Categorical Belief Updating Under Uncertainty

Stephen H. Dewitt[1], Carmen Li[1], Daniel Koh[1], Norman E. Fenton[2], David Lagnado[1]

[1]Department of Experimental Psychology, University College London, 26 Bedford Way, WC1H 0AP
[2]School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Rd, London E1 4NS

Abstract

The need to update our estimates of probabilities (e.g., the accuracy of a test) given new information is commonplace. Ideally, a new instance (e.g., a correct report) would just be added to the tally, but we are often uncertain whether a new instance has occurred. We present an experiment where participants receive conflicting reports from two early-warning cancer tests, where one has higher historical accuracy (HA). We present a model showing that while uncertain which test is correct, estimates of the accuracy of both tests should be reduced. However, among our participants, we find two dominant approaches: (1) participants increase the more HA test, reducing the other; (2) participants make no change to either. Based on mixed methods we argue that both approaches represent two sides of a 'binary' decision i.e., (1) update as if we have complete certainty which test is correct and (2) update as if we have no information.

Keywords: Categorical Reasoning; Bayesian; Uncertainty; Causal; Confirmation Bias

Introduction

The Problem

Reasoning under uncertainty has been a major focus within JDM for many decades. In the classic medical diagnosis problem (e.g., Casscells, Schoenberger & Graboys, 1978), below, participants are provided a population base rate for a disease and the FPR of a test to detect that disease. They are then asked to determine the chance that an individual who has had a positive test result actually has the disease:

.

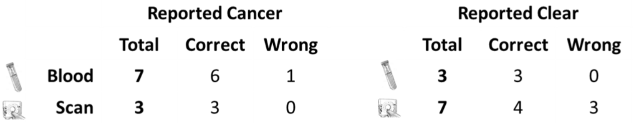
> *"If a test to detect a disease whose prevalence is 1/100 has a FPR of 5 per cent, what is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person's symptoms or signs?"*

A huge amount of research has been conducted examining why, when and for whom people can solve this problem (e.g., Gigerenzer & Hoffrage, 1995). However, in this paper we are interested in a secondary question that can be asked following the positive test result, but which has not been studied before. Instead of asking for the revised probability that the person has the disease, we are interested in participants' estimates of the revised accuracy of the test for positive reports. This question introduces a whole new set of dynamics, requiring participants to wrestle with higher orders of uncertainty (Kleiter, 2018).

Previous literature on this problem has assumed a fixed-point estimate for the FPR (e.g., 5%), with no variance, and thus no capacity for it to change. Indeed, it may well be reasonable to assume in the standard scenario that the 'test' has been extensively conducted such that our confidence in that 5% FPR being correct is so high that it is reasonable to treat it as a fixed-point estimate.

However, what about situations where we have very limited data on the accuracy of a test, and want to provide the best ongoing estimate possible, such as during the early stages of a clinical trial? How should the mere fact that the test reports a positive result affect our estimate of its accuracy? What if two such tests gave conflicting results?

For example, imagine we were trialing two early warning tests for cancer with patients in a high-risk group, one using blood, and another using a type of scan. We get the predictions from the two tests and monitor patients for 20 years to see if they develop cancer or not. We have so far had results from 10 patients, which can be seen below. For both tests you can see how often they reported 'cancer' ('positive') and how often they reported 'clear' ('negative'). Next to these numbers you can see how often their reports were correct or wrong. For example, the blood test reported that 7 out of the 10 patients had cancer, and 6 of these patients subsequently developed cancer (correct), while 1 did not (wrong). Similarly, the scan test reported that 7 out of the 10 patients did not have cancer, and 4 of these did not develop cancer (correct), while 3 did (wrong).



	Reported Cancer			Reported Clear		
	Total	Correct	Wrong	Total	Correct	Wrong
Blood	**7**	6	1	**3**	3	0
Scan	**3**	3	0	**7**	4	3

Figure 1. Test accuracy rate figures shown to participants in the 'blood' condition

The Model

Imagine now that we have a new patient from the same high-risk group (50% chance of developing this type of cancer over the next 20 years). We run both tests, and the blood test gives a positive result, while the scan test gives a negative. Before we potentially wait 20 years to find out the result, is it possible to adjust our estimates of the accuracy for these types of report (positive for the blood test, negative for the scan test)? We constructed a model of this scenario using Bayesian network software, which can be seen in Figure 2.

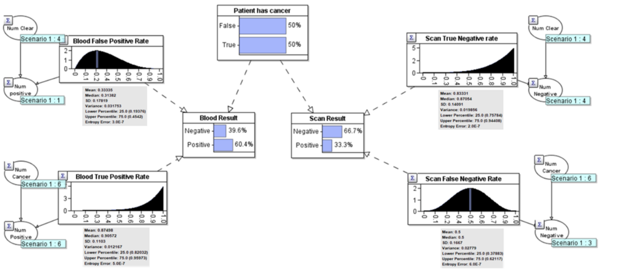


Figure 2. A Bayesian model of the 'two tests' problem with no observations made

While visually complicated, there is a lot of repetition in the model such that it is simpler than it may first appear. In the top center we have a node which represents the probability of the patient having cancer (50:50). On the left we have all the nodes relating to the blood test, and on the right all equivalent nodes relating to the scan test. For example, at the top left we have a distribution representing the false positive rate (FPR) of the blood test, based upon the above table (four patients were 'clear', and once it misreported one of them as having cancer). On the bottom left we have the true positive rate (TPR) for the blood test (six patients had cancer, and every time it correctly reported this).

When we move to the scan test on the right, we switch to using true negative rate (TNR) / false negative rate (FNR) labels instead of false positive / true positive. This is simply a re-framing as the TNR is the complement of the FPR (i.e., if a test has a 75% TNR, it has a 25% FPR), and the TPR is the complement of the FNR. We make this change because in this scenario the blood test provides a positive report, while the scan test provides a negative. Therefore, using the 'positive' framing of the rates for the blood test, and the 'negative' framing for the scan test makes the impact of those observations on both tests clearer and requires fewer mental gymnastics.

In this model each distribution begins as a uniform (0,1) distribution, updated based upon the observations so far. This does not produce a classical frequentist mean estimate. For example, the FPR of the blood test in Figure 2 provides a FPR mean for the blood test of 33.3…% rather than the classical frequentist estimate of 25% (1/4). This Bayesian approach takes sample size into account when estimating the mean, unlike the classical approach, and provides identical means to the values given by LaPlace's (1814) 'rule of succession':

Equation 1. LaPlace's (1814) 'Rule of succession' where x represents the number of observed instances, while n represents the number of trials

$$\frac{x+1}{n+2} = \frac{1+1}{4+2} = \frac{2}{6} = 0.333\,...$$

This approach is conservative, pulling mean estimates towards 50% and away from the extremes. As n grows, the estimate converges on the frequentist estimate. The benefit of this approach can be seen when estimating the TPR for the blood test. Here we have six correct out of six, providing a classical estimate of 100% and no variance, while the Bayesian mean is only 87.5% (using LaPlace's rule, 7/8 = 0.875). If the perfect accuracy were maintained with more trials, this estimate would converge on 100%, but never quite reach it, allowing that there is always the possibility of a future error. For example, with 60/60, the model produces a mean of 98.4% (61/62). We can also see that the TNR mean for the scan test (also perfect so far but based on the smaller sample size of 4/4) is a little further from 100%, at 83.33…% (5/6). Finally, as we will see, these specifics do not affect the main findings of the paper, as we are concerned with direction of change (decrease / no change / increase) rather than specific numerical values or precise magnitude.

Observations

We can make observations on the model, which effectively fix certain nodes at '100%' and the effects of these observations on other nodes in the model can be calculated. For example, when we observe that the blood test has reported a 'positive' (i.e., that the patient has cancer) the model increases both the FPR mean (to 36.0% [+2.66…%]) and TPR mean (to 88.5% [+1.0%]). This reflects the fact that we know the test has made a 'positive' claim, but we do not yet know which of these two it is (true or false), so the 'value' of this instance is spread between them based on this uncertainty. The probability that the patient has cancer rises to 72.4%, as can be seen in Figure 3.

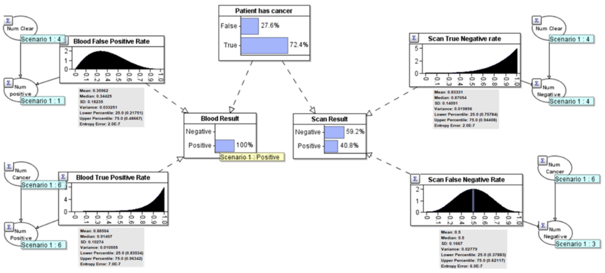


Figure 3. A Bayesian model of the problem with an 'observed' positive blood test

However, when we observe that the scan test has reported negative (Figure 4), the FPR mean of the blood test goes up further (to 37.0% [+1.0%]), and the TPR mean comes back down a little (to 88.3% [-0.2%]). This reflects the fact the positive report by the blood test is now more likely (but not certain) to be a false positive than a true positive. This is because the only other evidence we have (the scan test, which is still an informative test, even if not as accurate as the blood test) contradicts it. The probability that the patient has cancer comes down to 61.2%, which is still higher than the original 50% as we have two conflicting results but the test saying ‘positive’ (blood) has been more accurate in the past for these particular reports.

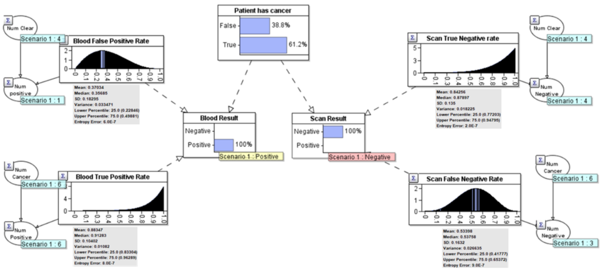


Figure 4. A Bayesian model of the problem with ‘observed’ positive blood test and negative scan test

From the point of view of the scan test, the initial TNR mean is 83.3% and the initial FNR mean is 50.0%[1] (Figure 2). When we observe the negative scan result before we make the blood test observation (not shown) these are revised to 84.8% [+1.5%] and 52.1% [+2.1%] respectively, as we do not know if this is a true negative or a false negative. When we observe that the blood test has produced a positive report (Figure 4), these are again revised to 84.3% [-0.5%] and 53.4% [+1.3%], indicating that the negative scan report is now more likely to be a false negative than a true negative.

Finally, we can see the effect once we find out for certain whether the patient really has cancer (Figure 5). If we observe that they really did not have cancer, the TPR mean of the blood test returns to its original level (87.5%), while the FPR mean increases further (to 42.9% [+5.9%]). This reflects our certainty that the blood test has produced a false positive: this node effectively now has 2/5 rather than 1/4 (and using LaPlace’s law, 3/7 = 0.429). The TPR however is now back to 6/6, so is unchanged from the baseline. Additionally, we now know that the negative scan test report was correct, so that TNR mean increases to 85.7% (now at 5/5 so with LaPlace’s law, 6/7 = 0.857), and the FNR returns to its original value (50%).

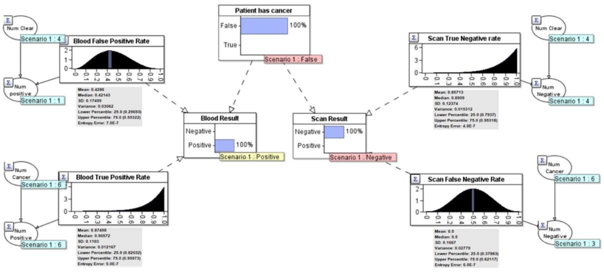


Figure 5. A Bayesian model of the problem with observations showing a positive blood test, negative scan test, and that the patient really does not have cancer

The initial figures provided set up the blood test as the more accurate of the two for the specific reports made (most notably, the FNR mean of the scan test is considerably higher than the FPR mean of the blood test). In the ‘conflict’ situation we set up, it is therefore entirely appropriate to conclude that the blood test is more likely to be correct on this occasion than the scan test, even though we cannot know for certain. However, when dealing with uncertainty in a range of situations, people are known to often treat uncertainties as categorical or digital values (0% or 100%) i.e., to make assumptions. This seems to be especially the case in multi-stage inferences, defined by Gettys, Kelly & Peterson (1973) as:

“…a series of single-stage inferences where the output of each previous stage becomes the input to the next stage” (Gettys, Kelly & Peterson, 1973 pp.364)

This is relevant to the current situation where participants may (stage 1) estimate which test is more likely to be correct on this occasion and (stage 2) use this assessment to update their estimates of the accuracy of each test for these reports (e.g., positive for the blood test, negative for the scan test). Both Gettys et al (1973) and Johnson, Merchant & Keil (2020) have shown how in such situations, presumably for the sake of computational simplification (e.g., Lieder & Griffiths, 2019), the initial estimate is often converted into categorical form for use in the next stage. In the current context that would mean assuming the blood test was correct this time when updating beliefs about how accurate the two tests are for these types of reports.

In fact, we have tentatively observed these cognitive processes in related work with different scenarios (Dewitt, Fenton, Liefgreen & Lagnado, 2020; Dewitt, Lagnado & Fenton, 2018). Dewitt et al (2018) presented participants with a scenario where two nations, X and Y, are the possible sources of a missile explosion where Y has a higher historical record of successfully exploding missiles (4/6 vs X’s 1/6). Participants were asked to update their estimates of the proficiency of X and Y at exploding missiles after the latest explosion. Dewitt et al (2020) presented participants with a modified version of the classic taxicab problem (Bar Hillel, 1980)[2], but where the witness’s accuracy at judging cab colours, rather than just being stated as 80%, was established by them being accurate 4 times out of 5 on a test. Furthermore, rather than focusing on participants’ estimates of the probability the cab involved in the hit and run scenario is blue, our focus was on their estimate of the witness’s accuracy after they report that the cab was blue. In each of these problems, as well as the current problem we

[1] Here we see that only at 50.0% do the Bayesian / LaPlacean and frequentist estimates completely match.

[2] A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data: 90% of the cabs in the city are Green and 10% are Blue. A witness identified the cab as Blue. The court tested the reliability of the witness under the circumstances that existed on the night of the accident and concluded that the witness correctly identified each of the two colours 80% of the time and failed 20% of the time. What is the probability that the cab involved in the accident was Blue rather than Green?

are asking participants to update their estimates of propensities in light of an ambiguous new instance.

Across both previous experiments we found two modal responses. First, around one third of participants appear to update based upon similar assumption-based or 'as if' reasoning as predicted by Gettys et al (1973). In Dewitt et al (2018) these participants increased their estimate of the proficiency of Y, leaving X unchanged (although normatively both should increase). In Dewitt et al (2020) these participants increased their estimate of the witness's accuracy, even though it should decrease (because their claim contradicts a strong base rate suggesting the cab is green). A second modal response also made by around one third of participants in each experiment was to make no change at all. In Dewitt et al (2018) this meant leaving both X and Y unchanged, and in Dewitt et al (2020) this meant making no change to the witness's accuracy.

These experiments in combination with the current represent the three basic causal structures: a simple chain (Dewitt et al [2020]), a 'common effect' (Dewitt et al [2018]: X and Y are both potential causes of the explosion) and now a 'common cause' (the patient's cancer status is a cause of both the blood and scan test reports). They also represent quite different scenarios across major domains of human endeavor: military reasoning, legal reasoning, and medical reasoning. Part of the purpose of the present experiment is therefore to add to the generalizability of this suite of experiments. This experiment also involves a more complex scenario, with the BN model having 15 nodes, rather than 8 (Dewitt et al [2018]) or 5 (Dewitt et al [2020]).

A further key aim of the present experiment is to test a theory that was developed over those two works about the relationship between the 'as if' response, and the 'no change' response. These two responses seem to represent two sides of a 'binary' choice: either update as if you were certain (e.g., that Y was responsible for the explosion / that the witness is correct this time) or update as if you have no information at all. What is lacking in these approaches is a graded / probabilistic approach. Indeed, we have theorized that these two response types in fact may have a similar representation of the problem, seeing these two approaches as the only two possible responses. The difference would then be how 'certain' they need to be (i.e., their certainty threshold) to make the 'as if' response. We have speculated that 'no change' responders have a higher threshold for this than 'as if' responders, and so withhold from making any update, seeing that as their only alternative.

As with previous experiments, our aim in the present is primarily to present participants with the current problem and observe their response patterns. Importantly, we are not interested in observing the magnitude of our participants estimates, as the mathematics requires sophisticated modelling, but instead are interested in their intuitions about the direction of change i.e., whether, at various stages they consider the tests to be more accurate than before (for the given type of report), less accurate, or the same. Observing their pattern of responses for both tests will allow us to determine if they are dealing with the problem in a categorical manner or in a graded / probabilistic manner. If, after being told the results of the two tests, but before being told if the patient has cancer or not, they operate categorically, assuming the blood test is correct (and therefore scan wrong), we expect them to increase their accuracy estimate of the blood test for positive reports and reduce their accuracy estimate of the scan test for negative reports. However, if they operate in a graded / probabilistic manner, avoiding assumptions, they will, like the model, decrease their accuracy estimates of both tests.

Finally, we will be providing our participants with a range of supplementary questions to give us as much data as possible on their cognitive processes, including most crucially, their certainty threshold for making the 'as if' response. In line with calls for more verbal protocol designs (McNair, 2015) we will also be asking our participants to explain their responses in open text boxes.

Method

Participants

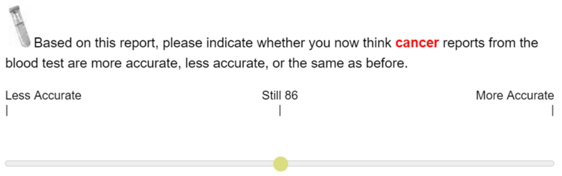
Participants (n = 225) were recruited from Prolific Academic and paid £9 per hour. Mean age was 28.8 (SD = 10.4), with a minimum of 18 and a maximum of 75.

Materials & Procedure

All materials and data can be found online at https://osf.io/ucg92/. Participants were presented with the scenario described in the introduction and given time with the numbers in Figure 1. They were then asked to make initial accuracy estimates using sliders of each test for each report type (Blood positive / negative; Scan positive / negative) based on that table to ensure comprehension.

Participants were then told about the new patient and the result of each test sequentially. After each result they were asked, as another comprehension check, to indicate any change in the patient's chance of having cancer.

Participants were then asked, in light of the two reports, to indicate on a sliding scale, whether their accuracy estimate for each test, for the type of report made (positive for the blood test, negative for the scan test) had increased, decreased, or not changed.



Based on this report, please indicate whether you now think cancer reports from the blood test are more accurate, less accurate, or the same as before.

Figure 6. The sliding mechanism for updating accuracy seen by participants

The original frequencies (Figure 1) for both tests were also provided on this page as a reminder. The participant's own initial estimate was also 'piped in' next to the word 'still'. We only recorded whether they increased, decreased or made no change. Participants could not just do nothing to be recorded as 'no change'. They had to move the slider to activate it and then put it back at 'Still X' for this to register.

Depending on which of the three response types participants made (Decrease; No change; Increase), they saw some identical, and some different follow up questions to probe their reasoning. Firstly, all participants were asked to indicate using a slider, their certainty that the blood test was correct on this occasion, and their confidence in that estimate. They were also asked 'When deciding how to update your accuracy estimates of the two tests, did you:'

1. Assume the blood test was correct this time
2. Assume the scan test was correct this time
3. Neither / Other [Plus accompanying open text box]

Following this, participants who made either no change, or increased their estimate of the blood test's accuracy for positive reports, were asked to provide the certainty threshold they would require for increasing their estimate. To provide this they were given a series of options from 50% to 100% in 5% intervals.

Participants who made no change to either test were asked an additional follow up question, requesting them to choose which of the following options most closely captured their reasoning:

1. Until I know for certain whether the blood test is correct this time (i.e. whether the patient really has cancer) it is incorrect to make any change to my accuracy estimates.
2. I saw my estimate of the test's accuracy as exactly equalling its true accuracy, which cannot change, whatever new information we get.
3. Although my accuracy estimate would change a little, it is a negligible change from only one extra observation.
4. Other [Plus accompanying open text box]

At the end of the experiment all participants were told that we eventually find out the patient really did not have cancer, contradicting the blood test's report. They were then asked again to update their accuracy estimate of the two tests for the report types made compared to their original (prior) estimate. This again acted as a comprehension check / comparison with their estimate under uncertainty.

Finally, the experiment was run in a second version with the figures for the blood and the scan test 'flipped' so that the scan test was the more historically accurate, to rule out possible confounding factors of world knowledge about these types of tests and between positive / negative reports. In the following results we combine both versions for analysis, referring to the test with the more reliable statistics (the blood test in the first version, the scan test in the second) as the more historically accurate (HA) test.

Results

Quantitative

In Table 1 the initial estimates provided by participants for each test (Blood / Scan) for each type of report (positive / negative) can be for both framings of the experiment (where blood was the more HA / where scan was the more HA). The reports made by the two tests in the scenario (positive for blood, negative for scan) are highlighted in grey.

Table 1. Participant estimates of the accuracy of the two tests when providing positive / negative reports for both framings of the experiment: [actual frequencies provided], **mean,** (standard deviation).

	Blood		**Scan**	
	Positive	**Negative**	**Positive**	**Negative**
Blood more HA	[6/7] **76.3** (21.3)	[3/3] **85.8** (25.5)	[3/3] **82.5** (26.9)	[4/7] **53.8** (18.1)
Scan more HA	[4/7] **56.8** (12.8)	[3/3] **84.8** (22.3)	[3/3] **86.1** (22.0)	[6/7] **80.2** (14.8)

Responses to the problem were coded according to how participants changed their accuracy estimates for both tests. Two major response types were found. The first was to make no change to either (NC-NC), and was made by 32.4% of the sample. The second, was to increase the more HA test, and reduce the other (INC-RED), and was made by 33.3% of the sample. No other response type was seen in more than 6% of participants (only 6 individuals reduced both), and the following analyses focus on understanding the cognitive processes lying behind NC-NC and INC-RED. In Table 2 we present the pattern of responses to a range of questions.

Table 2. Descriptive statistics for a range of questions divided by the two major response types

		NC-NC	**INC-RED**
	Total N (225)	73	75
Mean (SD)	**Certainty** more HA test is correct	74.7% (18.7)	76.3% (15.0)
	Confidence in above	66.0% (20.4)	72.8% (18.7)
	Certainty **threshold** for increasing accuracy estimate of more HA test	84.1% (12.4)	76.4% (12.1)
Proportion self-reporting…	Assumed more HA test correct	52.1%	81.3%
	Assumed less HA test correct	9.6%	6.7%
	Assumed neither / other	38.4%	12.0%

All following linear regressions use NC-NC vs INC-RED as independent variable (IV). Starting from the top of the table, three linear regressions were run with (1) 'Certainty' as dependent variable (DV) (74.7% vs 76.3%: B=1.6, F(1,146)=.334, p=.564), (2) 'Confidence' as DV (66.0% vs 72.8%: B=6.8, F(1,146)=4.5, p=.035) and (3) 'Threshold' as DV (84.1% vs 76.4%: B=7.7, F(1,146)=14.2, p<.001).

Three binary logistic regressions, again with NC-NC vs INC-RED as IV, were run to examine differences in terms of whether they self-reported (1) assuming the more HA test was correct (52.1% vs 81.3%: Wald(1) = 13.5, p<.001), (2) the less accurate test was correct (9.6% vs 6.7%: Wald(1) = .42, p=.517) and (3) whether they assumed neither / other (38.4% vs 12.0%: Wald(1) = 12.5, p<.001).

Participants making the NC-NC response were presented with a further multiple-choice question presenting three theorized cognitive processes as well as an 'other' option. Forty-four (60.3%) chose the option stating that it was incorrect to update accuracy estimates until we are certain whether the patient really has cancer or not. Fourteen (19.2%) chose the option stating that the accuracy rates were fixed and could not change, while a further 14 (19.2%) chose the option stating that the change was negligible. Only one individual chose 'other'.

Finally, when told the patient's result at the end of the study (which in both versions was in contradiction to the more HA test's report) 60.3% of NC-NC responders reduced their accuracy estimate for the more HA test and 56.2% increased their accuracy estimate for the less HA test.

Qualitative

INC-RED. Out of the 75 INC-RED individuals, 51 (68.0%) were coded as referencing the historical accuracy of the two tests in their explanation of their response (21 [28.0%] were coded as 'unclassified' as their reasoning could not be determined). For example, P62 said "[The] blood test when checking for cancer is more accurate. [The] scan test is less accurate in the case of a "clear" patient", P8 said, "The blood test is more precise than the scan test" and P10 said, "I think the blood tests are more accurate". The majority of these didn't explicitly state that they believed one or the other test was correct / incorrect this time. However some did, such as P12 "…the blood test has a better history of being accurate so it seems more likely that the blood test is right", and in the condition where the scan was more reliable, P216 said "Statistically the scan is less likely to be wrong on the diagnosis, that's why I think the blood test is wrong" and P152 said "Well the blood test was only right on 4 out of 7 positives. The scanner was pretty accurate (6/7) on clears so in this case I think the person is more likely to be clear than having cancer. So, the clear scanner makes me think it's a false positive on the blood test and that the blood test has 50% accuracy"

NC-NC. No single dominant code could be established for the 73 individuals who made no change (33 [45.2%] were coded as 'unclassified'), however some patterns of thinking emerged. Similar to INC-RED, 15 (20.5%) simply mentioned the historical accuracies e.g., P25 "I trust the blood test result more, because it has more accurate outcomes from the same number of trials.". However, nine individuals (12.3%) made clear that they couldn't change their estimates while we were unsure which test was correct e.g. P7 "We don't yet know whether the new participant has/will develop cancer or not, so we don't have any new data to influence the previous estimate.", P160 "We cannot change our predictions of the accuracies without knowing the result of this patient" and P162 "As I do not know if the patient does indeed have cancer, I can't update the probabilities of either of the tests being right or wrong." Four participants wanted more data e.g., P5 "More data is needed, I think, for me to be able to really adjust my estimates." Four participants simply stated that nothing had changed e.g., P111 "Nothing has changed as far as I can tell, and the statistics for the accuracy of the tests are still the same" and three saw the new information as irrelevant for updating propensities e.g., P21 "I don't see any change in the predictive nature of either test based on the circumstances."

Discussion

As we see in the model, the normative approach to the problem when both tests conflict with one another is to increase estimates of both their error rates i.e., to decrease estimates of their accuracy for the particular reports provided. Once that uncertainty is removed, and we know which test was correct, we would increase our estimate of the accuracy of that test, while decreasing our estimate of the accuracy of the one which was incorrect. However, while the outcome is uncertain, this is spread in a graded / probabilistic manner across the two tests according to how likely they are to be correct this time.

Our two majority responses, either making no change to the accuracy of either test (NC-NC), or to increase the accuracy of the more HA test while reducing the accuracy of the other (INC-RED) do not match the normative response. Instead, the INC-RED response matches the normative response when we know for certain that the more HA test was correct. We therefore suspect, in line with previous work, that these responders are using assumption-based thinking, acting 'as if' they knew that the more HA test was in fact was correct even while it is uncertain. Indeed, as can be seen in Table 1, 81.3% of INC-RED responders self-report as having made this assumption. While most responses from the qualitative data simply referenced the historical accuracy of the two tests, some explicit mention of this assumption was also seen there.

The NC-NC response makes a different error: it is the normative response if we had no information at all. Based on the pattern of responses to the follow up questions, we suspect that the NC-NC responders actually have a similar representation of the problem to INC-RED responders but are simply making a different choice of how to handle it. We can firstly see this in the large number of participants mentioning the historical accuracies of the two tests in the qualitative data. However, unlike INC-RED responders, many NC-NC responders say, even though they recognize that one test has been more accurate than the other, they cannot make change until they know for sure which one is correct. The NC-NC response may therefore be a 'conservative' choice of how to deal with uncertainty i.e., to wait until we know more and avoid updating based on assumptions. This fits with the fact that 60.3% of NC-NC responders chose the option stating that it was incorrect to make updates under uncertainty, and around 60.0% made the appropriate adjustment at the end of the study once they did know the patient's true status.

This general picture also fits with the quantitative data in table 1, demonstrating the difference in responses to a range of questions by the two approaches. Firstly, we can see that there is no difference in certainty: both were equally certain (around 75%) that the more HA test was correct on this occasion. There is a possible difference in confidence in this estimate, with INC-RED responders being more confident, but this should be considered tentative. Importantly however, there seems to be a clear difference (around 8%, p<.001) in the threshold that each response type reports they would require in order to increase their estimate in the more HA test. Notably, this threshold is lower among INC-RED responders and roughly matches their mean certainty (76.4% vs 76.3%), than among NC-NC responders, where it is considerably higher than their mean certainty (84.1% vs 74.7%). Connected to this, we can see that nearly 30% more INC-RED responders self-reported as 'assuming' the more HA test was correct. Correspondingly, around 25% more NC-NC responders reported not making any assumptions.

This tentatively presents a picture where both NC-NC and INC-RED responders have a similar representation of the problem. Both think the more HA test is likely to be correct this time, and the INC-RED responders are prepared to update accuracy estimates based upon this assumption (their certainty threshold has been met) while NC-NC responders are not prepared to, and prefer to make no change, either waiting for certainty, or for more data in general (until their certainty threshold is met). The two responses therefore seem to represent two sides of a 'binary' (or digital / categorical) choice – either make no change, as if we had no information at all, or make a change as if we were certain that the more HA test was correct. What is lacking in these responses representing two thirds of our sample therefore is a 'probabilistic' or 'graded' response to this problem.

While we do not know why participants make these responses, a general framework of 'resource rationality' is plausible, as the approach certainly simplifies the problem. However, as we have mentioned in previous work (Dewitt et al., 2018; Dewitt et al., 2020), the INC-RED 'simplification' of the problem produces a belief updating dynamic similar to that seen in the confirmation bias literature, where, because e.g., the blood test has been more accurate in the past, we update our beliefs based on the assumption that it is correct this time, producing a self-reinforcing dynamic. In each of our three experiments these assumptions are also based on a very small sample size. It would be interesting to see how this updating process is handled over multiple iterations i.e., if we saw several more patients with conflicting results each time. Participants may compensate e.g., after several times of assuming the blood test is correct, they may 'give one' to the scan test, in proportion to the priors. So, while they may not be using a graded approach within one individual instance, it is still possible that they may do so over time.

References

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. Acta Psychologica, 44(3052), 211–233.

Casscells, W., Schoenberger, A., & Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. The New England Journal of Medicine, 299(18), 999–1001.

Cook, J. (2016). Rational Irrationality: Modeling Climate Change Belief Polarization Using Bayesian Networks. 8, 160–179.

Dewitt, S. H., Fenton, N. E., Liefgreen, A., & Lagnado, D. A. (2020). Propensities and second order uncertainty: a modified taxi cab problem. Frontiers in Psychology, 11.

Dewitt, S. H., Lagnado, D., & Fenton, N. (2018). Updating prior beliefs based on ambiguous evidence. In COGSCI2018: Changing Minds (Vol. 40, pp. 2047-2052). Cognitive Science Society.

Gettys, C. F., Kelly, C., & Peterson, C. R. (1973). The best guess hypothesis in multistage inference. Organizational Behavior and Human Performance, 10(3), 364–373.

Gigerenzer, G., & Hoffrage, U. (1995). How to Improve Bayesian Reasoning Without Instruction: Frequency Formats. Psychological Review, 102(4), 684–704.

Johnson, Samuel G. B., Merchant, T., Keil, Frank C. (2020) Journal of Experimental Psychology: General, Vol 149(8), 1417-1434.

Kleiter, G. D. (2018). Imprecise Uncertain Reasoning: A Distributional Approach. 9(October), 1–16.

Laplace, Pierre-Simon (1814). Essai philosophique sur les probabilités. Paris: Courcier.

Lieder, F., & Griffiths, T. L. (2018). Resource-rational analysis: understanding human cognition as the optimal use of limited. Psychological Review, 85(4), 249-277.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. Journal of Personality and Social Psychology, 37(11), 2098–2109.

McNair, S. J. (2015). Beyond the status-quo: research on Bayesian reasoning must develop in both theory and method. Frontiers in Psychology, 6(February), 1–3.

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. Review of General Psychology, 2(2), 175–220.

Rabin, M., & Shrag, J. L. (1999). First Impressions Matter: A Model of Confirmatory Bias. 114(1), 37–82.