Bayesian modeling for the detection of adverse events underreporting in clinical trials

Yves Barmaz[1], PhD • Timothé Ménard[1], PharmD

1. F. Hoffmann-La Roche AG - CH-4070, Basel, Switzerland
corresponding author: timothe.menard@roche.com

Abstract

Introduction

Safety underreporting is a recurrent issue in clinical trials that can impact patient safety and data integrity. Clinical Quality Assurance (QA) practices used to detect underreporting rely on on-site audits, however adverse events underreporting remains a recurrent issue. In a recent project, we developed a predictive model that enables oversight of Adverse Event (AE) reporting for clinical Quality Program Leads (QPL). However, there were limitations to using solely a machine learning model.

Objective

Our primary objective was to propose a robust method to compute the probability of AE underreporting that could complement our machine learning model. Our model was developed to enhance patients safety while reducing the need for on-site and manual QA activities in clinical trials.

Methods

We used a Bayesian hierarchical model to estimate the site reporting rates and assess the risk of underreporting. We designed the model with Project Data Sphere clinical trial data that is public and anonymized.

Outcome

With the resulting reporting rates and metrics, we could focus quality assurance strategic activities for clinical trials on the investigator sites with the highest risk of underreporting.

Conclusion

The new model will be integrated into the current dashboard designed for clinical Quality Program Leads. This approach reduces the need for on-site audits, shifting focus from source data verification (SDV) to pre-identified, higher risk areas. It will enhance further quality assurance activities for safety reporting from clinical trials and generate quality evidence during pre-approval inspections.

1. Introduction

Adverse event (AE) underreporting has been a recurrent issue raised during Health Authorities (HA) Good Clinical Practices (GCP) inspections and audits [1]. Moreover, safety underreporting poses a risk to patient safety and data integrity [2]. The previous clinical Quality Assurance (QA) practices used to detect AE underreporting rely heavily on investigator site and study audits. Yet several sponsors and institutions have had repeated findings related to safety reporting, leading to delays in regulatory submissions.

In a previous project [3], we developed a predictive model that enables pharmaceutical sponsors oversight of AE reporting at the program, study, site and patient level. We validated and reproduced our model using a combination of internal data and an external dataset [4].

While the machine learning model has been successfully implemented since May 2019, there was a need to address calibration robustness. The first model relied on point predictions at the visit level and assumed Poisson distributed residuals. The decision to flag underreporting sites depended on the way these residuals were aggregated and we have been observing instabilities in longer running studies. This motivated us to tackle the problem by the other end and find a model for the distribution of adverse events reported by sites. The biggest value of point estimates from our initial machine learning model is for our stakeholders to direct their investigations, for instance which patient to target during an investigator site audit. On the other hand, an automated underreporting alert relies on well-calibrated probabilities for risk estimate. This was the main motivation for this project. These two solutions will be offered in parallel to quality professionals

at our organization, the probabilistic one to quantify the risk of underreporting, and the machine learning based model to provide a basis for in-depth investigations and audits.

The project has been conducted by a team of data scientists, in collaboration with clinical and QA subject matter experts (SMEs). This project was part of a broader initiative of building data-driven solutions for clinical QA to complement and augment traditional QA approaches and to improve the quality and oversight of GCP - and Good Pharmacovigilance Practices (GVP) - regulated activities.

2. Methods

2.1. Outline

The primary objective was to develop a robust methodology to assess the risk of adverse events underreporting from investigator sites. The scope remained focused on adverse events -not adverse drug reactions- that should occur in clinical trials. Good clinical practice requires all AEs, regardless of a causal relationship between the drug and events, to be reported timely to the sponsor [2]. Underreporting of safety events is a frequent and repetitive issue in clinical trials [1,5] with many consequences, e.g. delayed approval of new drugs [6-7] or amplifying shortcomings of safety data collection in randomized controlled trials [8].

The traditional way to detect AE underreporting in clinical trials is to conduct thorough site audits [9], on top of monitoring activities and through manual SDV. For sponsors with thousands of sites to audit, this is not manually scalable, hence the strong need for a data-driven approach.

Unfortunately, we will never know how many AEs should have been reported, it is something we have to infer from the data. In other words, we are dealing with an unsupervised anomaly detection problem and do not observe the true labels. The typical way to solve this type of problem is to fit a probability distribution to the data and assess individual data points with a metric like a tail area or a likelihood.

In a previous work [3], to infer the distribution of adverse events, we exploited the variety of covariates available at the patient and visit levels to estimate a conditional density $p(y_{visit} \mid x_{visit}\ ;\ \theta)$ via machine learning. For site-level estimates, we aggregated the visit-level distributions to patient-level and then site-level via successive summations. While this method tracks adverse event data generation at various resolutions, the aggregation introduced biases in the form of systematic over- or underestimation for certain sites, in particular in longer-running trials, probably due to the addition of non-independent errors. As a result, the risk assessment of safety underreporting from investigator sites was not well calibrated. This motivated the top-down approach presented here, as we were ultimately interested in the selection of sites for audits.

To further increase the robustness of the risk assessment, we adopted a Bayesian approach to quantify uncertainties through posterior probabilities. This is a very appealing property in sectors where risk management is essential such as healthcare or finance. In our case, a clear estimate of the probability of underreporting from the different sites enables targeting of the riskiest sites to audit, and, on the positive side, to gain confidence in the completeness of the collected safety data.

The general methodology of Bayesian data analysis is well described in the literature [10]. The main idea is to build a probabilistic model for the observed data, denoted by X, that contains unobserved parameters, collectively denoted by θ. This model relies on a subjective assessment of the distribution of the parameters in the form of a prior distribution $p(\theta)$, which is more or less sharp depending on the degree of certainty of the prior knowledge. The relation between the parameters and the observed data is expressed by the likelihood function $p(X \mid \theta)$. The goal of Bayesian inference is to refine the prior distribution, once the data is observed via the application of Bayes' theorem, and obtain the posterior distribution $p(\theta \mid X) = p(X \mid \theta)p(\theta)/p(X)$used to make decisions, estimate parameters, or assess risks. If the observed data is compatible with the prior distribution, the posterior distribution will typically have a smaller spread than the prior. If it is less compatible, then the likelihood and the prior will compete and the posterior distribution will represent a compromise.

In our problem, the observed data is numbers of AEs reported by the sites, grouped by patients, and parameters could be unobserved adverse event reporting rates from the individual sites. We

emphasize that there can be several competing models for the observed data, and the goal is to find one that is complex enough to capture the structures of interest, namely safety underreporting in our case, but as simple as possible to speed up computations and convey the clearest insights to stakeholders.

2.2. **Data**

We developed this project on our sponsored clinical trials, but this methodology is applicable to any trial. For illustration, we used public data from the project Data Sphere (PDS), "*an independent, not-for-profit initiative of the CEO Roundtable on Cancer's Life Sciences Consortium (LSC), operates the Project Data Sphere platform, a free digital library-laboratory that provides one place where the research community can broadly share, integrate and analyze historical, patient-level data from academic and industry phase III cancer clinical trials*" [11]. PDS data was fit for purpose to demonstrate the approach presented in this paper and are publicly available (lifting any concerns for data privacy and security). Specifically, we used the control arm of the registered clinical trial NCT 00617669 [12]. Of note, the data had been further curated to remove duplicate adverse events.

From the clinical trial data, we extracted for our analysis the count of adverse events reported by patient, grouped by investigator site (see Table 1).

Clinical Investigator Site	Observed AEs	Number of Patients
site_001	[4, 1]	2
site_002	[2, 2, 1, 2, 5, 5, 5, 1]	8
site_003	[7,4]	2
site_004	[3, 27 ,8]	3
site_005	[12, 6]	2
site_006	[2, 4]	2
site_007	[5]	1
site_008	[11, 4, 16, 31, 23, 23]	6
site_009	[6, 6]	2
site_010	[21, 10, 6, 17, 10, 26, 19, 18, 1, 18, 23,...	17

Table 1. - Sample of input data (the whole set includes 125 sites)

2.3. **Model**

We needed an observational model for this aggregated safety data that included variables describing the reporting behavior from sites. A hierarchical model was well indicated as it would capture the three levels structure of study, sites and patients. Concretely, we assumed that adverse event reporting by a given site could be modelled by a Poisson process. The observed numbers of adverse events for each of the n_i patients reported by the i-th site would then be realizations of the corresponding Poisson process,

$Y_1, Y_2, \ldots, Y_{n_i} \sim Poi(\lambda_i)$.

We further assumed that the reporting rates $\lambda_i, i = 1,2, \ldots, N_{sites}$ of sites were drawn from a single study-level Gamma distribution $\Gamma(\alpha_{study}, \beta_{study})$to model the variability of reporting behaviours among the sites, and we picked vague priors for the study parameters $\alpha_{site} \sim Exp(1)$ and $\beta_{site} \sim Exp(10)$ to account for uncertainty. These relations are summarized in a graphical representation (Fig. 1). The circles represent random variables, shaded when they correspond to observed data. Arrows indicate conditional dependencies, and plates represent repeated elements, with their labels indicating how many times.

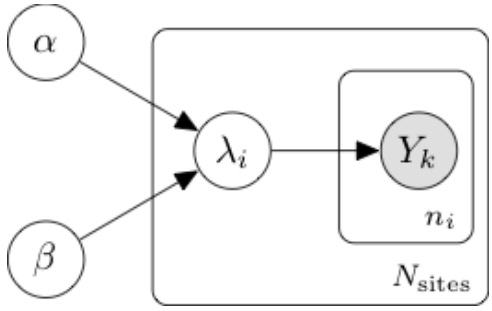


Figure 1. Graphical representation of the adverse event reporting model

In this hierarchical model, the posterior distribution of the site reporting rates λ_i given the observed numbers of AEs per patient reflects how many AEs per patient we expect to see at the individual sites, and the distribution width indicates the degree of certainty in these estimates. This width typically depends on the number of observations and how scattered they are. Even for sites with less patients, the mechanism of information borrowing enabled by the hierarchical structure leads to more robust estimates of the reporting rates.

The joint posterior distribution of the study level parameters α_{study} and β_{study} characterizes safety reporting patterns of a study and depends on the nature of the disease (e.g. cancers vs. cardiovascular diseases, etc.), the drug mechanism of action, the drug mode of administration, the design and execution of the clinical trial, and so on. In itself, the posterior distribution of these parameters is not very informative, but it can be used to build a posterior predictive distribution of the reporting rate of a yet unobserved site participating in the same trial. In particular, this predicted reporting rate provides a unique reference across the study by comparing the individual site reporting rates and assessing the risk of underreporting.

2.4. Inference and underreporting detection

Efficient sampling of the posterior distribution of hierarchical models requires specialized methods [13] such as the Hamiltonian Monte Carlo algorithm [14], which is readily implemented in modern probabilistic programming libraries. We used the TensorFlow-Probability library [15], and our code is available as a Jupyter notebook [16].

The samples of the posterior distribution returned by algorithms of the Markov Chain Monte Carlo family are typically used to compute expectation values with respect to that posterior. We started with the means and standard deviations of the site reporting rate samples to summarize these distributions, but we were ultimately interested in measuring the risk of underreporting. Since there is no single answer to our objective we picked two different metrics. In this model, it was natural to consider first the expected left tail area of the site rates under the posterior predictive distribution of reporting rates. In practice, this corresponds to the probability that a yet unseen reporting rate drawn randomly from the posterior predictive distribution falls below the inferred site reporting rates. We also computed the probability that the site rates fell below 75% of the posterior predicted reporting rate, as this quantity is easier to interpret.

The output is available in the code repository [16] and a sample of the sites with the top and bottom tail areas is presented in table 2. The inferred values of site 3046 illustrate interesting features of this model. Despite having a single observation of zero reported AEs, the inferred rate is still quite high, driven by information borrowed from the other sites, but with a certain uncertainty, characterized by a higher standard deviation than other sites with low numbers of reported AEs.

Clinical Investigator Site	Mean AE rate	Standard Deviation AE rate	Rate tail area	p75	Observed AE	Number of patients
site_3099	93.77	6.76	1.00	0.00	[140, 57]	2
site_3016	51.39	3.51	0.99	0.00	[17, 65, 75, 53]	4
site_3082	37.62	4.21	0.95	0.02	[45, 33]	2
site_3056	38.21	5.86	0.95	0.02	[41]	1
site_3054	33.76	5.53	0.92	0.03	[36]	1
...	...	...	...	...	...	...
site_3032	2.27	1.03	0.06	0.92	[3, 0]	2
site_3046	1.60	1.18	0.04	0.95	[0]	1
site_3037	1.32	0.51	0.02	0.96	[1, 0, 1, 3, 0]	5
site_3036	0.93	0.48	0.01	0.98	[0, 1, 0, 1]	4
site_3030	0.48	0.22	0.00	0.99	[0, 0, 0, 1, 0, 0, 0, 0, 0, 2]	10

Table 2 - Sample of the model output with statistics of posterior reporting rates, underreporting risk metrics and AE observations.

When it comes to deciding which sites to flag for under-reporting, a threshold has to be set by quality leads, as the values of the risk metrics cover a wide spectrum displayed in figure

The relationship between the posterior mean site rates and the corresponding expected rate tail areas follows the cumulative distribution function of the posterior predictive distribution of the reporting rates (Fig. 2). There is no clear rule to determine how low rate tail areas indicate underreporting. Low values might be due to the inherent variability of safety reporting. Nevertheless, auditing efforts should focus on the lowest values, for instance according to different alert levels at prespecified thresholds, e.g. 0.05 and 0.15, or up to a gap in the distribution of tail areas for more homogeneity in the QA process.

The user-friendly p75 metric is consistent with the more natural tail area (Fig. 3) and can help interpret the underreporting risk assessment.

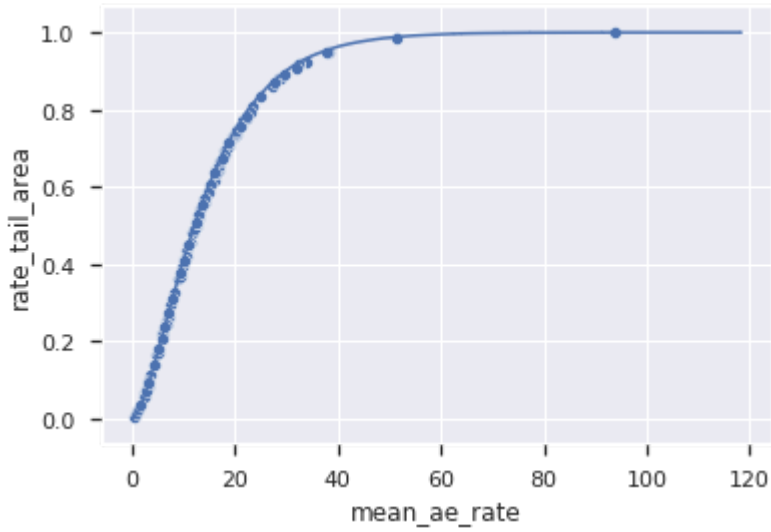


Fig. 2 - The rate tail area risk metric as a function of the posterior mean site rate.

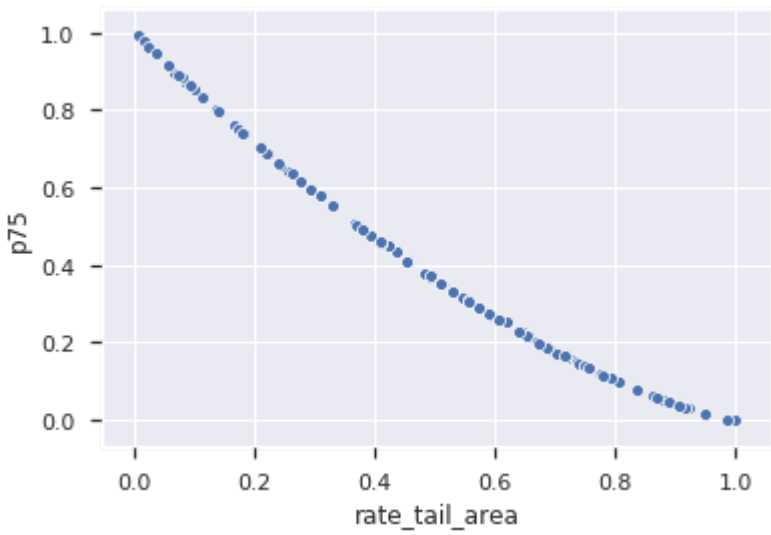


Fig. 3 - The relationship between the rate tail area and the p75 metric.

3. Discussion

The method presented here is applicable to completed studies to assess which sites might pose a risk of underreporting. In particular, it can demonstrate a degree of certainty in the completeness of collected safety data. In ongoing studies, patients do not enroll all at the same time which introduces more variability in the numbers of reported AEs. The longer a patient has been enrolled, the more AEs have been reported. We can still apply the same method, provided we select as observations the accrued number of adverse events for each patient up to the *n*-th visit and exclude

patients who have not reached that milestone. This analysis can be repeated for different values of *n*. In particular, when a study is close to a data-base lock (e.g. before performing an interim analysis), the model could be used to guide quality leads and/or clinical operations staff to detect underreporting sites and trigger queries and AE reconciliation. Hence, this gives reassurance to health authorities inspectors that AE underreporting had been detected, corrective actions had been implemented and integrity of the data had not been compromised [2].

An added benefit of the proposed Bayesian approach and the selected risk metrics is that the outputs are calibrated probabilities. The results of this underreporting risk assessment conducted on different studies and at different milestones are immediately comparable. A sponsor overseeing several studies can thus keep an overview of all of them, and monitor the evolution of the risk of underreporting over time.

In parallel to developing a new model for detection of AE underreporting, we have been piloting a machine learning model with Quality Assurance (QA) staff since May 2019. The outputs of the Bayesian approach (i.e. probabilities of site underreporting) will be integrated in the current QA dashboard, together with the outputs of the ML model [3] that are already available to Quality Program Leads. The advantage of combining both approaches is to have AE patient level predictions (from the ML model) and detection of AE underreporting at the site level (using the Bayesian approach). This will enhance further the quality assurance activities for safety reporting from clinical trials.

This model was developed during the Covid-19 pandemic where on-site audits could not be performed [17]. Having a data product enabling remote monitoring of safety reporting from investigator sites was essential to ensure business continuity for clinical quality assurance activities [18]. Our approach has the potential to reduce the need for on-site audits and thereby shift the focus away from source data validation and verification towards pre-identified, higher risk areas. It can contribute to a major shift for QA, where advanced analytics can detect and mitigate issues faster, and ultimately accelerate approval and patient access of innovative drugs.

Jianing Di *et al.* [19] also proposed the use of Bayesian methods for adverse events monitoring with a more clinical purpose. Their approach focused on the continuous monitoring of safety events to address the lack of knowledge of the full safety profile of drugs under clinical investigation. Their model could be applied for signal detection in early-phase trials and could also give further evidence to independent data monitoring committees for late stage studies. We developed a different model as our focus was on sites rather than patients. This illustrates one of the strengths of Bayesian data analysis, where different models of the same data can be optimized to answer different questions about the underlying process.

Clinical trials generate large amounts of data traditionally analysed with frequentist methods, including statistical tests and population parameter estimations, aimed at clinical questions related to efficacy and safety. There has been a push in recent years for Bayesian adaptive designs that have the potential to accelerate and optimize clinical trial execution. Examples include Bayesian sequential design, adaptive randomization, and information borrowing from past trials. For example, in a study redesigning a phase III clinical trial, a Bayesian sequential design could shorten the trial duration by 15 to 40 weeks and recruited 231 to 336 fewer patients [20]. Our approach for the detection of safety underreporting demonstrates the potential of Bayesian data analysis to address secondary questions arising from clinical trials such as quality assurance or trial monitoring. In clinical QA, where the majority of business problems are anomaly detection or risk assessment, there is a good rationale for exploring further applications of Bayesian approaches, for example for identification of laboratory data anomalies in clinical trials or in identifying issues with the number of unreported/reported protocol deviations by clinical study sites.

4. Conclusion

In this paper, we presented our approach to quantify the risk of AE underreporting from clinical trial investigator sites. We addressed a shortcoming of the model developed in our previous work that was good at predicting the evolution of safety reporting in clinical studies but failed to properly quantify the probabilities of quality issues.

The new model will be integrated into the current dashboard designed for quality leads. This is part of a broader effort at our Research and Development Quality organization. Similar approaches using statistical modeling and applied to other key risk areas (e.g. informed consent, data integrity) are being developed in order to provide a full set of advanced analytics solutions for clinical quality [3, 4, 17, 21]. We will also continue to explore the application of Bayesian methods to other datasets generated during the conduct of clinical study for QA purposes (e.g. protocol deviations).

However, in order to implement routine, remote, and analytics-driven QA operations, sponsors and agencies will have to continue to collaborate and address challenges such as fit-for-purpose IT infrastructures, automation, cross-company QA data sharing, QA staff data literacy and model validation [17, 20]. While the Covid-19 pandemic accelerated the adoption of new ways of working and pushed innovation further, it also brought new rationales for a change in the QA paradigm, i.e. where advanced analytics can help conducting QA activities remotely, detecting and mitigating issues faster, and ultimately accelerating approval and patient access of innovative drugs.

References

1. Medicine And Healthcare Products Regulatory Agency. Gcp Inspection Metrics Report. 2018. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/706356/GCP_INSPECTIONS_METRICS_2016-2017__final_11-05-18_.pdf. Accessed 18 Dec 2020.

2. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. E26(R2) Guideline for Good Clinical Practices. 2016. https://database.ich.org/sites/default/files/E6_R2_Addendum.pdf Accessed on 18 Dec 2020.

3. Ménard T, Barmaz Y, Koneswarakantha B, Bowling R, Popko L. Enabling Data-driven Clinical Quality Assurance: Predicting Adverse Event Reporting In Clinical Trials Using Machine Learning. Drug Saf. 2019;42(9):1045–53.

4. Ménard T, Koneswarakantha B, Rolo D, Barmaz Y, Bowling R, Popko L. Follow-up On The Use Of Machine Learning In Clinical Quality Assurance: Can We Detect Adverse Event under-reporting In Oncology Trials? Drug Saf. 2020;43(3):295–296.

5. Food and Drug Administration. Clinical Investigator Inspection List (CLIIL). https://www.fda.gov/drugs/drug-approvals-and-databases/clinical-investigator-inspection-list-cliil. Accessed and searched on 18 Dec 2020

6. Food and Drug Administration. Warning letter to AB Science 6/16/15 https://www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/warning-letters/ab-science-06162015. Accessed on 18 Dec 2020.

7. Sacks L, Shamsuddin H, Yasinskaya Y, Bouri K, Lanthier M, Sherman R. Scientific and regulatory reasons for delay and denial of fda approval of initial applications for new drugs, 2000–2012. JAMA. 2014;311(4):378–84.

8. Pitrou I, Boutron I, Ahmad N, Ravaud P. Reporting of safety results in published reports of randomized controlled trials. Arch Intern Med. 2009;169(19):1756–61.

9. Li H, Hawlk S, Hanna K, Klein G, Petteway S. Developing and implementing a comprehensive clinical QA audit program. Qual Assur J. 2007;11:128–37.

10. Gelman A, Carlin JB, Stern HS, Rubin DB. (2004). Bayesian Data Analysis, Second Edition. Chapman & Hall/CRC Texts in Statistical Science. 2004.

11. Project Data Sphere. https://www.projectdatasphere.org/projectdatasphere/html/home. Accessed on 18 Dec 2020.

12. Data For Clinical Trial Registered NCT00617669. Project Data Sphere. https://data.projectdatasphere.org/projectdatasphere/html/content/104. Accessed on 18 Dec 2020.

13. Betancourt M., Girolami M. Hamiltonian Monte Carlo for Hierarchical Models. 2013, arXiv 1312.0906. http://arxiv.org/abs/1312.0906.

14. Duane S., Kennedy AD., Pendleton BJ., Roweth D. Hybrid Monte Carlo. Physics Letters B. 1987, 195 (2): 216–222.

15. TensorFlow Probability library. https://www.tensorflow.org/probability/. Accessed on 18 Dec 2020

16. Barmaz Y. Code repository and Jupyter Notebook for the model. https://github.com/ybarmaz/bayesian-ae-reporting/. Accessed on 18 Dec 2020.

17. Food and Drug Administration. Guidance on Conduct of Clinical Trials of Medical Products during COVID-19 Public Health Emergency. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/fda-guidance-conduct-clinical-trials-medical-products-during-covid-19-public-health-emergency. Accessed on 18 Dec 2020.

18. Ménard, T., Bowling R., Mehta P. et al. Leveraging analytics to assure quality during the Covid-19 pandemic - The COVACTA clinical study example. Contemporary Clin Trials Comm. 2020; 20:100662.

19. Di J, Wang D, Brashear HR, Dragalin V, Krams M. Continuous Event Monitoring Via A Bayesian Predictive Approach. Pharmaceut. Statist. 2015, 15(2):109–122

20. Ryan, E.G., Bruce, J., Metcalfe, A.J. et al. Using Bayesian adaptive designs to improve phase III trials: a respiratory care example. BMC Med Res Methodol 2019, 19:99

21. Koneswarakantha, B., Ménard T., Barmaz Y., *et al*. Harnessing the Power of Quality Assurance Data: Can We Use Statistical Modeling for Quality Risk Assessment of Clinical Trials?. Ther Innov Regul Sci 2020; 54:1227–1235

Notes

Author's contributions

YB developed the model, wrote the code and produced the figures. YB and TM wrote the manuscript. TM quality checked the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgements

Content review was provided by Kelly Kwon who was employed by Roche/Genentech at the time this research was completed. Project Data Sphere data was curated by Donato Rolo and Björn Koneswarakantha (who were employed by Roche/Genentech at the time this research was completed).

Compliance with Ethical Standards

Funding

Funding for development and testing of the safety reporting model was supplied by Roche.

Conflict of Interest

Yves Barmaz and Timothé Ménard were employed by Roche at the time this research was completed.

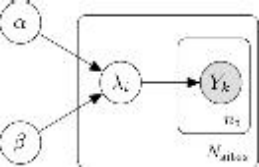
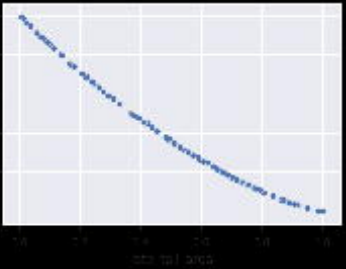
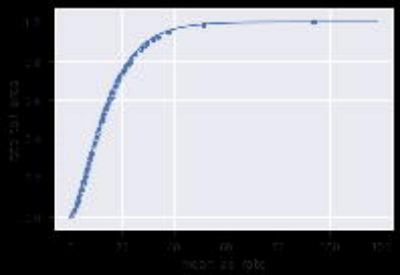
Ethics Statement

All human subject data used in this analysis were used in a fully de-identified format (see also the link to Project Data Sphere below)

Data and Code availability

The data can be accessed on Project Data Sphere - https://data.projectdatasphere.org/projectdatasphere/html/content/104

The code and the curated data are available on https://github.com/ybarmaz/bayesian-ae-reporting/


α
β
λ_i
Y_k
n_i
N_{sites}