Zhu H, Zhang S, Ahn C. Sample size considerations for historical control studies with survival outcomes. *J Biopharm Stat*. 2016;26: 657-671.

63. Hooper R. Versatile sample-size calculation using simulation. *Stata J*. 2013;13:21-38.
64. Food and Drug Administration (FDA). *Adaptive Design Clinical Trials for Drugs and Biologics*. Rockville, MD: FDA; 2010.
65. Food and Drug Administration (FDA). *Adaptive Designs for Medical Device Clinical Studies*. Rockville, MD: FDA; 2016.
66. Grieve AP. Idle thoughts of a well-calibrated Bayesian in clinical drug development. *Pharm Stat*. 2016;15:96-108.
67. Grieve AP. How to test hypotheses if you must. *Pharm Stat*. 2015; 14:139-150.
68. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Stat Med*. 2006;25: 4279-4292.
69. Gaydos B, Anderson KM, Berry D, et al. Good practices for adaptive clinical trials in pharmaceutical product development. *Drug Inform J*. 2009;43:539-556.
70. O'Kelly M, Anisimov V, Campbell C, Hamilton S. Proposed best practice for projects that involve modelling and simulation. *Pharm Stat*. 2017;16:107-113.
71. Smith MK, Marshall A. Importance of protocols for simulation studies in clinical drug development. *Stat Methods Med Res*. 2011;20:613-622.
72. Posch M, Maurer W, Bretz F. Type I error rate control in adaptive designs for confirmatory clinical trials with treatment selection at interim. *Pharm Stat*. 2011;10:96-104.
73. Summar ML, Dobbelaere D, Brusilow S, Lee B. Diagnosis, symptoms, frequency and mortality of 260 patients with urea cycle disorders from a 21-year, multicentre study of acute hyperammonaemic episodes. *Acta Paediatr*. 2008;97:1420-1425.
74. Food and Drug Administration Center for Drug Evaluation and Research (FDA CDER). Application Number: 20-645 Medical Review. http://www.accessdata.fda.gov/drugsatfda_docs/nda/2005/020645s000_MedR.pdf. Accessed July 19, 2017.
75. Ammonul [package insert]. *Scottsdale*, AZ: Medicis Pharmaceutical Corp; 2005.
76. Kishnani PS, Hwu WL, Mandel H, et al. A retrospective, multinational, multicenter study on the natural history of infantile-onset Pompe disease. *J Pediatr*. 2006;148:671-676.
77. Myozyme [package insert]. *Cambridge*, MA: Genzyme Corporation; 2014.
78. Genzyme Corporation. Myozyme approved in the US [International Pompe Assocation Web site]. http://www.worldpompe.org/index.php/news/170-myozyme-approved-in-the-us. Published April 28, 2006. Accessed July 19, 2017.
79. Quest Staff. FDA OKs lifesaving treatment for Pompe disease. Muscular Dystrophy Association (MDA) website. http://quest.mda.org/news/fda-oks-lifesaving-treatment-pompe-disease. Published April 27, 2006. Accessed July 19, 2017.
80. Food and Drug Administration Center for Drug Evaluation and Research (FDA CDER), Application Number: 125327Orig1s000 Summary Review. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2012/125327Orig1s000SumR.pdf. Accessed July 19, 2017.
81. Voraxaze [package insert]. West Conshohocken, PA: BTG International Inc; 2013.
82. Rattu MA, Shah N, Lee JM, Pham AQ, Marzella N. Glucarpidase (Voraxaze), carboxypeptidase enzyme for methotrexate toxicity. *PT*. 2013;38:732-744.
83. Blincyto [package insert]. Thousand Oaks, CA: Amgen Inc; 2017.
84. Gökbuget N, Dombret H, Ribera JM, et al. International reference analysis of outcomes in adults with B-precursor Ph-negative relapsed/refractory acute lymphoblastic leukemia. *Haematologica*. 2016;101:1524-1533.
85. Gökbuget N, Kelsh M, Chia V, et al. Blinatumomab vs historical standard therapy of adult relapsed/refractory acute lymphoblastic leukemia. *Blood Cancer J*. 2016;6:e473.
86. FDA approves new treatment for rare metabolic disorder. FDA website. https://wayback.archive-it.org/7993/20170404214922/https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm468836.htm. Published October 23, 2015. Accessed June 7, 2018.
87. Drug Trials Snapshots STRENSIQ. FDA website. https://www.fda.gov/Drugs/InformationOnDrugs/ucm476307.htm. Accessed July 19, 2017.
88. Strensiq [package insert]. Cheshire, CT: Alexion Pharmaceuticals, Inc; 2015.
89. FDA approves first drug to treat a rare enzyme disorder in pediatric and adult patients. FDA website. https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm476013.htm. Published December 8, 2015. Accessed July 19, 2017.
90. Kanuma [package insert]. Cheshire, CT: Alexion Pharmaceuticals Inc; 2015.
91. Food and Drug Administration Center for Drug Evaluation and Research (FDA CDER). Application Number: 761052Orig1s000 Summary Review. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2017/761052Orig1s000SumR.pdf. Accessed July 19, 2017.
92. Wandel S, Roychoudhury S. Designing and analysing clinical trials in mental health: an evidence synthesis approach. *Evid Based Ment Health*. 2016;19:114-117.
93. Dejardin D, Delmar P, Warne C, Patel K, van Rosmalen J, Lesaffre E. Use of a historical control group in a noninferiority trial assessing a new antibacterial treatment: a case study and discussion of practical implementation aspects. *Pharm Stat*. 2018;17:169-181.
94. Hugo V. https://www.brainyquote.com/quotes/quotes/v/victorhugo136258.html. Accessed July 19, 2017.