Mem Cogn (2017) 45:270–280
DOI 10.3758/s13421-016-0658-z

Physician Bayesian updating from personal beliefs about the base rate and likelihood ratio

Benjamin Margolin Rottman [1]

Published online: 17 October 2016

Abstract Whether humans can accurately make decisions in line with Bayes' rule has been one of the most important yet contentious topics in cognitive psychology. Though a number of paradigms have been used for studying Bayesian updating, rarely have subjects been allowed to use their own preexisting beliefs about the prior and the likelihood. A study is reported in which physicians judged the posttest probability of a diagnosis for a patient vignette after receiving a test result, and the physicians' posttest judgments were compared to the normative posttest calculated from their own beliefs in the sensitivity and false positive rate of the test (likelihood ratio) and prior probability of the diagnosis. On the one hand, the posttest judgments were strongly related to the physicians' beliefs about both the prior probability as well as the likelihood ratio, and the priors were used considerably more strongly than in previous research. On the other hand, both the prior and the likelihoods were still not used quite as much as they should have been, and there was evidence of other nonnormative aspects to the updating, such as updating independent of the likelihood beliefs. By focusing on how physicians use their own prior beliefs for Bayesian updating, this study provides insight into how well experts perform probabilistic inference in settings in which they rely upon their own prior beliefs rather than experimenter-provided cues. It suggests that there is reason to be optimistic about experts' abilities, but that there is still considerable need for improvement.

Keywords Bayesian reasoning · Probabilistic reasoning · Diagnosis

Whether humans can accurately make decisions in line with Bayes' rule has been one of the most important yet contentious topics in cognitive psychology (Barbey & Sloman, 2007; Koehler, 1996). Initial research, such as the famous mammogram task in which participants are told the result of a mammogram, the pretest probability of cancer, and the sensitivity and false positive rate of the test (likelihood ratio), has found very poor judgments, even among doctors (Casscells, Schoenberger, & Graboys, 1978; Eddy, 1982). One common mistake involves substituting the sensitivity, the probability of a positive test result given that the patient has the disease, for the posttest judgment, the probability that the patient has the disease given a positive test result. More broadly, human Bayesian reasoning has often been described as insufficiently sensitive to base rates, or even "neglectful" of base rates.

Subsequent work has found that a number of factors can lead to improved Bayesian reasoning (Gigerenzer & Hoffrage, 1995; Krynski & Tenenbaum, 2007). In particular, one setting that can lead to improved posterior probability judgments is when individuals experience the contingency between the two variables in question (e.g., mammogram result and having cancer) in a trial-by-trial learning paradigm (Christensen-Szalanski & Beach, 1982; Edgell, Harbison, Neace, Nahinsky, & Lajoie, 2004).

However, many cases of real-world reasoning match neither the "word problem" paradigm, in which the base rate and likelihood ratio are provided through textual instructions, nor the trial-by-trial learning paradigm, in which participants experience the contingency between the two variables of interest

✉ Benjamin Margolin Rottman
rottman@pitt.edu

[1] Department of Psychology, University of Pittsburgh, LRDC 726, 3939 O'Hara St., Pittsburgh, PA 15260, USA

within a short time frame. In many real-world cases, reasoners make posterior probability judgments based on existing beliefs about the base rate and likelihood ratio, and these beliefs may be informed by personal experiences with the probabilities over long periods of time, as well as by other sources of knowledge, such as socially communicated beliefs or instruction. For example, when interpreting the results of a diagnostic test, a physician has personal experience with the base rate of a disease and the sensitivity and false positive rate of the test, and the physician has likely read published estimates of these values. The question addressed in this study is how well individuals' own beliefs about of the base rate and likelihood ratio, developed from their prior experience, correspond to their judgments of posterior likelihood.

This goal has only been addressed a couple of times. In one study, participants were explicitly instructed about the base rate and used their own knowledge about the likelihood ratio, or vice versa (Evans, Handley, Over, & Perham, 2002). This study used a particular analytical technique that analyzes the influence of the logged prior beliefs and logged likelihood ratio belief on logged posterior judgment. Perfect Bayesian updating would result in regression weights for the prior and likelihood of 1; regression weights higher or lower than 1 reflect over-use versus under-use of the belief. The results of this study were complex. The regression weights on the log prior odds ranged from near zero (no influence at all) to .43, a significant influence, though still considerably less than optimal regression weight of 1. The regression weights for the log likelihood ratio ranged from .19 (underuse) all the way up to 1.97 (overuse). Perhaps the most interesting finding involved an experiment in which participants used their own beliefs for both the prior and likelihoods; the priors were not used at all (.03) and the likelihoods were used approximately normatively (.88). Overall, personal beliefs tended to dominate statistically presented information, which emphasizes the need for additional research into the use of personal beliefs in Bayesian reasoning.

Two other psychology studies investigated a similar phenomenon framed in terms of logical deduction (Evans, Thompson, & Over, 2015; Singmann, Klauer, & Over, 2014). However, unlike the study by Evans et al. (2002), which examined the relations between participants' beliefs about the base rate, likelihood ratio, and posterior, in these two studies not all these beliefs were assessed, so a single normative point estimate for the posterior could not be calculated from the other beliefs. Instead those studies assessed whether participants' posterior judgments fell within a "coherent" range that was probabilistically consistent with the other beliefs.

One of the most obvious instances of the need for Bayesian reasoning is within medical diagnosis—updating the believed probability of a disease after learning about a new symptom or a new test result. Only a couple studies have investigated Bayesian reasoning from personal beliefs in medical professionals.[1] Two studies found that posttest judgments were overestimated after a positive test result, relative to the posterior probability calculated by applying Bayes' rule to subjects' own beliefs about the likelihood and prior (Lyman & Balducci, 1993; Noguchi, Matsui, Imura, Kiyota, & Fukui, 2002). When the test result was negative, one study found that judgments were too low (Noguchi et al., 2002) and two found that the inferences were approximately normative (Bergus, Chapman, Gjerde, & Elstein, 1995; Lyman & Balducci, 1993). Unfortunately, because these studies did not use the analytical technique of Evans and colleagues (2002), it is impossible to know the reasons for the biased judgments, such as over-use versus under-use of the prior and/or likelihood beliefs.

This study tested whether physicians' judgments about the posttest probability of colorectal cancer in a clinical vignette cohere with their own beliefs about the pretest probability, sensitivity, and false positive rate of the test. The results analyzed whether the physicians are adequately sensitive to their own beliefs about the likelihood ratio and the base rate, which can provide insight into how experts with well-formed beliefs perform Bayesian updating.

Method

Participants

Residents and attending physicians in internal medicine and family medicine were recruited. Recruitment e-mails were sent to colleagues at our home and outside institutions, and we asked them to forward the e-mails to appropriate e-mail lists. Sixty-five attending physicians and 149 residents completed the study. One hundred and seven participants were male, 100 were female, and seven declined to specify. Five were Hispanic or Latino, and 23 did not specify whether or not they were Hispanic or Latino. In terms of nonmutually exclusive racial categories defined by the National Institutes of

[1] The most well-known study about Bayesian reasoning in physicians (Christensen-Szalanski & Bushyhead, 1981) unfortunately does not provide much insight into this issue. In this study, physicians were asked about the predictive value of various symptoms for pneumonia (e.g., the probability of pneumonia given that a patient has crackles during breathing), and their estimates were correlated with the actual objective predictive value in the patient population. Though this finding has often been cited as evidence that physicians are sensitive to the base rates of diseases that they experience, this conclusion is not justified for two reasons. First, the objective predictive value of the symptoms were calculated from patient data for an entire practice, not the patients that an individual physician treated, limiting the ability to conclude that the physicians were sensitive to their own experience. Second, making an estimate of predictive value in and of itself does not require the use of base rates (Kleiter et al., 1997).

Health of the United States, 135 identified as White, 60 as Asian, two as Black or African American, one as American Indian, and 19 did not answer. They were affiliated with 18 hospital systems; 80 % were affiliated with four primary hospital systems: 60 participants were from the University of Washington (Seattle) health network, 53 from the University of Chicago, 46 from Harvard, 14 from the University of Denver, and the rest were from other institutions with fewer than 10 participants per institution or did not provide information about their institution. Of the 65 physicians who had completed residency, they finished residency mean = 10, median = 8, standard deviation = 9 years before the study. Out of the 149 residents, 44 were in their third year, 42 were in their second year, and 63 were in their first year. Participants were paid $10 in an Amazon gift card or through an online money transfer.

Materials and design

The design of the experiment was 3 vignettes × 4 diagnostic tests × 2 test outcomes, entirely within subjects. The three vignettes were intended to convey a low, medium, or high pretest likelihood of colorectal cancer. The vignettes were all based around the same core case—the medium and high likelihood vignettes incorporated additional signs of colorectal cancer.[2] The low vignette was as follows:

> A 53-year-old woman comes to an outpatient clinic after requesting an urgent appointment. She has alternating constipation and diarrhea for 6 weeks and came in today because of severe abdominal pain. Her last bowel movement was 4 days ago, and she has not noticed bloody stool. CBC and iron studies are normal.
>
> The medium vignette added the following symptoms:
>
> She has experienced poor appetite and weight loss for the last 6 months.
>
> The high vignette added the following symptoms:
>
> She has had a narrowed diameter of stool for 3 months, and she has a family history of colorectal cancer.

For each vignette participants judged the pretest likelihood and made eight posttest likelihood judgments corresponding to positive and negative test results for four tests: fecal occult blood test (FOBT; guaiac, nonrehydrated stool sample), colonoscopy, sigmoidoscopy, and virtual colonoscopy (CT).

Procedure

The experiment was conducted online. Participants worked through the three vignettes in the order of low, medium, and high. Within each vignette they first made the pretest judgment, the likelihood that the patient has colorectal cancer, with the following question: What is your estimate of the likelihood that the patient has colorectal cancer on a scale from .01 %–99.99 %?

Then they made eight posttest judgments using the same scale (see Fig. 1 for the exact wording). Each judgment regarded one of the four tests with either a positive or negative result. To develop a common definition of a positive result across the three imaging studies, a positive result was defined as "any polyp(s) or other suspicious lesions(s)" and a negative result was defined as "no polyps or other suspicious lesions."

After finishing the three vignettes, participants reported their beliefs about the sensitivity and false positive rate for the four tests (see Fig. 2 for details). Sensitivity was defined as, "Suppose that there are 100 patients who *have colorectal cancer*. For how many of these patients would the test correctly detect the colorectal cancer?" The false positive rate was defined as, "Suppose that there are 100 patients who *do not* have colorectal cancer. For how many of these patients would the test be *falsely positive*?"

Results

The raw data and R code to reproduce the analyses can be found at https://osf.io/3t2xw/.

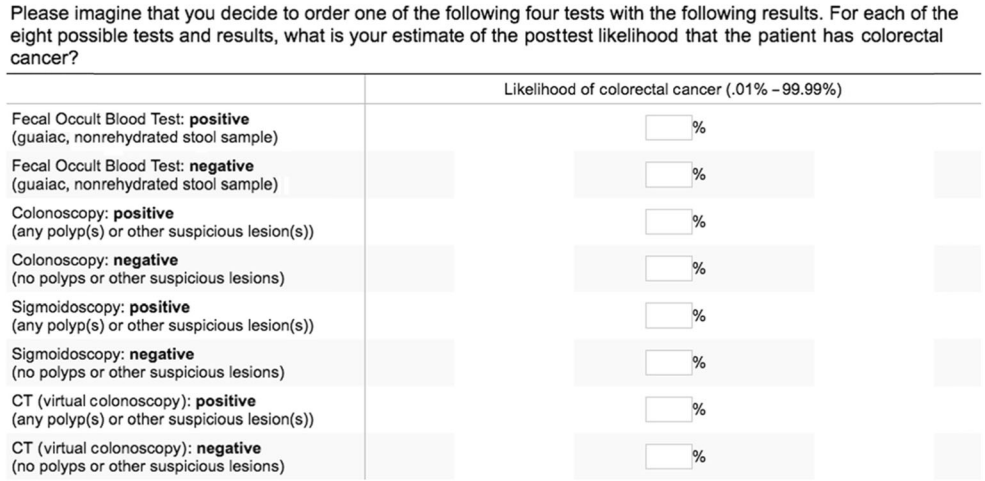
Descriptive results

On average, participants judged the three vignettes to have a 10 % (*SD* = 15 %), 29 % (*SD* = 23 %), and 53 % (*SD* = 27 %) pretest probability of colorectal cancer. This provides fairly wide range of prior probabilities from which to examine updating.

Table 1 presents the median judgments of the sensitivity and false positive rate for the four tests along with published estimates.[3] Overall, participants' judgments matched fairly closely with the published estimates.

[2] Participants reasoned about the three vignettes in the order of low, medium, and high, which introduces the possibility of an order effect. However, because there was not an obvious hypothesis for how such an order effect would influence the results of the study, and because using this increasing order was clearest for participants, a decision was made not to randomize the order of the three vignettes.

[3] To match the clinical vignette, the published estimates are for detecting colorectal cancer, not polyps, in symptomatic patients, when available. FOBT includes studies on Hemoccult II and Hemoccult Sensa. I could not find research on the false positive rate of a colonoscopy before biopsy; the false positive rate after biopsy should be negligible.



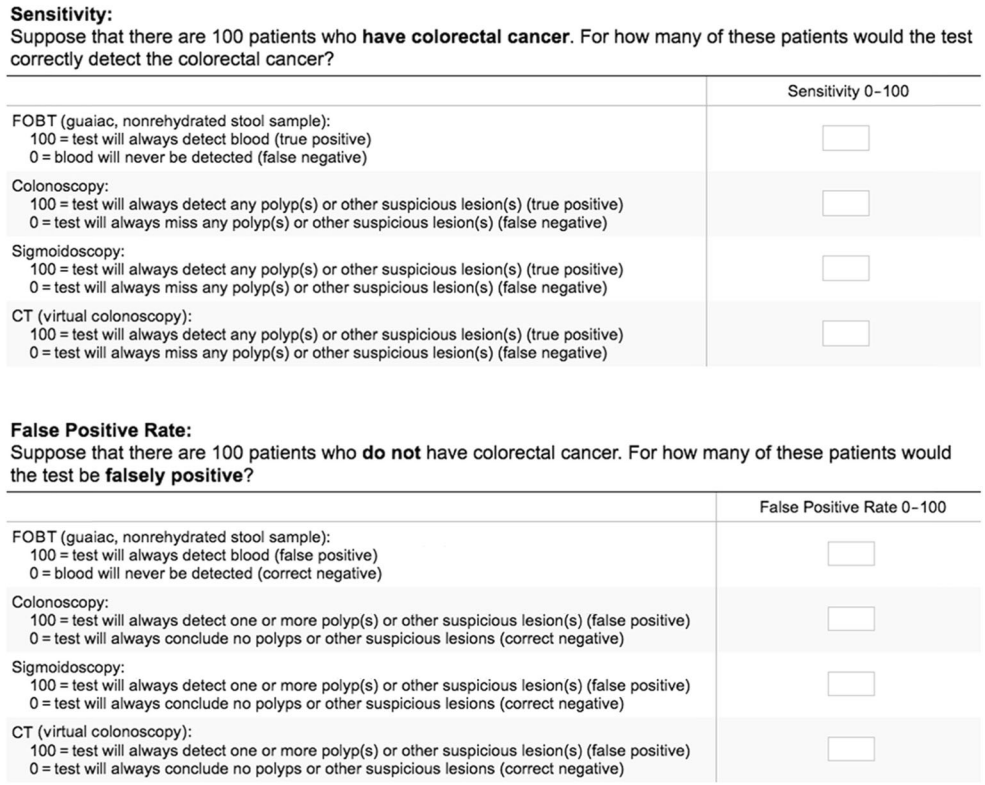
Please imagine that you decide to order one of the following four tests with the following results. For each of the eight possible tests and results, what is your estimate of the posttest likelihood that the patient has colorectal cancer?

	Likelihood of colorectal cancer (.01% – 99.99%)
Fecal Occult Blood Test: **positive** (guaiac, nonrehydrated stool sample)	%
Fecal Occult Blood Test: **negative** (guaiac, nonrehydrated stool sample)	%
Colonoscopy: **positive** (any polyp(s) or other suspicious lesion(s))	%
Colonoscopy: **negative** (no polyps or other suspicious lesions)	%
Sigmoidoscopy: **positive** (any polyp(s) or other suspicious lesion(s))	%
Sigmoidoscopy: **negative** (no polyps or other suspicious lesions)	%
CT (virtual colonoscopy): **positive** (any polyp(s) or other suspicious lesion(s))	%
CT (virtual colonoscopy): **negative** (no polyps or other suspicious lesions)	%

Fig. 1 Posttest questions for each of the four tests and two test results

Excluded observations

The following observations (individual posttest judgments) were excluded to aid interpretability; "t^+" versus "t^-" represents whether the test result is positive versus negative, and "c^+" versus "c^-" represent whether colorectal cancer is present versus absent. First, observations were dropped if a participant's sensitivity estimate, $P(t^+|c^+)$, was less than the false positive rate, $P(t^+|c^-)$, for a given test, which means that the test works in the wrong direction, and likely reflects an error (4.7 % of the observations). Second, observations were also dropped if participants judged that $P(t^+|c^+) = P(t^+|c^-)$ (5.6 % of the observations), which means that the test cannot discriminate colorectal cancer at all. This mainly occurred for the FOBT, which is a test that is widely viewed with skepticism by physicians because of its high false positive rate. Because the purpose of this study is to understand how physicians update their beliefs when updating is warranted, these observations were dropped. Third, there were 15 observations (0.3 %) for which participants gave sensitivity or false positive rates of exactly one or zero, which produced infinite log error values and had to be dropped for all analyses.



Sensitivity:
Suppose that there are 100 patients who **have colorectal cancer**. For how many of these patients would the test correctly detect the colorectal cancer?

	Sensitivity 0–100
FOBT (guaiac, nonrehydrated stool sample): 100 = test will always detect blood (true positive) 0 = blood will never be detected (false negative)	
Colonoscopy: 100 = test will always detect any polyp(s) or other suspicious lesion(s) (true positive) 0 = test will always miss any polyp(s) or other suspicious lesion(s) (false negative)	
Sigmoidoscopy: 100 = test will always detect any polyp(s) or other suspicious lesion(s) (true positive) 0 = test will always miss any polyp(s) or other suspicious lesion(s) (false negative)	
CT (virtual colonoscopy): 100 = test will always detect any polyp(s) or other suspicious lesion(s) (true positive) 0 = test will always miss any polyp(s) or other suspicious lesion(s) (false negative)	

False Positive Rate:
Suppose that there are 100 patients who **do not** have colorectal cancer. For how many of these patients would the test be **falsely positive**?

	False Positive Rate 0–100
FOBT (guaiac, nonrehydrated stool sample): 100 = test will always detect blood (false positive) 0 = blood will never be detected (correct negative)	
Colonoscopy: 100 = test will always detect one or more polyp(s) or other suspicious lesion(s) (false positive) 0 = test will always conclude no polyps or other suspicious lesions (correct negative)	
Sigmoidoscopy: 100 = test will always detect one or more polyp(s) or other suspicious lesion(s) (false positive) 0 = test will always conclude no polyps or other suspicious lesions (correct negative)	
CT (virtual colonoscopy): 100 = test will always detect one or more polyp(s) or other suspicious lesion(s) (false positive) 0 = test will always conclude no polyps or other suspicious lesions (correct negative)	

Fig. 2 Sensitivity and false positive rate judgments

Table 1 Comparison of sensitivity and false positive rates judged by the participants and from published research

Test	Sensitivity		False positive rate	
	Median Judgment	Research estimate	Median Judgment	Research estimate
FOBT (guaiac, nonrehydrated)	0.70	.69 (Niv & Sperber, 1995) .75 (Bjerregaard, Tøttrup, Sørensen, & Laurberg, 2009) .40–.70 (Zauber, Lansdorp-Vogelaar, Knudsen, Wilschut, van Ballegooijen, & Kuntz, 2008)	0.25	.27 (Niv & Sperber, 1995) .21 (Bjerregaard et al., 2009) .02–.08 (Zauber et al., 2008)
Colonoscopy	0.91	.95 (Zauber et al., 2008)	0.10	-
Sigmoidoscopy	0.75	.61–.72 (Zauber et al., 2008) .69 (Castiglione, Ciatto, Mazzotta, & Grazzini, 1995)	0.10	-
Virtual colonoscopy(CT)	0.85	0.90 (Johnson, Mei-Hsiu Chen, Toledano, Heiken, Dachman, Kuo, & Limburg, 2008)	0.15	0.14 (Johnson et al., 2008)

The analyses were also run excluding and including the following types of observations to test for the robustness of the effect. First, if a test result is positive (negative), the posttest judgment should be higher (lower) than the pretest judgment; 5.2 % of observations violated the normative direction of updating. Second, even though participants used a scale from 0.01 % to 99.99 %, some participants might have used 1 % to mean 1 % or less and 99 % to mean 99 % or more; 13.7 % of the observations normatively should have been <1 % or >99 %, and analyses are conducted with and without these observations.

Use of the likelihood ratio and the prior probability in updating

Analyzing whether participants' posttest judgments were sufficiently sensitive to the likelihood ratio (sensitivity divided by the false positive rate) and prior probability involved using the log odds form of Bayes' rule (Evans et al., 2002; Keren & Thujs, 1996; Lyman & Balducci, 1994). Equations 1a and 1b are used when a test result is positive or negative, respectively. The log odds form of Bayes' rule is useful because it permits the linear regression in Equation 2. Normatively, the regression weights for the log prior odds and the log likelihood ratio should be one, and the regression weight for the intercept should be zero. Because of the repeated measures, by-subject random effects on the intercept and the slopes of the log likelihood ratio and the log prior odds were included.

$$\log\left(\frac{P(c^+|t^+)}{1-P(c^+|t^+)}\right) = \log\left(\frac{P(t^+|c^+)}{P(t^+|c^-)}\right) + \log\left(\frac{P(c^+)}{1-P(c^+)}\right) \quad (1a)$$

$$\log\left(\frac{P(c^+|t^-)}{1-P(c^+|t^-)}\right) = \log\left(\frac{1-P(t^+|c^+)}{1-P(t^+|c^-)}\right) + \log\left(\frac{P(c^+)}{1-P(c^+)}\right) \quad (1b)$$

$$\log(\text{posttest odds}) = b_0 + b_1\cdot\log(\text{likelihood ratio}) + b_2\cdot\log(\text{pretest odds}) \quad (2)$$

Table 2 shows the results of the regressions; 95 % confidence intervals are provided to simultaneously compare the regression weights against both one and zero. Three regressions were run for both the positive and negative test results: one with all the exclusions mentioned in the previous section, one with all the exclusions except not excluding observations in which participants updated their judgments in the direction opposite to the normative direction, and one with all the exclusions except not excluding observations for which the normative answers are in the extreme parts of the scale (<.01 or >.99).

The fact that the regression weights for the likelihood ratio and the pretest odds were both significantly above zero for all six regressions implies that both beliefs were used. In all six regressions, the 95 % confidence interval for the log prior odds were below one, implying that participants did not use their beliefs in the prior sufficiently. Furthermore, in most of the regressions, the upper bound of the 95 % confidence interval for the log likelihood ratio was less than one, implying that participants did not use their beliefs in the likelihood (sensitivity and false positive rates of the test) quite enough. This was most obvious for a negative test result; however, after a positive test result, the 95 % confidence intervals were close to one and sometimes crossed over one, implying that participants nearly normatively used their likelihood beliefs after a positive test result.

The regressions for the most part had positive (negative) intercepts for the case of a positive (negative) test result. This can be interpreted to mean that the participants understood that they should increase (decrease) their judgments, but did not always understand that the updating should be tied directly to their beliefs in the likelihood ratio.

Table 2 Regression weights and 95 % confidence intervals of regressions testing for sensitivity to the log likelihood ratio and log pretest odds using different exclusions

Exclusions	*n*	Intercept	Log pretest odds	Log likelihood ratio
Positive test result				
All	2,163	.52, [.24, .81]	.66, [.61, .71]	.83, [.67, .997]
All but updating in the wrong direction	2,192	.46, [.18, .75]	.66, [.61, .71]	.85, [.69, 1.02]
All but normative inference >.99	2,259	.54, [.26, .82]	.65, [.61, .70]	.82, [.66, .98]
Negative test result				
All	1,636	-.18, [-.36, -.007]	.88, [.82, .94]	.75, [.65, .86]
All but updating in the wrong direction	1,738	-.10, [-.29, .09]	.81, [.75, .88]	.77, [.66, .88]
All but normative inference <.01	2,112	-.21, [-.37, -.05]	.88, [.83, .93]	.72, [.63, .81]

Note. n is number of observations

Follow-up regressions were run, testing whether there was a significant interaction between the log likelihood ratio and the log prior odds; normatively there should be no interaction. These regressions also included a by-subject random slope on the interaction. For the positive direction, the effect of the log likelihood ratio was weaker at higher pretest levels (see Table 3). This can be interpreted to mean that when the prior probability was fairly high, subjects refused to increase the probability as much as was warranted by their beliefs in the likelihood ratio, avoiding the upper end of the probability scale.

For the negative direction, the regressions would not converge; however, dropping the correlation between the random slopes permitted convergence in two out of the three regressions.[4] The effect of the log likelihood ratio was weaker at lower pretest levels (Table 3). This can be interpreted as subjects refusing to decrease the probability as much as warranted by their beliefs in the likelihood ratio when the prior probability was low.

In sum, though the use of the log prior odds and the log likelihood ratio were not quite normative, and there is evidence of a nonnormative interaction between the two, there is clear evidence that subjects used both the log prior odds and the log likelihood ratio.

Overestimation

Most studies on Bayesian reasoning have focused on whether the posterior judgments are close to the normative answer or whether they are too high or too low. Equation 3 was used to calculate the logged error in the posttest judgment (relative to a participant's own beliefs about the likelihood ratio and pretest odds), and analyses on the log error term were performed using the same three sets of exclusions used previously. Normatively, the logged error should be zero.

Overall, 63 % to 64 % of the inferences in the positive direction and 66 % to 69 % of inferences in the negative direction were above the normative calculation, depending on the particular set of exclusions. Regressions with a by-subject random intercept term found that the overestimation was significant for both the positive and negative test results (see Table 4). Follow-up regressions including a fixed effect for whether the physician was still a resident or had finished residency did not find significant effects.

$$\log(\text{posttest odds}) = \log(\text{likelihood ratio}) + \log(\text{pretest odds}) + \log(\text{error}) \quad (3)$$

Figures 3 and 4 show the relationship between the normative and actual posttest inferences when the test result was positive and negative. (Jitter was added to reduce overplotting.) The 45 degree line reflects perfect calibration. The thick black line shows the median inference at each level of the *x*-axis. The line connects 10 points, each of which

Table 3 Interaction results between log pretest odds and log likelihood ratio

Exclusions	*n*	Interaction
Positive test result		
All	2,163	-.11, [-.15, -.07]
All but updating in the wrong direction	2,192	-.10, [-.14, -.06]
All but normative inference >.99	2,259	-.10, [-.14, -.07]
Negative test result		
All	1,636	.09, [.05, .14]
All but updating in the wrong direction	1,738	.09, [.05, .14]
All but normative inference <.01	2,112	*

Note. n is number of observations. *Model would not converger

[4] The third regression, which included observations for which the normative value was <.01, was the one that would not converge. These inferences, when the extreme values were very low, are examined in the next section.

Table 4 Overestimation results

Exclusions	*n*	log(error)	% log (error) >0
Positive test result			
All	2163	.34, [.26, .42]	64 %
All but updating in the wrong direction	2192	.33, [.25, .41]	63 %
All but normative inference >.99	2259	.35, [.26, .43]	63 %
Negative test result			
All	1636	.09, [.03, .14]	66 %
All but updating in the wrong direction	1738	.14, [.08, .20]	68 %
All but normative inference <.01	2112	.15, [.08, .21]	69 %

Note. n is number of observations

represents a median containing one tenth of the data along the *x*-axis. The fact that the black line tends to be above the 45 degree line represents the overestimation.

Focusing on the inferences normatively less than 0.01, after the primary three reasons for excluding observations, 81 % of the remaining 590 inferences normatively less than 0.01 were too high. This likely reflects both a tendency to round to 0.01 and an overestimation bias. Sixty-one percent of the 70 inferences normatively greater than 0.99 were too *low*, suggesting that the overestimation bias does not extend to the very top of the scale, which can also be seen in Fig. 3.

The overestimation can be explained in terms of misuse of the components of Bayes' rule in the following ways. In the positive condition, there was a significant positive intercept in the regression analyses in the previous section, which contributes to overestimation. Additionally, the log pretest odds were generally negative (because most of the pretest probability judgments were less than .5); underuse of the log pretest odds (regression weight less than one) would thus inflate the inferences (bring them closer to .5).

The overestimation after a negative test result can be explained in the following way. First, the log likelihood ratio is negative for a negative test result. Underuse of the log likelihood ratio (regression weight less than one) implies that the participants did not reduce their estimates enough, resulting in overestimation. Second, the log pretest odds were generally negative, because most of the pretest probability judgments were less than .5. Underuse of the pretest odds would also inflate the inferences by bringing the posttest odds closer to zero (posttest probability closer to .5). Third, the intercept was negative, which suggests that participants tended to reduce the posttest estimates judgments in a way that is unassociated with

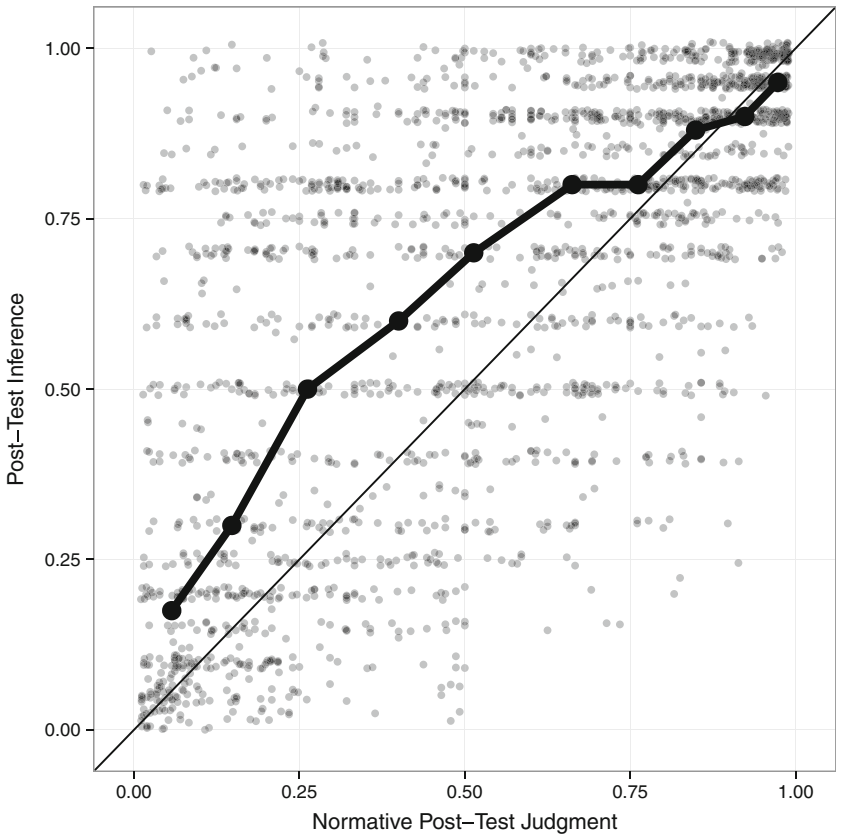


Fig. 3 Relationship between normative posttest calculations and posttest inferences after a positive test result

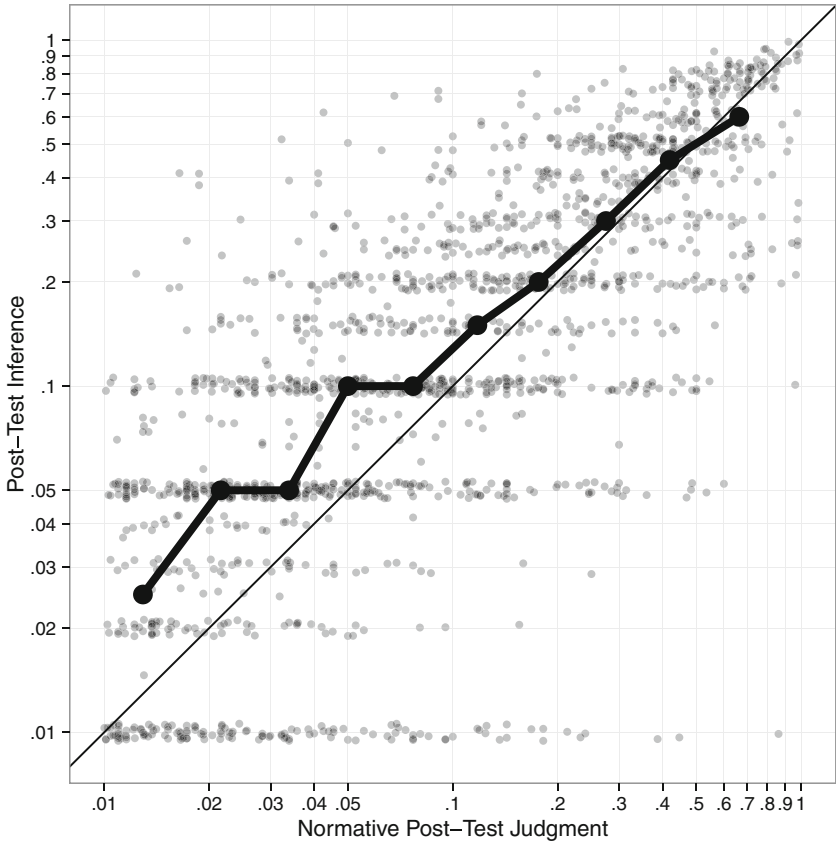


Fig. 4 Relationship between normative posttest calculations and posttest inferences after a negative test result. *Note*. This figure uses a log-log scale because many of the judgments are below .1

their belief in the likelihood ratio. All by itself, this would cause the inferences to be too low, but the underweighting of the likelihood ratio and the log prior odds had a bigger positive influence than the negative influence from the general tendency to decrease the estimates.

General discussion

Physicians' posttest judgments of colorectal cancer were compared to a normative calculation based on their own beliefs about the pretest probability and about the sensitivity and false positive rate of the test (likelihood ratio). On the one hand, the posterior judgments were nonnormative in a number of ways: the likelihood ratios and prior probabilities were not sufficiently used, there is evidence that subjects updated in a way somewhat independent from the likelihood ratio, and there was a nonnormative interaction between the beliefs about the likelihood ratio and the priors.

On the other hand, compared to the long history of findings of base rate underuse, use of the base rate beliefs were surprisingly accurate. Even in the study by Evans et al. (2002), subjects' use of their own base rate beliefs was much lower (.03 to .43, depending on the study) than in this study. Furthermore, the degree of use of the priors and the likelihoods were quite similar. Overall, there was a strong correlation between the inferred log posttest odds and the normative log posttest odds, $r^2 = 0.59$. In sum, this study presents some of the most optimistic data on the human ability for Bayesian reasoning in the literature.

There are three likely reasons why updating was fairly accurate in this study. First, this study examined updating based on participants' own beliefs about the prior rather than experimentally provided numbers (Christensen-Szalanski & Beach, 1982; Evans et al., 2002). Second, this study tested experts; it is possible that having extensive experience and knowledge could lead to stronger use of that knowledge. Third, in this study, subjects judged the prior before judging the posterior. Evans et al. (2002) found that use of priors can be improved by having subjects judge the prior before judging the posterior; however, in that study the use of the base rates was still much lower than in this study.

There are two limitations of this study. One weakness is that all subjects reasoned about the three vignette cases in the order of lowest to highest pretest probability. This feature of

the design was chosen to reduce the possible confusion of working with multiple similar cases, but has the weakness of being susceptible to order effects.

A second weakness is that the physicians might have had other beliefs about the uncertainty of a test result (e.g., that there is some low but nonzero probability that the test result was mistakenly reported for a different patient). How would beliefs about this sort of uncertainty influence the current study? One possibility is that when the participants reported their beliefs about the sensitivity and false positive rate of the tests, they incorporated this sort of uncertainty into those estimates. Doing so would result in a lower estimate of the sensitivity of the test and a higher false positive rate, capturing the fact that the test in realistic situations does not perform as well as it could in ideal situations. Incorporating such beliefs into the estimates of the sensitivity, false positive rate, and posttest would not be a problem for this study. Another possibility is that participants did not incorporate such uncertainty beliefs into their estimates of the sensitivity and false positive rate, but did incorporate such uncertainty beliefs into the posttest judgment. In this case, the normative way to calculate the posttest judgment would be to use Jeffrey conditionalization rather than Bayes' rule—this could account for a small amount of deviation from the normative calculations in the current study, making subjects' judgments appear less normative than they really are (Hadjichristidis, Sloman, & Over, 2014; Shafer, 1981; Talbott, 2015; Zhao & Osherson, 2010).

One open question is whether the current findings will translate into other settings. The overestimation findings can potentially be explained, in part, through a bias particular to the medical domain—a bias against "ruling out" or "missing" a potentially life-threatening diagnosis, and consequently inflating the posttest judgment. However, this explanation is not entirely convincing because what these results really show is an overestimation in the posttest judgment relative to the individual's own beliefs in the pretest probability and the likelihood ratio. (A bias against ruling out a disease could have been reflected in the prior and/or the likelihood rather than an inconsistency in how they are combined to form the posterior.) Furthermore, the simplest version of an overestimation bias would appear as a positive intercept in the regression analyses. A positive intercept explains some of the bias in the positive condition, but some of the bias in both conditions is explained by underuse of the prior odds and the likelihood ratio. Consequently, even in a setting in which there is no motivation to avoid low judgments, misuse of the likelihood ratio and the prior odds could still lead to bias.

Another idiosyncratic aspect of this study is that the prior probabilities were mainly in the bottom to middle of the probability scale. In this study there is evidence that the overestimation effect does not extend to the very top of the probability scale, and it is possible that a study that sampled higher probabilities would find more evidence for an underestimation effect. Even so, an underestimation effect at the top of the scale could still be explained through the same underlying processes—underuse of the likelihood ratio and underuse of the prior. For this reason, I find the regression analyses in which the likelihood ratio and the prior probabilities were somewhat underused to be the most important results when considering generalization to other settings.

Another open question is the process by which the participants actually made the posttest judgments. One possibility is that the participants used some sort of mental equation that is an imperfect approximation of Bayes' rule. For example, Gigerenzer and Hoffrage (1995) investigated the use of a number of simpler alternatives to Bayes' rule. Another option to explain the deviations from normality is that participants used the prior and likelihood to compute the posterior using mental math, but that there is "measurement error" or "noise" in participants' reporting of each of their beliefs about the pretest, sensitivity, false positive rate, and posttest (similar to Hilbert, 2012).[5] Some amount of noise is likely, given that participants reported their posttest judgments before they reported their beliefs about the prior and the likelihood, so it is possible that their beliefs about the prior and likelihood may not have been entirely stable across time. Though it is theoretically possible that the physicians used a mental math approach in this study, mental math seems most likely in situations in which reasoners are given a Bayesian problem (e.g., the mammogram problem) in a word-problem format, so that all the necessary mathematical components are clearly specified.

A second possibility that seems more plausible in this study is that participants do not "calculate" the posterior from their beliefs in the prior and likelihood, but rather that they rely upon preexisting experiences when judging the posterior. For example, when considering the probability that a patient has colorectal cancer after receiving a negative FOBT test, a physician could think back to a set of similar patients (who have a similar pretest likelihood or a similar set of demographic characteristics and symptoms) and who received a negative FOBT test, and estimate the percentage of these patients who eventually were found to have colorectal cancer. This process is believed to be the process participants use when they have access to extensive prior "natural" experience (Kleiter, 1994) or even lab-based trial-by-trial learning experience (Edgell et al., 2004; see also Barbey & Sloman, 2007; Gigerenzer & Hoffrage, 1995; Kleiter et al., 1997). Noisy recall of these quantities could also contribute to the imperfect coherence between the pretest, likelihood, and posttest judgments (Hilbert, 2012).

It is impossible to know which of these processes were used in this study, and it is possible that different participants used different processes or combinations of the processes (Singmann et al., 2014). Still, the current study makes a number of important contributions. This study contributes evidence that

[5] I thank a reviewer of this manuscript for suggesting this possibility.

probabilistic reasoning can be fairly accurate in situations when reasoners can rely upon their own beliefs. That said, Figs. 3 and 4 show that there is considerable variance in physicians' post-test judgments relative to the normative answer. And this variance has important consequences given that physicians need to make decisions about whether to stop testing versus continue to test versus start to treat based on probabilistic thresholds (Pauker & Kassirer, 1980; Warner, Najarian, & Tierney, 2010). Overestimating the posterior probability after a positive test result can lead to premature closure and missing the correct diagnoses. Overestimating the posterior probability after a negative test can also lead to refusing to stop testing a patient, potentially leading to false positives and side-effects from testing. In fact, premature closure and misjudging the usefulness of a finding are two of the most common cognitive errors believed to cause diagnostic mistakes (Croskerry, 2002; Croskerry, Singhal, & Mamede, 2013; Eva & Cunnington, 2006; Graber, Franklin, & Gordon, 2005; Graber et al., 2012; Reilly, Ogdie, Von Feldt, & Myers, 2013; Voytovich, Rippey, & Suffredini, 1985). Consequently, it is important to understand more broadly how well physicians and other experts perform Bayesian reasoning in more realistic situations with their own beliefs.

Acknowledgements Financial support for this study was provided by NIH Grant F32 1F32HL108711.

References

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *The Behavioral and Brain Sciences, 30*(3), 241–254. doi:10.1017/S0140525X07001653. discussion 255–297.

Bergus, G. R., Chapman, G. B., Gjerde, C., & Elstein, A. S. (1995). Clinical reasoning about new symptoms despite preexisting disease: Sources of error and order effects. *Family Medicine, 27,* 314–320.

Bjerregaard, N. C., Tøttrup, A., Sørensen, H. T., & Laurberg, S. (2009). Detection of colorectal cancer in symptomatic outpatients without visible rectal bleeding: Validity of the fecal occult blood test. *Clinical Epidemiology, 1,* 119–124. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2943167&tool=pmcentrez&rendertype=abstract

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

Castiglione, G., Ciatto, S., Mazzotta, A., & Grazzini, G. (1995). Sensitivity of screening sigmoidoscopy for proximal colorectal tumors. *The Lancet, 345,* 726–727. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Sensitivity+of+screening+sigmoidoscopy+for+proximal+colorectal+tumors#0

Christensen-Szalanski, J. J., & Beach, L. R. (1982). Experience and the base-rate fallacy. *Organizational Behavior and Human Performance, 29*(2), 270–278.

Christensen-Szalanski, J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance, 7*(4), 928–935. doi:10.1037//0096-1523.7.4.928

Croskerry, P. (2002). Achieving quality in clinical decision making: Cognitive strategies and detection of bias. *Academic Emergency Medicine : Official Journal of the Society for Academic Emergency Medicine, 9*(11), 1184–1204.

Croskerry, P., Singhal, G., & Mamede, S. (2013). Cognitive debiasing 1: Origins of bias and theory of debiasing. *BMJ Quality & Safety, 22,* ii58–ii64. doi:10.1136/bmjqs-2012-001712

Eddy, D. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249–267). Cambridge, UK: Cambridge University Press.

Edgell, S. E., Harbison, J. I., Neace, W. P., Nahinsky, I. D., & Lajoie, A. S. (2004). What is learned from experience in a probabilistic environment? *Journal of Behavioral Decision Making, 17*(3), 213–229. doi:10.1002/bdm.471

Eva, K. W., & Cunnington, J. P. W. (2006). The difficulty with experience: Does practice increase susceptibility to premature closure? *The Journal of Continuing Education in the Health Professions, 26*(3), 192–198. doi:10.1002/chp

Evans, J. S. B. T., Handley, S. J., Over, D. E., & Perham, N. (2002). Background beliefs in Bayesian inference. *Memory & Cognition, 30*(2), 179–190.

Evans, J. S. B. T., Thompson, V. A., & Over, D. E. (2015). Uncertain deduction and conditional reasoning. *Frontiers in Psychology, 6,* 1–12. doi:10.3389/fpsyg.2015.00398

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*(4), 684–704. doi:10.1037//0033-295X.102.4.684

Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic error in internal medicine. *Archives of Internal Medicine, 165*(13), 1493–1499. doi:10.1001/archinte.165.13.1493

Graber, M. L., Kissam, S., Payne, V. L., Meyer, A. N. D., Sorensen, A., Lenfestey, N., . . . Singh, H. (2012). Cognitive interventions to reduce diagnostic error: A narrative review. *BMJ Quality & Safety, 21*, 535–557. doi:10.1136/bmjqs-2011-000149

Hadjichristidis, C., Sloman, S. A., & Over, D. E. (2014). Categorical induction from uncertain premises: Jeffrey's doesn't completely rule. *Thinking & Reasoning, 20*(4), 405–431. doi:10.1080/13546783.2014.884510

Hilbert, M. (2012). Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychological Bulletin, 138*(2), 211–237. doi:10.1037/a0025940

Johnson, C., Mei-Hsiu Chen, M. M. M., Toledano, A. Y., Heiken, J. P., Dachman, A., Kuo, M. D., . . . Limburg, P. J. (2008). Accuracy of CT colonography for detection of large adenomas and cancers. *The New England Journal of Medicine*, *359*(12), 1207–1217. Retrieved from http://www.nejm.org/doi/full/10.1056/nejmoa0800996

Keren, G., & Thujs, L. J. (1996). The base rate controversy: Is the glass half-full or half-empty? *Behavioral and Brain Sciences, 19,* 26.

Kleiter, G. D. (1994). Natural sampling: Rationality without base-rates. In G. H. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 377–388). New York, NY: Springer-Verlag.

Kleiter, G. D., Krebs, M., Doherty, M. E., Garavan, H., Chadwick, R., & Brake, G. (1997). Do subjects understand base rates? *Organizational Behavior and Human Decision Processes, 72*(1), 25–61. doi:10.1006/obhd.1997.2727

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences, 19*(01), 1. doi:10.1017/S0140525X00041157

Krynski, T. R., & Tenenbaum, J. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology. General, 136*(3), 430–450. doi:10.1037/0096-3445.136.3.430

Lyman, G., & Balducci, L. (1993). Overestimation of test effects in clinical judgment. *Journal of Cancer Education*, *8*(4).

Retrieved from http://www.tandfonline.com/doi/full/10.1080/08858199309528246

Lyman, G., & Balducci, L. (1994). The effect of changing disease risk on clinical reasoning. *Journal of General Internal Medicine, 9*(9), 488–495. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7996291

Niv, Y., & Sperber, A. (1995). Sensitivity, specificity, and predictive value of fecal occult blood testing (Hemoccult II) for colorectal neoplasia in symptomatic patients : A prospective study with total colonoscopy. *The American Journal of Gastroenterology, 90*(11), 1974–1977. Retrieved from http://cat.inist.fr/?aModele=afficheN&cpsidt=2902479

Noguchi, Y., Matsui, K., Imura, H., Kiyota, M., & Fukui, T. (2002). Quantitative evaluation of the diagnostic thinking process in medical students. *Journal of General Internal Medicine, 17*(11), 839–844. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12406355

Pauker, S., & Kassirer, J. (1980). The threshold approach to clinical decision making. *The New England Journal of Medicine, 302*(20), 1109–1117. Retrieved from http://europepmc.org/abstract/MED/7366635

Reilly, J. B., Ogdie, A. R., Von Feldt, J. M., & Myers, J. S. (2013). Teaching about how doctors think: A longitudinal curriculum in cognitive bias and diagnostic error for residents. *BMJ Quality & Safety, 22,* 1044–1050. doi:10.1136/bmjqs-2013-001987

Shafer, G. (1981). Jeffrey's rule of conditioning. *Philosophy of Science, 48,* 337–362.

Singmann, H., Klauer, K. C., & Over, D. (2014). New normative standards of conditional reasoning and the dual-source model. *Frontiers in Psychology, 5,* 1–14. doi:10.3389/fpsyg.2014.00316

Talbott, W. (2015). *Bayesian epistemology.* Retrieved from http://plato.stanford.edu/entries/epistemology-bayesian/

Voytovich, A. E., Rippey, R. M., & Suffredini, A. (1985). Premature conclusions in diagnostic reasoning. *Journal of Medical Education, 60*(4), 302–307. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3981589

Warner, J. L., Najarian, R. M., & Tierney, L. M. (2010). Perspective: Uses and misuses of thresholds in diagnostic decision making. *Academic Medicine : Journal of the Association of American Medical Colleges, 85*(3), 556–563. doi:10.1097/ACM.0b013e3181ccd59b

Zauber, A., Lansdorp-Vogelaar, I., Knudsen, A. B., Wilschut, J., van Ballegooijen, M., & Kuntz, K. M. (2008). Evaluating test strategies for colorectal cancer screening: A decision analysis for the U.S. Preventive Services Task Force. *Annals of Internal Medicine, 149*(9), 659–668. Retrieved from http://annals.org/article.aspx?articleid=743580&atab=11

Zhao, J., & Osherson, D. (2010). Updating beliefs in light of uncertain evidence: Descriptive assessment of Jeffrey's rule. *Thinking & Reasoning, 16*(4), 288–307. doi:10.1080/13546783.2010.521695