Inferring Causality from Finite Data using Conditional Independence

Daniel Speyer

May 13, 2017

Abstract

There exist many techniques for inferring causal structure from conditional independence, but they all assume establishing conditional independence is trivial, which from a finite sample it is not. Also, all return a graph, rather than a graph and a probability that it is the correct one. In this paper I offer two techniques of limited scope that address these problems. I then discuss the techniques limitations, and apply them to finding probiotic treatments for Crohn's Disease.

1 Introduction

Inferring causal graphs from observational data is a widely-sought goal in statistics. The most common tool for it is conditional independence. The specifics of a causal graph determine which node are independent conditioned on which others by the "bayes ball" rule, and therefore it should be possible to observe the independences and work backward to the causal graph.

This has proved to be more difficult in practice. In particular, observing independence is not as simple as it sounds. Pearl suggests we test joint conditional probability distributions for equality.[1] Leaving aside the curse of dimensionality, this assumes we have the exact distributions. If all we have is a finite random sample from those distributions, we can construct posteriors, but we cannot perform an equality test. Spirtes, Glymour and Shine encourage us to assume that "the statistical decisions required by the [causal inference] algorithms are correct for the population", while admitting that this "is often not met in practice".[2] Others have described this as "possessing an independence oracle".[3,4] Many simply include a "test if $a \perp\!\!\!\perp b|s$" step in their algorithms, assuming the reader will already know how.

In practice, the usual solution is to treat independence as a null hypothesis, try to reject it at some p threshold, and treat any failure as establishing it. Needless to say, this is incorrect.

Dealing with finite data means the possibility of dealing with too little data. The elegant solution is to give some numerical expression of confidence which becomes "I don't know" when the data gets too small. This solution has another benefit: in the biomedical context, it is routine to test thousands of equally plausible hypotheses at once. A flat accuracy of 99% is unhelpful in the face of this, but a numeric expression of confidence allows proper compensation.

1.1 Motivating Problem: Crohn's Disease

This paper is optimized around a specific practical problem: untangling the microbiome's role in Illial Crohn's Disease. It is well established that there are many differences in the intestinal bacteria of healthy people and of people with the disease,[5] but it is not established whether the differences of bacteria *cause* the disease. If we can find a set of bacteria such that p(disease | do(bacteria)) is low, we will have a cure for the disease. There have been attempts to determine this by randomized controlled trial, but early results are discouraging[6] and the number of species, combined with other relevant variables, make exhaustive RCTs impractical.

Data is available on this problem,[7] including genetic, microbiome and health information, but for only 58 patients. The microbiome information is a series of 16S reads, but can generally be described as "present" or "absent", with very low concentrations of a species rounded off to "absent". This comes much closer to fitting the empirical distributions than any convenient scalar formula (see figure 1), which is why this paper will use binary variables.

2 Extending Graphs

Let us begin with the simplest case. Suppose we have a known cause, a known effect, and a variable which connects to the effect in an unknown way, that is $A \rightarrow B - C$. We do not observe a correlation between A and C, but that might only mean our test is underpowered. For Crohn's Disease, A would be mutations in the NOD2 gene,[8] B would be the disease, and C would be each of 222 species of plausibly relevant bacteria.

For now, we will assume that there is no direct causal link between A and C, and furthermore that there are no unobserved confounders or selection effects. We will consider these later. This leaves us only two models, $A \rightarrow B \rightarrow C$ or $A \rightarrow B \leftarrow C$. We can call these "chain" and "collide" models, or m_{ch} and m_{co} for short. Can we distinguish between them?

Yes. Let us consider $p(A, C)$:

$$\begin{aligned} p(A, C|m_{ch}) &= \sum_B p(A)p(B|A)p(C|B) \\ p(A, C|m_{co}) &= p(A)p(C) \end{aligned}$$

We do not actually know the terms on the right side of those equations, so let us parameterize both models with θ and rewrite the equations as:

$$p(A, C|m_{ch}) = \int \left(\sum_B p(A|\theta)p(B|A, \theta)p(C|B, \theta) \right) p(\theta)d\theta$$

$$p(A, C|m_{co}) = \int p(A|\theta)p(C|\theta)p(\theta)d\theta$$

We can learn $p(\theta)$ from available data using dirichlet priors. The integrals would be difficult algebraically, but they can be adequately approximated with monte-carlo sampling.

Once we have these, let $n_{a,c}$ be the count of datapoints with A=a,C=c and we can use:

$$p(n_*|m) \propto \prod_{a,c} p(a, c|m)^{n_{a,c}}$$

Proportional instead of equal because there is a combinatoric term, but it cancels when taking odds ratios so it can be safely ignored.

From here, we can apply standard bayesian updating. Empirically, this works well. With all probabilities chosen at random, the mean log probability it assigns to the wrong model is only -0.62. Furthermore, if the output is bucketed by posterior, the fraction correct

forms a graph very similar to a straight line (see figure 2). It does show a minor bias toward chain models, but it goes away at the edges, and collider models are the useful ones, so this is a safe bias.

While the test is well-calibrated, it returns bayes factors very close to one most of the time. For random probabilities and 58 datapoints, it only gives a clear answer (i.e. a bayes factor outside the range [0.1,10] 8% of the time). This includes cases in which one of the causal effects is very weak. Altering the generative function to more closely resemble the Crohn's data (by having the real p(NOD2) and p(Crohn's|NOD2) and detectable $B \not\perp\!\!\!\perp C$) brings this up to 13%. Even with realistic data, 500 datapoints are needed to gain conclusive answers more than half the time (see figure 3). The realistic data does make the test somewhat less calibrated (figure 4), which probably reflects a collide model being more likely to produce Crohns-like data in the first place, but the errors remain small and on the side of safety.

2.1 Non-Simply-Connected Graphs

What if there is a causal effect $A \to C$? It must be weak enough not to be detected, but that doesn't say much. Let us consider it by cases.

2.1.1 False Colliders

Can a true graph of $\underbrace{A \to B \to C}$ produce a $p(A, C)$ more similar to a graph $A \to B \leftarrow C$? This would require the direct and indirect influence of A on C to cancel out rather precisely. If the indirect influence is stronger, we will pick the correct model, albeit underconfidently. If the direct influence is too strong, we will see a correlation between A and C. And if the direct influence is in the same direction as the indirect, the correct model will remain the better fit (though both models will be worse).

2.1.2 False Chains

Similarly, can a true graph of $\underbrace{A \to B \leftarrow C}$ produce a $p(A, C)$ more similar to a graph $A \to B \to C$? Again, this is possible, but there is no reason for $p(C|A)$ to resemble $\sum_B p(B|A)p(C|B)$. If $p(C|A) - p(C|\bar{A})$ is too large, $A \not\perp\!\!\!\perp C$ will be detected initially, whereas if it's too small, the correct model will be chosen. Furthermore, what shrinks the upper bound is for the predicted A-C relation in the chain model to be small, which also means any erroneous bayes factor will be small.

2.1.3 Empirical Examination

These examinations are difficult to make rigorous, but simulation suggests there is little to worry about. If an $A \to C$ link is added but bounded by a χ^2 test, there is a tendency toward underconfidence, but little other effect (see figure 5. This applies to the Crohns-like case, and does not fully generalize.

Overall, this technique is not *brittle* in the face of an $A \to C$ link, but it is not robust against an arbitrarily strong one either. Domain knowledge or other tools must be used to establish safety. In the case of Crohn's disease, the generally understood role of NOD2, mediates *untargetted* immune responses,[8] makes an $A \to C$ link unlikely

2.2 Unobserved Confounders

This technique is vulnerable to confounders. Specifically, if the $B-C$ link is the product of a common cause, that will generate no $A-C$ dependence, exactly like a collider.

2.3 Building a Graph

This technique may be iterated to expand a graph, though whether it will suffice to describe an entire graph depends on the specifics. Since it does offer bayes factors, these can be used in the iteration, and provide a metric of when one has extrapolated too far from too little domain knowledge.

3 Severing

Suppose we do not have a known causal effect to expand from. The standard oracle-requiring algorithms do not need one. It is probably impossible for a formula to say that $A \not\perp\!\!\!\perp B$ in full generality, but it may be possible to say $A \not\perp\!\!\!\perp B|S$ sufficiently for our purposes.

Suppose A, B and S are already known to correlate. The simplest associated models are S-A-B, A-S-B and A-B-S, where S severs only in the A-S-B case (the directions on the arrows are not important at the moment, except that there can be no colliders). For the first two models:

$$\begin{aligned} p(A, B|S, m_{\not{sever}}) &= p(A|S)p(B|A) \\ p(A, B|S, m_{sever}) &= p(A|S)p(B|S) \end{aligned}$$

And there is a symmetric formula for the last one. Since there are two alternative hypotheses, we take whichever produces the greater probability. Somewhat surprisingly, there is no advantage here in using monte-carlo posteriors of simple maximum likelihood estimators for the conditional probabilities.

There is another simple causal graph that can cause three variables to correlate: all could be influenced by a common confounder H (for "hidden). This case is impossible to test for in full generality, because S could follow H so closely as to be an acceptable proxy, and therefore controlling for S effectively controls for H. In theory, it should be possible to infer H and its conditionals usings a Gibbs-sampler-like system. In practice, there is more room to overfit this more complex model and no straightforward way to compensate. What does work surprisingly well is to ignore this case and apply the exact same test as before.

Empirically, severing vs. chain has mean log p(false) of -.41 and vs. common cause has of -.49. Both follow close to a straight line (see figure 6).

3.1 Applying Severing

If we are willing to trust the severing test, can we use it to find causes of a variable of interest? Deducing the entire causal graph is a brittle endevour, but there is a local solution.

Suppose there exist three variable with a common cause, such that one causes the variable of interest, that is $\underbrace{H \rightarrow A \rightarrow D \; B \; C}$. The trio A, B, C can be identified because they all correlate and remain correlated when controlling for any of them. No special test is needed for this, standard χ^2 can produce a p value for all of this. The role of A can be identified because it severs B and C from D (standing for "disease", remembering the application). No other simply connected graph has these properties.

This technique does not try to find *all* causes of D, only those which happen to fit this motif and are strong enough for a χ^2 test to pick up. As such, it is unclear what it would mean to test it in simulation. Certainly with parameters chosen for the purpose, it will work. With random parameters, it will generally fail to find trios.

3.2 Unobserved Confounders

If the links between H and A, B, C in the preceeding graph are actually mutual effect

from a confounder, then that confounder can be thought of as a part of H (which is itself unobserved). If the link between A and D is via confounder, then there will be no $B - D$ or $C - D$ link to sever.

3.3 Non-Simply-Connected Graphs

These techniques depend on simply-connected graphs. Both the severing test and its application can say nothing of significance in the face of multiple connections.

4 Parsimony

Given that it will be necessary to assume either "no unobserved confounders" or "simple connection", can we at least say that these are the most parsimonious explanations for our data?

From a graphical model perspective, they are not. It is the absence of a link which is a statement about the world.

From a biochemical perspective, they are. Each link asserts an interaction, and the default for organic molecules or microorganisms is not to interact with eachother. As for unobserved confounders, the presence of unobserved variables in biology is a given, but for them to be confounders requires *two* interactions that relate to eachother.

Fields other than biology will need to reconsider this.

5 Crohn's Disease

Returning to the original motivating problem, what do these techniques show for Crohn's Disease?

The bacterial concentrations can be converted to boolean values by placing thresholds on them. For each species, a threshold was chosen that maximized the number of Crohn's Disease statuses that could be correctly predicted by that bacterium alone. These boolean values can then be thought of as "has a clinically relevant dosage of the bacterium". Once the values were booleanized, all those with entropies less than 0.5 were dropped on the logic that a species that's always the same can't have a clinical effect. This left 222 species.

Of these the direction technique found 8 which show signs of impacting Crohn's Disease, though only one (Lactobacillus acidophilus) has a bayes factor greater than 10. The 8 are shown in figure 7.

Inconveniently, that species appears to be *harmful*, which is surprising given the usual role of lactobacilli but less so given that two RCTs found this same result (albeit non-significantly).[9,10] It may be relevant that L. acidophilus is usually found in the *jejunum*,[11] but these samples are from *ileum* biopsies. Perhaps the cause of Crohn's Disease is not which bacteria are in the intestine, but where in the intestine they are. This result is inconvenient because removing or moving bacteria is far more difficult than adding them.

The species without strong bayes factors are still encouraging. The p-value confirms they relate to Crohn's Disease in some way, in which case "causes" and "is caused by" are the most likely models by simple priors. These small values bump our posterior belief some. This is not enough to recommend to patients, but it is a better place to start RCTs than anything we had before.

References

[1] Pearl "Causality: Models, Reasoning and Inference" Chapter 2 *Springer* 2008

[2] Spirtes, Glymour & Scheines "Causation, Prediction, and Search" Section 5.4 *MIT Press* 2001

[3] Peters, Mooij, Janzing & Schölkopf "Causal Discovery with Continuous Additive Noise Models" *Journal of Machine Learning Research* 2015

[4] Kalisch, Mächler, Colombo, Maathuis & Bühlmann "Causal Inference Using Graphical Models with the R Package pcalg" *Journal of Statistical Software* 2012

[5] Hofer "Bacterial imbalance in Crohn's disease" *Nature Reviews Microbiology* 2014

[6] C Prantera "Probiotics for Crohns Disease: What Have We Learned?" *Gut* 55.6(2006):757759, 2016

[7] E Li, D Frank & R Sartor, "Effect of Crohns Disease Risk Alleles on Enteric Microbiota" *Encyclopedia of Metagenomics*, 1-5, 2014

[8] Philpott, Sorbara, Robertson, Croitoru & Girardin, "NOD proteins: regulators of inflammation in health and disease" *Nature Reviews: Immunology* 2014

[9] C Prantera, M Scribano, G Falasco, A Andreoli, C Luzi "Ineffectiveness of probiotics in preventing recurrenece after curative resection for Crohn's disease: a randomised controlled trial with Lactobacillus GG." *Gut* 200251405409.409, 2002

[10] A Bousvaros, S Guendalini, R Baldassano, C Botelho, J Evans, G Ferry, B Goldin, L Hartigan, S Kugathasan, J Levy, K Murray, M Oliva-Hemker, J Rosh, V Tolia, A Zholudev, J Vanderhoof & P Hibberd "A randomized, double blind trial of Lactobacillus GG versus placebo in addition to standard maintenance therapy for children with Crohn's disease." *Inflamm Bowel Dis* 200511833839.839, 2005

[11] J Walter "Ecological Role of Lactobacilli in the Gastrointestinal Tract: Implications for Fundamental and Biomedical Research" *Appl. Environ. Microbiol.* 74(16):49854996, August 2008

6 Figures

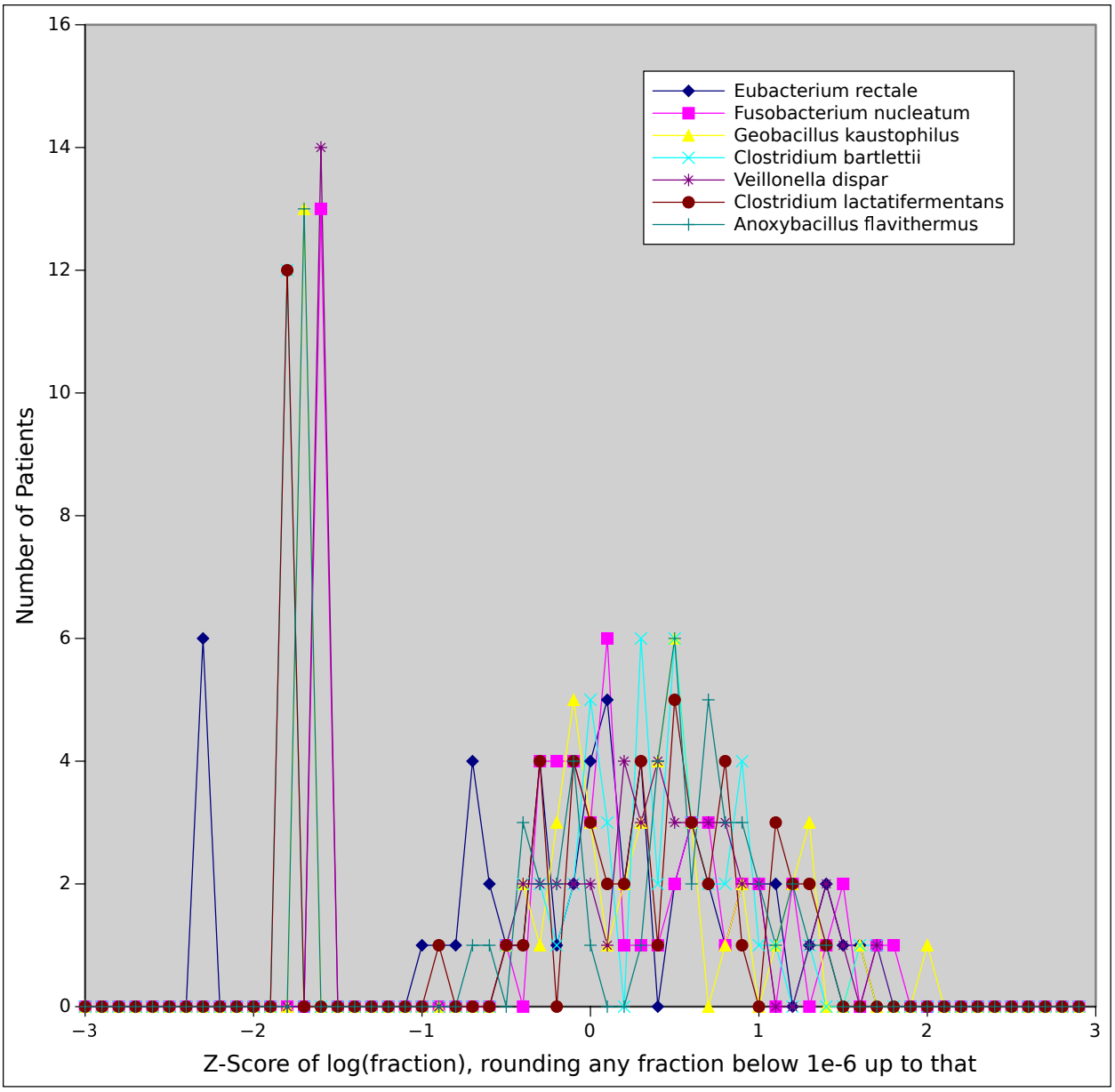

Figure 1: Histogram of representative interesting bacterial species. The fractions are rounded up before taking logs to avoid log(0). These could be approximated as normal, albeit not especially well, were it not for the giant spikes on the left. Perhaps a compound model would be possible, but so many degrees of freedom become tricky to work with.

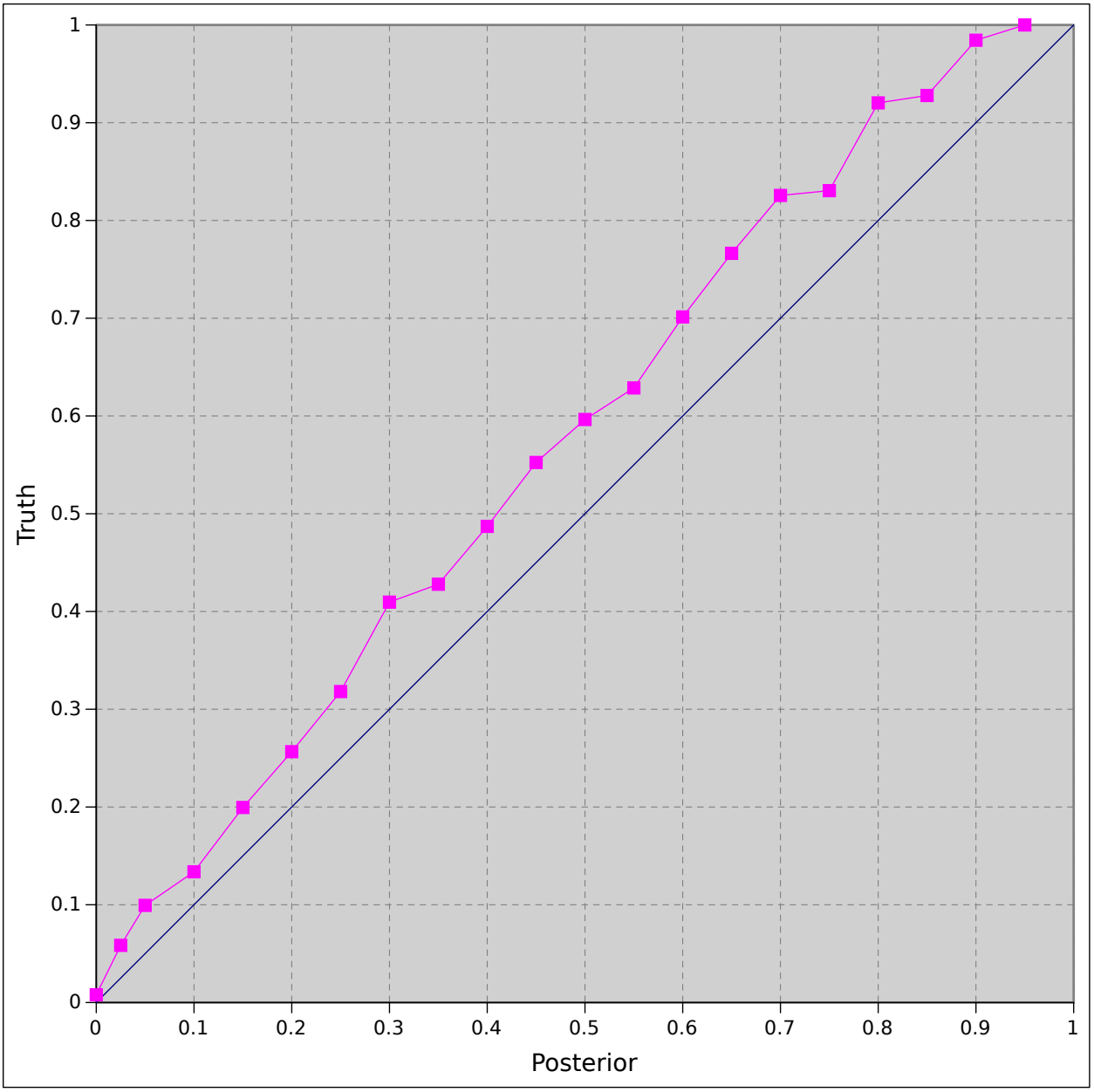

Figure 2: Simulated data using 10k each chain and collide models with flatly random parameters, using 0.05-sized buckets for the posterior. Higher numbers indicate collider.

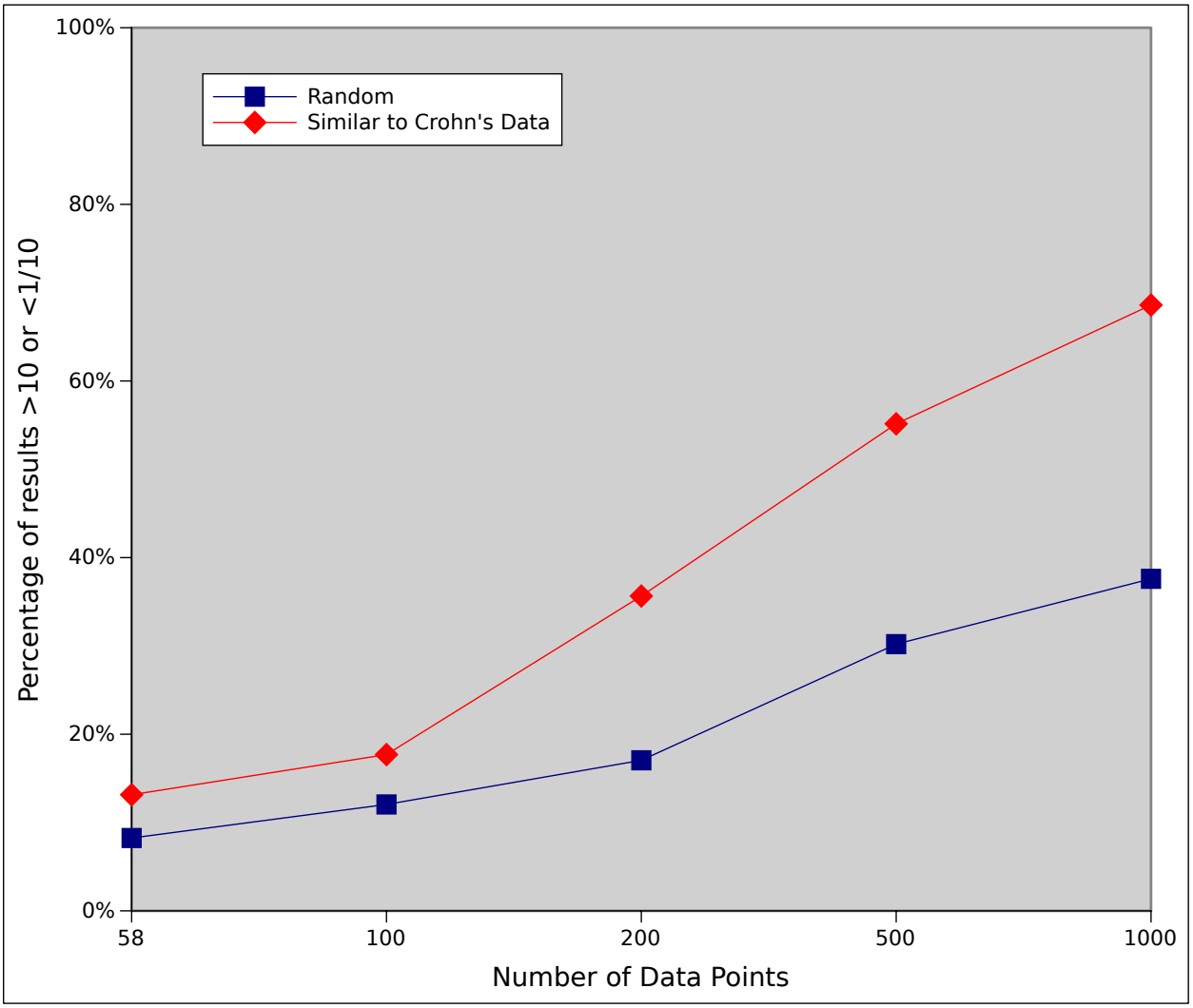

Figure 3: Fraction of bayes factors that were at least 10:1 one way or the other as a function of dataset size

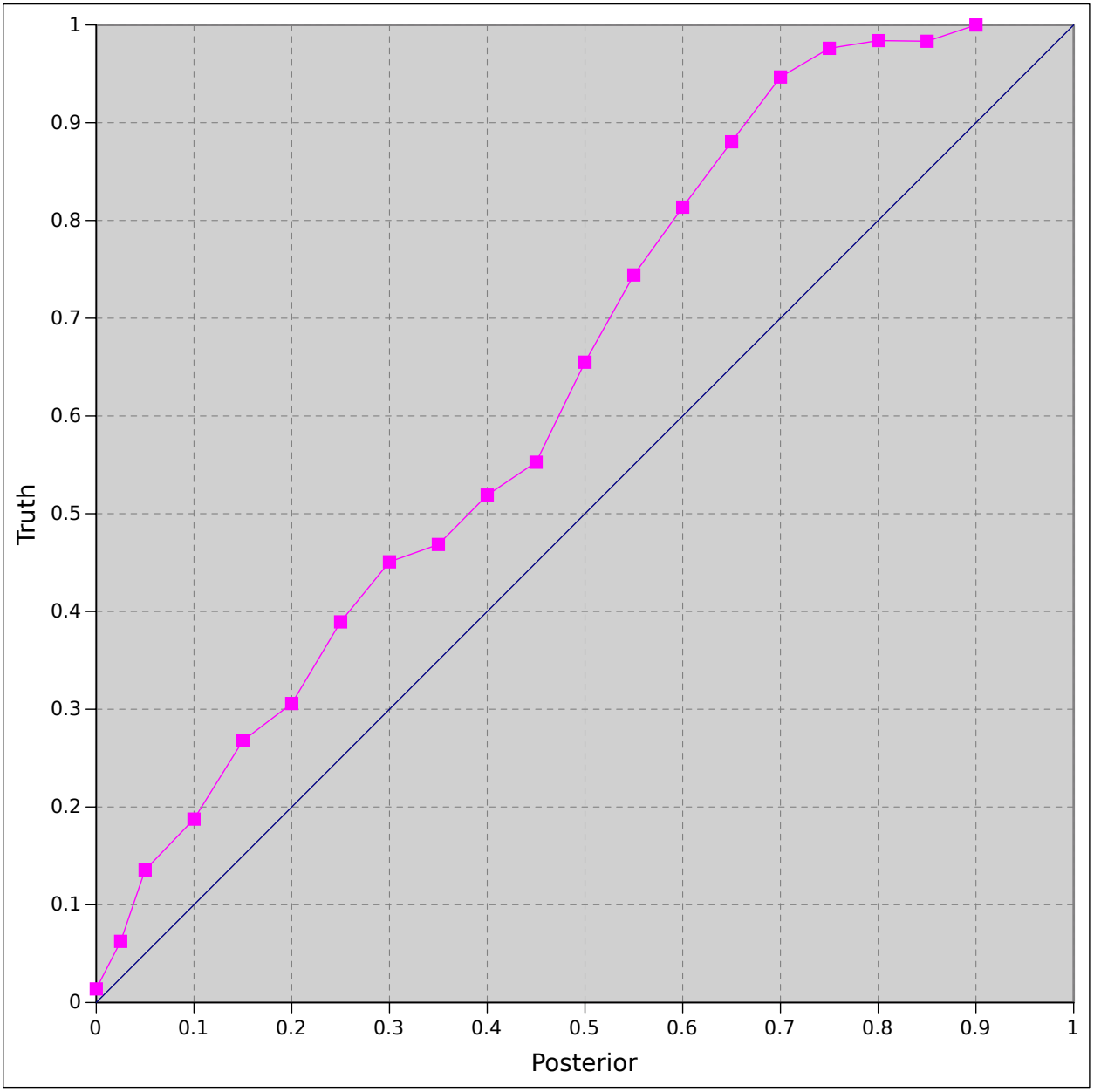

Figure 4: The same simulation as 2, but with p(NOD2) and p(ICD|NOD2) taken from data, and the independences measurable by χ^2 test.

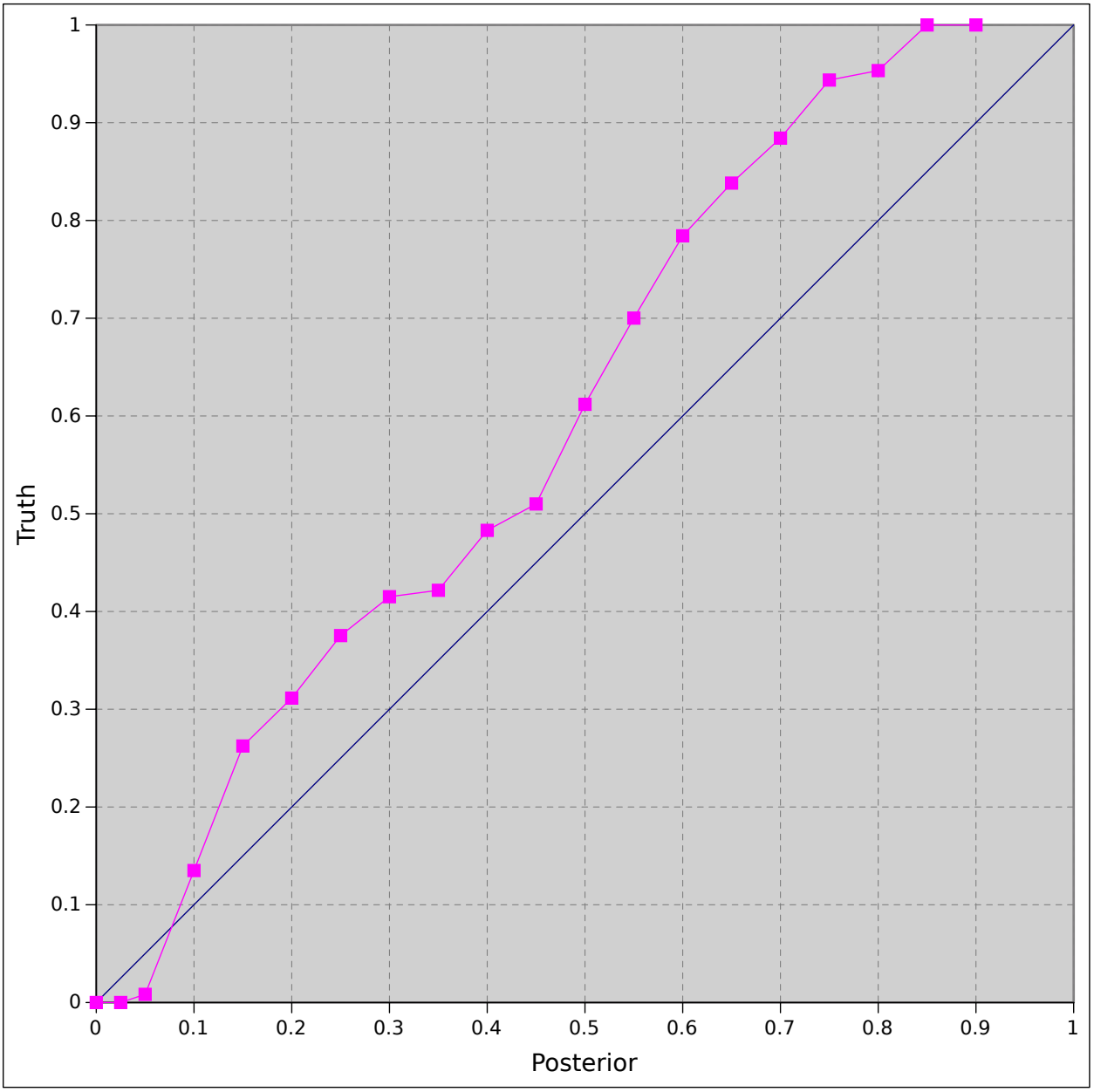

Figure 5: The same simulation as 4, but with allowing an $A \to C$ causal link, albeit one which does not show up as $A \not\perp\!\!\!\perp C$ on a χ^2 test.

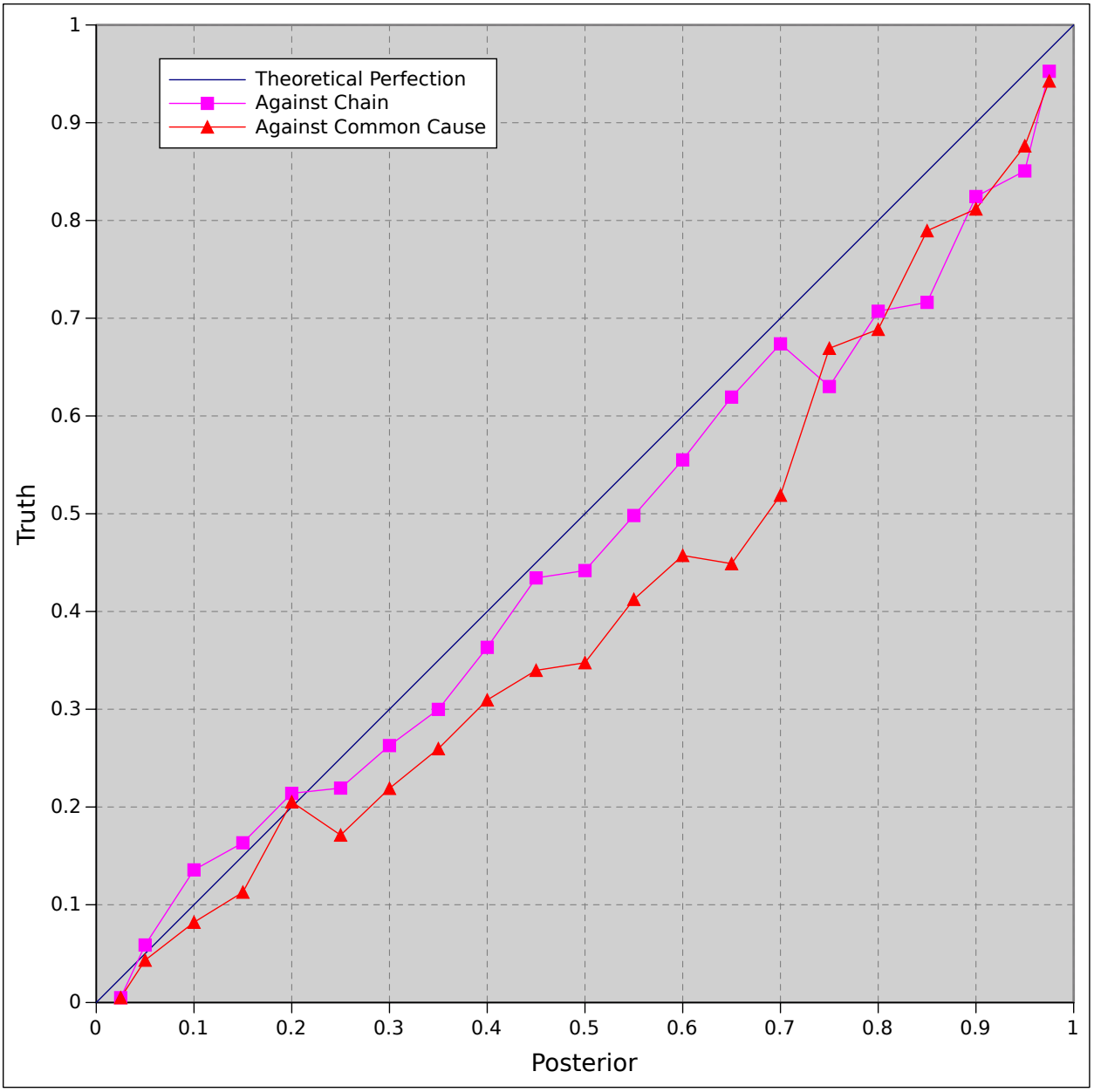

Figure 6: Calibration for the severing test, both in the case it was designed for and in the case of comparing severing to common cause. Higher probabilities indicate severing.

Species	Sick When	P-Value ICD link	Bayes Factor Causality
Streptococcus pseudopneumoniae	>6.36E-05	3.1e-05	5.15
Streptococcus infantis	present	0.0014	2.55
Lactobacillus acidophilus	>8.15E-05	0.00017	10.64
Sphingopyxis alaskensis	present	0.0004	2.43
Clostridium methylpentosum	≤6.31E-04	0.00085	2.53
Roseiflexus castenholzii	present	7.3e-05	3.30
Ruminococcus faecis	≤1.07E-03	0.0022	2.40

Figure 7: Results of the direction test on 222 interesting species, using χ^2 to check a relationship to Crohn's Disease and the direction test described here to establish that it causes (or prevents) the disease.