Fallacies of Reasoning

How We Fool Ourselves

Wüst, Valentin

December 18, 2019

Contents

1 Introduction

There is a lot to be said about how others can try to fool us, and misrepresent facts to suit their narrative. However, what is equally interesting is how humans can fool themselves. One of the ways of fooling oneself are common fallacies of reasoning, of which I will present some in the following.

2 Base-Rate / Prosecutor's Fallacy

In the presentation that was given just before mine, there was a statement concerning an accuracy: "Humans are 54% accurate in recognizing lies." This immediately provoked a question about the corresponding base-rate of lies with which this value was obtained.

Now, in this concrete example, one could argue that an accuracy of 54% is unimpressive regardless of the base rate. Always guessing "lie" will achieve an accuracy that is equal to the base-rate of lies, while always guessing "no lie" will achieve an accuracy of one minus the base-rate of lies. Therefore, by choosing one of those strategies, it is always possible to achieve an accuracy of at least 50% by guessing in a manner that is completely uncorrelated to the way in which the lie is presented. One might further assume that humans are not actively worse at spotting lies than random chance, which would mean that the base-rate of lies should lie between 46% and 54%.

But in general, accuracies, relative changes, and percentages are meaningless unless the corresponding base-rate is known. What is usually called the base-rate fallacy is that humans tend to ignore base-rates when presented with relative information.

2.1 HIV Self-Test

One topic where the base-rate fallacy can be observed is the presentation of medical tests, in this example an HIV self-test. It works by using blood from a finger-prick, and the test result will either be positive or negative. The specificity, i.e. the conditional probability of getting a positive result if you are truly HIV positive p(P|HIV), of those test is practically one. This is of course intentional, since the case where an infected person is not recognized by the test should be avoided. A high sensitivity usually comes at the price of a lower specificity, i.e. the conditional probability to get a negative result if you are in fact HIV negative p(N|¬HIV). However, for this test it is still about 99.8%, which seems to be reasonably close to one on first glance. This information can also be

presented in a table, as is done in Table 1, where the missing values have been calculated by using that all columns have to add to one.

	HIV positive	HIV negative
test positive	$p(P|HIV) = 1$	$p(P|\neg HIV) = 0.002$
test negative	$p(N|HIV) = 0$	$p(N|\neg HIV) = 0.998$

Table 1: Probabilities for getting negative and positive results if you are HIV negative or positive.

Let us now take a look at how one of those tests is promoted, the following examples are taken from the web page and instructions of use of the INSTI HIV test:

- "Accuracy greater than 99%."
- "Two dots means your test result is positive. You are probably HIV positive."
- "Specificity is calculated by dividing the number of INSTI negative test results by the number of truly HIV negative persons that were tested. The higher the specificity the better the test is at correctly identifying truly non-infected persons. The INSTI HIV Self Test has a specificity of 99.8% in a performance evaluation conducted by untrained lay users. This means a positive result will be correct 998 out of every 1000 tests."

The first statement is true, since the accuracy is defined as the share of true results among all results. As both HIV positive and HIV negative persons will get a true result in at least 99.8% of all cases, a population with any prevalence of HIV will get almost only true results. But is this actually the number we care about? Additionally, if the prevalence of HIV is lower than 0.2%, even a "test" that only returned negative results regardless of actual infection status would achieve this accuracy.

The second statement is closer to what we would actually want to know, that is, how high the probability of having HIV is if the test result is positive. We have already established that negative results rule out HIV, so the only remaining question is how accurate positive results are, not the overall accuracy of the test. The third example tries to answer this question, but confuses the specificity with the probability that a positive result will be a true positive. However, using Bayes' theorem, we can calculate this probability $p(HIV|P)$:

$$\begin{aligned}
\mathrm{p(HIV|P)} &= \frac{\mathrm{p(P|HIV)\,p(HIV)}}{\mathrm{p(P)}} \\
&= \frac{\mathrm{p(P|HIV)\,p(HIV)}}{\mathrm{p(P|HIV)\,p(HIV) + p(P|\neg HIV)\,p(\neg HIV)}} \\
&= \left(1 + \frac{\mathrm{p(P|\neg HIV)}}{\mathrm{p(HIV)}}\,(1 - \mathrm{p(HIV)})\right)^{-1}
\end{aligned}$$

Here, we used that $\mathrm{p(P|HIV)} \approx 1$ and $\mathrm{p(\neg HIV)} = 1 - \mathrm{p(HIV)}$. A plot of the resulting function of the prevalence of HIV p(HIV) is shown in Fig. 1. We can now insert actual prevalences into this function to see how accurate positive results of this test actually are. The prevalence of HIV in Germany is about one in 1000, which means that positive results are only 33% accurate. However, in South Africa about 20% of people are infected with HIV. In this high-prevalence population, the accuracy of positive results is almost perfect at 99%.

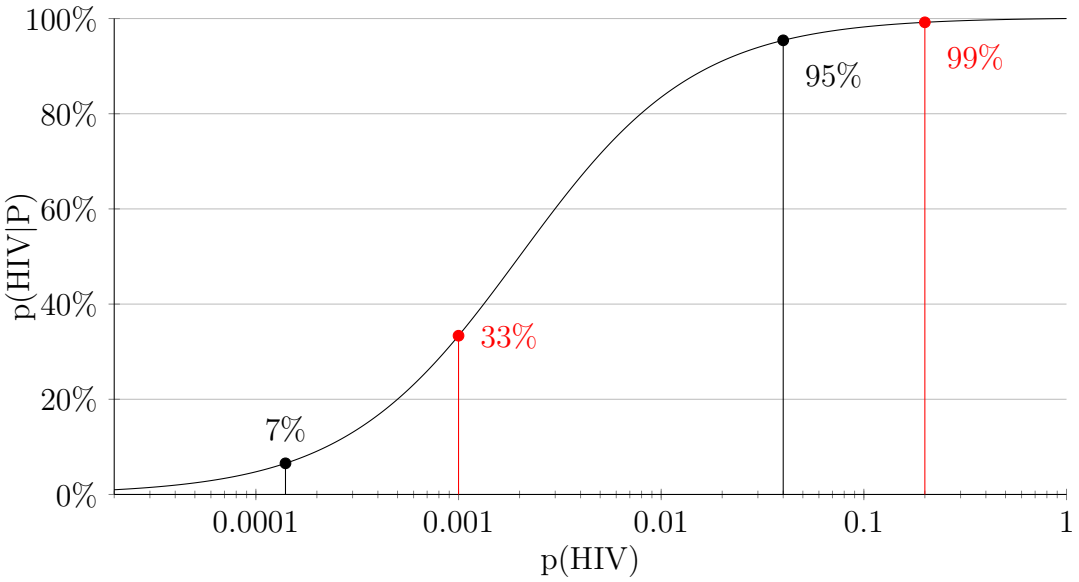


Figure 1: True positives for different prevalences. The two high values are for South Africa, the two low values for Germany. The red ones use the total prevalence of HIV, the black ones the prevalence of undiagnosed HIV among those that do not have a diagnosis of HIV.

In practice, however, those who are already diagnosed with HIV should be excluded from this calculation, since they are not the target audience for an HIV test. Since in both Germany and South Africa about 87% of all infected are already diagnosed, the

prevalence of undiagnosed HIV among the remaining population is about 1 in 7000 for Germany, and about 4% for South Africa, which gives an accuracy of 7% and 95%, respectively. So for the still relatively high prevalence of 4%, positive results usually indicate an infection, whereas in Germany they are close to meaningless. The problem with communicating those numbers is that this calculation crucially depends on the prevalence in the target population. If this test is marketed for certain high-risk groups, the results are quite reliable. For example, there are homosexual communities in Germany where the prevalence of undiagnosed HIV is probably in the percentage range, and for them, it is justified to claim that a positive result probably indicates an infection. But for the larger population, this is certainly not true, as we have just seen.

`http://www.rwi-essen.de/unstatistik/86/`
`https://www.instihivtest.com/en/`

2.2 Facial Recognition

A similar example concerns facial recognition. A joint project of the German ministry of the interior, the federal police and the Deutsche Bahn tested the feasibility of facial recognition in train stations as a means of detecting certain targets, such as Islamists who might be prone to terrorism. There are a host of other issues regarding how they obtained their results, for example the testing conditions were far from realistic. But even the results they obtained under those conditions would make for a rather useless system, although they were communicated as a huge success. They used a list of target persons, and measured the sensitivity and specificity with which they could be identified among non-targeted persons. The results are displayed in Table 2.

	target	not target
alarm	0.8	0.001
no alarm	0.2	0.999

Table 2: Probabilities of raising an alarm for persons that were either targeted or not.

Again, one could argue that the overall accuracy of their test is at least 80%. However, as before, the accuracy of positive results is what is more relevant in practice. We can again calculate it using Bayes' theorem, but first we need the base-rate of targets. This technology was promoted as a way of alerting the police when potential Islamist terrorist frequent train stations, so we will estimate the corresponding base-rate. In Germany, about 12 million people take the train every day. On the other hand, there are only about

600 known Islamists who are on what amounts to a terrorist watch-list. Assuming that they take the train about as often as everybody else, about 100 of them will frequent a train station on any given day, which amounts to a base-rate of only $p(T) \approx 8 \times 10^{-6}$, or one in 120,000. This means that the accuracy of alarms will only be $p(A|T) = 1/125$, so for every alarm that indicates an Islamist, 124 false alarm will be raised. In practice, this accuracy would be too small for any meaningful reaction to alarms, as you would get 80 justified alarms every day, but also 12,000 false alarms. Again, the low base-rate ruins everything.

`http://www.rwi-essen.de/unstatistik/84/`

2.3 Hormonal Birth Control

Another topic where the base-rate is usually ignored is hormonal birth control. Periodically, the high relative increase in thrombosis risk for women who use hormonal birth control is mentioned, but the small absolute risk for young women is usually ignored.

Susan Solymoss. "Risk of venous thromboembolism with oral contraceptives". In: *CMAJ : Canadian Medical Association journal* 183.18 (2011)

2.4 SIDS

A fallacy that is closely related to the base-rate fallacy is the prosecutor's fallacy, which is again best illustrated with an example.

A British lawyer named Sally Clark had two children, in 1996 and 1998. Both of them died from SIDS, i.e. sudden infant death syndrome. This is a diagnosis that is made if an infant dies before its first birthday, and if no cause of death can be established, which is not very common but does happen. A prosecutor then calculated the probability of two children dying from SIDS to be one in 73 million and proceeded to indict her for murder, arguing that this is so improbable that she must have killed her children. She was subsequently convicted by a jury, with the sole evidence against her being this probability, and spent three years in prison until she was eventually acquitted.

However, the prosecutor made two crucial mistakes. First, he assumed that the death of her children were uncorrelated events, when in reality the chance of dying from SIDS is much higher for children who already lost a sibling to it. A more reasonable estimate

of the probability for two sibling to both die of SIDS is one in 300,000, but still this number in itself does not tell us anything. It expresses the conditional probability of both her children dying, given that they died of SIDS p(both siblings died|SIDS), when in reality, the probability that their children died from SIDS given that they are dead p(SIDS|both siblings died) is the relevant quantity. We are not interested what the prior probability for the death of her children is, since this event has already taken place. To argue that she killed them, what is actually required is the chance that this was not due to SIDS, and therefore due to murder, given that it already happened.

But in the example in subsection 2.1, we have seen that one cannot simply assume that p(A|B) = p(B|A), since to get from one to the other we have to use Bayes' theorem. Crucially, the result will depend on the base-rates of the different conditions! So the correct calculation for this example would compare the probability of her children dying from SIDS with the probability of the alternative, i.e. that she killed them. As double infanticide is comparatively unlikely, the result will be completely different from the prior probability of either of them happening. One can try to estimate the probability of her guilt, given that her children died, as was done in [Hil04], and it is probably closer to 10%, a far cry from the one in 70 million chance implied by the prosecutor, and certainly not large enough to convict her without any further evidence towards her guilt. It is probably even smaller, since there are reasons to believe that the pathologist actually overlooked evidence of an infection in her second child.

This is, however, a commonly used argument by prosecutors. Every time the chance of a random agreement, e.g. for DNA tests, is mentioned, it is strongly implied to correspond to the probability for the defendants innocence. In reality, one also needs to consider the prior probability, since for example applying a DNA test to a large enough population of innocent people will invariably produce false positives.

Intuitively, most people understand this. Nobody would automatically prosecute lottery winners for fraud on the sole basis that winning the lottery is very unlikely. But as with the base-rate fallacy, it is still easy to ignore the pitfalls of Bayesian statistics.

Ray Hill. "Multiple sudden infant deaths – coincidence or beyond coincidence?" In: *Paediatric and Perinatal Epidemiology* 18.5 (2004)

2.5 p-Values

The prosecutor's fallacy is also commonly applied to p-values. The p-value is defined as the conditional probability of getting results that are as, or more, extreme as the

actual results, given that the null hypothesis it true, i.e. that the postulated effect does not exist. However, we have just seen that it is, in principle, impossible to infer the probability of the null hypothesis being false from just the p-value, as the prior probability of the null hypothesis needs to be taken into account.

$$p(\text{Data}|\text{Null}) \neq p(\text{Null}|\text{Data})$$

If the prior probability for an effect is very small, even a small p-value is not sufficient to significantly improve the posterior probability, while hypotheses that are almost certainly true should not be discarded even if a small p-value is not achieved. This is just a formal version of saying that "extraordinary claims require extraordinary evidence".

This can also be used as an argument for why most published research findings might be false. If one assumes that most tested hypotheses are unlikely to be true, and then uses $p < 0.05$ to either confirm or discard them, one would expect to get many more false positives than true positives.

One illustrative example concerns epidemiology. Consider an epidemiologist who has randomly selected 50 participants for his study, and one of them happens to be an albino. Now, being a good scientist, the researcher might calculate a p-value for this observation, assuming that the test subjects are humans. Since albinism is quite rare, the probability of getting one albino among 50 randomly selected humans is only about one in 400, with gives a p-value of $p \approx 0.0025 < 0.05$, which would, following the conventional logic, be considered significant enough to reject the null hypothesis. But of course the null hypothesis, i.e. that the subjects really are humans, is overwhelmingly likely, so this p-value is not nearly small enough to lower the posterior probability by a large amount.

Jordan Ellenberg. *How not to be wrong.* London: Penguin Books, 2015, p. 136

3 Gambler's / Hot Hand Fallacy

3.1 Monte Carlo 1913

On August 18, 1913, roulette was played in the Casino in Monte Carlo, when black came up fifteen times in a row. At this point, gamblers, who thought that surely red was now long overdue, began to bet large amounts of money on red. However, the streak continued for another eleven rounds, bringing the total to twenty-six times black. At this point, the Casino had made some millions of francs.

Figure 2: Roulette.

The gamblers in his example exhibited the classic gambler's fallacy, they assumed that the past occurrence of black or red influences their future likelihood. There are multiple arguments for why this must be false. If this were indeed the case, the probability of future events would depend on the amount of observations one takes into account. As the roulette wheel should not care how many past results we observed, it cannot depend on our observations. Also, probabilities are only guaranteed to hold in the limit of infinitely many observations. Streaks like the above are not actively compensated for, but in this limit they are diluted until they no longer matter.

Darrell Huff and Irving Geis. *How to take a chance.* Gollancz, 1960, pp. 28-29

3.2 Would You Pay for Transparently Useless Advice?

Humans are prone to paying for objectively useless advice. Obvious examples for this are fortune-tellers and astrologists, but also economic forecasts and investment advice. For example, it has been shown that, on average, financial "experts" at most barely outperform the market, and usually not to a degree that justifies their high salaries.

However, humans tend to infer agency when presented with random occurrences. This causes us to overestimate the influence that skill has on those predictions, and underestimate the influence of chance. In particular, humans tend to attribute streaks of good and bad performance to the skill of the predictor, when in reality they are to a large degree due to chance. This is connected to the fact that humans tend to vastly underestimate the occurrence of streaks in random data, since those simply do not "feel" random.

Powdhavee et al. investigated whether this effect also persists for obviously impossible predictions in [PR15]. They used 378 undergraduate students from Thailand and Singapore for their study, and asked them to bet on the outcome of coin flips. It was made completely clear that the outcome of those coin flips was random. The coins were provided by participants, flipped by randomly drawn participants, and both the coins and the person who flipped them were repeatedly exchanged. In short, the outcomes were clearly not predictable. The researchers then predicted them.

Before the experiment started, every participant was supplied with a random prediction for all five coin tosses. They could pay money before every toss to open the relevant prediction, if they did not do so they were still instructed to open them after every coin toss. By random chance, some predictions were right, but of course the number of participants who observed only correct predictions halved with every round. They observed that those participants for whom all previous predictions were right were significantly more likely to buy the next prediction (see Fig. 3), even though correct predictions were obviously due to chance alone. This fallacy is usually called the hot-hand fallacy, where a streak of correct predictions, bets etc. is assumed to be more likely than not to continue, when in reality of course past random occurrences do not influence future probabilities. In some sense, this is the opposite of the gambler's fallacy.

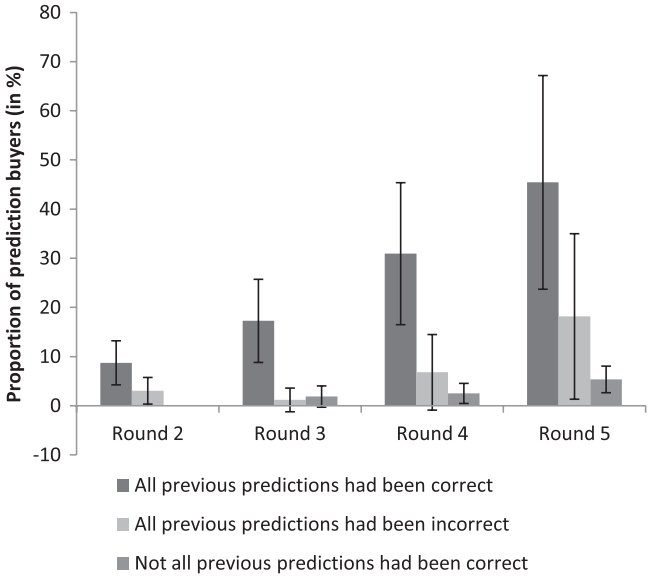


Figure 3: Results of Powdhavee et al.

However, there are some points of their analysis that can be criticized. I observed that the confidence intervals of their results approximately doubled between the pre-print of the paper [PR12] and the eventual (after three years) published version, so I am probably not the first person to criticize them. In general, those confidence intervals are quite large for the relevant groups. This is of course caused by the real randomness in their experiment, since the number of people who observe only correct predictions decays exponentially. They are still small enough to support the general conclusion, but concrete values should be taken with a grain of salt.

They do a rather elaborate statistical analysis, which results in about 160 fit parameters, for each of which they indicated whether it was significant with $p < 0.1$, $p < 0.05$ and $p < 0.01$. I counted them, Fig. 4 shows the percentage of parameters in each p-value range.

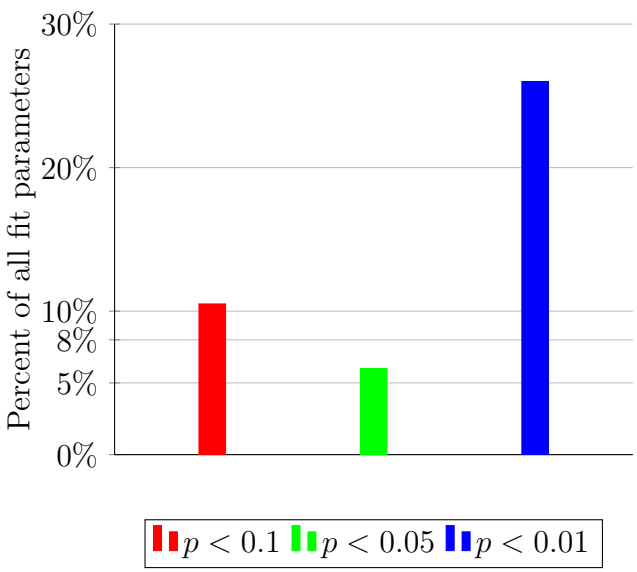


Figure 4: Percent of parameters they indicated as significant below a certain threshold, $p < 0.1$ presumably means $0.1 > p > 0.5$ etc.

The problem with using the normal significance threshold here is that a lot of false positives are expected by chance alone, since they have so many parameters that are tested for significance. This is known as the problem of multiple comparisons. There are many ways of dealing with this, the easiest version of course being to simply ignore it, as the authors did here. A way of fixing the α error, i.e. the probability of getting any false positives, is the Bonferroni correction. If one wants an alpha error α, dividing it by the number of parameters m yields a new significance threshold $\gamma = \alpha/m$, using this results in the overall alpha error α. In this case, this would result in the condition $p < 0.0003$ for significance, which most of the indicated parameters probably would not have crossed. However, this correction is quite drastic, and therefore also strongly increases the β error, i.e. the probability of false negatives. Nevertheless, it is usually used for studies that do huge numbers of comparisons, for example in genome-wide association studies. A less drastic way is the Benjamini-Hochberg procedure, which uses the largest p value for which $p_k \leq \frac{k}{m}\alpha$ is still true as the threshold. It also controls the α error if one assumes that the comparisons are not independent, which is probably better in an analysis like this, and here results in the condition $p < 0.01$. But even with this, many of the parameters they indicated should not be counted as significant.

Nattavudh Powdthavee and Yohanes E Riyanto. "Would you Pay for Transparently Useless Advice? A Test of Boundaries of Beliefs in The Folly of Predictions". In: *Review of Economics and Statistics* 97.2 (2015)

4 Hindsight Bias

4.1 Clinicopathologic Conferences

In the USA, it is traditional that hospitals hold clinicopathologic conferences. A presenter, usually a young physician, is given all the relevant medical information for an old case in which the patient ultimately died, except the final diagnosis. The presenter then has to present the case to an audience, go through the possible diagnoses, and choose the most likely one. Afterwards, the pathologist who did the actual autopsy announces the correct diagnosis. As those cases are usually chosen for their difficulty, the presenter frequently chooses an incorrect diagnosis.

Dawson et al., [Daw+88], went to four of those conferences, with a total number of 160 attending physicians. They then divided them into two groups, a foresight (N=76) and a hindsight (N=84) group. The foresight group was asked to assign probabilities to the possible diagnoses after the presenter finished, the hindsight group was given the same task after the pathologist had already stated the correct diagnosis. Two of those cases were deemed harder than the other two by expert physicians, and for their analysis they further divided the attendees into more (N=75) and less (N=85) experienced physicians. Fig. 5 is the original graph of the results from their paper. Three of the groups show a hindsight bias, but the group of more experienced physicians for the harder cases actually shows a negative bias.

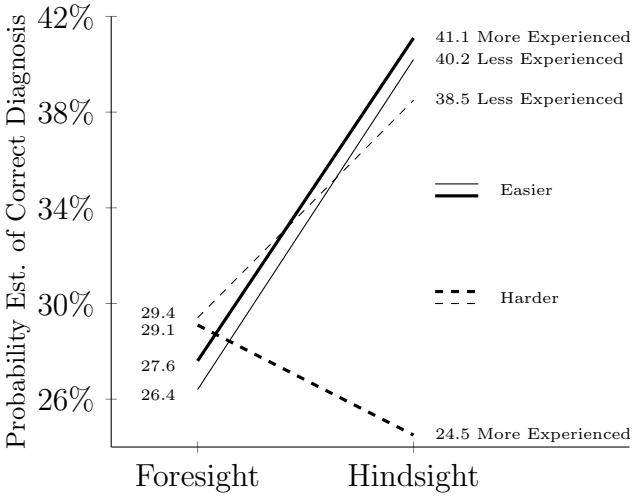
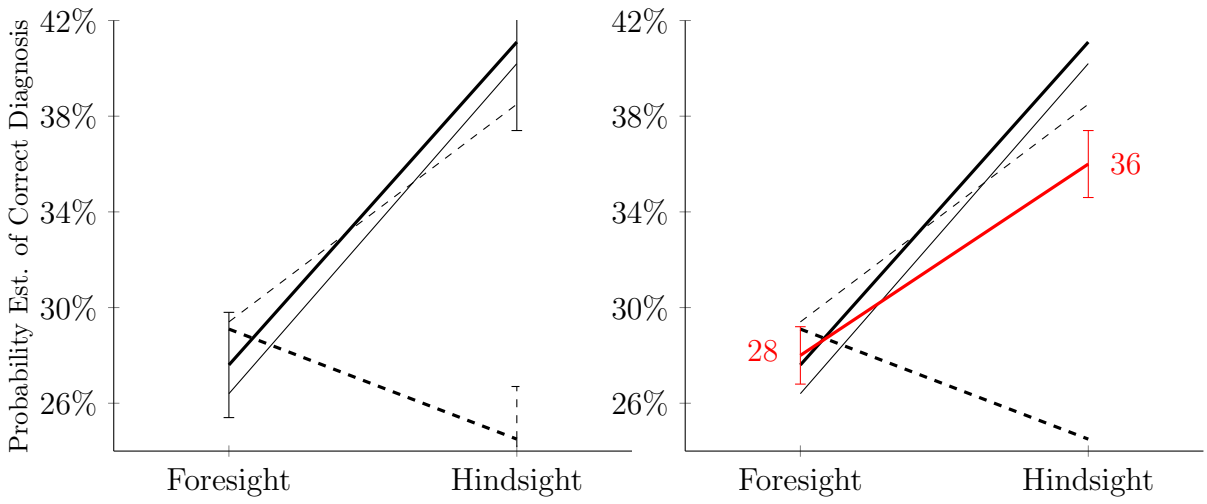


Figure 5: Mean estimated probabilities of the correct diagnosis by timing of the estimates (foresight vs. hindsight), experience of the estimators (less, thin lines, vs. more, bold lines), and case difficulty (easier, solid lines, vs. more difficult, broken lines).

The difference in mean assigned probabilities for the actual diagnosis for the easier cases was significant with $p < 0.05$. The same difference for the harder cases had $p = 0.06$, "which fell just short of the traditionally accepted significance level". However, due to the multiple comparisons problem, $p < 0.05$ is not actually the correct threshold.

For the three groups that showed the hindsight bias, 30% ranked the correct diagnosis first in foresight, and 50% in hindsight. Even though this analysis excludes one quarter of their data, it is presented in the abstract without this caveat. They later say that 35% of all foresight groups assigned the correct diagnosis the highest probability. It is not given anywhere in the paper, but I would assume the corresponding number from the hindsight group is probably closer to 45%, considering that one forth actually showed a negative bias.

They did not provide any confidence intervals. They also did not mention the concrete N for every group, it is probably about N=20 for each of them. Assuming a Poisson distribution, which usually gives a reasonable estimate, one can calculate the error of the mean as $\Delta = \sqrt{\mu/N}$. Fig. 6a contains confidence intervals estimated in this way.



(a) With 95% Confidence Intervals.

(b) Mean of all groups.

Figure 6: Improved versions of Fig. 5.

Due to the small size of each group, the confidence intervals are quite large. It is also instructive to show the mean results for all groups, as in Fig. 6b, because in general, one has to be very careful about subdividing the participants and analysing the subgroups independently. This does not only result in the usual problem of multiple comparisons, but one also strongly increases the number of comparisons beyond the number that is actually done. Since the same arguments that the authors used to rationalize that one subgroup did not show the bias could also be adapted to explain why any other group did not show the bias, what is now relevant for judging the significance of the found pattern is not simply its probability, but the probability of finding any similar pattern

in the data. In this case, to judge the effect size they found when excluding one group, one would have to calculate the probability of finding such an effect size when one is allowed to exclude any similar group, e.g. the less experienced physicians etc., and when the definitions of those groups are slightly altered. As this probability is usually rather high, and as they did not specify the criteria by which they actually divided the groups, one has to be very careful about analyses of this kind. In general, the mean results for all participants are the relevant ones, but in this case they also show a significant hindsight bias, i.e. the result seems more obvious in hindsight than in foresight.

Neal V. Dawson et al. "Hindsight Bias: An Impediment to Accurate Probability Estimation in Clinicopathologic Conferences". In: *Medical Decision Making* 8.4 (1988)
Hal R. Arkes. "The Consequences of the Hindsight Bias in Medical Decision Making". In: *Current Directions in Psychological Science* 22.5 (2013)

4.2 Determinations of Negligence

In [LL96], the researchers sent out case studies to randomly drawn people, of which about 300 replied. The case studies contained one out of six different cases, in all of which a therapist had to decide how to handle a potentially violent patient. They were then told to imagine themselves as jurors in a malpractice lawsuit, and were asked whether they would convict the therapist for criminal negligence. Crucially, they sent out three different versions of the six cases. For each of them, they varied whether any outcome of the case was reported, and if, whether the patient ultimately did or did not get violent and harmed another person. The participants were then asked to answer eleven questions. The percentages of participants who agreed with two statements, that violence was, a priori, likely, and that the therapist was criminally negligent,

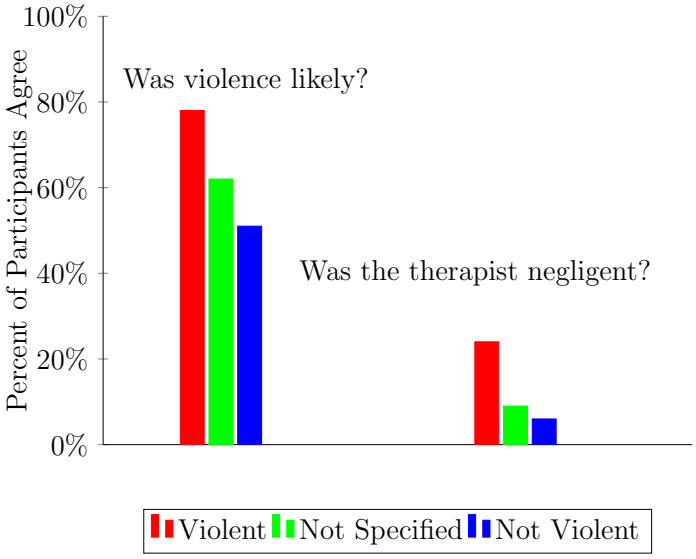


Figure 7: Percentages of participants who affirmed the above questions.

are shown in Fig. 7.

They observed that telling the participants that the patient got violent significantly influenced their judgement of the case study, and that they were more likely to convict the therapist for bad outcomes, irrespective of the actual reasonable behaviour.

Numbers for the different outcome groups were not provided, I had to calculate them myself to be about 97, 94 and 102 for violent outcome, no outcome mentioned, and non-violent outcome, respectively. They also did not include any error bars in their results. However, since they should follow a Poisson distribution, they can be easily calculated as $\Delta\mu = \sqrt{\mu/N}$, and they seem to be small enough (Fig. 8). The problem of multiple comparisons was again ignored, but this is not so crucial here as they did a more modest number of significance tests.

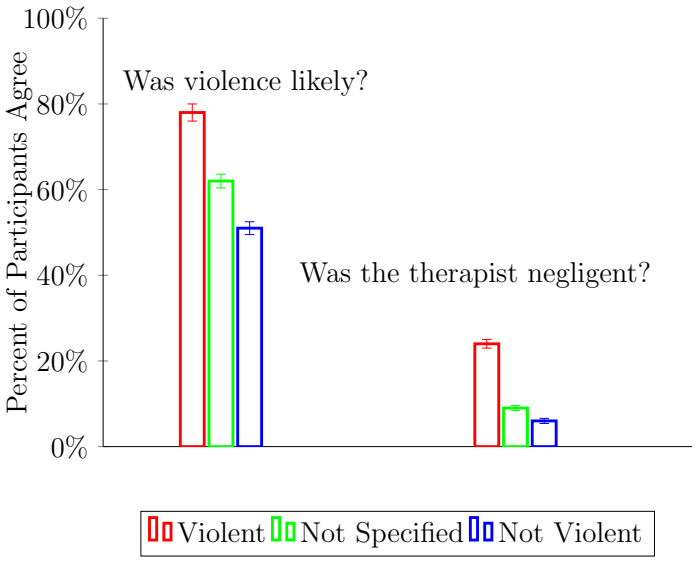


Figure 8: Results from Fig. 7 with error bars (1σ).

Susan J LaBine and Gary LaBine. "Determinations of Negligence and the Hindsight Bias". In: *Law and Human Behavior* 20.5 (1996)

References

[Ark13] Hal R. Arkes. “The Consequences of the Hindsight Bias in Medical Decision Making”. In: *Current Directions in Psychological Science* 22.5 (2013).

[Axe00] Stefan Axelsson. “The base-rate fallacy and the difficulty of intrusion detection”. In: *ACM Transactions on Information and System Security (TISSEC)* 3.3 (2000).

[BGK14] Thomas K. Bauer, Gerd Gigerenzer, and Walter Krämer. *Warum dick nicht doof macht und Genmais nicht tötet. über Risiken und Nebenwirkungen der Unstatistik*. Campus-Verl., 2014.

[Cha+18] Lakshmi Chaitanya et al. “The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation”. In: *Forensic Science International: Genetics* 35 (2018).

[Daw+88] Neal V. Dawson et al. “Hindsight Bias: An Impediment to Accurate Probability Estimation in Clinicopathologic Conferences”. In: *Medical Decision Making* 8.4 (1988).

[Ell15] Jordan Ellenberg. *How not to be wrong*. London: Penguin Books, 2015.

[HG60] Darrell Huff and Irving Geis. *How to take a chance*. Gollancz, 1960.

[Hil04] Ray Hill. “Multiple sudden infant deaths – coincidence or beyond coincidence?” In: *Paediatric and Perinatal Epidemiology* 18.5 (2004).

[How19] Jonathan Howard. *Cognitive Errors and Diagnostic Mistakes*. Springer International Publishing, 2019.

[Krä15] Walter Krämer. *So lügt man mit Statistik*. Frankfurt am Main [u.a.]: Campus-Verl., 2015.

[LL96] Susan J LaBine and Gary LaBine. “Determinations of Negligence and the Hindsight Bias”. In: *Law and Human Behavior* 20.5 (1996).

[PR12] Nattavudh Powdthavee and Yohanes E. Riyanto. *Why Do People Pay for Useless Advice? Implications of Gambler's and Hot-Hand Fallacies in False-Expert Setting*. IZA Discussion Papers 6557. Institute of Labor Economics (IZA), 2012.

[PR15] Nattavudh Powdthavee and Yohanes E Riyanto. “Would you Pay for Transparently Useless Advice? A Test of Boundaries of Beliefs in The Folly of Predictions”. In: *Review of Economics and Statistics* 97.2 (2015).

[Sol11] Susan Solymoss. “Risk of venous thromboembolism with oral contraceptives”. In: *CMAJ : Canadian Medical Association journal* 183.18 (2011).