

AR06: Predictive Analytics for Adverse Event Under-reporting

US CONNECT 2025

Steven A. Gilbert PhD

Pfizer

steven.a.gilbert@pfizer.com



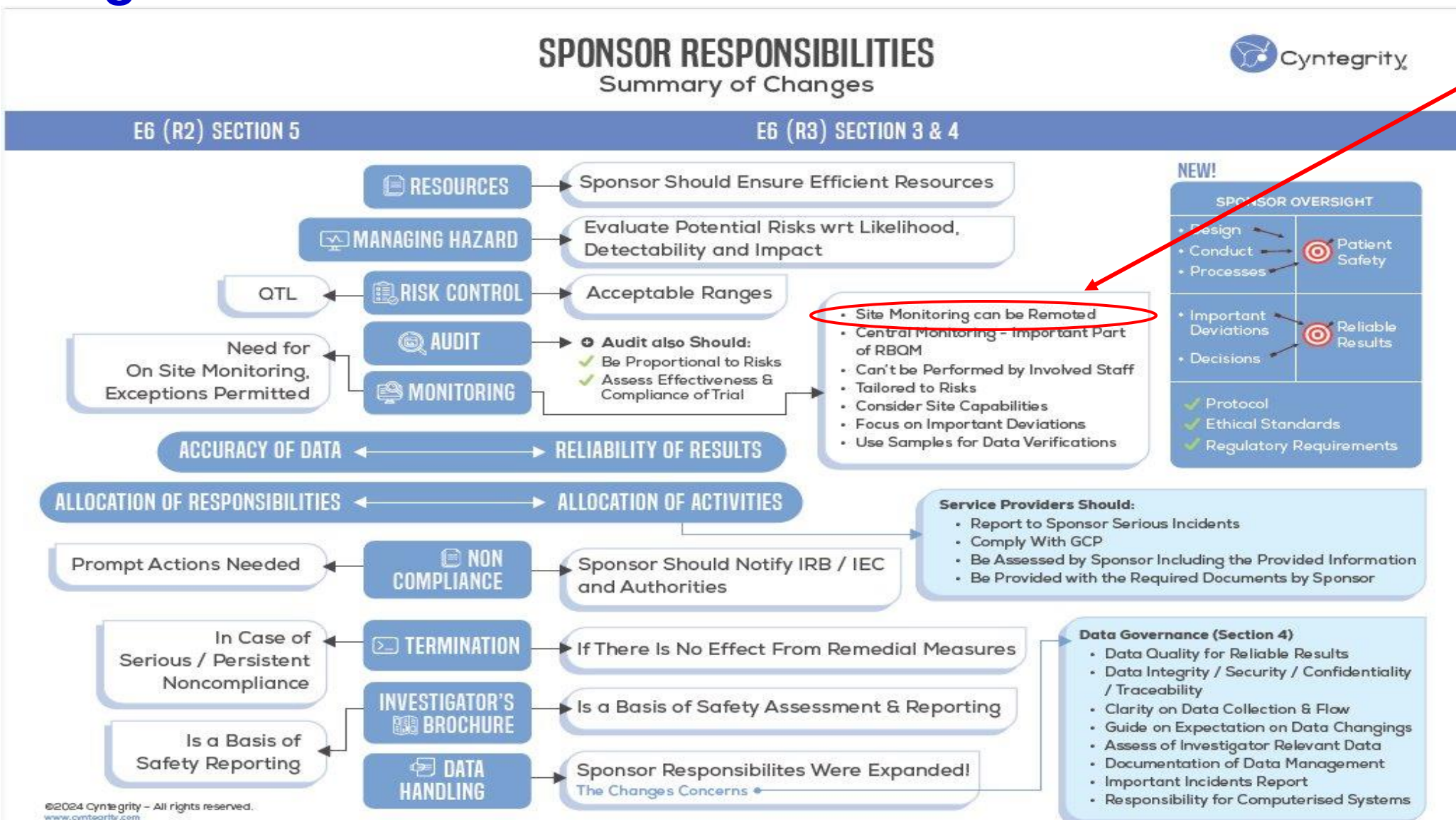
March 16-19

Disclaimer

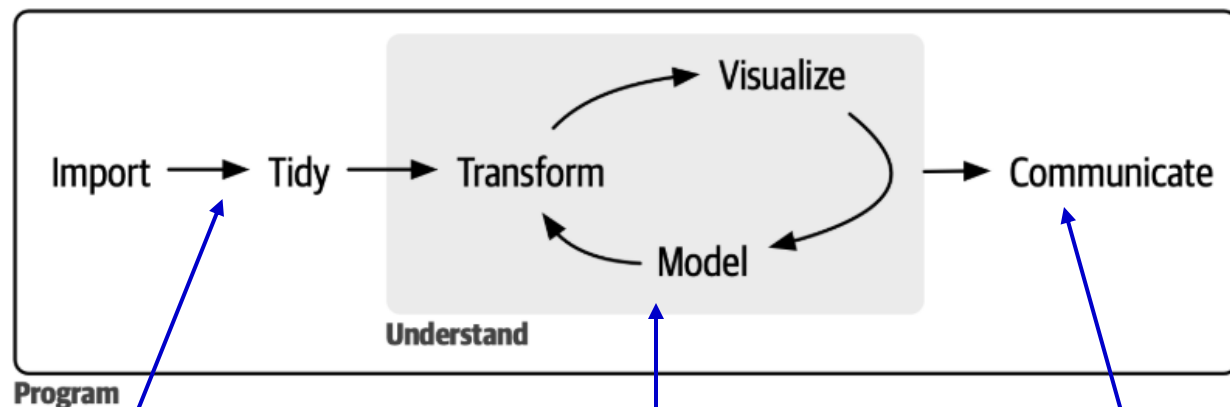
- All opinions are those of Steve Gilbert and not Pfizer
- All examples use completely synthetic data and do not represent any real trial or trial participants

Big Picture

Today's
Talk:
One aspect
of this
bullet point



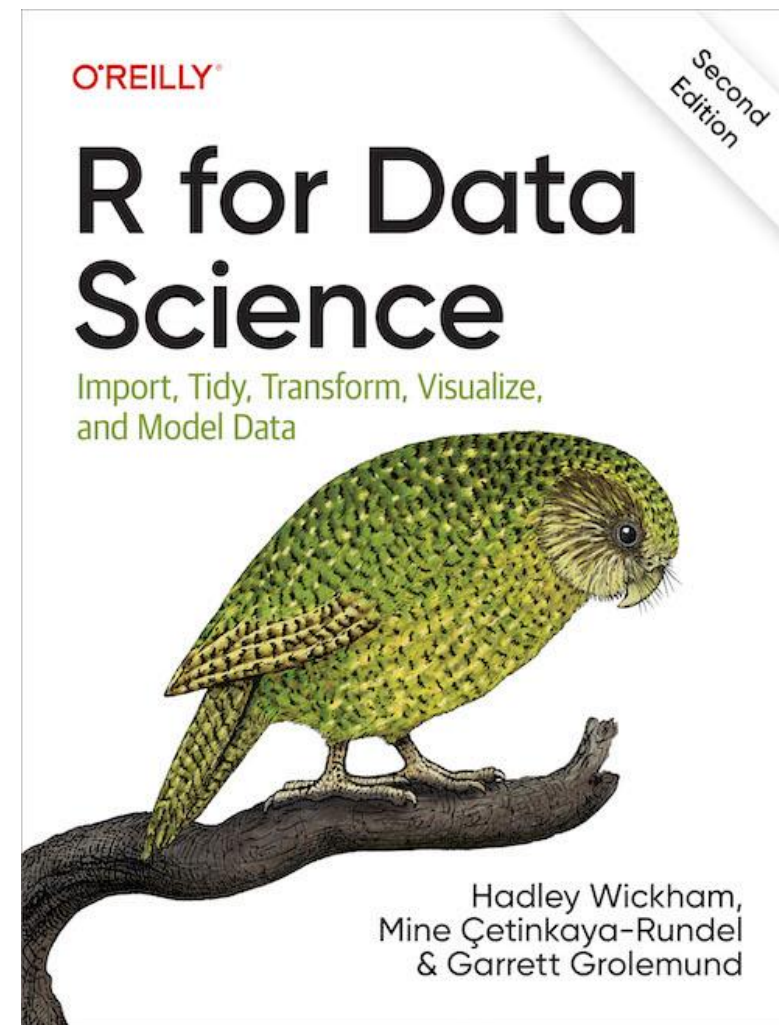
Data Workflow



**Unavoidable but
can be
automated.**

**Use as much or as
little of this step
as necessary to
reach your end
goal**

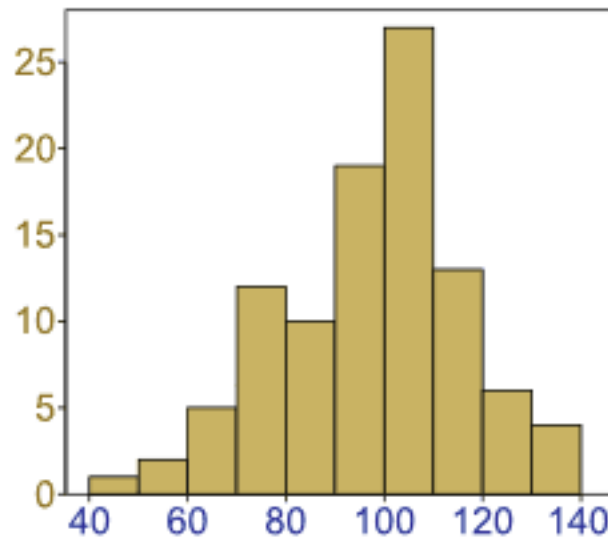
**The end goal.
Communicate
actionable
results to
study teams.**



Quick Level-Setting

Histograms

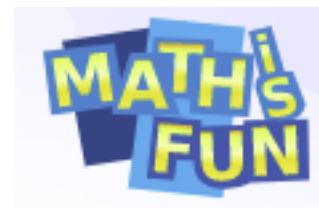
Histogram: a graphical display of data using bars of different heights.



It is similar to a [Bar Chart](#), but a histogram groups numbers into **ranges**.

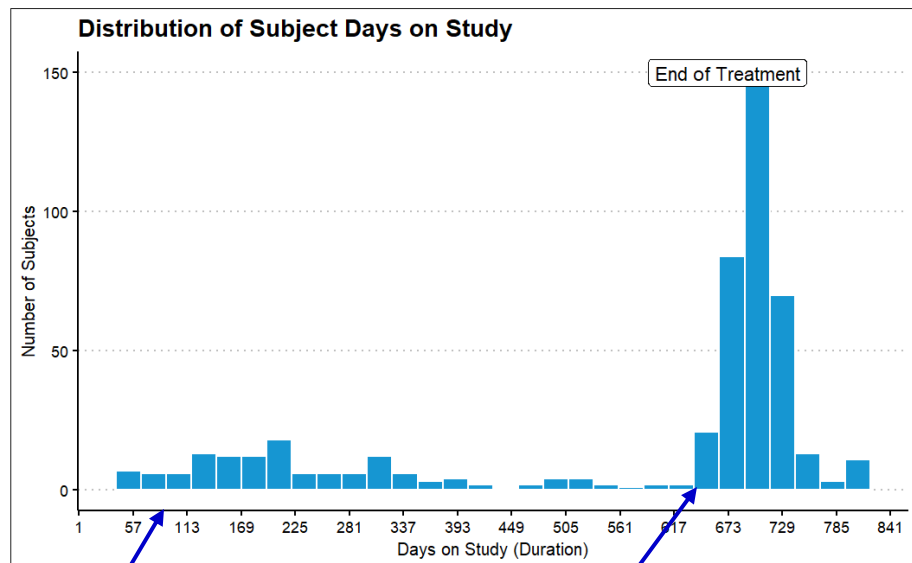
The height of each bar shows how many fall into each range.

And you decide what ranges to use!



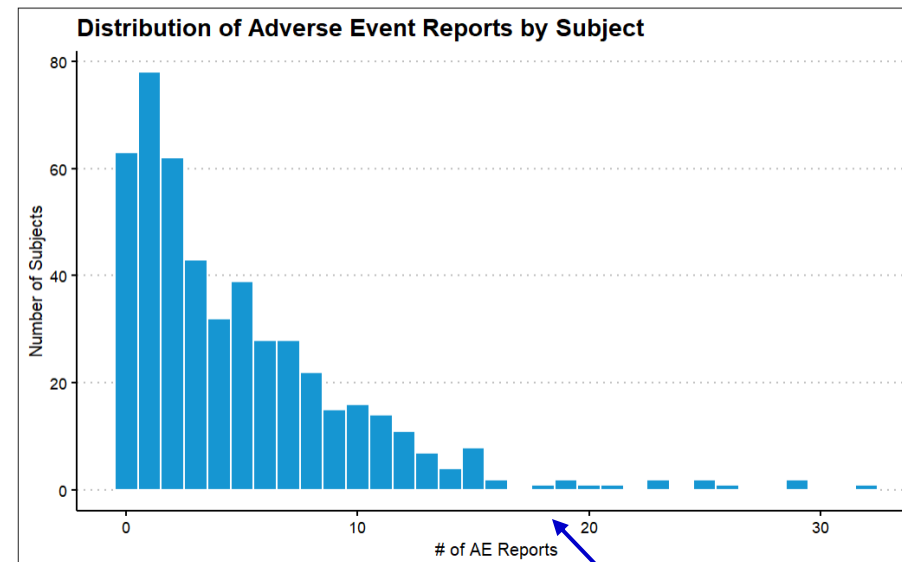
[Histograms \(mathsisfun.com\)](https://www.mathsisfun.com)

How long do subjects stay on study and how many AE reports do they have?

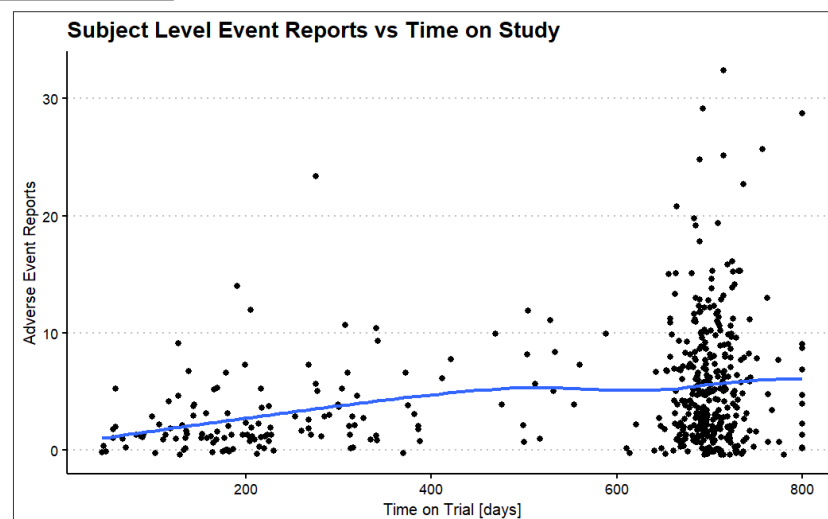


Early dropout

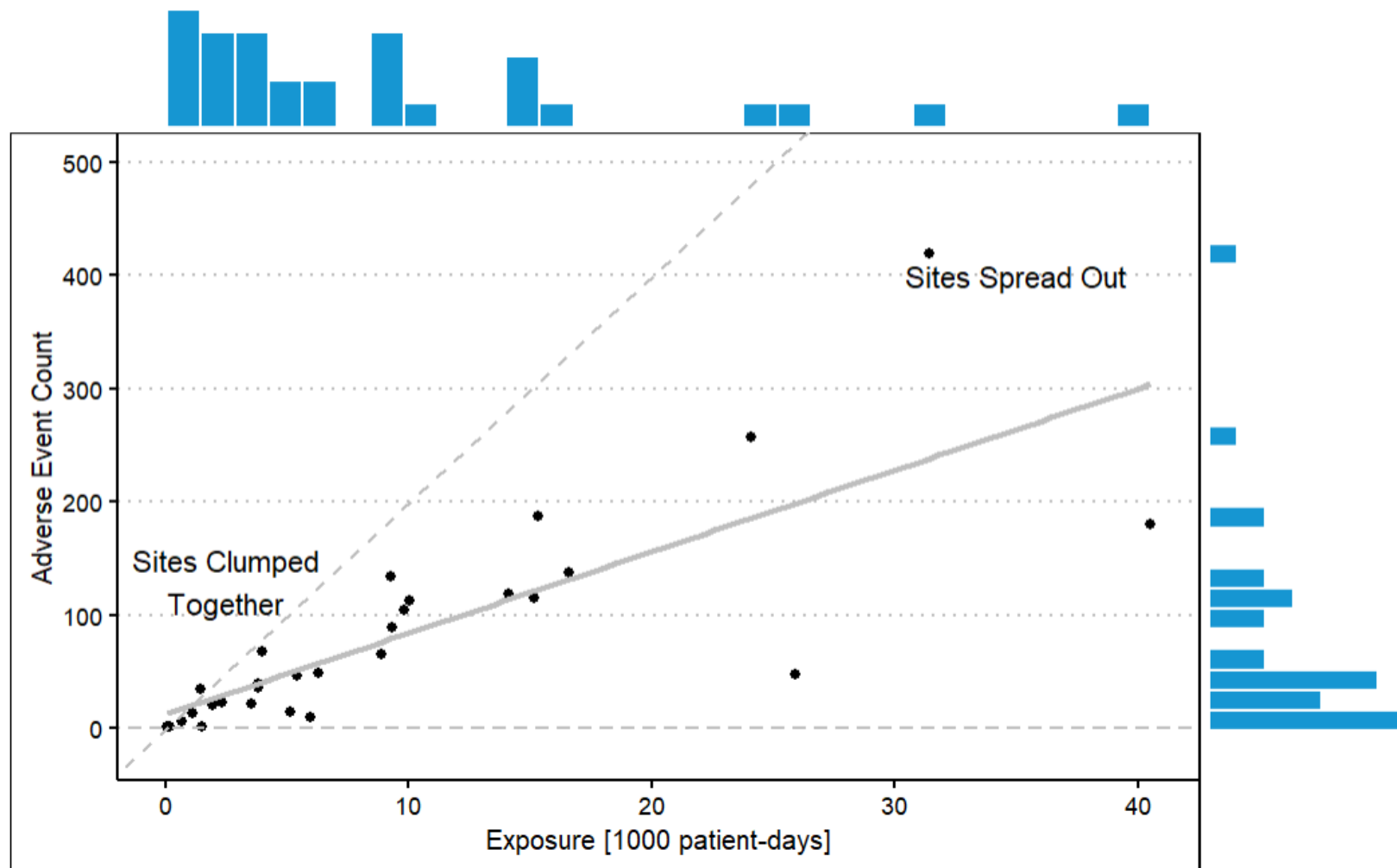
Scheduled end of study



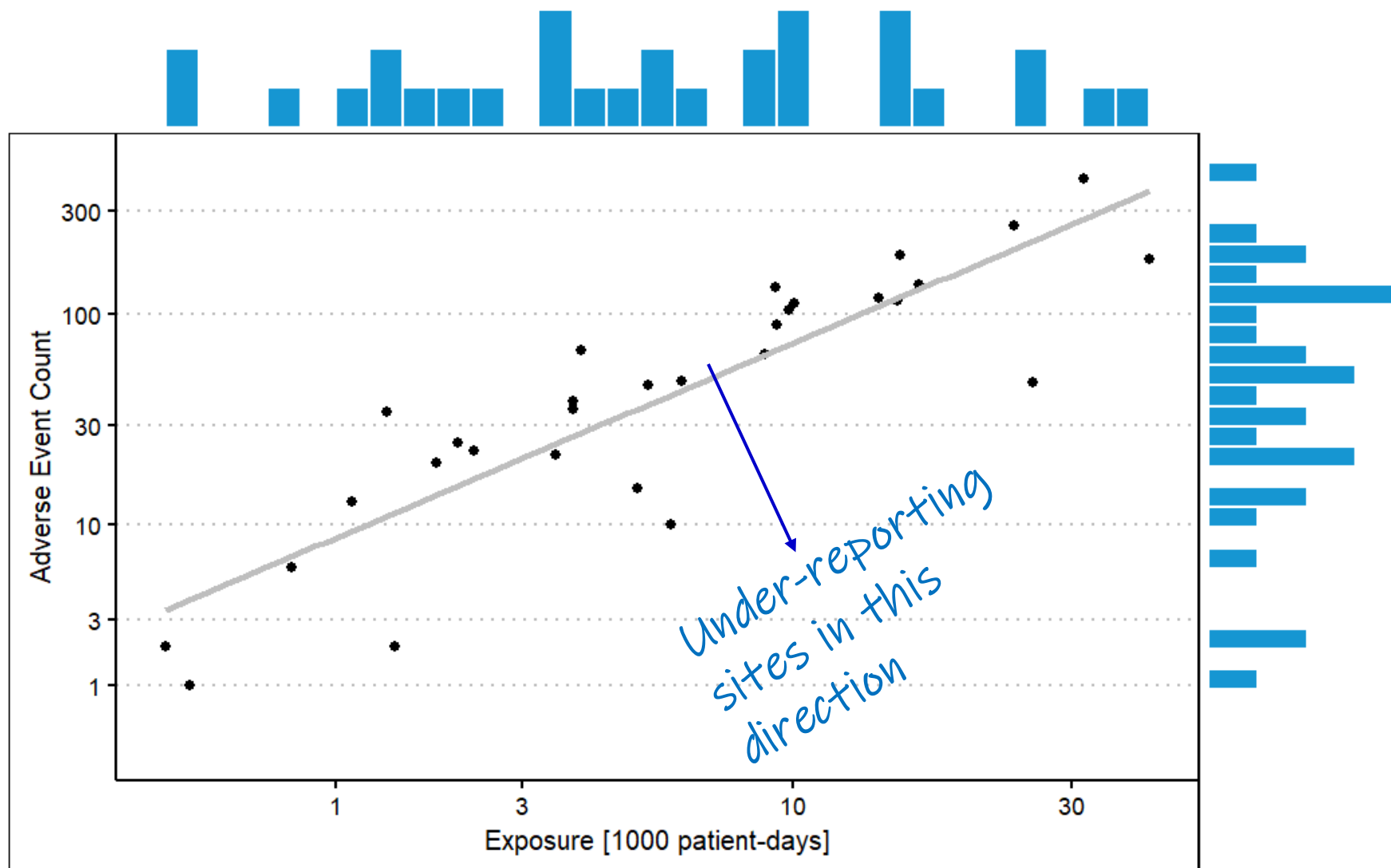
Right skewed count data distribution



What happens when we aggregate data to a site level?



First Fix, Transform the Scale of the Data



Transformation:
 $\text{Log1p}(x) = \log(1 + x)$

Useful trick

1. Spreads out data from lower end of graph
2. Shrinks data from higher end
3. Works on zero counts

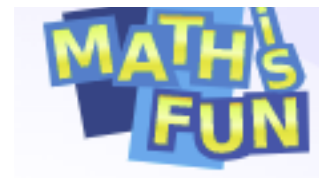
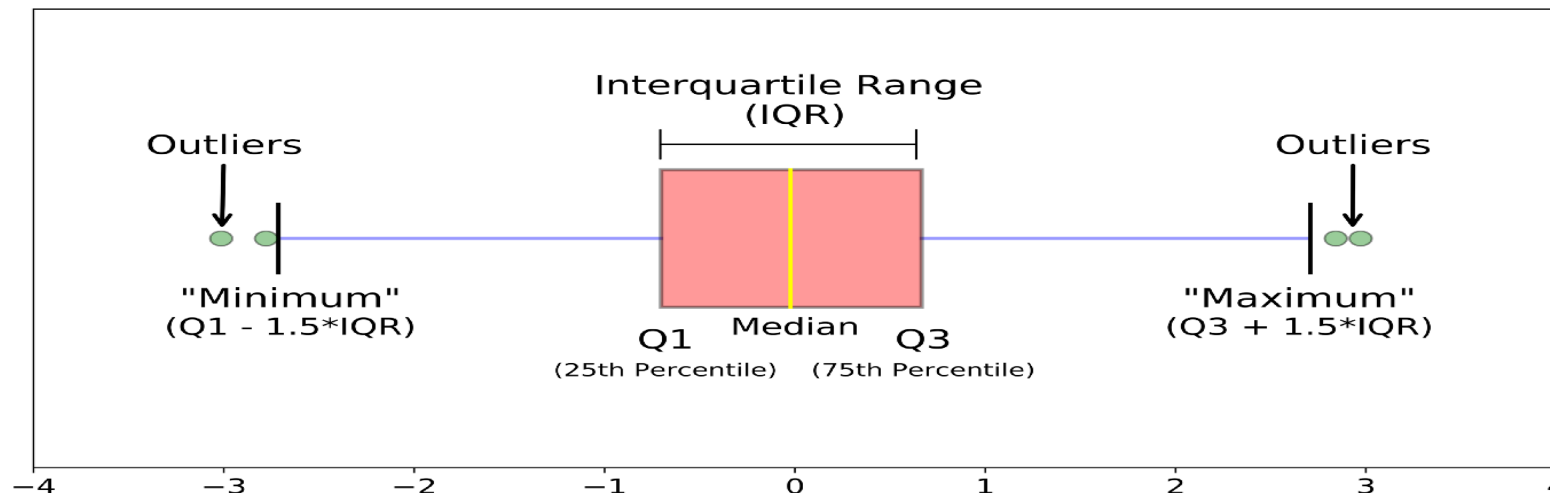
Quick Level - Setting

Definition of

Box and Whisker Plot

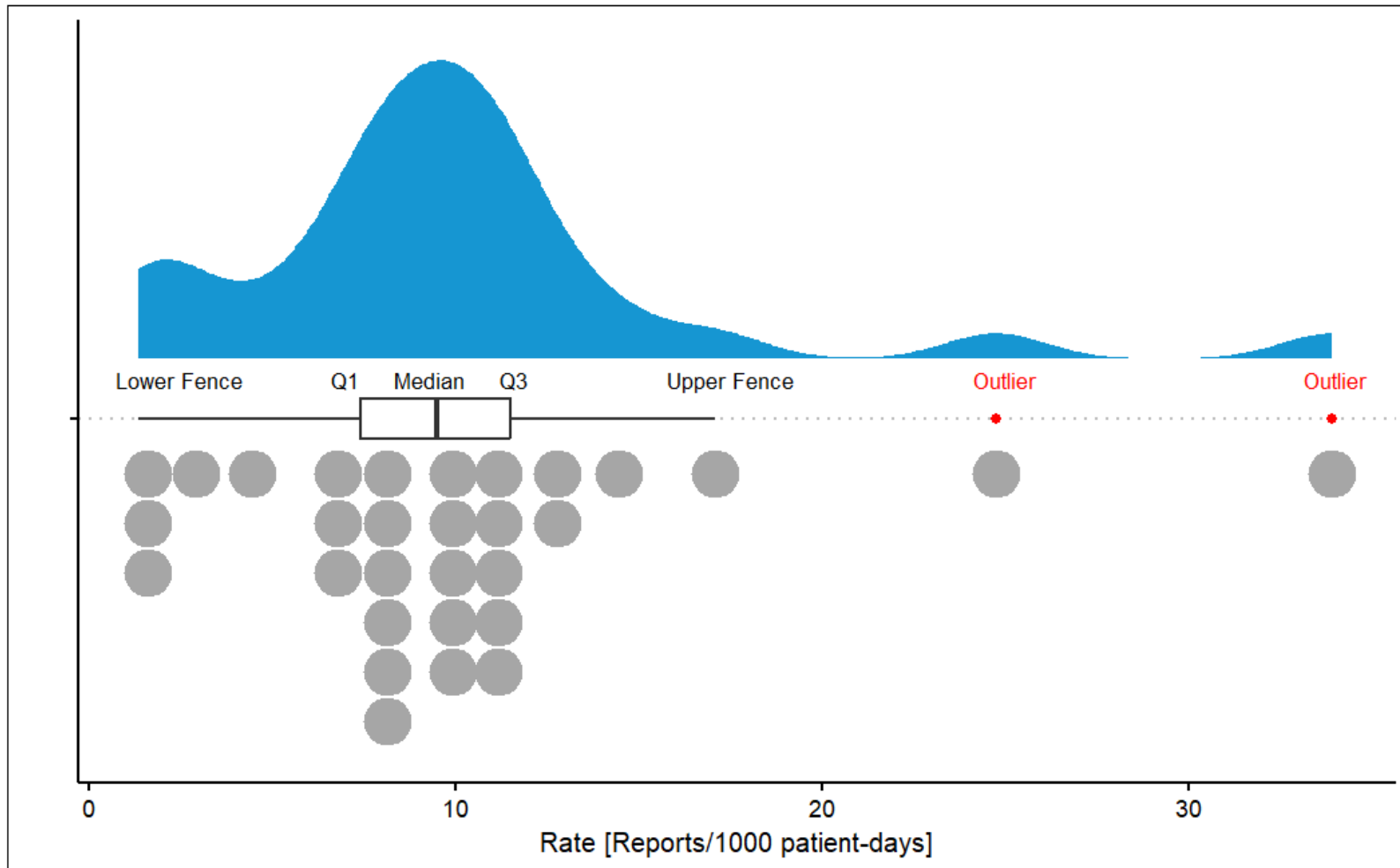
[more ...](#)

A special type of diagram showing Quartiles 1, 2 and 3 (where the data can be split into quarters) in a box, with lines extending to the lowest and highest values, like this:



[Definition of Box and Whisker Plot \(mathsisfun.com\)](https://mathsisfun.com)

Second Fix, Normalize the Data To Make 2 Variables Into 1

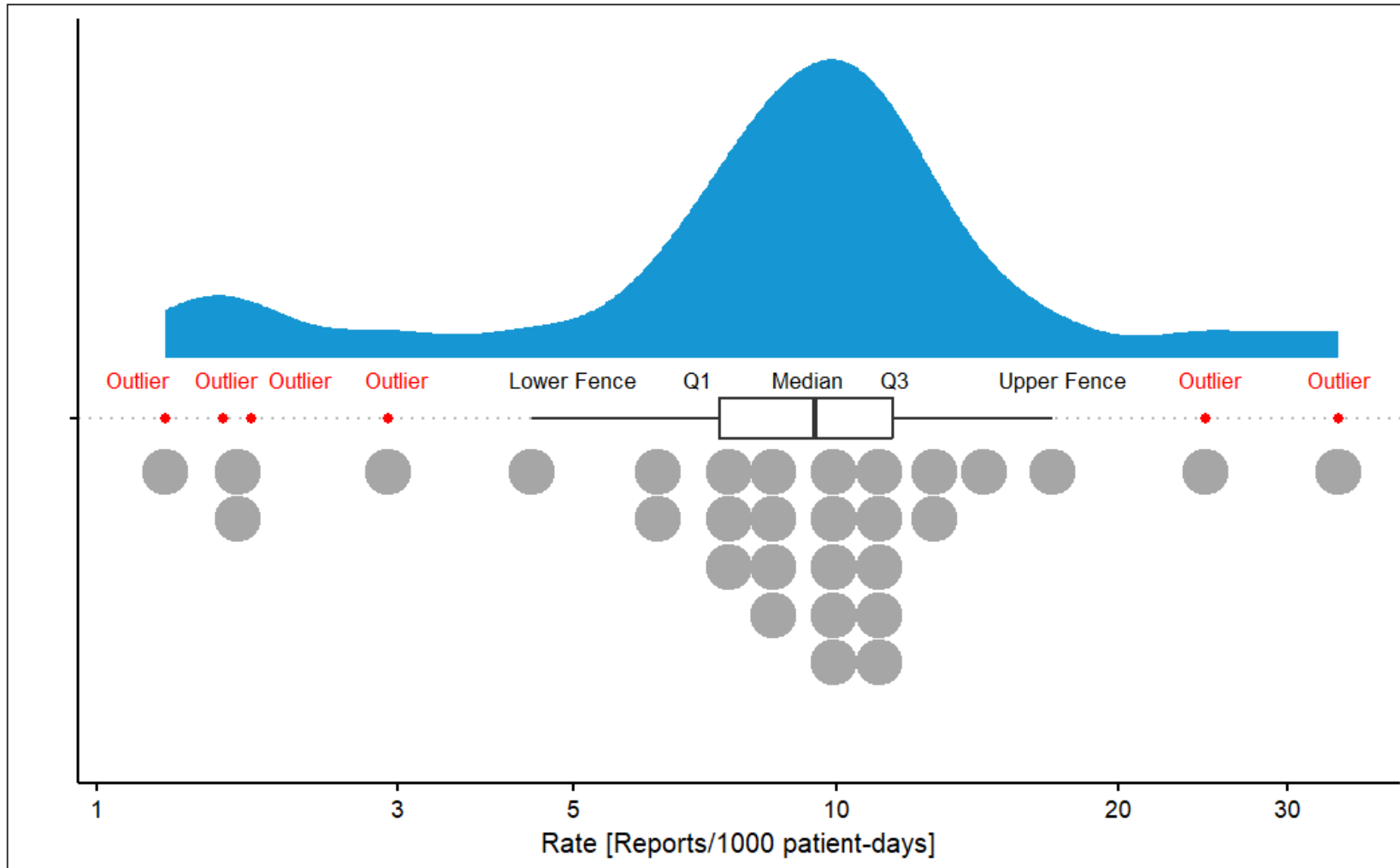


Normalization:

Rate= Count/Duration

Still not very good at finding under reporting sites

Third Fix, Apply a Transformation to the Normalized Data

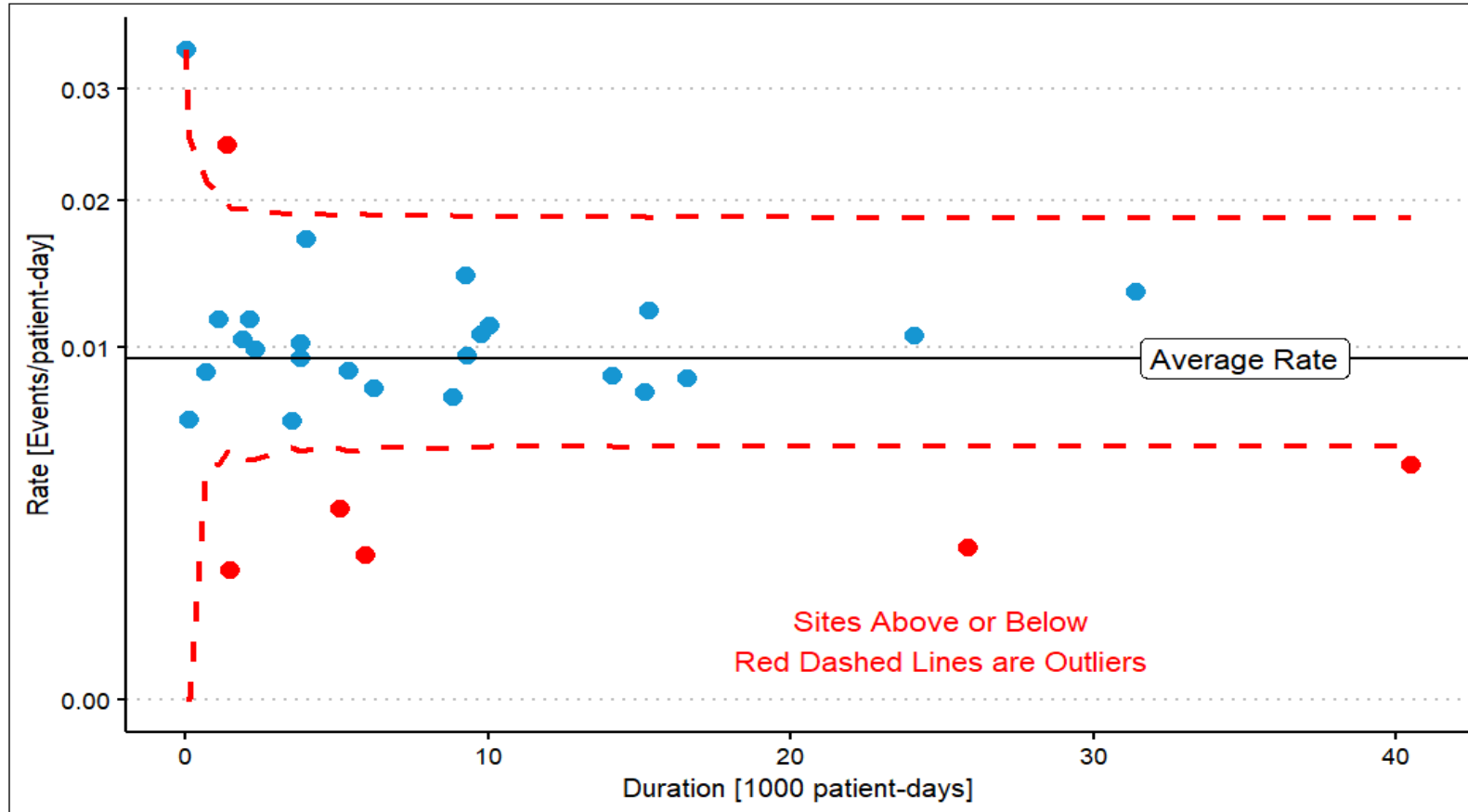


$\text{Log}(\text{Rate}+1)$

Can still benefit from a data transformation.

Better at identifying sites with low reporting rates

Fourth Fix, Plot Results from a Statistical Model



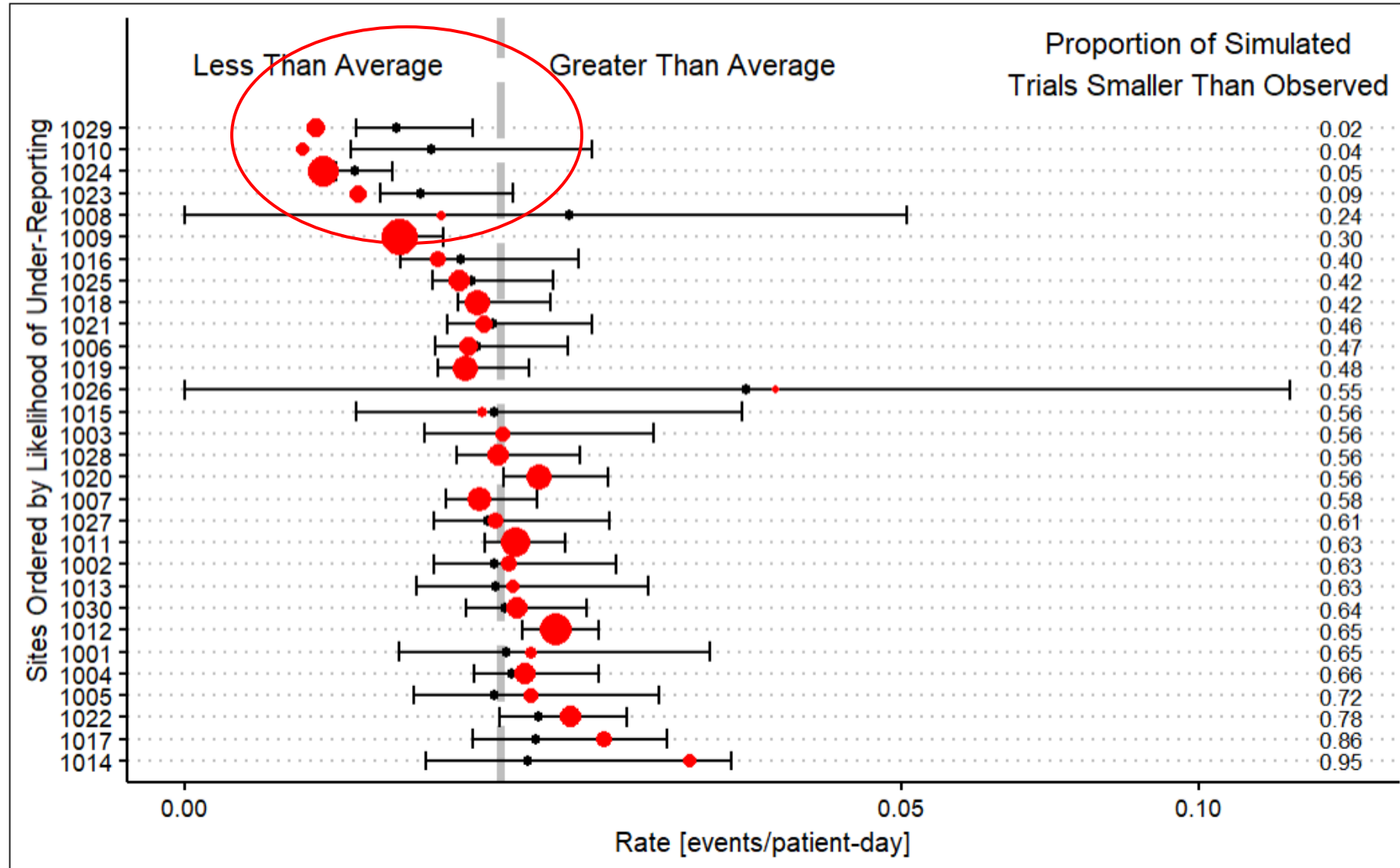
```
Call: MASS::glm.nb(formula = AES ~ 1 + offset(log(TotalDuration)),  
  data = sitedf, init.theta = 3.667842165, link = log)
```

```
Coefficients:  
(Intercept)  
-4.669
```

```
Degrees of Freedom: 29 Total (i.e. Null); 29 Residual  
Null Deviance: 32.18  
Residual Deviance: 32.18 AIC: 278.8
```

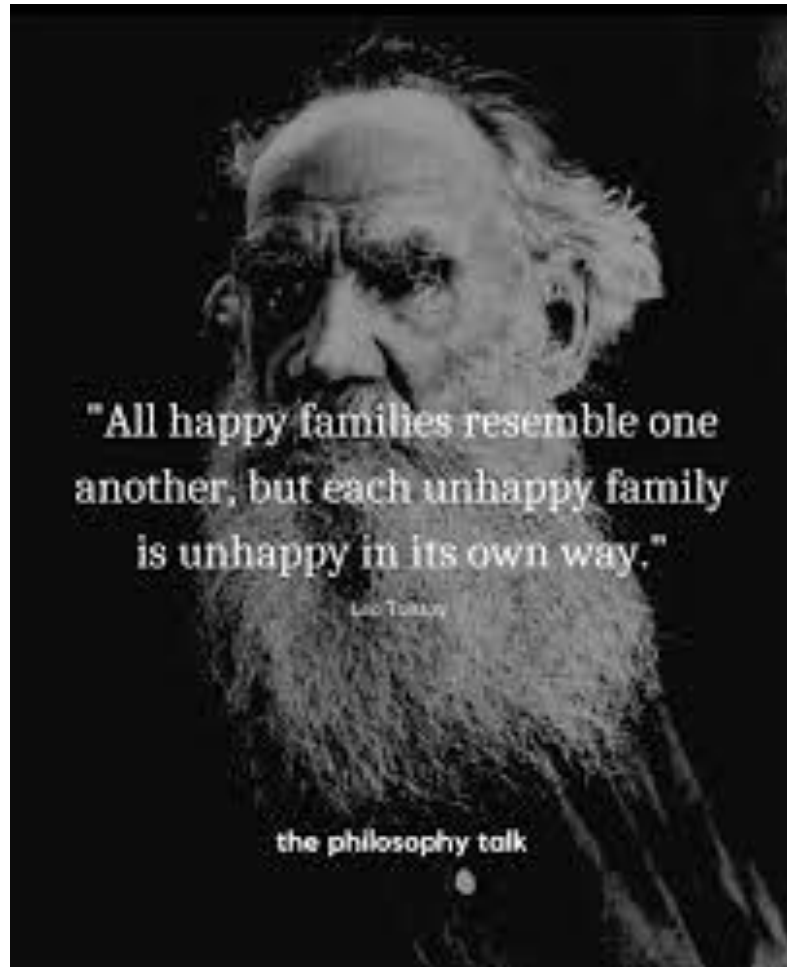
Graph produced from site level summaries and calculations using only 2 numbers from a statistical model.

Fifth Fix, Plot Results of Complex Model



Monitor does not need to fully understand the model. Only how to read the graph and know that the results have been adjusted for age, sex and time on study.

Anna Karenina Principle



There are many ways that data can be recorded poorly but only 'one' good way.

We model the single 'good' way and check if the data is consistent with that model.

Basic principle behind many statistical test.

Some Model Details

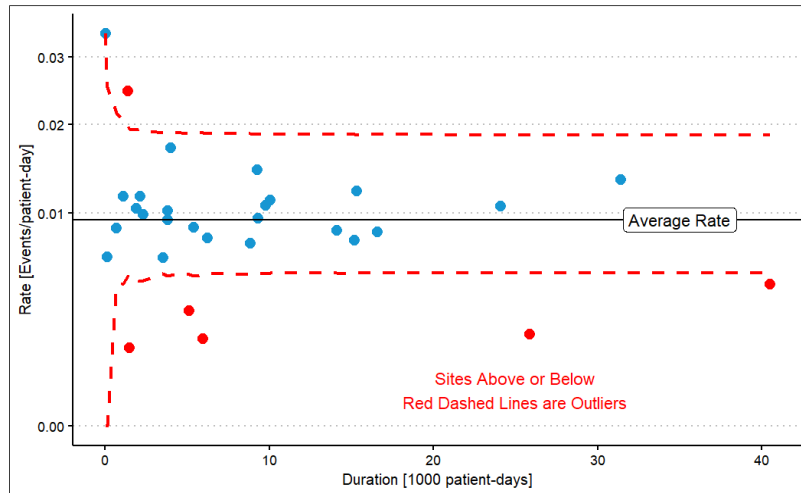
Adjustments for time
on study, Age and Sex.
Other variables could
be included

Allows individual
sites to deviate
from the "average
site" but not by too
much.
Said to "Borrow
Strength" or partial
pooling. Helps with
small sites.

```
brm(AEcount ~ log(duration) + Age + Sex + (1 | SITE),  
    data=tib4,  
    family=negbinomial,  
    cores=2, chains=2,  
    prior= set_prior("gamma(2,66)", class="sd"))
```

**R code for predictive
model.**

Pooling

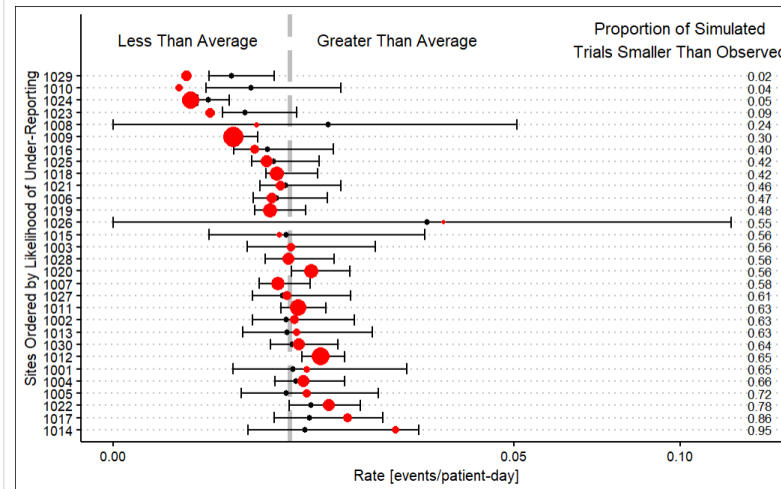


Complete Pooling

Assumes all sites are identical

Example: Funnel Plot Tosses all data into a single group and calculates average

Cons: “One size fits all”, one ‘bad’ outlier can mask other outliers.



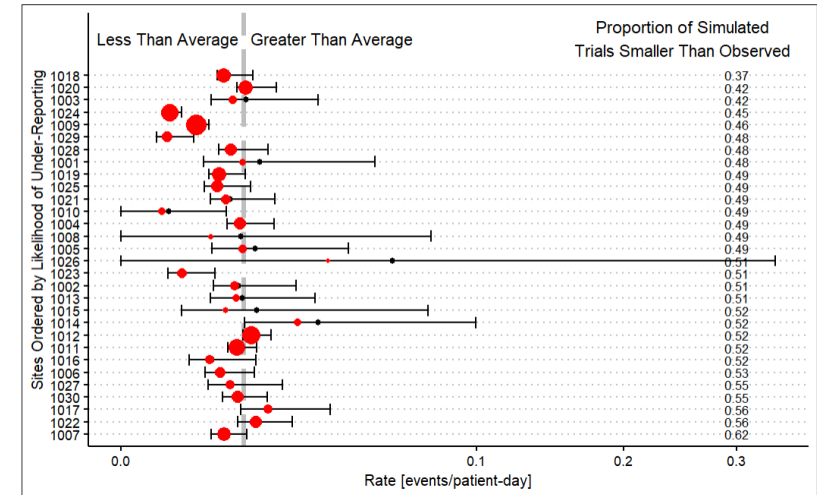
Partial Pooling

Assumes sites are similar but not identical

Example: Predictive model using hierarchical Bayesian methods

Pros: Sensibly includes information from small sites.

Cons: Requires more expertise to fit and troubleshoot.



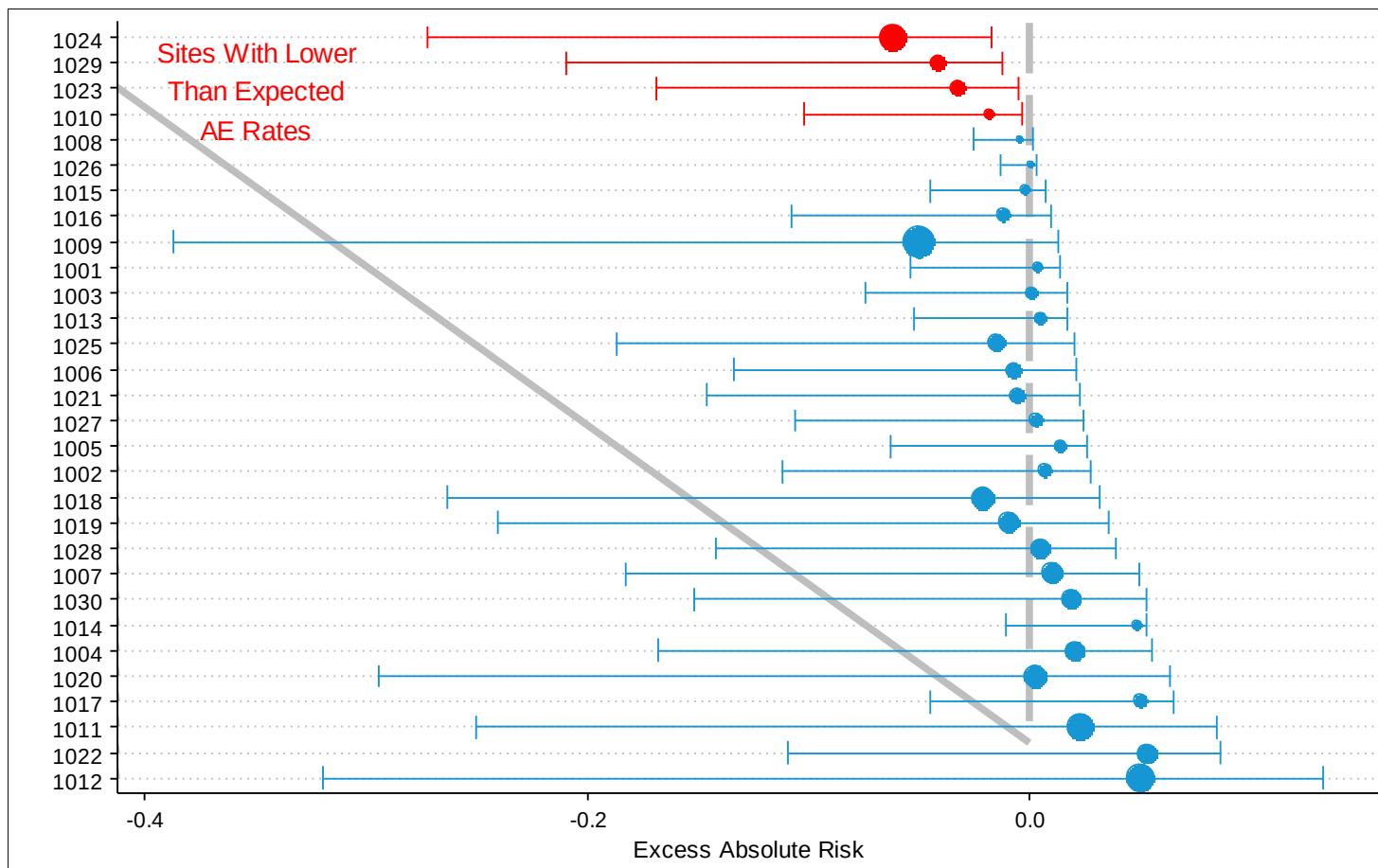
No Pooling

Assumes all sites are unique

Examples: Simple summary or model with fixed effect for each site

Cons: A priori says there are no outliers

Results Can Be Presented in Alternative Formats



Graph of difference between observed rates and model predictions. Look for large difference from zero.

Final Thoughts

- Depending on your needs you can choose a simple visualization or a complex model.
- Simple figures require more user interpretation. What looks too extreme?
- Complex methods can account for many variables at once as well as statistical uncertainty, allowing a single “bright line” decision to flag a site.
- Simple figures may be much better for communication than complex models if they tell the same story.
- Better to flag too many sites than not enough.
- This is part of a much larger monitoring process.

Thank You

