Journal of Data Science 23 (1), 38–47
January 2025

DOI: 10.6339/25-JDS1161C
Discussion

Discussion of "Power Priors for Leveraging Historical Data: Looking Back and Looking Forward"[☆]

Margaret Gamalo[1,*], Heliang Shi[1], Yuxi Zhao[1], and Maria Kudela[1]

[1] *Inflammation and Immunology Biometrics Strategy, Pfizer Research and Development, USA*

The integration of high-performance computing has revolutionized the application of Bayesian models, enabling the practical implementation of advanced algorithms such as Gibbs sampling, Metropolis-Hastings, and Hamiltonian Monte Carlo (HMC). These computational advances have transformed Bayesian methods into routine tools for data analysis, particularly in estimating parameters and their associated probabilities based on all available evidence. These advancements have also facilitated the development of many data-based priors techniques enhancing the flexibility of Bayesian approaches to incorporate historical or external evidence systematically. Additionally, this evolution has expanded the potential of Bayesian techniques even in handling complex and large-scale data.

One prominent family of data-based priors, the power prior (Ibrahim and Chen, 2000), has become a foundational tool for integrating external or reference data into Bayesian models since its introduction and serves as a central focus of Chen et al. (2025). The method utilizes a power parameter α to influence the amount of borrowing from reference data, where $\alpha = 0$ implies no influence from the reference data, and $\alpha = 1$ equates the importance of reference and target data. The Normalized Power Prior (NPP) (Duan et al., 2006), one of the first and better known extensions of the traditional power prior method, incorporates a normalization factor that, in addition, scales the influence of external data relative to target data using a power parameter. Many further variations of the method were proposed highlighting the flexibility of this family of methods that makes it well-suited for applications such as clinical trials and epidemiological studies, where leveraging historical or reference data is critical for robust statistical modeling.

In this discussion, we will focus on specific aspects of Chen et al. (2025) while also providing broader commentary on the Bayesian paradigm as a whole. It is important to note that many considerations extend beyond the specifics of the method itself and must be evaluated within the context of sound analytical practices, whether adopting a Bayesian or frequentist approach.

Systematic and Systemic Integration The power prior provides a systematic framework for incorporating reference data, aligning with Bayesian principles through explicit manipulation of the likelihood with the parameter α. To provide context to subsequent discussions, we review power priors based on normalized power priors by considering two distinct datasets: the target dataset consists of outcomes $\boldsymbol{Y}_t = [Y_{1t}, \ldots, Y_{n_t,t}]$ while the reference dataset is $\boldsymbol{Y}_r = [Y_{1r}, \ldots, Y_{n_r,r}]$, where n_t and n_r are number of observations in target and reference data, respectively. For each outcome Y_{it} in the target dataset, the outcome follows a distribution F parameterized by θ_{it}. The parameter ϑ_{it} for each patient is connected to the parameter θ_{it} via a link function g and is characterized by a distribution $H(\boldsymbol{\eta})$. The population parameter, $\boldsymbol{\eta}$, is

[☆] Main article: https://doi.org/10.6339/24-JDS1161.

[*] Corresponding author. Email: Margaret.Gamalo@pfizer.com.

initially informed by a prior distribution. The NPP has the form:

$$\pi_{\mathrm{NPP}}(\boldsymbol{\eta}|\boldsymbol{Y}_r, \alpha) \propto \left[\prod_{i=1}^{n_r} L(\vartheta_i|Y_{ir}, \boldsymbol{\eta})\right]^{\alpha} \cdot \frac{\pi_0(\boldsymbol{\eta})}{C(\alpha)}, \tag{1}$$

where $C(\alpha) \equiv \int \prod_{i=1}^{n_r} L(\vartheta_i|Y_{ir}, \boldsymbol{\eta})^{\alpha} \pi_0(\boldsymbol{\eta}) d\boldsymbol{\eta}$ is a normalization constant that ensures the prior distribution remains properly scaled. It is crucial to highlight that the non-normalized version of Equation (1) remains proper provided that $C(\alpha) < \infty$ is satisfied.

In the general scenario, a joint prior distribution for $\boldsymbol{\eta}$ and α is specified instead of just conditioning on a fixed value of α in Equation (1). This is then factorized as $\pi_0(\boldsymbol{\eta}, \alpha) = \pi_0(\boldsymbol{\eta}|\alpha)\pi_0(\alpha)$. Given that the exact value of α is often ambiguous and not prespecified in real-world scenarios, a comprehensive Bayesian approach treats α as a random variable, assigning it an appropriate prior distribution. A natural choice for this prior is the beta distribution, $Beta(a, b)$, due to its definition over the [0,1] interval; while $\pi_0(\boldsymbol{\eta}|\alpha)$ simplifies to $\pi_0(\boldsymbol{\eta})$. Upon integrating out the parameter α, we derive the posterior distribution for $\boldsymbol{\eta}$:

$$\pi(\boldsymbol{\eta}|\boldsymbol{Y}_t, \boldsymbol{Y}_r) \propto \int_0^1 \frac{1}{C(\alpha)} \prod_{i=1}^{n_t} L(\vartheta_i|Y_{it}, \boldsymbol{\eta}) \prod_{i=1}^{n_r} L(\vartheta_i|Y_{ir}, \boldsymbol{\eta})^{\alpha} \pi_0(\boldsymbol{\eta})\pi_0(\alpha) d\alpha. \tag{2}$$

From this formulation, it is clear that the power prior, governed by the power parameter α, offers a systematic (i.e., a framework for data integration) and systemic (i.e., affects all the moments of the distribution through the likelihood) approach for incorporating reference data into the analysis, allowing explicit control over the weight of reference data based on their relevance and reliability. With the common parameter of interest, i.e., $\boldsymbol{\eta}$, it assumes *exchangeability* (Bernardo, 1996) and its direct manipulation of the likelihood ensures alignment with Bayesian updating principles and provides a clear mechanism to fine-tune prior influence, making it well suited for scenarios where reference data are compatible, well understood and scientifically justified. However, the approach also assumes that a single parameter α can adequately reflect the characteristics/behavior and relevance of reference data, which becomes problematic when the quality or compatibility of reference data varies. Furthermore, interpreting and justifying α can be challenging, and computational stability issues can arise when α is considered random (Pawel et al., 2023). In contrast, robust priors (Berger and Berliner, 1986; Schmidli et al., 2014), though less transparent and more reliant on distributional assumptions, have gained broad applicability in real-world settings with heterogeneous or conflicting data, as they naturally downweight extreme values and uncertainties through heavy-tailed distributions or hierarchical models. While power priors offer balanced and interpretable adjustments across all observations, their performance may vary in highly diverse data contexts. In such settings, methods designed to address variability and outliers, like robust priors, may offer complementary advantages depending on the nature of the data.

The concept of systematic integration in models involving multiple parameters, such as regression models, presents unique complexities, particularly when interest lies in a subset of parameters. Three variations of the power prior also implemented in Chen et al. (2025) that attempt to address this challenge are *partial borrowing power prior*, *borrowing-by-parts power prior*, and *partial borrowing-by-parts power prior*. While these methods offer flexibility, they also introduce concerns that merit closer examination.

Partial borrowing power prior is conceptually linked to regression standardization or *g-computation*, where the effects of parameters not of interest are integrated out. In the Kociba and NTP study (Chen et al., 2025), the resulting power prior can be expressed as:

$$\pi(\boldsymbol{\eta}|\boldsymbol{Y}_r,\alpha) \propto \int \prod_{i=1}^{n_r}\left[\frac{\exp\{y_{ir}(\eta_{0r}+\eta_1 x_{ir})\}}{1+\exp(\eta_{0r}+\eta_1 x_{ir})}\right]^{\alpha}\pi_0(\eta_{0r})d\eta_{0r}\pi_0(\eta_0,\eta_1),$$

where $\boldsymbol{\eta} = (\eta_0, \eta_1)$ represents the intercept and the slope associated with the exogenous dose covariate x_i. In this framework, integrating out the effects of non-interest parameters assumes that exchangeability can be induced selectively by marginalizing these effects. However, this assumption hinges critically on the alignment of covariate distributions between the reference and target populations. If covariate distributions differ significantly, the integration process might account only for the covariate distribution in the reference population, leading to bias or misalignment in the parameter $\boldsymbol{\eta}$. Moreover, when the power prior is combined with the target population data $\boldsymbol{Y}_t$, the borrowing parameter α becomes implicitly dependent on the distribution of η_{0r} in the reference data and η_1 in both the reference and target populations. This dependence raises concerns about the ability of α to adequately adjust for such differences. A possible improvement could involve integrating covariate effects in both populations before borrowing to harmonize the process.

In the normalized power prior setting, on the other hand, the posterior distribution can be expressed as:

$$\pi(\boldsymbol{\eta},\alpha|\boldsymbol{Y}_r) \propto \frac{\int \pi_0(\eta_{0r})\pi_0(\eta_0,\eta_1)\prod_{i=1}^{n_r}\left[\frac{\exp\{y_{ir}(\eta_{0r}+\eta_1 x_{ir})\}}{1+\exp(\eta_{0r}+\eta_1 x_{ir})}\right]^{\alpha}d\eta_{0r}\pi_0(\alpha)}{\int\int\int \pi_0(\eta_{0r})\pi_0(\eta_0^*,\eta_1^*)\prod_{i=1}^{n_r}\left[\frac{\exp\{y_{ir}(\eta_{0r}^*+\eta_1^* x_{ir})\}}{1+\exp(\eta_{0r}^*+\eta_1^* x_{ir})}\right]^{\alpha}d\eta_{0r}d\eta_0^*d\eta_1^*\pi_0(\alpha)}.$$

In this formulation, it is unclear whether α implicitly adjusts for covariate differences (for those integrated out). This potential lack of clarity underscores the need for additional research to verify whether the parameter α can appropriately account for such differences.

Borrowing by parts allows for information to be borrowed for different subsets of parameters independently, with each subset having its own discounting parameter. This approach is particularly appealing in regression models where borrowing is intended for the control arm but not the treatment arm. For instance, in models like Cox regression, analyzing treatment arms independently may not be analytically valid for making inference on effects of treatments, making partial borrowing a practical alternative. This method also addresses the criticism whether a single power parameter can adequately reflect the characteristics/behavior and relevance of historical data. However, this method assumes that the segmentation of parameters is appropriate from a scientific and clinical perspective and that the discounting parameters are adequately tuned to reflect differences between populations.

Partial borrowing-by-parts combines the ideas of integrating out certain parameters and segmenting others for selective borrowing. While this approach offers flexibility, it also inherits potential pitfalls from both partial borrowing and borrowing-by-parts. Specifically:

- **Covariate Mismatch**: If covariate distributions differ significantly between reference and target populations, integrating out certain parameters could introduce bias.
- **Parameter Dependence**: As in partial borrowing, the dependence of α on the reference and target covariate distributions may complicate the interpretation and reliability of the borrowing process.
- **Complexity in Model Specification**: The dual process of integrating out some parameters while applying borrowing discounts to others adds complexity, increasing the risk of model misspecification.

While these power prior variations provide valuable tools for selective borrowing and integration of historical data, their application requires careful consideration of covariate alignment, parameter dependencies, and computational feasibility. Future research on these can explore methods for better aligning covariate distributions and improving the interpretability and robustness of discounting parameters. These improvements are critical for ensuring reliable and meaningful borrowing in complex regression settings.

Core Data Quality The use of Bayesian methods varies significantly between *regulatory* and *peri-regulatory* contexts. Regulatory applications require rigorous justification for prior selection and emphasizing transparency and reproducibility. Conversely, peri-regulatory contexts allow greater flexibility, enabling tailored priors for exploratory decision-making frameworks. These distinctions highlight the need for stakeholder alignment when integrating Bayesian methods into drug development and decision-making processes.

The broader applicability of Bayesian methods is often framed by parallels with observational studies as it should be. Both approaches require scrutiny of biases in data and validation of model assumptions to ensure validity and applicability. Data-based priors, like observational studies, face the same challenges as well as in generalizability when the reference populations differ significantly from the target context (Izem et al., 2022). This underscores the importance of aligning priors with the target population's characteristics and ensuring compatibility in terms of inclusion criteria, treatment regimens, and endpoints (Lin et al., 2022). Sensitivity analyses and robustness checks are vital for mitigating biases and validating conclusions, mirroring practices in causal inference and observational research.

In the context of the described applications, differences in study contexts and timelines can introduce unique challenges. For instance, in the example of KOCIBA and NTP datasets discussed in Chen et al. (2025), differences in exposure time could significantly impact adverse toxicological effects, particularly for outcomes with long latency periods. Similarly, in the ADNI and ADNI-GO2 studies, differences in the time periods when the studies were conducted may reflect variations in standards of care, potentially influencing outcomes such as ADAS and MMSE scores. These differences necessitate careful assessment, as their impacts can be difficult to quantify. Addressing these challenges requires a thorough understanding of the study context and its potential implications for generalizability and interpretation, as well as robust sensitivity analyses to quantify and account for these effects.

Interpolation vs. Extrapolation Bayesian methods should be carefully tailored to address scientific questions within the distinct contexts of *interpolation* and *extrapolation*, two areas that remain underexplored and insufficiently differentiated in the Bayesian statistical literature. Interpolation involves predicting outcomes within the range of observed data, under the assumption that no intrinsic differences exist between the reference and target populations. For example, determining an initial pediatric dose can leverage pharmacokinetic (PK) and pharmacodynamic (PD) data from adults or older pediatric cohorts. This approach is supported by a substantial body of evidence linking dosing to factors such as body weight, organ maturation, and developmental changes in drug metabolism and clearance (Job et al., 2019). Applications like those described in the KOCIBA and NTP datasets or the ADNI and ADNI-GO2 datasets fall squarely within the realm of interpolation, as they involve extending insights within a relatively well-defined and aligned data space.

Extrapolation, by contrast, involves predicting outcomes beyond the range of observed data and relies on additional assumptions about the similarity of disease progression and treatment response across populations. This context is particularly relevant in pediatric drug development, where assumptions about comparability between adults and children are often made *de facto* to support label extensions for pediatric use based on adult data (Gamalo et al., 2022). In these cases, the validity of extrapolation depends on the extent to which the adult data can reliably inform pediatric outcomes. The degree of borrowing in extrapolation depends on the assessed similarity between populations, going beyond simple comparisons of baseline characteristics or routine outcome measures often used in interpolation. Concepts like translatable information, which assess deeper structural or mechanistic similarities between populations, can help guide the extent of borrowing. Importantly, borrowing should be reduced or negated entirely when substantial differences between populations are detected that undermine the assumptions of similarity required for extrapolation.

The distinction between interpolation and extrapolation plays a critical role in determining appropriate methods for integrating data-based priors from a reference population into a target population. Interpolation can be viewed as requiring robust statistical methods to account for potential heterogeneity within the same population, often manifesting through heavy-tailed distributions in the likelihood. This heterogeneity arises when data from the target population exhibit variability that challenges assumptions of homogeneity.

In contrast, extrapolation represents a continuum from no borrowing of information from the reference population to full borrowing. Under this perspective, the degree of integration reflects the gradual incorporation of information from the reference into the target population. One could further argue that extrapolation aligns conceptually with interpolation when the reference population resides in the tails of the target population's data distribution. Thus, even under extrapolation, the reference data can be treated as an extreme subset of the target distribution, maintaining a conceptual bridge between the two frameworks.

Model Selection, Type I Error, and Effective Sample Size In regulatory applications of Bayesian methods and with the numerous types of data-based priors available in literature, three measures — *bias*, *type I error*, and *effective sample size (ESS)* — as well as the topic of *model selection* have gained critical importance and evaluated differently in interpolation and extrapolation contexts. In interpolation, model selection and goodness-of-fit are assessed through purely quantitative methods, emphasizing statistical accuracy within the same population. In contrast, extrapolation requires models to not only be statistically accurate but also tailored robustly to the target population, ensuring they are scientifically fit for purpose.

Type I error thresholds are stricter in interpolation due to its reliance on direct evidence from the same population for drug approval. For extrapolation, however, type I error is nuanced by the degree of similarity between the reference and target populations. For example, in pediatric drug development, type I error tolerance may depend on unmet medical needs or the therapeutic landscape. In such cases, greater flexibility may be acceptable, particularly when the drug has already been approved in the reference population. Meanwhile, ESS quantifies the degree of borrowing from reference data. In interpolation, ESS ensures that external information is appropriately weighted without overshadowing the target data. In extrapolation, ESS becomes critical to assess how differences in translatable information between populations influence the robustness of conclusions. Constraints that prevent ESS from exceeding the size of the target population are vital, particularly when reference and target population data differ significantly. Advanced methodologies, such as sensitivity and tipping point analyses, enhance Bayesian modeling by identifying thresholds where conclusions may shift and evaluating the influence of prior distributions on posterior results, thereby supporting robust regulatory decision-making.

To better understand and illustrate how metrics such as ESS and type I error calibrate different Bayesian methods for regulatory application, the performance of NPP and robust priors are compared using two datasets — one binary and one continuous (normal). Simulation results based on 1,000 iterations highlight how these methods behave under varying conditions of reference and target data characteristics. For binary data, suppose the reference data size is $n_r = 80$ or $n_r = 300$, with a fixed reference population proportion of $\tilde{\theta}_r = 0.8$. The target data size is either $n_t = n_r$ or $n_t = \frac{1}{2}n_r$, and the true target population proportion $\tilde{\theta}_t$ that takes values from 0.5 to 0.8. The null hypothesis is $H_0 : \tilde{\theta}_t \leqslant 0.5$ against the alternative $H_1 : \tilde{\theta}_t > 0.5$, with a pre-specified threshold of $P(\tilde{\theta}_t > 0.5 \mid D_r, D_t) > 0.975$. Results presented in Table 1 indicate that when the true target mean $\tilde{\theta}_t$ is close to the observed reference mean $\tilde{\theta}_r$ (e.g., $\tilde{\theta}_t = 0.7$), the power priors method demonstrates higher statistical power than the robust priors method. However, type I error is inflated for power priors when $\tilde{\theta}_t$ deviates significantly from $\tilde{\theta}_r$ (e.g., $\tilde{\theta}_t = 0.5$). When $n_r = 300$, $n_t = 150$, and $\tilde{\theta}_t = \overline{\theta}_r = 0.8$, both methods perform comparably. However, the robust priors method borrows an effective sample size notably larger than the target sample size.

For normal data, assume the reference data has a sample size of $n_r = 80$ or $n_r = 300$, with a fixed sample mean $\mu_r = 1$ and standard deviation $\sigma_r = 1$. The target data has a sample size $n_t = n_r$ or $n_t = \frac{1}{2}n_r$, with target variance σ_t^2 equal 0.25, 1, or 4 and true mean $\tilde{\mu}_t$ equal zero or one. The null hypothesis is $H_0 : \tilde{\mu}_t \leqslant 0$ against the alternative $H_1 : \tilde{\mu}_t > 0$, with a pre-specified threshold of $P(\tilde{\mu}_t > 0 \mid D_r, D_t) > 0.975$. Results presented in Table 2 indicate that when $\sigma_t \geqslant \sigma_r$ and $\tilde{\mu}_t = \mu_r$, both methods perform comparably. However, robust priors outperform power priors in controlling type I error when $\tilde{\mu}_t = 0$ differs from μ_r, regardless of the target data variance. Robust priors consistently borrow a larger ESS than power priors, which influences the weighting of reference data in target population inferences.

These results underscore the importance of aligning Bayesian methods with metrics such as type I error and ESS and that while one metric is satisfied, the others may not. It may happen that the efficiency of a method will be severely constrained by other metrics. In general, the method that will be selected is the one that balances all the metrics optimally. Along with sensitivity analyses, these metrics play a critical role in ensuring the validity and robustness of conclusions. While power priors offer higher power in some scenarios, robust priors may provide stronger type I error control and reliable borrowing across heterogeneous datasets. Selecting an appropriate prior depends on the scientific context, the degree of similarity between reference and target populations, and the regulatory requirements for evidence-based decision-making.

Ordering and Extent of Borrowing The technique of alignment of outcomes to determine appropriate weights in the case of multiple data sets from the reference population, requires careful evaluation. As highlighted, the relevance and quality of data is a critical factor that must be rigorously assessed before exchangeability can be applied. This consideration becomes even more vital in cases of extrapolation, where data from related conditions or populations are integrated. Such scenarios demand a careful and deliberate approach to ensure meaningful integration.

In both interpolation and extrapolation, the correct sequencing of the borrowing process is paramount. Prioritizing datasets that are most similar to the target population ensures systematic and meaningful borrowing based on data quality and relevance. The *ordering* of prior data, as discussed in Gamalo et al. (2014), plays a crucial role in achieving this prioritization. The

Table 1: Simulation results for binary data. n_r: reference population size; $\tilde{\theta}_r$: reference mean; n_t: target population size; $\tilde{\theta}_t$: true target mean; $RMSE_{\mathrm{NPP}}$, $RMSE_{\mathrm{RP}}$: root mean squared error for normalized power prior and robust prior; ESS_{NPP}, ESS_{RP}: effective sample size for normalized prior and robust prior; OC_{NPP}, OC_{RP}: operating characteristics expressed as probability of passing significance criterion under various treatment effects for normalized power prior and robust prior. Red color highlights notable differences.

n_r	$\tilde{\theta}_r$	n_t	$\tilde{\theta}_t$	$RMSE_{\mathrm{NPP}}$	$RMSE_{\mathrm{RP}}$	ESS_{NPP}	ESS_{RP}	OC_{NPP}	OC_{RP}
80	0.8	40	0.5	0.107	0.091	0	0	0.145	0.049
			0.6	0.090	0.091	10	0	0.592	0.331
			0.7	0.057	0.070	26	14	0.931	0.795
			0.8	0.035	0.036	31	34	0.999	0.994
		80	0.5	0.067	0.057	0	0	0.081	0.036
			0.6	0.066	0.066	3	0	0.627	0.464
			0.7	0.046	0.054	27	12	0.991	0.962
			0.8	0.030	0.028	36	48	>0.999	>0.999
300	0.8	150	0.5	0.045	0.040	0	0	0.056	0.024
			0.6	0.052	0.043	0	0	0.800	0.640
			0.7	0.045	0.049	36	0	>0.999	0.999
			0.8	0.017	0.015	125	177	>0.999	>0.999
		300	0.5	0.030	0.028	0	0	0.043	0.025
			0.6	0.032	0.029	0	0	0.956	0.930
			0.7	0.031	0.034	8	0	>0.999	>0.999
			0.8	0.015	0.013	141	222	>0.999	>0.999

resulting prior is then expressed as:

$$\pi(\boldsymbol{\eta}|\boldsymbol{Y}_r,\boldsymbol{\alpha}) \propto \frac{\prod_{k=1}^{K}\prod_{i=1}^{n_k} L(\vartheta_{ik}|\boldsymbol{Y}_{rk},\boldsymbol{\eta})^{\alpha_{(k)}}\pi_0(\boldsymbol{\eta})}{\int \prod_{k=1}^{K}\prod_{i=1}^{n_k} L(\vartheta_{ik}|\boldsymbol{Y}_{rk},\boldsymbol{\eta})^{\alpha_{(k)}}\pi_0(\boldsymbol{\eta})d\boldsymbol{\eta}}\pi_0(\boldsymbol{\alpha}), \tag{3}$$

where $\alpha_{(1)} \leqslant \alpha_{(2)} \leqslant \cdots \leqslant \alpha_{(K)}$ are ordered power parameters that control the amount of borrowing from corresponding K reference studies. This ordering involves prioritizing datasets that are not only similar to the target population but also of the highest quality, ensuring that the borrowing process maintains rigor and relevance. Data could also be blocked under a similar weight.

Further research is needed to investigate how the ordering of prior data impacts the weighting process and to develop optimized methodologies for effectively leveraging prior data. These methodologies must maintain the relevance and integrity of conclusions while maximizing the utility of available information, especially in cases requiring extrapolation or complex integrations.

Consistency of Borrowing in Multiple Endpoints The application of Bayesian methods to multiple endpoints presents distinct challenges in interpolation and extrapolation, particularly when the endpoints are correlated and critical for decision-making, such as co-primary endpoints or those essential for labeling. In interpolation, borrowing should be influenced not only by the proximity of the target outcome to the reference data but also by the variability of the endpoint. This approach aligns with the principles of population similarity and data quality,

Table 2: Simulation results for normal data. n_r: reference population size; μ_r: reference mean; σ_r: reference standard deviation; n_t: target population size; σ_t: target standard deviation; $\tilde{\mu}_t$: true target mean; $RMSE_{\text{NPP}}$, $RMSE_{\text{RP}}$: root mean squared error for normalized power prior and robust prior; ESS_{NPP}, ESS_{RP}: effective sample size for normalized power prior and robust prior; OC_{NPP}, OC_{RP}: operating characteristics expressed as probability of passing significance criterion under various treatment effects for normalized power prior and robust prior. Red color highlights notable differences.

n_r	μ_r	σ_r	n_t	σ_t	$\tilde{\mu}_t$	$RMSE_{\text{NPP}}$	$RMSE_{\text{RP}}$	ESS_{NPP}	ESS_{RP}	OC_{NPP}	OC_{RP}
80	1	1	40	1	0	0.227	0.170	0	1	0.122	0.035
					1	0.105	0.064	35	70	>0.999	>0.999
				2	0	0.321	0.527	6	3	0.065	0.140
					1	0.288	0.103	16	177	0.924	0.974
				0.5	0	0.087	0.081	0	2	0.044	0.035
					1	0.070	0.052	0	22	>0.999	>0.999
			80	1	0	0.128	0.109	0	2	0.053	0.018
					1	0.085	0.058	38	76	>0.999	>0.999
				2	0	0.221	0.307	6	0	0.040	0.055
					1	0.204	0.071	14	240	0.995	0.996
				0.5	0	0.057	0.054	0	2	0.033	0.020
					1	0.054	0.045	0	22	>0.999	>0.999
300	1	1	150	1	0	0.090	0.082	0	2	0.061	0.025
					1	0.053	0.030	130	274	>0.999	>0.999
				2	0	0.164	0.166	5	0	0.042	0.030
					1	0.156	0.043	12	831	>0.999	>0.999
				0.5	0	0.041	0.040	0	2	0.033	0.025
					1	0.042	0.029	0	75	>0.999	>0.999
			300	1	0	0.060	0.057	0	2	0.040	0.026
					1	0.043	0.028	143	290	>0.999	>0.999
				2	0	0.113	0.113	6	3	0.028	0.022
					1	0.115	0.034	14	991	>0.999	>0.999
				0.5	0	0.029	0.029	0	1	0.020	0.016
					1	0.029	0.024	0	76	>0.999	>0.999

ensuring robust inference across correlated endpoints. In extrapolation, the complexity increases as borrowing decisions are shaped by the translatability of the disease and the variability of the measure. These factors introduce additional uncertainty, necessitating careful alignment of priors to account for variations in endpoint behavior between populations. Bounded priors may be essential for maintaining consistency, ensuring that borrowing is balanced by the variability and importance of each endpoint. This systematic approach is particularly critical in extrapolation, where predictions extend beyond the observed data, emphasizing the need for coherence across endpoints in the face of greater uncertainty.

Upstream and Downstream Borrowing Guardrails Recent advancements have applied the principle of exchangeability by incorporating propensity score methods, effectively aligning Bayesian borrowing with causal inference principles. Unlike traditional causal models, which adjust effect size estimation using baseline characteristics, Bayesian methods leverage propensity scores to refine the likelihood function. This approach facilitates conditional exchangeability by proactively addressing dissimilarities between datasets, thereby strengthening causal inferences (Lin et al., 2022). Furthermore, integrating propensity scores into Bayesian models enhances the robustness of results, reduces sensitivity to model assumptions, and improves the reliability of conclusions.

The combination of propensity scores and Bayesian methodologies offers a powerful and complementary approach that should be regarded as a fundamental component of Bayesian analysis. However, it has been noted that the iptwPP-based approach may result in over-borrowing, leading to biased parameter estimates when the current data and historical data are not sufficiently similar (Chen et al., 2025). It is important to recognize that several factors, including unmeasured confounding and model misspecification, may also contribute to this observed bias, underscoring the need for careful consideration and methodological rigor in the application of such approaches.

In conclusion, Bayesian methods, particularly power priors, provide a robust framework for integrating historical data into statistical analyses. However, the selection of the right method among different variations of Bayesian priors remain a challenging task as was also noted in Chen et al. (2025) based on their empirical analyses. Bayesian methods effectiveness hinges on thoughtful evaluation of data quality, contextual relevance, and the accurate quantification of translatability between or within populations. Advances in computational capabilities and methodological innovations have broadened the scope of Bayesian frameworks, underscoring the necessity of transparency, rigorous validation, and continuous refinement to align with scientific objectives. As these methods evolve, addressing considerations presented above, they offer significant potential for enhancing statistical inference and facilitating evidence-based decision-making across diverse applications in drug development.

References

Ibrahim JG, Chen M-H (2000). Power prior distributions for regression models. *Statistical Science*, 15(1): 46–60. https://doi.org/10.1214/ss/1009212673

Chen M-H, Guan Z, Lin M, Sun M (2025). Power priors for leveraging historical data: looking back and looking forward. *Journal of Data Science*, 23(1): 1–30. https://doi.org/10.6339/24-JDS1161

Duan Y, Ye K, Smith EP (2006). Evaluating water quality using power priors to incorporate historical information. *Environmetrics: The Official Journal of the International Environmetrics Society*, 17(1): 95–106. https://doi.org/10.1002/env.752

Bernardo JM (1996). The concept of exchangeability and its applications. *Far East Journal of Mathematical Sciences*, 4: 111–122.

Pawel S, Aust L, Held L, Wagenmakers E-J (2023). Normalized power priors always discount historical data. *Stat*, 12(1): e591. https://doi.org/10.1002/sta4.591

Berger J, Berliner LM (1986). Robust bayes and empirical bayes analysis with ε-contaminated priors. *The Annals of Statistics*, 14(2): 461–486.

Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4): 1023–1032. https://doi.org/10.1111/biom.12242

Izem R, Buenconsejo J, Davi R, Luan JJ, Tracy L, Gamalo M (2022). Real-world data as external controls: practical experience from notable marketing applications of new therapies. *Therapeutic Innovation & Regulatory Science*, 56(5): 704–716. https://doi.org/10.1007/s43441-022-00413-0

Lin J, Gamalo-Siebers M, Tiwari R (2022). Ensuring exchangeability in data-based priors for a bayesian analysis of clinical trials. *Pharmaceutical Statistics*, 21(2): 327–344. https://doi.org/10.1002/pst.2172

Job KM, Gamalo M, Ward RM (2019). Pediatric age groups and approach to studies. *Therapeutic Innovation & Regulatory Science*, 53(5): 584–589. https://doi.org/10.1177/2168479019856572

Gamalo M, Bucci-Rechtweg C, Nelson RM, Vanh L, Porcalla A, Thackray H, et al. (2022). Extrapolation as a default strategy in pediatric drug development. *Therapeutic Innovation & Regulatory Science*, 56(6): 883–894. https://doi.org/10.1007/s43441-021-00367-9

Gamalo MA, Tiwari RC, LaVange LM (2014). Bayesian approach to the design and analysis of non-inferiority trials for anti-infective products. *Pharmaceutical Statistics*, 13(1): 25–40. https://doi.org/10.1002/pst.1588