Received: 25 July 2019 | Revised: 22 November 2019 | Accepted: 1 December 2019

DOI: 10.1002/sim.8456

RESEARCH ARTICLE

Statistics in Medicine WILEY

Bayesian leveraging of historical control data for a clinical trial with time-to-event endpoint

Satrajit Roychoudhury[1] | **Beat Neuenschwander**[2]

[1]Pfizer Inc, New York, New York

[2]Novartis Pharma AG, Basel, Switzerland

Correspondence
Satrajit Roychoudhury, Pfizer Inc, New York, NY.
Email: satrajit.roychoudhury@pfizer.com

The recent 21st Century Cures Act propagates innovations to accelerate the discovery, development, and delivery of 21st century cures. It includes the broader application of Bayesian statistics and the use of evidence from clinical expertise. An example of the latter is the use of trial-external (or historical) data, which promises more efficient or ethical trial designs. We propose a Bayesian meta-analytic approach to leverage historical data for time-to-event endpoints, which are common in oncology and cardiovascular diseases. The approach is based on a robust hierarchical model for piecewise exponential data. It allows for various degrees of between trial-heterogeneity and for leveraging individual as well as aggregate data. An ovarian carcinoma trial and a non-small cell cancer trial illustrate methodological and practical aspects of leveraging historical data for the analysis and design of time-to-event trials.

KEYWORDS

hierarchical model, Historical data, meta-analysis, piecewise exponential model, prior distribution, time-to-event data

1 | INTRODUCTION

Historical data often provide valuable information for the design of a new study. For example, sample size calculations depend on variability and effect sizes from previous trials. However, current practice usually ignores historical data in the analysis of a new trial. In recent years, however, leveraging historical data in the analysis of clinical trials has been encouraged by the European Medicines Evaluation Agency (EMEA),[1] the US Food and Drug Administration (FDA) including the Prescription Drug User Fee Act VI,[2] and the 21st Century Cures Act.[3] Recent examples in various phases of drug and medical device development include French et al,[4] Hueber et al.,[5] and Campbell.[6]

Leveraging historical data is appealing to practitioners and regulators for reasons of improved efficiency (smaller and therefore faster trials) and ethics (fewer patients assigned to a less effective treatment). Yet this may be challenging and requires care. First, identifying relevant historical data should build on good judgment and collaboration by the various stakeholders of a clinical trial. A thorough review of the literature and other resources (eg, registry data) in a specific disease (eg, study population, definition of endpoint, data collection, etc.) is important. This is ideally done using systematic reviews techniques (Egger et al,[7] Cochrane Collaboration,[8] Rietbergen et al.[9]), and involving a third party or independent group may be useful.

Second, a principled statistical approach, which leverages the historical data while allowing for potential differences between historical and actual data, is needed. Various statistical methods have been proposed in this context: Pocock's bias model,[10] power priors (Ibrahim and Chen[11]), commensurate priors (Hobbs et al[12]), and meta-analytic-predictive (MAP)

priors (Spiegelhalter et al,[13] Neuenschwander et al.,[14] Schmidli et al[15]); for an overview, see Viele et al[16] and Lewis et al.[17] The methods are very similar and use hierarchical models that allow for different parameters in the historical and actual trial, which will imply discounting of the historical data when analyzing the new trial.

Here, we will be concerned with leveraging historical control data for time-to-event endpoints (eg, time-to-disease progression, time to death). Such endpoints are the primary outcome in various therapeutic areas, including oncology and cardiovascular diseases. Among the various methods we will use the MAP approach. This is essentially a Bayesian random-effects meta-analysis of the historical data with the prediction of the parameter in the new trial. The basic model, which assumes exchangeable parameters for the new and the historical trials, can be extended in various ways. We will extend the *MAP* methodology for one-dimensional parameters to piecewise exponential (PWE) time-to-event data. And, to hedge against potential prior-data conflict, we propose an extension of the basic exchangeability assumption by adding a robust mixture component.

Although using historical data in clinical trials has gained increasing interest recently, applications in the time-to-event setting are sparse. A major challenge is the need of patient-level data for most of the statistical time-to-event approaches (Murray et al.,[18] Bertsche et al,[19] and Hobbs et al [20]). Our methodology is applicable for both patient-level and summary data (eg, Kaplan-Meier curves), the latter often being available from publications.

The paper is structured as follows. In Section 2 we introduce the basic *MAP* methodology and extend it to time-to-event data. Two applications are discussed in Section 3, with emphasis on the analysis of trial data in the presence of historical data, and the design of a new trial leveraging historical data, respectively. Section 4 concludes with a discussion.

2 | METHODS

In this section we discuss a MAP approach to leveraging historical data for a time-to-event outcome. We consider the randomized setting, with a control and test treatment for which the use of historical data is confined to the control group. That is, while the prior distribution for the control (baseline) hazard will be informed by historical data, the prior for the treatment contrast (proportional hazards parameter) will be weakly informative. This methodology is also applicable when the proportional hazard assumption is questionable.

2.1 | The MAP prior distribution

The *MAP* approach uses the historical control data from J trials $Y_1, \dots, Y_J$ to obtain the *MAP* prior distribution for the control parameter $\theta_\star$ in the new study

$$p(\theta_\star | Y_1, \dots, Y_J), \tag{1}$$

The derivation of Equation (1) is meta-analytic, using a hierarchical model with trial-specific parameters $\theta_\star, \theta_1, \dots, \theta_J$. The simplest hierarchical model assumes exchangeable parameters across trials, which is usually represented as

$$\theta_j = \mu + \epsilon_j, \quad \epsilon_j \sim N(0, \tau^2), \quad j = \star, 1, \dots, J. \tag{2}$$

While Equation (2) allows for biases, they are assumed nonsystematic. If needed, the basic model can be extended by replacing μ by $X_j\beta$ to allow for systematic biases explained by trial-specific covariates X_j.

2.2 | A hierarchical model for PWE time-to-event data

We assume PWE data and the time axis partitioned into K intervals,

$$(I_{k-1}, I_k], \quad k = 1, \dots, K, \quad I_0 = 0. \tag{3}$$

For each historical trial and time interval, there are n_{jk} control patients at risk (with a corresponding total exposure time E_{jk}), of which r_{jk} experience an event. For each interval, the data model is Poisson

$$r_{jk}|\lambda_{jk} \sim \text{Poisson}(\lambda_{jk}E_{jk}), \quad j = 1, \ldots, J; \; k = 1, \ldots, K, \tag{4}$$

where λ_{jk} is the hazard in interval k of trial j. Of interest is the prior for the control hazards $\lambda_{\star 1}, \ldots, \lambda_{\star K}$ in the new trial. The similarity of the new and the historical trials is captured in a parameter model. The simplest model assumes normally distributed log-hazard parameters $\theta_{jk} = \log(\lambda_{jk})$,

$$\theta_{\star k}, \theta_{1k}, \ldots, \theta_{Jk}|\mu_k, \tau_k \sim N(\mu_k, \tau_k^2), \quad k = 1, \ldots, K. \tag{5}$$

For the across-trial mean parameters μ_k, different implementations are possible. First, ignoring the time structure, the parameters may be assumed unrelated, with independent prior distributions

$$\mu_k \sim N(m_{\mu k}, s_{\mu k}^2), \quad k = 1, \ldots, K. \tag{6}$$

Second, the time structure of $\mu_1, \ldots, \mu_K$ may be modeled using, for example, a multivariate normal distribution with a structured covariance matrix (eg, AR(1)), or a dynamic linear model

$$\mu_k \sim N(\eta_k, \sigma_k^2), \quad k = 1, \ldots, K. \tag{7}$$

$$\eta_k = \eta_{k-1} + \rho_{k-1} \quad k = 2, \ldots, K. \tag{8}$$

The latter will be used in the applications of Section 3, which gives more details for the prior distributions.

Finally, prior distribution for the across-trial SDs τ_k on the log-hazard scale are required. We will use half-normal priors

$$\tau_k \sim \text{half-normal}(s_{\tau k}) \tag{9}$$

with scale parameter $s_{\tau k}$ representing anticipated heterogeneity. A weakly informative half-normal prior with $s_{\tau k}$=0.5 puts approximately 5% probability for values of τ_k greater than 1 (95% interval (0.02, 1.12)). This allows a wide (small to large) range of heterogeneity scenarios a priori. As for the mean parameters μ_k, rather than assuming unrelated τ_k parameters, a multivariate normal distribution or dynamic linear model could be used for the $\log(\tau)$ parameters.

From the above data model, parameter model, and prior distributions, the *MAP* prior for the vector of log-hazards in control group then follows as the conditional distribution of $\theta_{\star_1}, \ldots, \theta_{\star_K}$ given historical data,

$$p(\theta_{\star_1} \ldots, \theta_{\star_K}|r, E), \tag{10}$$

where r and E denote the number of events and exposure times across all historical trials and time intervals. The *MAP* prior can be obtained via markov chain monte carlo (MCMC) method; for a R and WinBUGS implementation, see Appendix B.

2.3 | Prior effective number of events

When desiging a new trial, knowing the amount of information introduced by the *MAP* prior is useful. It is sometimes expressed as the prior effective sample size (ESS), or, in the time-to-event setting, the effective number of events (ENE). Various methods have been suggested by Malec,[21] Morita et al,[22] Neuenschwander et al,[14] Pennello and Thompson.[23] They compare the information of the prior (variance or precision) to the one from one observation (or event). While conceptually similar, they can lead to suprisingly different *ESS* or *ENE*.

We will use an improvement, the *expected local-information-ratio* ESS_{ELIR} as introduced by Neuenschwander et al.[24] It is defined as the expected (under the prior) ratio of prior information $i(p(\theta))$ to Fisher information $i_F(\theta)$

$$\text{ESS}_{\text{ELIR}} = E_\theta \left\{ \frac{i(p(\theta))}{i_F(\theta)} \right\} \tag{11}$$

where $i(p(\theta)) = -d^2 \log p(\theta)/d\theta^2$ and $i_F(\theta) = -E_{Y_1|\theta}\{d^2 \log p(Y_1|\theta)/d\theta^2\}$.

It can be shown that, like the other methods, ESS_{ELIR} fulfills the necessary condition of being consistent with the well-known *ESS* for conjugate one-parameter exponential families. Unlike previous methods, however, it also fulfills a basic predictive criterion. That is, for a sample of size N, the expected posterior *ESS* (under the prior distribution) is the sum of the prior *ESS* and N.

For the PWE model with log-hazard parameters $\theta_{\star k}$ $(k = 1, \dots, K)$ in the new trial, the Fisher information $i_F(\theta_{\star k})$ for one event is 1; note that due to the possibility of censoring, the ENE rather than the ESS is used here. Since the *MAP* priors for the log-hazard parameters are available only as an MCMC sample, approximations for $p(\theta_{\star k})$ will be needed. Mixtures of standard distributions are convenient and can approximate these distributions with any degree of accuracy (Dalal and Hall,[25] Diaconis and Ylvisaker[26]). Fitting mixture distributions can be done by various procedures (eg, SAS,[27] R package RBesT[28,29]). Finally, after having obtained the ENE for each log-hazard parameter, the total ENE of the *MAP* prior follows as $\text{ENE} = \sum_{k=1}^{K} \text{ENE}_{\text{ELIR}}(\theta_{\star k})$.

In general, while the prior *ESS* may be useful to design a study with historical control data, the posterior *ESS* may be of interest as well. It follows the definition (11), with the prior $p(\theta)$ replaced by the posterior $p(\theta|Y)$. For PWE time-to-event analyses in particular, the posterior *ENE* is the sum of the interval-specific posterior *ENE*; for an example, see Section 3.1.

2.4 | Analysis of the data in the new trial

We now turn to the analysis of the new, randomized trial, where the control treatment (C) will be compared to a test treatment (T). For the PWE model, the control and treatment data in interval k follow Poisson distributions

$$r_{C\star k}|\lambda_{C\star k} \sim \text{Poisson}(\lambda_{\star k} E_{C\star k}), \quad r_{T\star k}|\lambda_{T\star k} \sim \text{Poisson}(\beta\lambda_{\star_k} E_{T\star k}), \tag{12}$$

where $\lambda_{\star k} = \exp(\theta_{\star k})$ are the control hazards, and β is the Cox proportional hazards parameter. Prior information for the control log-hazards $\theta_{\star k}$ are given by the *MAP* prior Equation (10), whereas the prior for β will usually be weakly informative.

When leveraging historical data, the analysis for the new trial can be done in two ways.

- Following the *MAP* approach, one would combine the *MAP* prior (10) from historical controls with the likelihood of the new data (12). This is complicated because the *MAP* prior for the K control log-hazards $\theta_\star$ is not known analytically but only available as an MCMC sample. Approximations of the *MAP* prior could be used; for example, by matching a multivariate normal distribution with the same means, SD, and correlations; or, mixture approximations to the *MAP* priors could be used. The latter would allow for accurate approximations but would also be technically more complex.
- Alternatively, the meta-analytic-combined (MAC) approach can be used. It consists of the hierarchical analysis of all the data and has been shown to be equivalent (Schmidli et al.[15]) to the *MAP* approach. Importantly, it is computationally easier than the two-step *MAP* approach and is therefore used here. For applications and their implementation, see Section 3 and Appendix A, respectively.

2.5 | Robust meta-analyses

To hedge against potential prior-data conflict, Schmidli et al.[15] proposed robust versions of *MAP* or *MAC* analyses. The idea builds on extending the standard parameter model (exchangeable parameters), allowing the control parameter in the new trial to be nonexchangeable with the historical parameters. The implementation uses a mixture distribution with weights w and $1 - w$ for exchangeability and nonexchangeablity, respectively.

This idea can be easily extended to the time-to-event setting as follows: for each interval k the robust version assumes exchangeablity, $\theta_{\star k} \sim N(\mu_k, \tau_k^2)$, with probability w_k and nonexchangeability with probability $1 - w_k$. For the latter, interval-specific prior distributions

$$\theta_{\star k} \sim N(m_{\theta k}, s_{\theta k}^2), \tag{13}$$

are needed. To achieve good robustness, weakly informative priors should be used (eg, approximate unit-information priors, ie, $s_k = 1$ for exponential data on the log-scale). Of note, even though the w_k are fixed, the Bayesian calculus ensures dynamic updating. That is, the prior weights w_k will be updated to posterior weights depending on the similarity of the new and historical data. Eventually one may be interested in extending this idea to all trials by introducing trial-specific mixture weights $w_{jk}, j = 1, \dots, J, k = 1, \dots, K$.

3 | APPLICATION

We now discuss two applications of leveraging historical time-to-event data. The first illustrates the joint analysis of historical and new data with the MAC approach of Section 2.4, whereas the second emphasizes the design of a new trial in the presence of historical data.

3.1 | Analysis with historical data

Voest et al[30] and Fiocco et al[31] investigated survival data from 10 studies on patients with advanced epithelial ovarian carcinoma. For our purpose, we assume the last study in Table 1 as the study of interest and the data of the remaining nine studies as the historical data. Thus, we want to infer the survival curve of the last study while leveraging the data from the other studies. The Kaplan-Meier (KM) plots in Figure 1 show considerable heterogeneity across the nine historical trials (left panel), with median survivals ranging from 1 to 2.9 years.

Fiocco et al[31] used the following 12 invervals (in years): [0-0.25], (0.25-0.50], (0.50-0.75], (0.75-1], (1-1.25], (1.25-1.50], (1.50-1.75], (1.75- 2.08], (2.08-2.50], (2.50-2.92], (2.92-3.33], and (3.33-4]. For the following meta-analytic analyses, the number of deaths and total exposure time per interval have been extracted from the published KM plots using the Parmar et al[32] approach: the total exposure time E_{jk} for interval $(I_{k-1}, I_k]$ and trial j is calculated as

$$E_{jk} = \frac{L_{jk}}{2} \times (r_{jk} + c_{jk}) + L_{jk} \times (n_{jk} - r_{jk} - c_{jk}); \qquad j = 1, \dots, 10, k = 1, \dots, 12$$

Here, L_{jk} is the interval length, and n_{jk}, r_{jk}, and c_{jk} are the number of patients at risk, dead, and censored, respectively. The data extraction assumes constant censoring rates for each interval and all deaths at mid-interval times. Table 1 summarizes the data for the 12 intervals in the 10 studies.

TABLE 1 Application 1: Number of deaths/exposure time for 12 intervals in ten studies

	Historical studies									
Interval (years)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**New study**
0.00-0.25	1/9.4	9/21.1	1/21.9	1/5.6	5/6.4	0/17.8	2/8.0	0/9.2	2/5.2	1/23.4
0.25-0.50	3/8.8	1/19.9	3/21.4	2/5.2	3/5.4	6/17.0	2/7.5	1/9.1	0/5.0	5/22.6
0.50-0.75	3/7.9	0/19.8	5/20.4	2/4.8	6/4.2	3/15.9	5/6.6	3/8.6	3/4.6	17/19.9
0.75-1.00	4/7.0	10/18.5	7/18.9	4/4.0	2/3.2	12/14.0	3/5.6	4/7.8	1/4.1	0/17.8
1.00-1.25	3/6.1	6/16.5	9/16.9	3/3.1	3/2.6	8/11.5	3/4.9	1/7.1	4/3.5	2/17.5
1.25-1.50	0/5.8	6/15.0	4/15.2	1/2.6	3/1.9	2/10.2	3/4.1	1/6.9	0/3.0	7/16.4
1.50-1.75	0/5.8	5/13.6	5/14.1	3/2.1	0/1.5	3/9.6	2/3.5	4/6.2	1/2.9	8/14.5
1.75-2.08	2/7.3	9/15.7	10/16.2	0/2.3	2/1.7	2/11.9	3/3.8	1/7.4	1/3.5	4/17.2
2.08-2.50	0/8.8	9/16.2	0/18.5	0/2.9	1/1.5	11/12.4	3/3.6	6/8.0	0/4.2	0/21.0
2.50-2.92	6/7.6	3/13.6	0/18.3	0/2.9	1/1.0	1/9.9	0/2.9	0/6.7	0/4.2	6/19.7
2.92-3.33	0/6.2	0/12.5	3/17.0	0/2.9	1/0.6	0/9.4	0/2.9	0/6.6	0/4.1	2/17.4
3.33-4.00	0/10.0	0/20.1	7/24.5	0/4.7	0/0.7	10/12.1	0/4.7	0/10.7	0/6.7	0/27.5

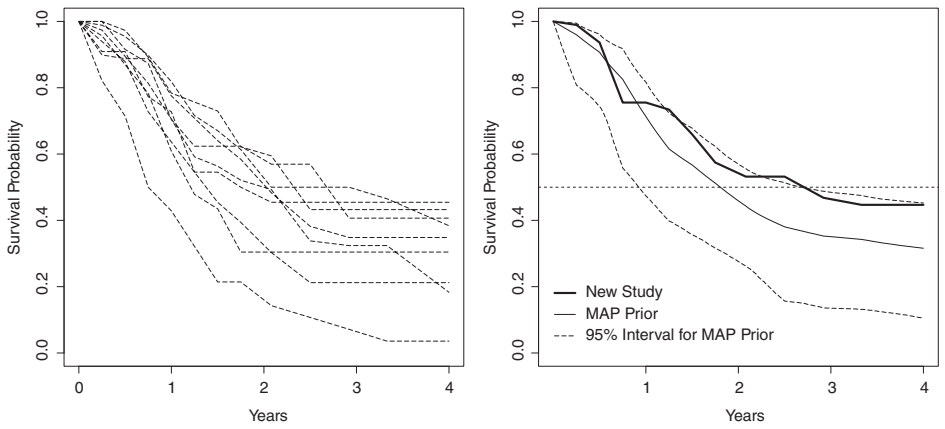

FIGURE 1 Application 1: Kaplan-Meier curves for the control group in nine historical studies (left panel), meta-analytic-predictive (MAP) prior for the new study (median (solid line), 95% interval (dashed line) in right panel), and Kaplan-Meier curve for the new study (thick solid line, right panel)

The prior information about the hazards in study ten is captured by the *MAP* prior. The respective prior for survival (median and 95%-intervals in the right panel of Figure 1) shows a median of approximately 1.8 years (95% interval 0.9-2.7 years). The prior ENE (Section 2.3) is 58. Figure 1 shows the KM plots of the nine historical studies (left panel) and MAP prior (right panel). The KM curve of the tenth trial (thick solid line in the right panel) is also shown in Figure 1.

After having access to the data from trial 10, for illustration we will compare two *MAC* analyses and the stratified analysis:

(i) *EX*: Full exchangeability *MAC* analysis. This is model (5) with between-trial standard deviations τ_k following half-normal priors with scale 0.5, which cover small to large heterogeneity (95% interval (0.02,1.12)). For η_1, a unit-information prior $N(-1.171, 1)$, centered at log(0.31) (the overall estimated log-hazard for death), is used. $N(0, 1)$ priors are used for ρ_i, $i = 1 \ldots k$.

(ii) *EXNEX*: Exchangeability-nonexchangeability *MAC* analysis. This is the robust analysis of Section 2.5, where for the tenth trial we assume weight 0.5 for exchangeability EX (5) and 0.5 for nonexchangeability *(NEX)* (13). For the remaining nine trials full exchangeability is assumed (weight=1). For *EX*, the prior assumptions in (i) are used, whereas for *NEX* the priors are weakly informative with means centered at the mean of the MAP prior and variances approximately worth one observation.

(iii) *STRAT*: This is the analysis of the data from the tenth study only, ignoring the data from the historical trials. For η_1, a noninformative prior $N(0, 10^2)$ is used. Similar to the EX model, $N(0, 1)$ priors are used for ρ_i, $i = 1 \ldots k$.

For the Fiocco data (Figure 2), the EX analysis shows the strongest borrowing, while under EXNEX borrowing is less because of the additional robust NEX component. Compared with STRAT, EX, and EXNEX deliver considerable precision gains except for the intervals [2.08-2.50) and [3.33-4.00), which may be due to the fact no events occurred during these intervals in the tenth trial.

For the three analyses, the posterior summaries for yearly survival rates in trial 10 are summarized in Table 2. The *EX* analysis provides smaller survival rates compared to the stratified analysis, and the estimated median survival is 2.01 years, considerably different from the observed median (2.7 years). On the other hand, the *EXNEX* analysis is more robust, with an estimated median survival of 2.51 years. The posterior ENEs are 96 and 76 for the EX and EXNEX analyses, respectively.

3.2 | Study design with historical data

The second application is concerned with the design of a randomized phase II trial for non-small cell lung cancer with historical control data. As of 2018, lung cancer is the most common cause of cancer-related death in men and women, responsible for 1.76 million deaths annually worldwide (WHO Cancer Factsheet[33]).

The aim of the phase II proof-of-concept study was to compare a new treatment *(T)* against an active control treatment *(C)* for patients with locally advanced recurrent or metastatic lung cancer. The primary endpoint was progression-free survival *(PFS)*, defined as the time from treatment assignment to disease progression or death from any cause. Superiority

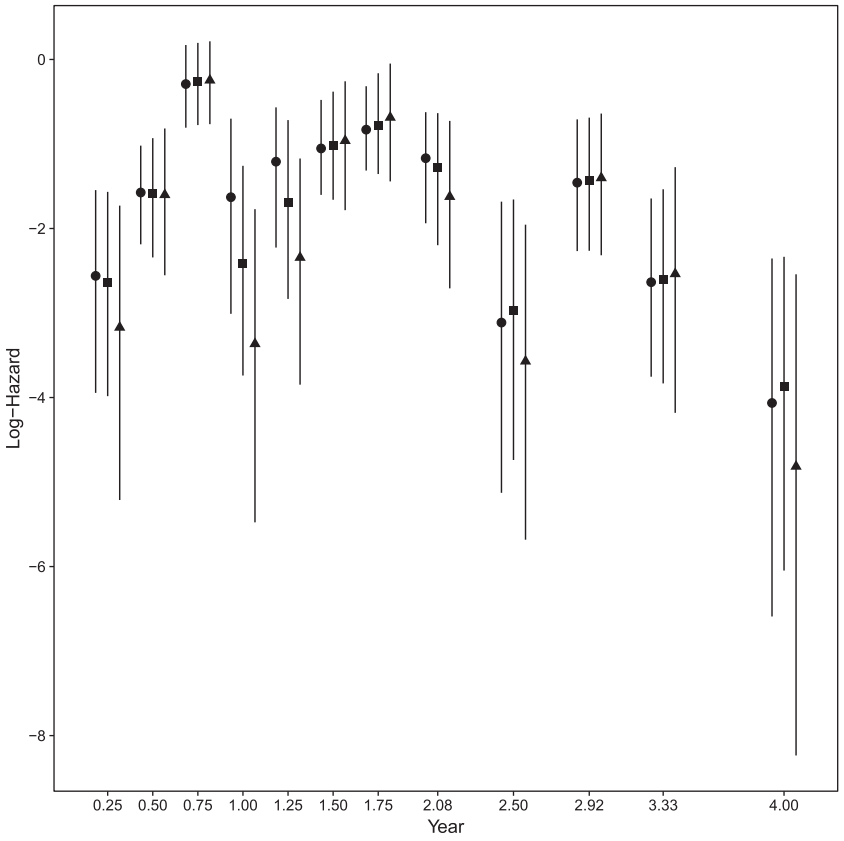

FIGURE 2 Application 1: Posterior medians and 95% intervals of log-hazard for 12 intervals for EX (circle), EXNEX (square), and stratified analysis (triangle) in the new study

TABLE 2 Application 1: Posterior median (95% interval) of yearly survival rates and median survival for EX, EXNEX, and STRAT in the new study

	EX	EXNEX	STRAT
Survival rate			
1 year	0.72 (0.57, 0.86)	0.74 (0.59, 0.86)	0.75 (0.60, 0.87)
2 year	0.50 (0.31, 0.71)	0.53 (0.31, 0.75)	0.54 (0.32, 0.77)
3 year	0.43 (0.22, 0.67)	0.45 (0.22, 0.71)	0.47 (0.23, 0.73)
4 year	0.41 (0.19, 0.67)	0.44 (0.20, 0.71)	0.44 (0.20, 0.71)
Median survival (years)	2.01 (1.59, 3.19)	2.51 (1.68, 5.05)	2.71 (1.69, 39.86)

of the treatment T is typically established by a statistically significant log-rank test or an upper confidence limit of the Cox proportional hazard-ratio (HR) less than 1.

For the control treatment, the median *PFS* for this population is approximately 5 months. Assuming a 45% reduction in *PFS* ($HR = 0.55$) for the new treatment, a 2:1 (T:C) randomization, an enrollment rate of 30 patients per month, a one-sided 2.5% level of significance and 90% power, the study would require 133 events (230 patients). Due to limited resources, however, it was decided to leverage historical data and enroll only 130 patients. For this modified design, the final analysis was planned after 110 events.

After an extensive search, seven historical studies with data for the active control were identified by the clinical team and disease area experts. Only KM plots were available from the literature. Figure 3 shows the KM plots as well as the number of events and exposure time for the following intervals (in days): (0-30], (30-60], (60-90], (90-120], (120-150], (150-180], (180-240], (240-300], and (300-360]. The data extraction from the KM plots was done by a method similar to Parmar et al[32] (for details see Appendix A). Figure 3 shows the rather heterogeneous historical studies (medians vary between 3.5 and 6 months) and the MAP prior (median= 5.12 months; 95% interval (1.6, 6.2 months)). The prior effective number of events (Section 2.3) is 36.

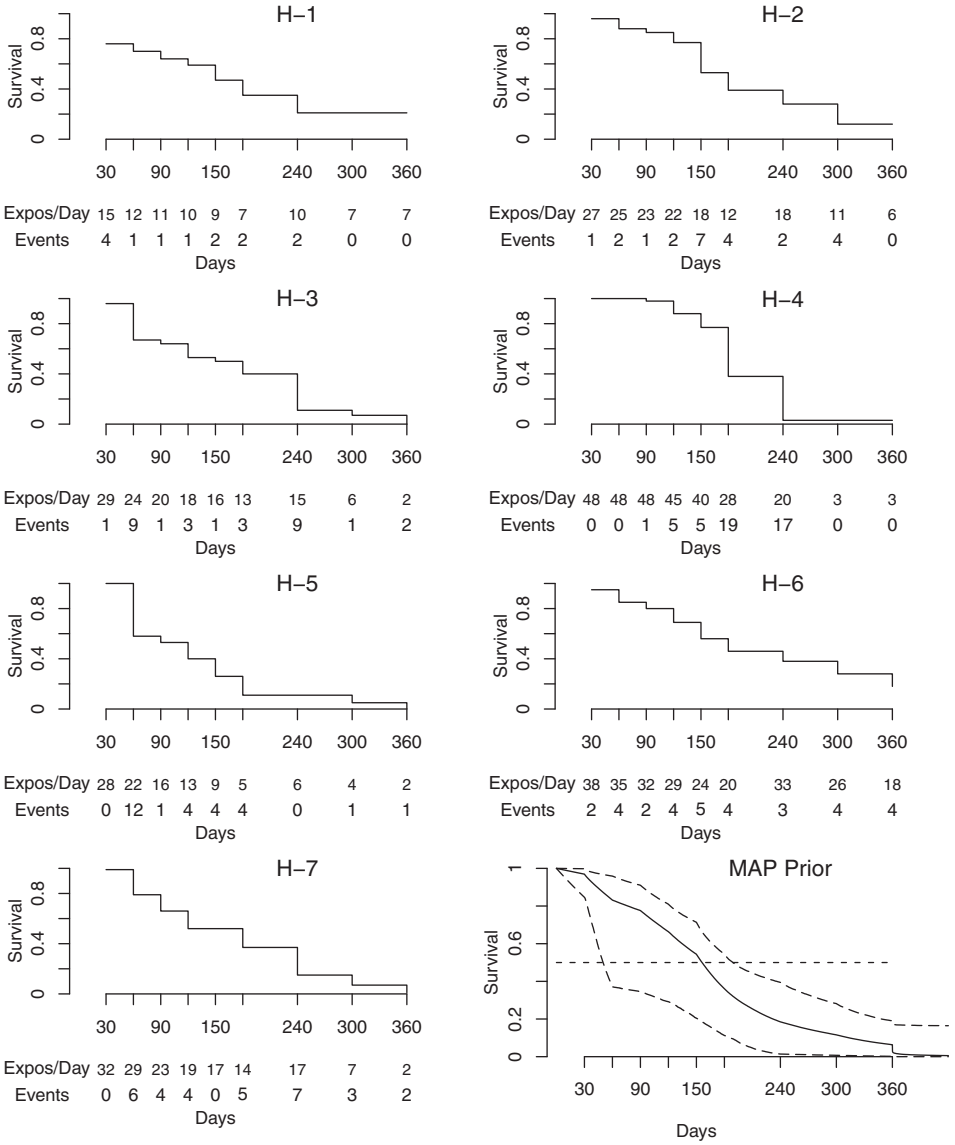

FIGURE 3 Application 2: Kaplan-Meier curves for seven historical studies and meta-analytic-predictive (MAP) prior (median and 95% intervals) for the new study

The frequentist operating characteristics (type-I error, power, bias, and root-mean-square error of the log-hazard ratio) of the design were assessed for exponential data, two scenarios for the HR (1 and 0.5), and various scenarios for the control median (3.5, 4.5, 5, 5.5, 6.5, and 7.5 months). Study success was defined by the Bayesian criterion $P(\text{HR} < 1|\text{data}) > 0.975$ or equivalently $P(\beta < 0|\text{data}) > 0.975$. For 2000 simulated trials, the above metrics were assessed for four models:

(i) *EX*: Full exchangeability *MAC* analysis. This is model (5) with between-trial SDs τ_k following half-normal priors with scale 0.5, which cover small to large heterogeneity (95% interval (0.02,1.12)). For η_1, a unit-information prior $N(-5.167, 1)$, centered at log(0.00575) (the overall estimated log-hazard for death), was used. $N(0, 1)$ priors were used for ρ_i, $i = 1 \ldots k$.

(ii) *EXNEX90*: exchangeability-nonexchangeability *MAC* analysis. This is the robust analysis, with weights 0.9 for exchangeability EX (5) and 0.1 for nonexchangeability *(NEX)* (13), which reflects high confidence in the historical data. For *EX*, the prior assumptions in (i) were used, whereas for *NEX* the priors were assumed as weakly informative with means centered at the mean of the MAP prior and variances equal to 1 (worth one observation).

(iii) *EXNEX50*: same as *EXNEX90* but with 50-50 weights for exchangeability and nonexchangeability.

(iv) *STRAT*: the analysis of the data from the new study only, ignoring the data from the historical trials and assuming $N(0, 10^2)$ priors for the log-hazards in each interval.

TABLE 3 Application 2: Type-I error and power (%) for different control medians (3.5 to 7.5 months) and treatment effects (hazard ratio = 1 or 0.55) for EX, EXNEX90, EXNEX50, and STRAT analysis

Scenarios	Control	Treatment	EX	EXNEX90	EXNEX50	STRAT
	Median	Median				
	HR=1 (type-I error)					
1	3.5	3.5	< 0.01	< 0.01	< 0.01	2.6
2	4.5	4.5	3.7	2.7	2.1	2.5
3	5.0	5.0	5.3	5.1	4.0	2.5
4	5.5	5.5	9.0	8.3	5.9	2.8
5	6.5	6.5	16.3	14.5	9.4	2.7
6	7.5	7.5	23.7	20.9	13.2	2.6
	HR=0.55 (power)					
7	3.5	6.4	91.7	90.9	88.6	85.2
8	4.5	8.2	97.3	96.9	94.5	84.9
9	5.0	9.1	97.0	96.8	96.1	84.6
10	5.5	10.0	98.9	98.7	97.1	85.0
11	6.5	11.8	99.3	99.2	98.0	84.8
12	7.5	13.6	99.7	99.5	98.5	84.5

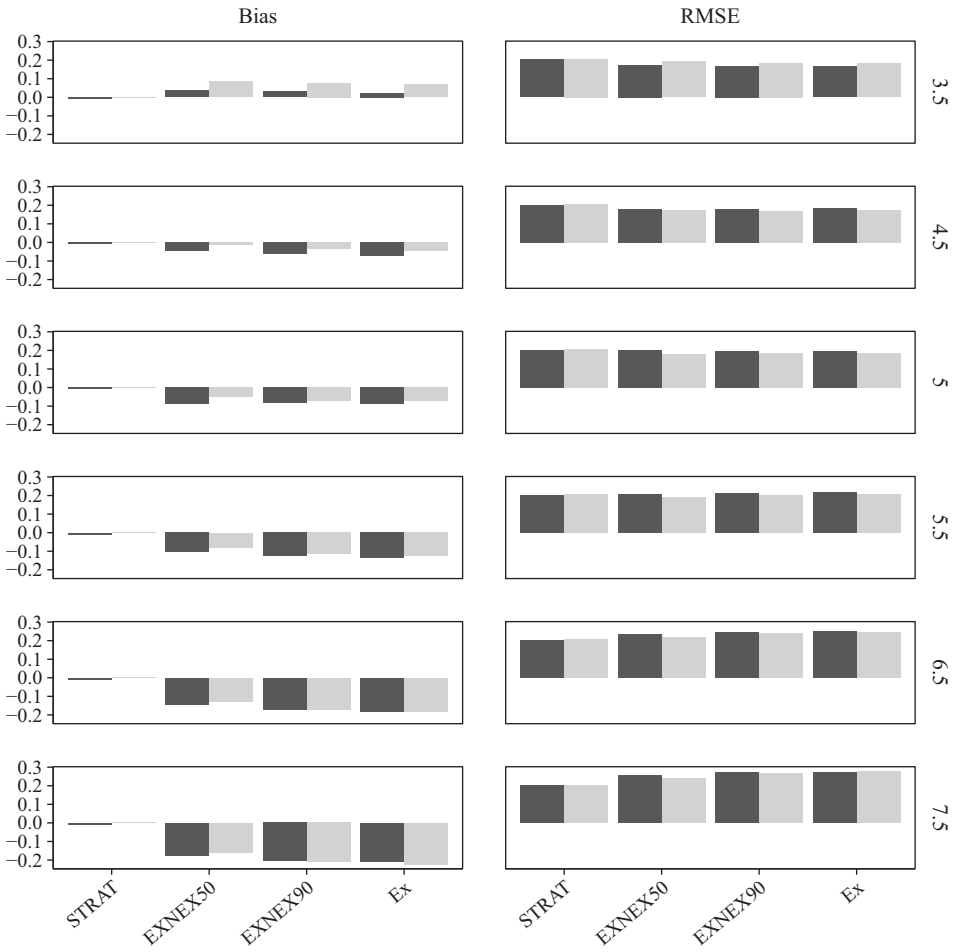

FIGURE 4 Application 2: Bias (left panel) and root-mean-square error (RMSE) (right panel) of log(HR) for different control medians (3.5 to 7.5 months) and hazard-ratios for STRAT, EXNEX50, EXNEX90, and EX analysis; black and gray bars show results for HR=1 and HR=0.55, respectively

Table 3 and Figure 4 summarize the operating characteristics for the four models.Table 3 shows the type-I error and power for different scenarios. For the type-I error (HR=1), *EX* and *EXNEX90* show fairly similar results. If the assumed control median is considerably smaller (3.5 months) than the one from the *MAP* prior ($\approx$ 5 months, Figure 3), the type-I error is much smaller than 2.5%. On the other hand, if the control median is much larger (6.5 or 7.5 months), the type-I error is larger than 10%. For the more robust *EXNEX50* model, however, the type-I error is much less affected. Finally, for the Bayesian stratified analysis, the type-I error is close to the expected 2.5%. For power (HR = 0.55), the analyses that leverage the historical data exhibit substantial gains compared to the stratified analyses, and power gains increase with increasing true control medians.

Regarding estimation of the log-HR parameter, the bias (left panel of Figure 4) increases with increased leveraging (from stratified to fully exchangeable) and true control medians much smaller (3.5 months) or much larger (7.5 months) than suggested by the historical data (5 months). For the nonrobust *EX* model, the bias is approximately –20% for the worst case scenario (true control median 6.5 months) but is generally much smaller for the robust analyses and true control medians closer to the expected 5 months.The root-mean-square error (on log-HR scale) (Figure 4) shows gain in efficiency in EX, EXNEX50, and EXNEX90 when the new trial data are aligned with historical data.

4 | DISCUSSION

Clinical trial results are usually interpreted in the context of other relevant data. In addition to such informal considerations of historical data, or more generally co-data (Neuenschwander et al.[34]), we have considered leveraging historical data in the Bayesian analysis of a new clinical trial with a time-to-event endpoint. Historical control data are increasingly used in earlier stages of drug development. As these studies primarily inform company-internal decisions, the advantages are often considered to outweigh potential risks. In phase III clinical trials, however, historical controls are currently rare. Yet the regulatory environment is evolving in special areas such as rare diseases, pediatric populations, and noninferiority trials (EMEA[35-37] and FDA[38]).

Leveraging historical control data has advantages. First, and most importantly, it allows randomizing fewer patients to the control group. This will shorten the duration of a clinical trial and hence lead to faster decisions. It will also decrease trial costs, as fewer patients will be needed. Second, if the control is ineffective (eg, for placebo), historical data designs may also be more ethical because fewer patients receive the ineffective treatment (Berry[39]).

The selection of historical data requires care and should follow recommendations for systematic reviews. This minimizes the risk of systematic biases, which can arise as a result of, for example, changes in standard of care over time, differences in inclusion/exclusion criteria, confounding environmental factors, or the evolution of diagnostic tools. The robust mixture approach proposed here mitigates some of these problems but cannot compensate for using biased data.

When analyzing the data of the new trial, a suitable model that allows for different degrees of similarity between historical and new data, is needed. Hierarchical models are the most obvious choice. For PWE time-to-event data, we have discussed a flexible model with interval-specific exchangeable hazard parameters and between-trial SDs. For few trials, which is typical for most settings, information on between-trial variability will usually be sparse. Yet valuable information may be available from similar disease settings, which may lead to more informative prior distributions for the between-trial SDs (Turner et al[40,41]). In the applications of Section 3, due to lack of additional information, we have used prior distributions that cover the range of small to large heterogeneity. Finally, the selection of mixture weights for exchangeability should be justified by judging the relevance of the historical data or by tuning the type-I error and power of the design. The former may be sufficient in exploratory trials whereas the latter appears more relevant for trials with more stringent error control requirements.

Widening the scope to real-world evidence, leveraging more but possibly less relevant data may be of interest. This will require adjustments to the methodology of Section 2 with regard to bias and heterogeneity. Having access to relevant predictors that explain anticipated biases will be key, and including these predictors (via meta-regression or propensity score methods) will be needed. Moreover, if historical data are of different quality, this may be accounted for by using different between-trial standard deviations. Adjustments for anticipated systematic biases (via meta-regression or propensity score methods Lim et al[42]) will require having access to relevant predictors as well as individual patient data. Further research will be needed to incorporate such extensions in the time-to-event setting.

ORCID

Satrajit Roychoudhury https://orcid.org/0000-0003-4001-3036

REFERENCES

1. European Medicines Agency. Innovative medicines initiative 2: Europe's fast track to better medicines; 2014.
2. US Food and Drug Administration. *PDUFA Reauthorization Performance Goals And Procedures Fiscal Years 2018 Through 2022*. 2018.
3. US Congress. 21st Century Cures Act (Public Law 114–255, 130 STAT 1033–1344); 2016.
4. French Jacqueline A, Temkin Nancy R, Shneker Bassel F, et al. Conversion to monotherapy: first study using a historical control group. *Neurotherapeutics*. 2012;9(1):176-184.
5. Hueber Wolfgang SBE, Steve L, et al. Secukinumab, a human anti-IL-17A monoclonal antibody, for moderate to severe Crohn's disease: unexpected results of a randomised, double-blind placebo-controlled trial. *Gut*. 2012;61(12):1693-1700.
6. Campbell G. Bayesian methods in clinical trials with applications to medical devices. *Commun Stat Appl Methods*. 2017;24(6):561-581.
7. Egger M, Smith GD, Altman DG. *Systematic Reviews in Health Care: Meta-Analysis in Context*. London: BMJ Publishing Group; 1995.
8. European Medicines Agency: Committee for Medicinal Products for Human Use (CHMP). *Points to Consider in Application with 1. Meta-analysis; 2. One Pivotal Study; CPMP/EWP/2330/99*; 2001.
9. Rietbergen C, Klugkist I, Janssen KJM, Moons KGM, Hoijtink HJA. Incorporation of historical data in the analysis of randomized therapeutic trials. *Contemp Clin Trials*. 2011;32(6):848-855.
10. Pocock SJ. The combination of randomized and historical controls in clinical trials. *J Chronic Dis*. 1976;29:175-188.
11. Ibrahim JG, Chen M-H. Power prior distributions for regression models. *Stat Sci*. 2000;15(1):46-60.
12. Hobbs BP, Carlin BP, Mandrekar SJ, Sargent DJ. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*. 2011;67(3):1047-1056.
13. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-care Evaluation*. New York, NY: John Wiley & Sons; 2004.
14. Neuenschwander B, Capkun-Niggli G, Branson M, Spiegelhalter DJ. Summarizing historical information on controls in clinical trials. *Clin Trials*. 2010;7(1):5-18.
15. Heinz S, Sandro G, Satrajit R, Anthony O'H, David S, Beat N. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*. 2014;70(4):1023-1032.
16. Kert V, Scott B, Beat N, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharm Stat*. 2014;13(1):41-54.
17. Lewis CJ, Sarkar S, Zhu J, Carlin BP. Borrowing from historical control data in cancer drug development: a cautionary tale and practical guidelines. *Stat Biopharm Res*. 2019;11(1):67-78.
18. Murray TA, Hobbs BP, Lystig TC, Carlin BP. Semiparametric Bayesian commensurate survival model for post-market medical device surveillance with non-exchangeable historical data. *Biometrics*. 2014;70(1):185-191.
19. Bertsche A, Fleischer F, Beyersmann J, Nehmiz G. Bayesian phase II optimization for time-to-event data based on historical information. *Stat Methods Med Res*. 2017;4:1-18.
20. Hobbs BP, Carlin BP, Sargent DJ. Adaptive adjustment of the randomization ratio using historical control data. *Clin Trials*. 2013;10(3):430-440.
21. Malec D. A closer look at combining data among a small number of binomial experiments. *Stat Med*. 2001;20(12):1811-1824.
22. Morita S, Thall PF, Müller P. Determining the effective sample size of a parametric prior. *Biometrics*. 2008;64(2):595-602.
23. Pennello G, Thompson L. Experience with reviewing Bayesian medical device trials. *J Biopharm Stat*. 2007;18(1):81-115.
24. Neuenschwander B, Weber S, Schmidli H, O'Hagan A. Predictively consistent prior effective sample sizes. *Biometrics*. 2019; (forthcoming).
25. Dalal SR, Hall WJ. Approximating priors by mixtures of natural conjugate priors. *J R Stat Soc B*. 1983;45(2):278-286.
26. Diaconis P, Ylvisaker D. Quantifying prior opinion. Bayesian Statistics 2. In: Bernardo J, DeGroot M, Lindley D, Smith A, eds. *Proceedings of the Second Valencia International Meeting September 6-10 1983*. North-Holland, Amsterdam: Elsevier; 1985:133-156.
27. SAS Institute. *SAS User Guide: Statistics. The FMM Procedure*. Cary, NC: SAS Institute Inc; 2014.
28. Weber Sebastian. RBesT: R Bayesian evidence synthesis tools. R package version 1.4-0; 2019.
29. Sebastian W, Yue L, John S, Tomoyuki K, Heinz S. Applying meta-Analytic predictive priors with the R Bayesian evidence synthesis tools. *arXiv*. 2019; e-prints.
30. Voest EE, Houwelingen Van JC, Neijt JP. A meta-analysis of prognostic factors in advanced ovarian cancer with median survival and overall survival (measured with the log (relative risk)) as main objectives. *Eur J Cancer Clin Oncol*. 1989;25:711-720.
31. Fiocco M, Putter H, Houwelingen Van JC. Meta-analysis of pairs of survival curves under heterogeneity: a Poisson correlated gamma-frailty approach. *Stat Med*. 2009;28(30):3782-3797.
32. Parmar MKB, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Stat Med*. 1998;17(24):2815-2834.
33. *WHO cancer factsheet*. *World Health Organization*; 2018.
34. Neuenschwander B, Roychoudhury S, Schmidli H. On the use of co-data in clinical trials. *Stat Biopharm Res*. 2016;8(3):345-354.
35. European Medicines Agency. Committee for medicinal products for human use (CHMP). *Guideline on Clinical Trials in Small Populations*; 2006.
36. European Medicines Agency: Committee for Proprietary Medicinal Products (CPMP). *Note for Guidance on Clinical Investigation of Medicinal Products in the Pediatric Population*. 2001.

37. European Medicines Agency: Committee for Medicinal Products for Human Use (CHMP). *Guideline on the Choice of the Non-Inferiority Margin*. 2006.
38. US Food and Drug Administration. *Draft Guidance for Industry: Interacting with the FDA on Complex Innovative Trial Designs for Drugs and Biological Products*. 2019.
39. Berry DA. Bayesian statistics and the efficiency and ethics of clinical trials. *Stat Sci*. 2004;19(1):175-187.
40. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgin PT. 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.
41. Turner RM, Jackson D, Wei Y, Thompson SG, Higgins PT. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Stat Med*. 2015;34:984-998.
42. Jessica L, Rosalind W, Jiacheng Y, et al. Minimizing patient burden through the use of historical subject-level data in innovative confirmatory clinical trials: Review of methods and opportunities. *Ther Innov Regul Sci*. 2018;52(5):546-559.
43. Mark Mitchell, Baurzhan Muftakhidinov, Winchen Tobias, et al. *Engauge Digitizer Software*. 2019.
44. Sturtz S, Ligges U, Gelman A. R2WinBUGS: a package for running WinBUGS from R. *J Stat Softw*. 2005;12(3):1-16.
45. Martyn P, Nicky B, Kate C, Karen V. CODA: convergence diagnosis and output analysis for MCMC. *R News*. 2006;6(1):7-11.
46. Therneau Terry M. A Package for Survival Analysis in S; 2015.
47. Venables WN, Ripley BD. *Modern Applied Statistics with S*. 4th ed. New York, NY: Springer; 2002.

How to cite this article: Roychoudhury S, Neuenschwander B. Bayesian leveraging of historical control data for a clinical trial with time-to-event endpoint. *Statistics in Medicine*. 2020;1–12. https://doi.org/10.1002/sim.8456

APPENDIX A: EXTRACTION OF DATA FROM PUBLISHED KM PLOTS

For the seven historical trials in Application 2 (NSCLC example), we could only get the KM plots from the published articles. The data extraction comprises the following steps:

1. The time axis is divided into intervals of 30 days up-to 180 days and 60 days afterward. Based on historical data the maximum follow up time is restricted to 360 days.
2. Probabilities of remaining progression-free at prespecified time points are extracted using Engauge Digitizer,[43] a freeware that converts an image to numbers.
3. The number of patients at risk in a specific interval is calculated by multiplying the total number of patients with the probability of remaining progression-free at the beginning of the interval.
4. The number of events (progression or death) within a time interval are approximated by using the probabilities of remaining progression-free in the current and next interval as follows;

$$\log(S(\hat{t_{i+1}})) - \log(S(\hat{t_i})) = 1 - \frac{d_i}{n_i},$$

   where d_i is the number of events and n_i is the risk set for interval $[t_i, t_{i+1})$
5. The exposure time for an interval is the interval length if the patient did not have an event or was not censored. Otherwise, the exposure of the patient is half the interval length. The total exposure time for an interval is the sum of the individual exposure times at the risk set.

APPENDIX B: WINBUGS CODE FOR MAC ANALYSIS

The R and WinBUGS code are available at https://github.com/roychs04/MAPSURV. It contains the main WinBUGS code and R wrapper function to reproduce the results for Fiocco data example in Section 3.1. The material also contains detailed illustrations of MAP prior derivation, EX, and EXNEX analysis. Note that the R-packages RBesT,[28] R2WinBUGS,[44] CODA,[45] survival,[46] and MASS[47] are required.