Supplementary material for "Including historical data in the analysis of clinical trials: Is it worth the effort?"

Joost van Rosmalen*[1], David Dejardin[2], Yvette van Norden[3], Bob Löwenberg[4], and Emmanuel Lesaffre[5]

[1] *Department of Biostatistics, Erasmus University Medical Center, Rotterdam, the Netherlands*

[2] *F. Hoffmann-La Roche, Basel, Switzerland*

[3] *HOVON Data Center, Department of Hematology, Erasmus MC Cancer Institute, Rotterdam, the Netherlands*

[4] *Department of Hematology, Erasmus University Medical Center, Rotterdam, the Netherlands*

[5] *Interuniversity Institute for Biostatistics and Statistical Bioinformatics, KU Leuven, Leuven, Belgium*

Calculating the scaling constant of the modified power prior

In this appendix we describe the algorithm that was used to obtain the posterior results of the modified power prior (MPP). This algorithm is based on the principle of path sampling [1].

Friel and Pettitt [2] describe algorithms for calculating the marginal likelihood of a model using *power posteriors*. The goal of their method is to calculate the marginal likelihood (i.e. $\int_\theta L(\theta|H)^\alpha p(\theta)\,\mathrm{d}\theta$, with $\alpha = 1$). In their algorithm, they introduce the weight α as an auxiliary variable, only

*Correspondence to: Department of Biostatistics, Erasmus MC, P.O. Box 2040, 3000 CA Rotterdam, the Netherlands. E-mail address: j.vanrosmalen@erasmusmc.nl.

for computational purposes. It is straightforward to adapt their algorithm to also calculate $C(\alpha) = \int_\theta L(\theta|H)^\alpha p(\theta)\,\mathrm{d}\theta$ for all other values of α.

The main idea behind the algorithm of Friel and Pettitt is that the logarithm of $C(\alpha)$ is equal to an integral of the log-likelihood, where the integration is done with respect to α. Friel and Pettitt (p. 594) showed that $\frac{\mathrm{d}}{\mathrm{d}\alpha}\log(C(\alpha)) = \mathbf{E}_{p(\theta|\alpha,H)}\left(\log(L(\theta|H))\right)$, i.e. the expected log-likelihood of the historical data as a function of parameters that are sampled from the power prior $L(\theta|H)^\alpha p(\theta)$. This result yields that

$$\log\left(C(\alpha)\right) = \int_{\alpha^*=0}^{\alpha} \mathbf{E}_{p(\theta|\alpha^*,H)}\left(\log(L(\theta|H))\right)\,\mathrm{d}\alpha^*. \tag{1}$$

Our path sampling algorithm for calculating $C(\alpha)$ based on (1) consists of the following steps:

1. Choose Δ_α, the increase in α per iteration and n_{iter}, the number of MCMC samples per iteration. Initialise $\alpha = 0$, and initialise the model parameters using a draw from the prior $p(\theta)$.

2. Repeat the following until $\alpha \geq 1$:

   (a) Increase the value of α by Δ_α.

   (b) Sample n_{iter} MCMC iterations from the power prior distribution with weight α.

   (c) Calculate the average log-likelihood of the historical data using all the parameter sets sampled for the current value of α

3. Calculate the cumulative sum of the average log-likelihood values that were calculated in the last step, as a function of α.

4. $C(\alpha)$ is now proportional to the exponential of the cumulative sum calculated in the previous step.

This algorithm efficiently calculates $C(\alpha)$ for a number of values of α between 0 and 1. A prerequisite for this algorithm is that it is possible to sample from the power prior $L(\theta|H)^\alpha p(\theta)$, e.g. using MCMC sampling. In the MCMC sampling of Step 2b, the last MCMC sample of the previous value of α is used for the initial values of the parameters. If the step size of α is sufficiently small, the power prior distribution $L(\theta|H)^\alpha p(\theta)$ should remain approximately stable between successive values of α and a short burn-in phase may suffice for the MCMC sampler in Step 2B.

We implemented this path sampling method using a step size for α equal to $\Delta_\alpha = 0.01$ and $n_{iter} = 5000$ MCMC samples per iteration. After running the algorithm for $C(\alpha)$, the posterior results of the power prior can be computed by sampling from

$$p(\alpha, \theta|D, H) \propto \frac{1}{C(\alpha)} L(\theta|D) L(\theta|H)^\alpha p(\theta) p(\alpha), \tag{2}$$

using a Metropolis-Hastings algorithm and looking up the value of $C(\alpha)$ for every candidate value of α using the results of the first algorithm. This Metropolis-Hastings sampler requires that the value of $C(\alpha)$ is available for every α in [0,1], though the path sampling algorithm described above only calculates $C(\alpha)$ for a number of fixed values of α. However, inspection of the values of $\log(C(\alpha))$ showed that this function can be accurately approximated by a linear function. To calculate $C(\alpha)$ for values of α in between the points used in the path sampling method, we use linear interpolation of $\log(C(\alpha))$.

Model specification for the baseline hazard

To model the baseline hazard, we used the piecewise exponential specification given by Murray et al. [3]. The follow-up time was divided into K intervals and a constant baseline hazard was assumed within each interval. The likelihood of the data was given by

$$L(\theta|D, H) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[\exp\left(d_{ik}\mu_{ik} + (\min(\kappa_k, y_i) - \min(\kappa_{k-1}, y_i))\, e^{\mu_{ik}} \right) \right], \tag{3}$$

where y_i is the follow-up time (until death or censoring) for patient i, κ_k is the end of the interval k, $d_{ik} = 1$ if patient died in interval k and 0 otherwise, and μ_{ik} is the hazard rate for person i in interval k. Following Murray et al., we used $K = \max\left(5, \min(\frac{r}{8}, 20)\right)$ intervals, where r denotes the total number of deaths in the observed data (including both current and historical data), and the cut-offs between intervals were chosen using the empirical distribution of the observed event times. Cut-off κ_k, $k = 1, \ldots, K-1$ was chosen as the $100 \times \frac{k}{K}$th percentile of the observed follow-up times for patients who have died, with κ_0 chosen as 0 and κ_K as the maximum follow-up time. A correlated prior process was used for the baseline hazard, with $p(\alpha) = p(\alpha|\eta)p(\eta)$, where α describes the logarithm of the baseline hazard and η describes the smoothness of the baseline hazard. The prior $p(\alpha|\eta)$ is given

by

$$\alpha_1 \sim N(0, 10^4)$$
$$\alpha_k|\eta, \alpha_{k-1} \sim N(\alpha_{k-1}, \eta^2), k = 2, \ldots, K \quad (4)$$

and $p(\eta)$ is uniform on [0.01,100].

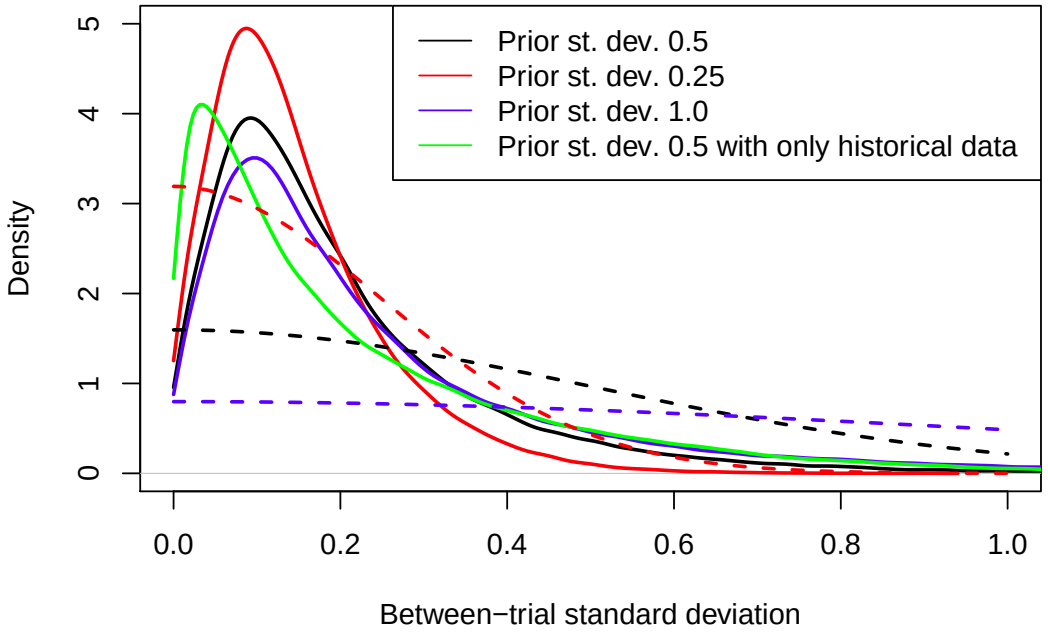
Additional tables and figures

The tables and figures below present additional results of the analysis of the HOVON data and the simulation study. Further explanation of these tables and figures is given in the main text of the article. The estimated sample size reduction in Table S4 is calculated as

$$1 - \frac{(\Phi(0.025) + \Phi(1 - P_c))^2}{(\Phi(0.025) + \Phi(1 - P_s))^2}, \quad (5)$$

where P_s denotes the estimated power of the specific method, P_c denotes the estimated power of the "Current data" analysis, and $\Phi()$ is the quantile function of the standard normal distribution. This formula is based on standard formulas for sample size calculations, which depend on the power and the significance level via the term $(\Phi(\alpha/2) + \Phi(1 - \beta))^2$, where α is the significance level and β the probability of a type II error. The estimated sample size reduction in Equation 5 is obtained by comparing the results of this term between the borrowing method and the "Current data" analysis, under the assumption that the type error rate is controlled at a 5% level. Note that these sample size reductions are only valid if the type I error rate is adequately controlled.

Figure S1: HOVON data: sensitivity of the posterior distribution (solid lines) of σ_η in the MAP approach to the prior distribution (dashed lines). The green line gives the posterior of σ_η using only the data of HOVON 29 and 42, which is the informative prior for the analysis of HOVON 42A.



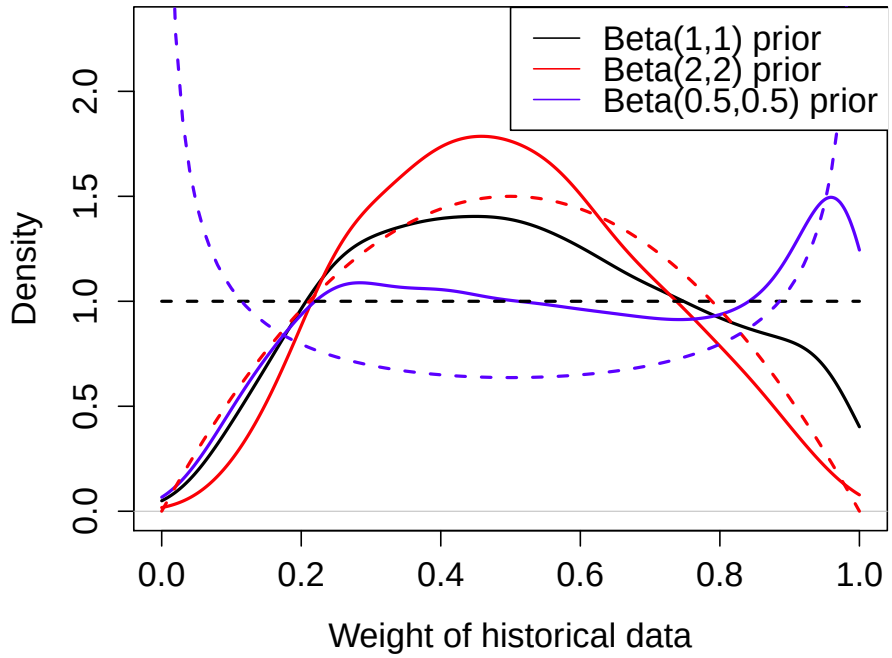
References

[1] A. Gelman and X.-L. Meng, "Simulating normalizing constants: From importance sampling to bridge sampling to path sampling," *Statistical science*, pp. 163–185, 1998.

[2] N. Friel and A. N. Pettitt, "Marginal likelihood estimation via power posteriors," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 3, pp. 589–607, 2008.

[3] T. A. Murray, B. P. Hobbs, T. C. Lystig, and B. P. Carlin, "Semiparametric Bayesian commensurate survival model for post-market medical

Figure S2: HOVON data: sensitivity of the posterior distribution (solid lines) of α in the MPP to the prior distribution (dashed lines)



device surveillance with non-exchangeable historical data," *Biometrics*, vol. 70, p. 185191, 2014.

Figure S3: Simulation study: Kaplan-Meier curves of a simulated data set in Scenario 1, including the population survival curves corresponding with the 2.5th and 97.5th percentile of the trial-specific effect

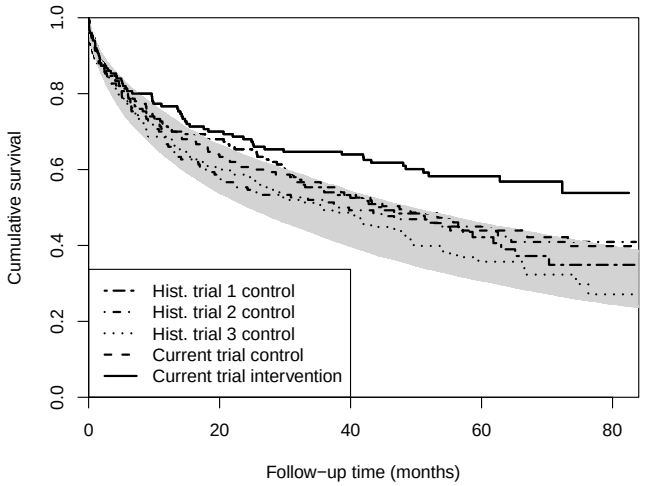


Figure S4: Simulation study: Kaplan-Meier curves of a simulated data set in Scenario 2, including the population survival curves corresponding with the 2.5th and 97.5th percentile of the trial-specific effect

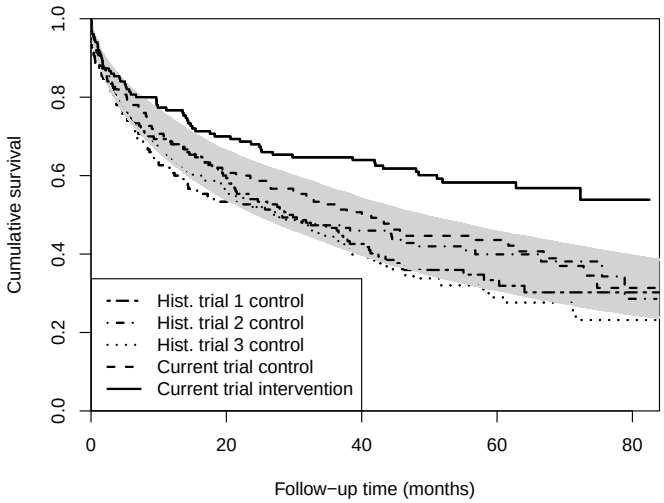


Figure S5: Simulation study: Kaplan-Meier curves of a simulated data set in Scenario 3, including the population survival curves corresponding with the 2.5th and 97.5th percentile of the trial-specific effect

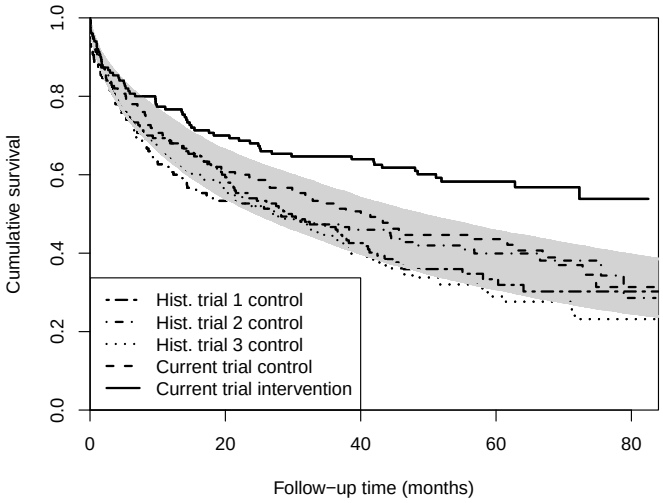


Figure S6: Simulation study: Kaplan-Meier curves of a simulated data set in Scenario 4, including the population survival curves corresponding with the 2.5th and 97.5th percentile of the trial-specific effect

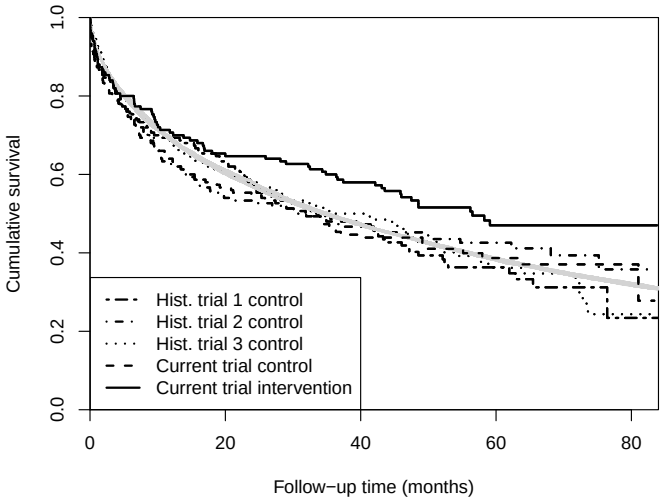


Figure S7: Simulation study: Kaplan-Meier curves of a simulated data set in Scenario 5, including the population survival curves corresponding with the 2.5th and 97.5th percentile of the trial-specific effect

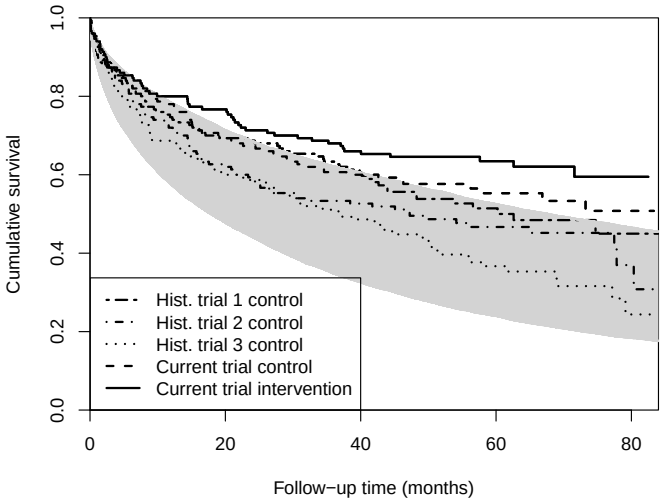


Figure S8: Simulation study: Kaplan-Meier curves of a simulated data set in Scenario 6, including the population survival curves corresponding with the 2.5th and 97.5th percentile of the trial-specific effect

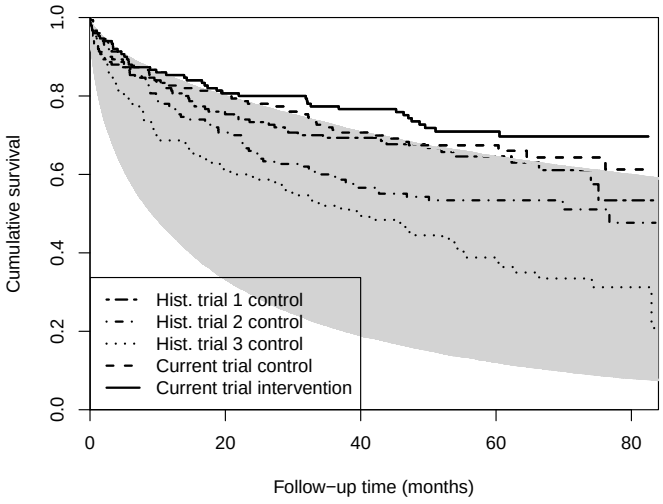


Table S1: Average bias (95% confidence interval) of estimated treatment effect in simulation study, based on 500 simulated data sets

Scenario	1	2	3	4	5	6
Time trend	No	Yes	Yes	No	No	No
Heterogeneity	Small	Small	Small	None	Moderate	Large
Current data	-0.018 (-0.031 - -0.005)	-0.018 (-0.031 - -0.005)	-0.018 (-0.031 - -0.006)	-0.019 (-0.032 - -0.005)	-0.017 (-0.030 - -0.004)	-0.018 (-0.031 - -0.005)
Pooled data	-0.018 (-0.030 - -0.005)	-0.095 (-0.108 - -0.083)	-0.016 (-0.029 - -0.003)	-0.020 (-0.031 - -0.009)	-0.018 (-0.036 - 0.000)	-0.018 (-0.049 - 0.014)
Pocock's method	-0.018 (-0.030 - -0.006)	-0.082 (-0.094 - -0.070)	-0.017 (-0.030 - -0.004)	-0.020 (-0.031 - -0.008)	-0.020 (-0.036 - -0.005)	-0.028 (-0.055 - -0.002)
Power prior with $\alpha = 0.5$	-0.017 (-0.029 - -0.006)	-0.080 (-0.092 - -0.068)	-0.017 (-0.029 - -0.004)	-0.019 (-0.030 - -0.008)	-0.018 (-0.033 - -0.002)	-0.019 (-0.045 - 0.008)
MPP	-0.016 (-0.028 - -0.004)	-0.069 (-0.081 - -0.058)	-0.015 (-0.028 - -0.003)	-0.018 (-0.030 - -0.007)	-0.015 (-0.029 - -0.001)	-0.017 (-0.033 - -0.001)
MAP approach	-0.018 (-0.030 - -0.006)	-0.043 (-0.055 - -0.031)	-0.017 (-0.030 - -0.005)	-0.020 (-0.032 - -0.008)	-0.018 (-0.030 - -0.005)	-0.020 (-0.033 - -0.007)
Robust MAP approach	-0.018 (-0.030 - -0.006)	-0.036 (-0.049 - -0.024)	-0.017 (-0.029 - -0.004)	-0.019 (-0.032 - -0.007)	-0.018 (-0.030 - -0.005)	-0.020 (-0.033 - -0.007)
Method of Murray et al.	-0.017 (-0.030 - -0.005)	-0.084 (-0.097 - -0.072)	-0.016 (-0.030 - -0.003)	-0.020 (-0.031 - -0.009)	-0.016 (-0.032 - -0.000)	-0.017 (-0.035 - 0.001)
Test-then-pool method	0.014 (0.002 - 0.026)	-0.008 (-0.021 - 0.004)	-0.009 (-0.022 - 0.004)	-0.001 (-0.013 - 0.010)	0.013 (-0.000 - 0.026)	-0.011 (-0.024 - 0.002)

Table S2: Average posterior standard deviation (95% confidence interval) of estimated treatment effect in simulation study, based on 500 simulated data sets

Scenario	1	2	3	4	5	6
Time trend	No	Yes	Yes	No	No	No
Heterogeneity	Small	Small	Small	None	Moderate	Large
Current data	0.158 (0.157 - 0.158)	0.158 (0.157 - 0.159)	0.158 (0.157 - 0.159)	0.158 (0.158 - 0.159)	0.158 (0.157 - 0.159)	0.160 (0.158 - 0.162)
Pooled data	0.129 (0.129 - 0.130)	0.129 (0.128 - 0.129)	0.147 (0.146 - 0.147)	0.129 (0.129 - 0.130)	0.129 (0.129 - 0.130)	0.131 (0.129 - 0.132)
Pocock's method	0.135 (0.134 - 0.135)	0.134 (0.134 - 0.135)	0.151 (0.151 - 0.152)	0.135 (0.135 - 0.135)	0.135 (0.134 - 0.136)	0.137 (0.135 - 0.138)
Power prior with $\alpha = 0.5$	0.135 (0.135 - 0.136)	0.135 (0.134 - 0.135)	0.151 (0.150 - 0.151)	0.135 (0.135 - 0.136)	0.135 (0.135 - 0.136)	0.137 (0.135 - 0.138)
MPP	0.137 (0.137 - 0.138)	0.139 (0.138 - 0.140)	0.152 (0.151 - 0.152)	0.136 (0.136 - 0.137)	0.143 (0.142 - 0.144)	0.153 (0.151 - 0.155)
MAP approach	0.148 (0.148 - 0.149)	0.149 (0.149 - 0.150)	0.155 (0.154 - 0.155)	0.145 (0.144 - 0.145)	0.153 (0.152 - 0.154)	0.158 (0.157 - 0.160)
Robust MAP approach	0.151 (0.150 - 0.152)	0.151 (0.151 - 0.152)	0.157 (0.156 - 0.157)	0.149 (0.149 - 0.150)	0.154 (0.153 - 0.155)	0.158 (0.157 - 0.160)
Method of Murray et al.	0.133 (0.132 - 0.134)	0.133 (0.132 - 0.134)	0.148 (0.147 - 0.149)	0.132 (0.132 - 0.133)	0.139 (0.137 - 0.141)	0.150 (0.147 - 0.153)
Test-then-pool method	0.143 (0.141 - 0.144)	0.150 (0.149 - 0.152)	0.155 (0.154 - 0.156)	0.137 (0.136 - 0.138)	0.151 (0.149 - 0.152)	0.159 (0.157 - 0.161)

Table S3: Average root mean square deviation (95% confidence interval) of estimated treatment effect in simulation study, based on 500 simulated data sets

Scenario	1	2	3	4	5	6
Time trend	No	Yes	Yes	No	No	No
Heterogeneity	Small	Small	Small	None	Moderate	Large
Current data	0.208 (0.203 - 0.213)	0.208 (0.203 - 0.214)	0.208 (0.203 - 0.213)	0.211 (0.205 - 0.217)	0.209 (0.204 - 0.214)	0.211 (0.205 - 0.217)
Pooled data	0.183 (0.178 - 0.189)	0.199 (0.191 - 0.206)	0.202 (0.197 - 0.207)	0.173 (0.168 - 0.177)	0.221 (0.212 - 0.230)	0.333 (0.316 - 0.350)
Pocock's method	0.183 (0.178 - 0.189)	0.194 (0.188 - 0.201)	0.201 (0.196 - 0.206)	0.178 (0.173 - 0.183)	0.210 (0.202 - 0.217)	0.295 (0.281 - 0.309)
Power prior with $\alpha = 0.5$	0.183 (0.178 - 0.189)	0.194 (0.187 - 0.200)	0.201 (0.196 - 0.206)	0.178 (0.174 - 0.183)	0.209 (0.201 - 0.216)	0.291 (0.278 - 0.305)
MPP	0.185 (0.179 - 0.190)	0.194 (0.187 - 0.200)	0.201 (0.196 - 0.206)	0.179 (0.174 - 0.184)	0.202 (0.196 - 0.208)	0.224 (0.217 - 0.231)
MAP approach	0.193 (0.188 - 0.198)	0.198 (0.193 - 0.203)	0.203 (0.198 - 0.208)	0.190 (0.185 - 0.194)	0.202 (0.197 - 0.207)	0.209 (0.203 - 0.215)
Robust MAP approach	0.196 (0.192 - 0.201)	0.199 (0.194 - 0.204)	0.206 (0.201 - 0.211)	0.196 (0.191 - 0.201)	0.202 (0.197 - 0.208)	0.209 (0.203 - 0.215)
Method of Murray et al.	0.185 (0.179 - 0.190)	0.197 (0.190 - 0.204)	0.202 (0.197 - 0.207)	0.175 (0.170 - 0.180)	0.213 (0.205 - 0.220)	0.236 (0.228 - 0.244)
Test-then-pool method	0.190 (0.185 - 0.196)	0.197 (0.191 - 0.203)	0.205 (0.199 - 0.210)	0.179 (0.174 - 0.184)	0.206 (0.200 - 0.211)	0.212 (0.206 - 0.217)

Table S4: Estimated percentage reduction in required sample size for future study, based on increase in power in simulation study compared with "Current data" analysis

Scenario	1	2	3	4	5	6
Time trend	No	Yes	Yes	No	No	No
Heterogeneity	Small	Small	Small	None	Moderate	Large
Pooled data	31.3	52.4	8.5	41.7	14.2	-4.7
Pocock's method	31.3	48.7	4.1	37.7	17.4	-2.3
Power prior with $\alpha = 0.5$	30.1	47.8	3.6	36.3	15.8	-5.7
MPP	27.1	43.6	1.8	35.2	14.2	-0.5
MAP approach	14.3	21.6	4.5	25.4	10.1	2.7
Robust MAP approach	11.9	17.2	-0.9	16.2	7.1	2.7
Method of Murray et al.	28.2	47.4	7.2	38.5	13.0	0.5
Test-then-pool method	5.9	7.6	-1.9	27.2	-9.7	-2.3

Table S5: Average power and type I error rate of the 95% credible interval of the treatment effect in data sets with 1 historical trial, based on 500 simulated data sets

Scenario	1	2	3	4	5	6
Time trend	No	Yes	Yes	No	No	No
Heterogeneity	Small	Small	Small	None	Moderate	Large
Power						
Current data	0.644	0.650	0.650	0.628	0.634	0.628
Pooled data	0.762	0.788	0.642	0.770	0.708	0.644
Pocock's method	0.720	0.774	0.642	0.730	0.688	0.642
Power prior with $\alpha = 0.5$	0.716	0.758	0.642	0.720	0.692	0.636
MPP	0.714	0.744	0.640	0.726	0.682	0.634
MAP approach	0.668	0.684	0.644	0.670	0.642	0.640
Robust MAP approach	0.654	0.652	0.642	0.640	0.632	0.636
Method of Murray et al.	0.752	0.766	0.644	0.752	0.688	0.650
Test-then-pool method	0.672	0.674	0.646	0.698	0.622	0.602
Type I error rate						
Current data	0.050	0.052	0.050	0.050	0.058	0.052
Pooled data	0.088	0.082	0.052	0.052	0.198	0.440
Pocock's method	0.064	0.062	0.052	0.044	0.112	0.274
Power prior with $\alpha = 0.5$	0.064	0.064	0.052	0.046	0.104	0.262
MPP	0.062	0.058	0.052	0.046	0.086	0.080
MAP approach	0.054	0.054	0.052	0.044	0.056	0.050
Robust MAP approach	0.050	0.050	0.054	0.050	0.052	0.052
Method of Murray et al.	0.080	0.076	0.052	0.056	0.110	0.108
Test-then-pool method	0.058	0.054	0.052	0.046	0.072	0.048

Note: the width of each side of the 95% binomial proportion confidence interval (not shown in the table) is approximately 2% to 3% for the type I error rate and 4% for the power.