Paper AR06

Predictive Analytics for Adverse Event Under-Reporting

Steven A. Gilbert, Phd, Cambridge, MA, US

ABSTRACT

In this paper we provide a framework for incorporating predictive analytics into central monitoring to identify clinical study sites which under-report adverse events. Under-reporting can risk patient safety as well as overall trial integrity and is more difficult to identify in an ongoing trial than overly high adverse event rates. Common study characteristics that contribute to the difficulty in identifying under-reporting include; a large numbers of sites with many sites have only a small number of subjects, staggered entry of sites into the study, site-to-site differences in patient demographics and regional differences in health care. We explain how to identify under (over) - reporting sites starting with simple graphics and progressing naturally to predictive modeling. The paper is intended to be pedagogic with methods explained in a plain language and references as necessary to software are provided for those looking to implement these methods.

INTRODUCTION

Predictive analytical models can help central monitors identify clinical trial sites that under-report adverse events. Methods for identifying under-reporting sites range from simple exploratory graphs to advanced predictive models using statistical and epidemiological methods. Exploratory analyses can be done with commonly available office software like Excel or Power BI and are useful for both identifying suspicious sites and communicating findings to study teamd. More advanced methods require less common specialty software like R, Python, and probabilistic programming languages (PPLs) such as Stan or PyMC. However, central monitors do not need expertise in the specialty software since they will repeat very similar analyses from trial to trial allowing the coding to be done once by a smaller team of data scientists or statisticians; all they need is training in how to interpret the results of the analyses.

In this paper, we use a fictitious trial of 30 sites and 485 subjects to demonstrate methods from exploratory graphs to predictive models.

THE CHALLENGES AND OPPORTUNITIES OF AE UNDER REPORTING

Before introducing methods, it is helpful to enumerate the challenges these methods must meet to adequately identify under-reporting sites. The primary challenges that confront such analyses are:

1. Under-reporting is not operationally obvious. Teams and medical reviewers see individual adverse event reports and not necessarily aggregated summaries that highlight lack of adverse event reports. That is, a reviewer is given adverse event reports to check for anomalies, not reports of non-events.
2. Under-reporting is statistically challenging to identify; in some therapeutic areas and study designs with low adverse event rates, a small site may plausibly have zero or a low number of event reports due to random variation alone.
3. Site to site comparisons can be difficult due to uneven subject enrollment ranging from 1 or 2 subjects at a small site to hundreds or more at a large site.
4. Site to site comparison can also be difficult due to differences in the ‘patient-mix’ from site to site (eg, sicker or healthier subjects).
5. The distribution of event counts and rates do not follow a normal or gaussian (bell shaped) distribution making many standard visualization tools such as histograms and scatter plots difficult to use.
6. Little data at the beginning of a trial making it difficult to visualize patterns or fit complicated statistical or machine learning models, the so called “cold-start” problem.

HOW TO START

"Begin at the beginning," the King said, very gravely, "and go on till you come to the end: then stop."
- Lewis Carroll, Alice in Wonderland

We follow Hadley Wickham's data analysis workflow [5]. It starts by importing and tidying the data, turning it into a clean file like an Excel worksheet with rows for observations and columns for variables (e.g., subject number, dates, event descriptions). Rows can represent individual reports, subject summaries, or site summaries. This is often time-consuming, so developing an efficient data pipeline with IT is crucial to standardize trial monitoring.
This paper focuses on the next step after data import: an iterative loop of visualization, modeling, and transforming (creating new variables, normalizing data, feature engineering). Visualization is key as it highlights data errors, aids quality control, identifies missing data, and checks if the assumptions needed for advanced model are met.

VISUALIZATION AND TRANSFORMATION BASICS: SCATTER PLOT VERSUS NORMALIZED DATA

Adverse event reporting assumes that longer exposure to a drug in a trial leads to more reported events, and sites with more subjects will have more reports. Combining exposure time and subject count gives total patient-days, summarizing the study duration of all subjects at a site. This is a crude summary that assumes the probability of reporting an event is the same for each day a subject is observed. For those familiar with survival analysis these ng are the assumptions a homogenous Poisson process[10], the simplest model to account for time as well as events. More realistic assumptions where rates can vary for subject over time are also possible, for example in vaccine trials most subjects have adverse events very early in the trial due to injection site reactions (eg pain, swelling etc) and very few later on, this is beyond the scope of the present paper.

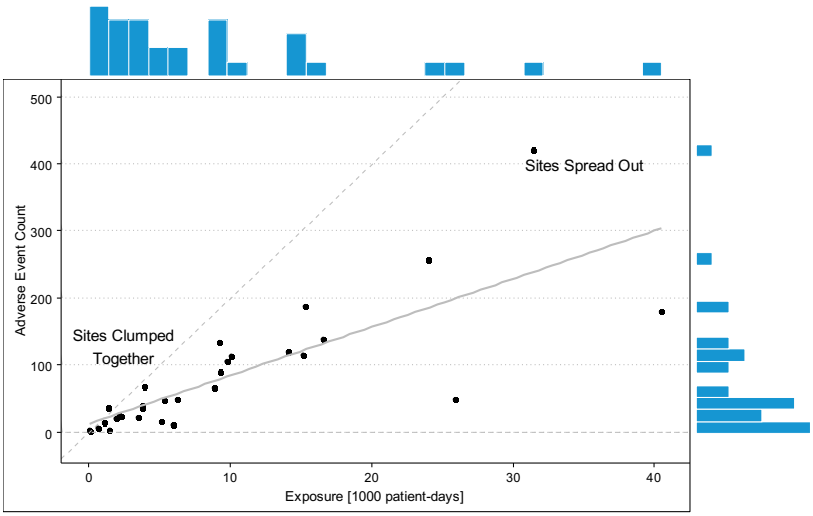

Figure 1 AE Counts vs Exposure

Figure 1 presents the total number of Adverse Event counts per site on the y-axis versus exposure time on the x-axis. The plot is enhanced with histograms of the variables in the top and right margins. The histogram at the top indicates that exposure times are right-skewed, with most exposures concentrated at the lower end of the axis and several larger observations extending to the right. This skewness is attributable to the study's distribution of site sizes. Distributions of site sizes have been well studied in the literature; site sizes are commonly modeled for planning purposes using a Poisson-gamma (Negative Binomial) model (Anisomov, 2011 [6]). This model aptly describes the f the histograms shape, the mathematical details are not necessary at this point, other than to note that intuitions for identifying patterns and outliers in data based on reviewing normal/gaussian distributions (the so-called bell shaped curve) will not work well for these types of data.

The histogram on the right of the chart illustrates the distribution of event counts which is also skewed. This should not be surprising since we assumed the number of reports is related (correlated) with the exposure time at the site which in this case is highly correlated with the size of the site. Therefore, the right skewed distribution of exposure times naturally carries forward into the event counts. The scatter plot in the center of the figure validates this assumption showing that sites with small exposure times have low counts and those with high exposure times have higher counts (the regression line in grey has a positive/upward slope highlighting this pattern).

Unfortunately the graph is not helpful in identifying under (over)-reporting sites. The sites with low reporting rates requiring further examination are all clumped into the lower left corner of the plot making it difficult to distinguish between sites. This is a consequence of the skewed distributions putting most of the observations in that lower-left corner. There is also a statistical issue making the figure hard to read. Mathematically distributions of counts (in this case report counts) have variability related to their average. That is illustrated with the dashed grey lines fanning out from the origin, visually the spread of the points increases as we move to the right. This spread is mathematically predictable but difficult to adjust for visually when looking at the plot to identify outliers.

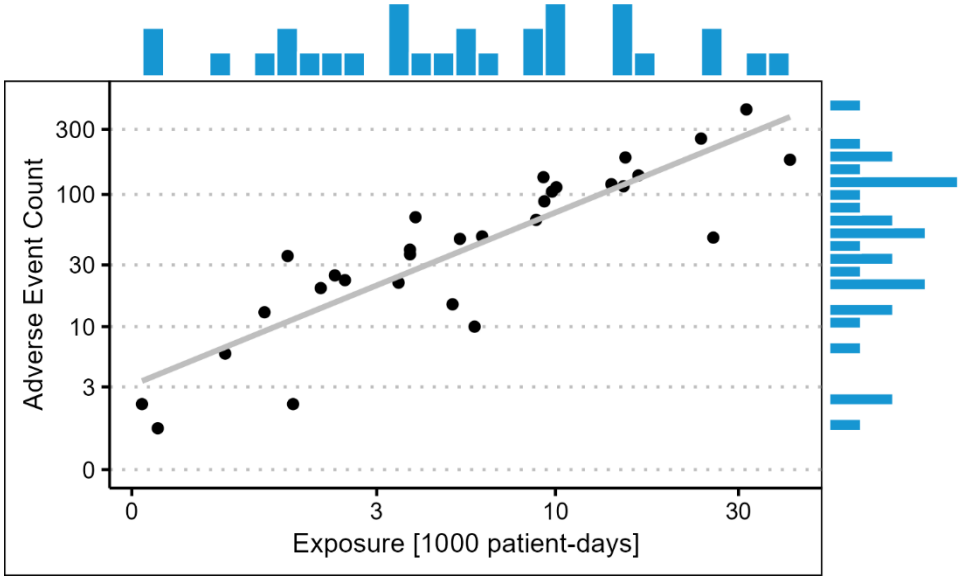

Figure 2 Adverse Events vs. Exposure Log-Log Scale

Figure 2 has a small change to reduce distortion from skewed data and increasing variability. The data was transformed using a modified logarithmic transformation to spread out sites clumped in the lower corner. The marginal histograms now show more evenly distributed data and the scatter plot is easier to read with points closer to the grey regression line, forming a stable band above and below it. However, identifying under-reporting remains challenging due to difficulty in visually assessing the importance of points far below the regression line.

Some key points for implementation should be noted. The modified transformation, which adds 1 before taking the logarithm, ensures zeros remain zeros. A pure logarithmic transformation would drop zero counts, distorting data interpretation by only showing positive counts. Although this dataset has no zero counts, they are common in real data, especially early in trials.

For presentation, axis labels are shown on the original untransformed scale. For instance, the x-axis tick marks at 0, 3, 10, and 30 [1000 patient-days] translate to 0, 1.39, 2.40, and 3.43 on the transformed scale, which is harder to interpret.

Lastly, a logarithmic transformation makes data appear linear even if event counts are proportional to some power of exposure time (e.g., counts proportional to the square root of exposure time). This generalizes our homogeneous Poisson assumption, making the plot more universally applicable without extra effort from the monitor.

COMPOSITE TRANSFORMATION OF VARIABLES

Event counts at different sites cannot be directly compared without considering the exposure time or site size. This section discusses transforming two variables, count and exposure, to create a single normalized variable. The simplest method is to normalize site counts to rates by dividing the event count by the exposure time, resulting in a single number that can be compared between sites (with some caveats explained later).

Figure 3 shows a raincloud plot of site counts divided by the exposure time in days (and then divided by 1,000 for

readability). The blue “cloud” at the top is a density estimate, essentially a smoothed-out histogram. A boxplot is displayed in the middle, and the “raindrops” at the bottom provide detailed information about the observed data. The boxplot, also known as a box and whisker chart [7], is a schematic representation of the data. The box extends from the first quartile (Q1), which separates the lowest 25% of the values, to the third quartile (Q3), which separates the lowest 75% from the highest 25%. By definition, 50% of the observed data falls within this box. The box has a line at the median, separating the lower and upper 50% of the data. The black lines extending from the box are the whiskers.

Depending on the software used, there may be different rules for the length of the whiskers. Data points falling far from the whiskers may be outliers, plotted in red. The rule for identifying outliers assumes the data are normally distributed, flagging approximately 6 out of 1,000 observations sampled from a normal distribution. In this case, two sites are flagged as outliers for potentially having too many reports, while no sites are flagged for under-reporting.

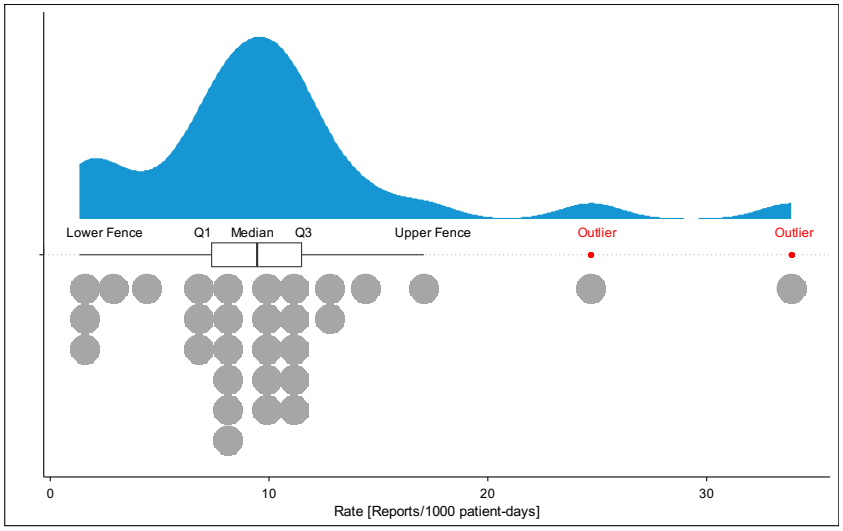

Figure 3 Untransformed Rates

Simple methods such as standardizing, graphing and outlier heuristics can also benefit from further transforming the data as was done in the last section. In Figure 3 the data are still skewed and clumped up towards the left; what if we apply a transformation to the rates to spread the data out? In Figure 4 we again use the modified logarithmic transformation to the rates and create the same type of plots. The distribution of the variables are now more symmetric around the median (more like a gaussian distribution) and therefore the outlier rules now have a better chance to find under-reporting sites, in this case 4.

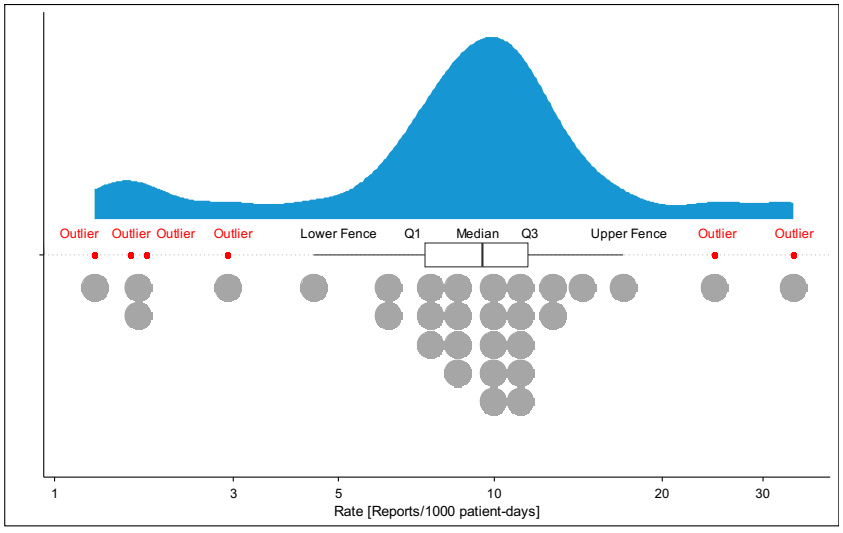

Figure 4 Transformed Rates

FUNNEL PLOTS

We present a plotting method using a statistical regression model for count data to better identify outliers. Figure 5 is a funnel plot [11], where the y-axis shows the event rate in events/patient-day on a square root scale for symmetry (this step is not necessary) in the figure. The solid black line in the middle is the estimated rate for all sites combined labelled the "Average Rate" on the plot to help guide the eye to an approximate "center" of the plot. The red dashed lines are confidence limits. Funnel plots get their name from wide limits on the left that narrow as they move right, reflecting increased precision with larger sample sizes (or longer durations).

The figure also makes allowances for the fact that the non-normal distribution of the data even after transforming counts to rates. Asymmetric confidence intervals are used because it is harder to find under-reporting than over-reporting (a small site can easily have zero events due to random luck); the upper red line is a 95% confidence limit and the lower red line is a 20% limit. The limits are calculated using a negative binomial regression model. A special regression model for data suitable when the responses are counts. The model is readily fitted in many software packages including R (MASS package glm.nb for the example, see appendix for more detail), Python (statsmodels) or SAS. The underlying model is "simple" in that it assumes all sites have exactly the same event rate and that differences in counts between sites are due only to differences in exposure and to random sampling.

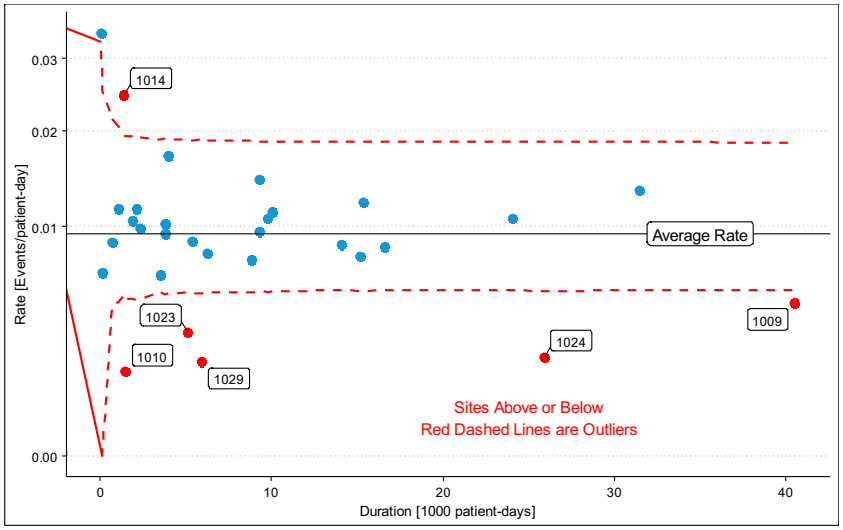

Figure 5 Funnel Plot

In Figure 5, outliers are identified by looking for sites that are either above or below the red dashed lines. It appears that site 1014 may be over-reporting while sites 1010, 1023, 1029, 1024 and 1009 may be under-reporting. If this were presented to a study team, they may have questions, such as, "Have you accounted for the high number of older subjects at site 1014?" To answer that question, we need methods that can adjust for patient-mix as well at duration, we turn to these methods in the next section.

PREDICTIVE ANALYTICS AND ANOMALY DETECTION

"All happy families are alike; each unhappy family is unhappy in it is own way."

- Leo Tolstoy, Anna Karenina

"Predictive analytics is the process of using data to forecast future outcomes. The process uses data analysis, machine learning, artificial intelligence, and statistical models to find patterns that might predict future behavior. Organizations can use historic and current data to forecast trends and behaviors seconds, days, or years into the future with a great deal of precision."[1] We are going to do this while keeping in mind Tolstoy's observation and complication number 6; a minimal amount of data for model fitting at the beginning of the trial.

The strategy is to fit a predictive model that includes historical knowledge when appropriate and predict how many AEs a site would report in a large number of future hypothetical (aka simulated) trials where sites follow the protocol and good clinical practice. Thus, creating a reference set of what a good trial should look like; if any of the sites fall too far out in the extremes of this reference set of hypothetical trial results, the site is flagged for further inspection. The focus is on simulating the data under the assumptions of a well-run trial, the happy families in Tolstoy's quote and identifying unhappy families/sites without specifying exactly why they are unhappy. This is simpler and far more practical than creating a highly complex model that can account for all of the ways data collection can go wrong, all that is needed is to find a discrepancy between the simulated results and the observed data, the exact nature of the discrepancy can be left for a further root-cause analysis.

The tool used here is a random effects negative binomial regression fitted in R using the brms package. Regression methods have been used in epidemiology to standardize rates for multiple variables simultaneously instead of just a single variable (eg duration in the last section). Bayesian statistics is one of two major approaches to statistics, Frequentist and Bayesian software for both approaches are readily available. The differences between the approached is beyond the scope of this paper but we note that the Bayesian approach was used for this analysis primarily for computational reasons: 1) The software is more flexible allowing for custom models appropriate for this use-case to be quicky prototyped and fitted; 2) Once a model is fitted, downstream activities including making predictions and transforming variables to different scales for graphing and presenting are straightforwad. The hard work is the initial fitting which is done by complex algorithms already built into the software and need not be re-invented by the user; 3) The use of prior distributions [8] a feature of Bayesian methods provide a way of incorporating historical knowledge of site behavior into the modeling, above and beyond the decision of whether to include a covariate. There are some drawbacks however, the level of expertise required to build and fit these models is higher than more standard Frequentist models and they have a higher computational burden requiring some time to produce results for very large trials.

Figure 6 displays the output of the model run on our dataset. The model adjusts for Age, Sex and duration on trial at a patient level. That is, every row in the dataset represents a single subject and Site is a covariate. Each site will have a separate row in the data for each subject. If the subject has no event, the event count variable will be zero.Age, sex and duration (actually the logarithm of the duration for technical reasons) are covariates in this regression model. The inclusion of covariates should be chosen on a trail by trial basis with input from clinical colleagues. The purpose of the covariates is to adjust site rates for known factors (eg. Older sicker patients are expected to have more adverse events than younger healthier patients) that may affect the number of adverse events at a site due to patient characteristics, but do not reflect on the underlying quality of the site. For example, when presenting results to a study team they can be told the results have already accounted for the patient-mix and the apparent under (over) -reporting is not due to overly healthy (sick) subjects.

On the left-hand side of Figure 6 is a 'forest' plot (if you tilt your head some people think the black bars look like tree trunks). The results for the sites are arranged vertically. At the top of the plot the first red dot represents the observed event rate for site 1029, with the size of the red dot proportional to the total exposure at the site; small dots represent small sites and large dots represent large sites. The grey dashed line running from the top to bottom of the figure shows the overall site average rate, anything to the left is less than average and anything to the right is greater than average. The horizontal black lines and dots show a model based prediction interval for each site; the black dot is the predicted rate and the lines show the width of the interval (again 20% on the left and 95% on the right). Visually look for red dots (observed) values that are to the left or right of the associated interval to identify outliers. For example, the first 4 sites at the top have observed values lower than the left end of the intervals indicating suspiciously low event rates. Lastly, there is a set of number that more succinctly show how low the event rate is compared to the model predictions. For example, site 1029 had 0.02 or 0.2% of the simulated trials with a lower event rates than what was observed, indicating this is an extreme observation under the "happy family" model. For sites with high reporting rates (eg 1014 at the bottom of the figure) most simulated trials are less than the observed and the proportions are large, in this case 0.95 or 95%. The reader is cautioned not to treat these numbers as traditional p-values with guaranteed false-positive rates but rather a quantitative summary of how extreme a site is and a way to prioritize sites for further follow-up.

Looking again at Figure 6 note that not only do the site have different observed rates, they have different predicted rates (the black dots). This is different from the funnel plot Figure 5 where the model was much more rigid and predicts all sites have the same rate. The differences in the predictions come from two sources. First the covariates and duration, since each site has a different patient-mix and total exposure time. Second, the model includes a "random" site effect allowing for some site- to- site differences not accounted for by the covariates. Site behavior depends on many factors including, investigator experience, site coordinator experience, type of site (eg hospital, clinic, academic center) and others. It is not practical or possible to gather all of this information accurately for trial monitoring purposes. Instead, the random effect with constraints on how much extra variability (this was done with a "strong" prior distribution [9]) was used.

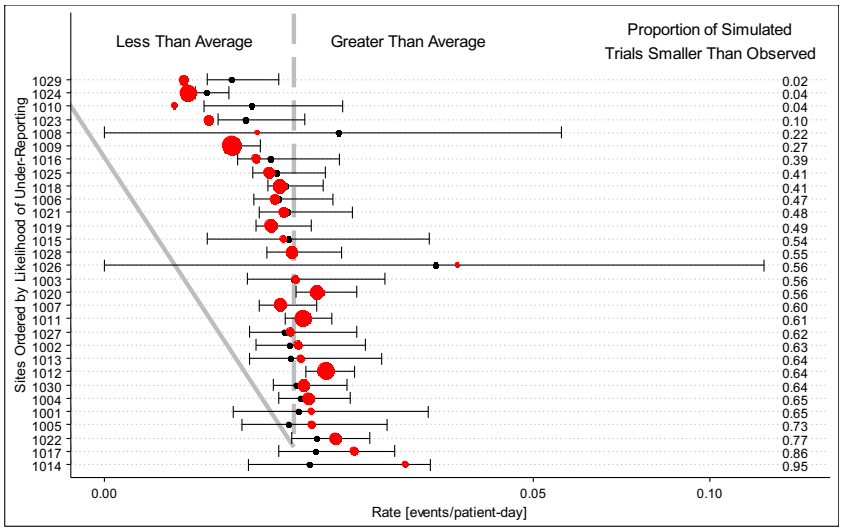

Figure 6 Predictive Model Forest Plot

Figure 7 shows an alternative plot of the same analysis. Using the fact that once a Bayesian model is fitted, it is straightforward to produce any desired summary from the results. In this plot instead of plotting the predictions the dots represent the difference between the observed and predicted values. All sites are expected to have a zero difference, the dashed grey line. Visually sites are now compared by seeing how far to the left or right they fall from the zero line. The predictions are hidden from view making for a simpler but slightly less informative plot. All sites with dots to the left are reporting less than expected and those to the right more than expected. Due to random variability we do not expect the numbers to ever be exactly zero and allow for some slack in the estimate. This is done with the prediction intervals, which are still asymmetrical. In this plot, any site with its interval entirely to the left of the zero line is suspiciously low and intervals entirely to the right are suspiciously high.

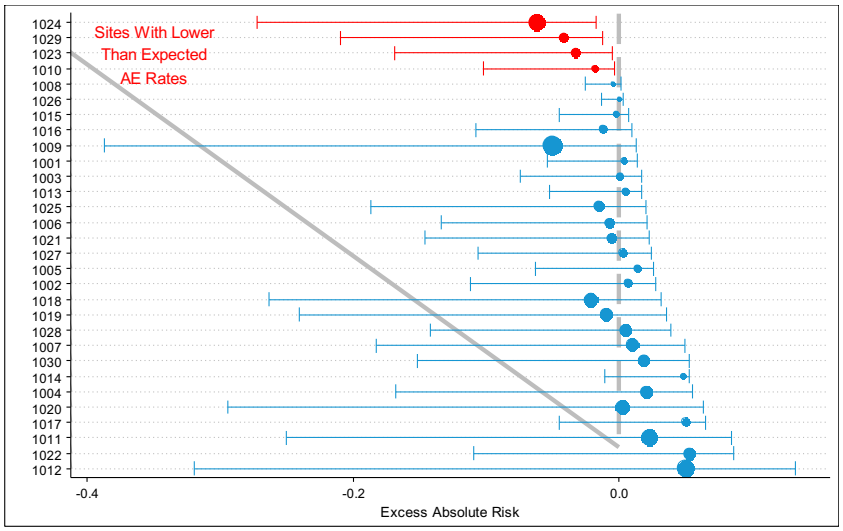

Figure 7 Excess Risk Plot

CONCLUSION

We have demonstrated a range of progressively more intricate methods, spanning from simple data transformations and visualizations to advanced models. How effective were these methods? In the simulation of data, four sites (1009, 1023, 1024, and 1029) were randomly selected to have their expected counts reduced by 50%.
In Figure 3, using a box plot with untransformed event rate data, no under-reporting sites were identified, and two over-reporting sites (1014 and 1026) were flagged. When the box plot was applied to transformed data in Figure 4, the results were improved; it correctly identified sites 1023, 1024, and 1029, though it missed site 1009 and falsely flagged sites 1010, 1014, and 1026. The funnel plot in Figure 5 performed even better, accurately identifying all four under-reporting sites and only flagging two additional sites (1010 and 1014). Lastly, the predictive model successfully

identified three out of four under-reporting sites (missing 1009 but flagging 1010) and correctly determined that no sites were over-reporting.

It is important to note that this analysis is based on a single simulation. Comprehensive comparison of these methods would require thousands of simulations and aggregation of the results. While no predictive method will work flawlessly every time—a limitation inherent to any predictive approach—they can all be adjusted to be more or less permissive in flagging sites for further review. Unlike medical trial outcomes where regulators are primarily concerned with false positives (approving a drug that does not benefit patients), the primary concern in monitoring is failing to flag an issue before it becomes irreparable, which could compromise the trial's validity. Although false positives are still problematic, leading to increased monitoring costs and workload for sponsors and sites, they are considerably less detrimental than having to repeat a study or failing to secure approval for a safe and effective drug.

Sponsors must carefully consider the level of effort they wish to devote to this process. Risk-based monitoring is not conducted for a single parameter alone, such as event rates, but across multiple risk indicators and data quality checks, under the assumption that poorly performing sites will likely exhibit deficiencies across various measures.

There are numerous avenues for future research. These include the analysis of historical data to identify better covariates and models, the utilization of alternative, more efficient computational or machine learning methods, and further investigation into the relative success (operating characteristics) of these methods across trials typical of the organization conducting the monitoring.

APPENDICES

FUNNEL PLOT

The negative binomial model is fitted using the following line of code:

```
nbml <- MASS::glm.nb(AEs ~ 1 + offset(log(TotalDuration)),  data=SiteLevelDf
```

where SiteLevelDF is the data set with 1 row per site, AEs are the count of reports a the site, TotalDuration is the sum of patient-days at the site and log() is the natural (base e) logarithm.

PREDICTIVE MODEL

The Bayesian predictive model is fitted with the following R code using the brms package [12],

```
myreg <-  brm(AEcount ~ (1|SITE) + log(duration)+Age + Sex,
              data= PatientLevelDF,
              family=negbinomial,
              cores=2, chains=2,
              prior= set_prior("gamma(2,66)", class="sd")),
```

where PatientLevelDF is a dataset with one record per subject (if they had no AE reports they are still included in the dataset with a count of zero), AEcount is the number of adverse events reported for a single patient, SITE is the clinical site number, duration is the days of exposure for the patient, Age is age in years and Sex is an indicator of male or female. The family=negbinomial telss brms to fit a count data regression for negative binomial data. Prior puts an informative prior on the variability from site to site stopping the model from over-fitting and obscuring outliers.

REFERENCES

[1] Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. Journal of Statistical Software, 76, 1–37.

[2] PyMC: A Modern and Comprehensive Probabilistic Programming Framework in Python, Abril-Pla O, Andreani V, Carroll C, Dong L, Fonnesbeck CJ, Kochurov M, Kumar R, Lao J, Luhmann CC, Martin OA, Osthege M, Vieira R, Wiecki T, Zinkov R. (2023)

[3] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[4] Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

[5] Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). R for Data Science (2nd ed.). O'Reilly Media

[6] Anisimov, V. (2011). Statistical Modeling of Clinical Trials (Recruitment and Randomization). Communications in Statistics – Theory and Methods, 40: 3684-3699

[7] Tukey, J., (1977) Exploratory data analysis. Addison-Wesley

[8] Google Cloud What is predictive analytics and how does it work? | Google Cloud

[9] Gelman, A., Carlin, B., Stern, H., unson, D., Vehtari, A., and Rubin, D., (2025) BDA3.pdf

[10] Ross, S., (1970) Applied Probability Models with Optimization Applications, Dover.

[11] Spiegelhatler, D., (2005). Funnel plots for comparing institutional performance. Statistics in Medicine, 24: 1185-1202

[12] Paul-Christian Bürkner (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software, 80(1), 1-28.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:
Author Name: Steven A. Gilbert
Company: Pfizer
Email: steven.a.gilbert@pfizer.com