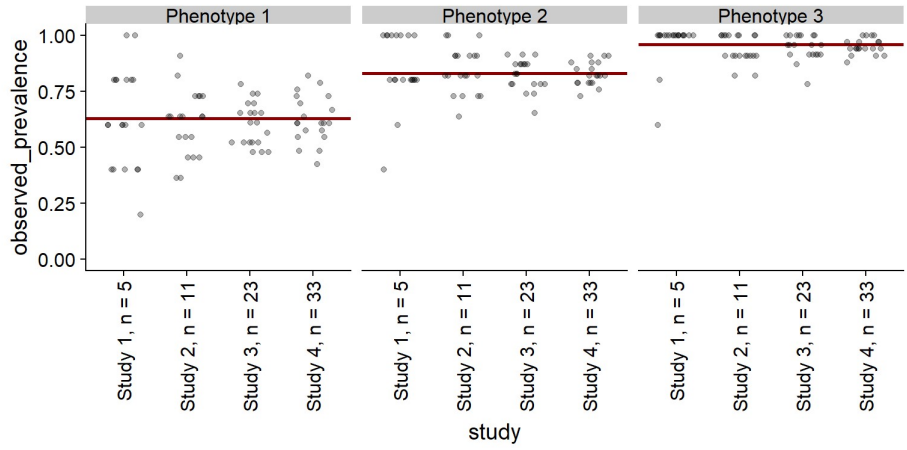
Accessible explanation

The model - an accessible explanation

As all models, the model we use simplifies and abstracts the medical reality in hope we can arrive at useful conclusions. Our model is a member of generalized linear model family, using logit link and hierarchical terms in a fully Bayesian treatment. Let's unpack this a little, starting with what a logit link does. In the following, we will describe how we handle a single phenotype as the estimation for individual phenotypes is mostly independent.

Our model tries to estimate theoretical *true prevalence* of the phenotype in a population - i.e. the probability that a randomly selected patient from the population will exhibit this phenotype. But all we observe is that each individual either exhibits the phenotype or not. Depending on the number of individuals enrolled in a study, the *observed prevalence* will jump more or less around this true prevalence - let's have a look at an example:



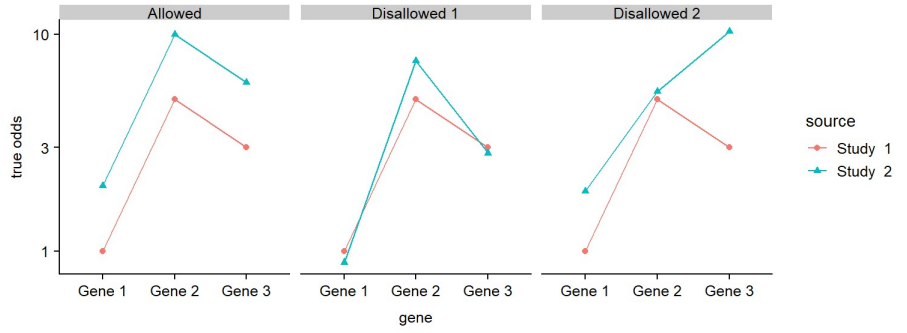
The red line shows the true prevalence of a phenotype (assuming it is the same across multiple studies). Each point shows the observed prevalence of a single possible realization of a study. We see that the observed prevalence can differ substantially from the true value and that the spread decreases with increasing n. In our data we can't however expect large n as the biggest n we have is 33 patients. In 74% of cases there are less than 5 patients with a given mutation in the same study. Also note that the observed values are clustered at discrete "levels" because e.g., among 5 patients you only can have prevalence of 0.2 or 0.4 and nothing in between. Further, you can see that with high prevalence, there is less variation in the observed prevalence.

For mathematical convenience, the model does not work with prevalence directly, but with *odds*. Odds are just another way to express prevalence - for example, when the prevalence is 20%, we expect one patients to exhibit the phenotype for each four patients not exhibiting the phenotype, leading to odds $1:4$ or 0.25 and log odds (base 10 here) of roughly -0.6. Unlike prevalence, which is constrained between 0 and 1, odds can be any non-negative number.

Most of the results of the model are reported as comparisons of odds in different populations. For this we use the ratio of the corresponding odds. E.g., when the odds ratio is 2, we expect the first population to have twice as much patients exhibiting the phenotype for each healthy patient than the second population.

Technically the model works with the logarithm of odds. The logit function transforms a probability (prevalence) to log odds, hence the logit "link" used in the model. What is the linear part?

Our model assumes that the true odds of a phenotype is a function of four numbers (*coefficients*): the overall odds of the phenotype in the population, a modifier for the gene the patient has damaged and a modifier for the study the patient is enrolled in. One additional modifier is added when the mutation is a certain loss-of-function (cLOF). These four numbers are multiplied to arrive at the final odds for the patient. Assuming only cLOF mutations for simplicity, this means that while the odds of a phenotype are allowed to vary between genes and the overall rate of a phenotype may vary between studies, the odds ratio of different genes is the same across all studies. Let's look at an example:



Above on the leftmost panel we see that the two studies differ in the odds of the phenotype for each gene, but the ratio of odds for Gene 1 to Gene 2 (and 3) is the same in both studies (since the odds are shown on log scale this manifests as a constant gap between the two lines). This type of between study variation is allowed. On the other two panels, the odds ratio for Gene 1 to Gene 2 (and 3) differs between studies - this type of variation is not allowed by the model.

Another way to describe the allowed case is that, for both studies a mutation in Gene 2 makes a patient five times more likely to exhibit the phenotype than a patient with a mutation in Gene 1, although the base rate of the phenotype may vary between studies.

The cLOF coefficient is held constant for all genes in a given phenotype, meaning that odds for any given phenotype are multiplied by a small number when the mutation is cLOF. Once again this means that relative odds are the same among cLOF and other mutations, but absolute odds can be higher (or lower) in cLOF mutations.

This is the "linear" part of the model - we multiply odds, which is the same as adding the logarithm of the odds and addition is a neatly linear thing.

Now the "hierarchical" part. This ties the coefficients in the model in two important ways: i) it assumes small differences in odds across genes and studies are more likely than large differences and updates the estimates accordingly ii) *partial pooling*: the degree to which odds are allowed to

“jump around” across both genes and studies is informed by the data, e.g., if the odds are similar for all genes except one, the model will put higher weight on the possibility that the difference in the last mutation is just noise and shrink its estimate towards the average for other genes. On the other hand, if the odds vary wildly across all genes, the model will assume it is more likely this is a true variation and not shrink the estimate much. The amount of shrinkage also depends on the number of observations as estimates where there is a larger number of observations are shrunk less. The variability across studies is pooled in a similar way.

Together those two features result in low risk of overfitting the data, even though we have very little observations for most study - gene - phenotype combinations.

We also allow for a correlation between phenotypes, e.g., that some phenotypes occur frequently together while others rarely manifest in the same patient. Once again the amount of correlation is estimated from data.

Finally, the “Bayesian” part: We follow the Bayesian paradigm, so our estimates of the model coefficients are not a single number, but rather a distribution - some values are more likely than others, but the data are insufficient to let us determine the coefficients with high certainty. Therefore, we never report exact numbers but rather 50% and 95% *credible intervals* of the distribution. Unlike confidence intervals in frequentist analysis, we can directly interpret the 95% credible interval as the interval that contains the true value with 95% probability - assuming our model is correct (which it is not, but we hope it is still a useful abstraction).