Original Article

Medical Decision Making

Medical Decision Making
2018, Vol. 38(3) 390–399

DOI: 10.1177/0272989X18754508
journals.sagepub.com/home/mdm
SAGE

Natural Frequencies Do Foster Public Understanding of Medical Tests: Comment on Pighin, Gonzalez, Savadori, and Girotto (2016)

Michelle McDowell, Mirta Galesic, and Gerd Gigerenzer

Abstract

Patients and doctors often need to make decisions based on the results of medical tests. When these results are presented in the form of conditional probabilities, even doctors find it difficult to interpret them correctly. There is over 20 y of research supporting the finding that people are better able to calculate the correct positive predictive value of a test when given information in natural frequencies, as opposed to conditional probabilities. Natural frequencies are one of a few psychological tools that have made it into evidence-based medicine. Recently, Pighin and others (Med Decis Making 2016;36:686–91) argued that natural frequencies could hinder informed decision making, a critique based on a single task and a crude scoring criterion we refer to as the 50%-Split. Our commentary addresses these criticisms based on three analyses. First, we show how the 50%-Split scoring used by Pighin and others misclassifies known errors, such as solely attending to the hit rate (true-positive rate) of the test, as strategies that support understanding. Second, we reanalyze data from 21 additional problems completed by various participant groups to show that their scoring criterion does not support their results in 19 out of 21 cases. Third, we apply the mean deviation scoring method and show that, when given information in natural frequency formats, participants provide estimates that are closer to the correct Bayesian solution than for conditional probability formats. In each analysis, natural frequencies lead to more correct judgements and therefore promote informed decision making relative to conditional probabilities. We welcome further discussions of performance metrics that can provide insight into how the public and therefore patients understand the implications of medical test results.

Keywords

conditional probabilities, medical test, natural frequencies

Date received: June 8, 2017; accepted: October 14, 2017

Natural frequencies are one of a few psychological tools that have made it into evidence-based medicine.[1,2] Natural frequencies have been recommended for communicating the results of diagnostic tests or screenings to patients and the public, as they have been repeatedly shown to facilitate calculations associated with inferring outcomes from jointly occurring events.[3,4] Take, for example, the problem of inferring the probability that a baby has Down syndrome given that the mother tested positive on the prenatal chorionic villus sampling (CVS)

Harding Center for Risk Literacy, Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Berlin, Germany (MM, GG); Cowan Chair in Human Social Dynamics, Santa Fe Institute, Santa Fe, New Mexico, USA (MG). The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. Financial support for this study was provided entirely by the Harding Center for Risk Literacy, Max Planck Institute for Human Development. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. MM and GG are employed by the sponsor.

Corresponding Author

Michelle McDowell, Harding Center for Risk Literacy, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. (mcdowell@mpib-berlin.mpg.de)

test, known in medicine as the positive predictive value of the test, illustrated in Figure 1.[i] Both Figure 1A and 1B present information about the base rate of Down syndrome (prevalence), the hit rate (true-positive rate or sensitivity) and the false-positive rate (1-specificity) of the CVS test. Yet, in Figure 1A, the information is presented in natural frequencies and 1B as conditional probabilities. Natural frequencies present the joint frequencies that result from a process of natural sampling: a sequential process where information about events and their classes are acquired naturally from experience.[5,6] When compared to conditional probability formats that normalize information (compare Figure 1A and 1B), natural frequencies facilitate computations and thereby improve the number of people who can calculate the correct solution. Medical professionals, medical students, adults from the general population, and even fourth-graders are better able to provide a correct estimate on such problems when presented as natural frequencies as opposed to conditional probabilities.[7–10] These studies have informed guidelines about how to present information about medical test results to improve patient understanding.[3,4]

Pighin and others[11] question the recommendation that natural frequencies should be used to communicate the results of medical tests to patients. They argue that they have "replicated the only study reporting that natural frequencies foster public understanding of medical test results" (p 690), and claim that the criterion for classifying estimates as correct Bayesian solutions in this cited study (Galesic and others[7]) was too lenient. Pighin and others[11] claim that this scoring criterion allowed participants who erroneously estimated the base rate as the positive predictive value, or gave another erroneous estimate, to be classified as correct, errors that they claim were more common in natural frequency formats. In their second study, Pighin and others[11] provide test statistics for the CVS test for Down syndrome, shown in Figure 1, with a positive predictive value of 60% (the probability that a baby has Down syndrome given that the mother tested positive on the CVS diagnostic test) to participants from the general public, and propose to classify participant's estimates according to a new scoring criterion that we term the "50%-Split." The 50%-Split classifies participant's estimates as supporting the Down syndrome hypothesis (that the baby has Down syndrome) if their estimate was above 50% and not supporting the Down syndrome hypothesis if their estimate was below 50%. As more participants gave higher estimates of the positive predictive value for conditional probability formats and were thus classified as supporting the Down syndrome hypothesis, Pighin and others[11] concluded that natural frequencies therefore did not foster understanding of medical test results.

The present commentary addresses the critique by Pighin and others.[11] First, we show that numerous studies have reported that natural frequencies foster understanding of medical test results, correcting their claim that they have replicated "the only study." We document over 20 y of research that consistently demonstrates that natural frequencies facilitate understanding, using standard and more stringent scoring procedures, in both lay and expert samples. Second, Pighin and others[11] based their critique on their new scoring approach, the 50%-Split, on a single study with a single problem (CVS test), and on a problem with an unusual feature: a 100% hit rate (the probability that a woman pregnant with a child with Down syndrome will test positive on the CVS test). We reanalyze a set of 19 problems studied by Hafenbrädl and Hoffrage[12] with university students or executive managers, 17 of which do not have a 100% hit rate and 2 that do, and find that, even with the 50%-Split, natural frequencies lead to more correct judgements than do conditional probabilities, and also produce estimates that are closer to the correct Bayesian solution in almost all cases. We show the same for the 2 problems from Galesic and others,[7] who tested older and younger adults from the general public, and argue that the 50%-Split therefore systematically misclassifies known errors, such as erroneously providing the hit rate, as strategies that facilitate understanding. Third, we highlight research showing that conditional probability formats are more affected by variations in the specific numerical values used across problems (e.g., a higher or lower hit rate) than are natural frequency formats. On this basis, we emphasize the need to focus not only on outcomes to infer understanding but also on process using write-aloud protocols,[5] for example, to determine where errors in understanding lie.

We conclude that the arguments by Pighin and others, based on a single problem with an unusual hit rate of 100% and a 50%-Split, do not stand up when many problems are analyzed. Further, a reanalysis of their own data shows that when the crude 50%-Split is replaced by an analysis of actual estimates, natural frequencies were in fact better than conditional probabilities. To their credit, the authors put their fingers on the general problem of how to classify judgements as Bayesian or non-Bayesian, which deserves more analysis. Yet the various scoring criteria used in our present study—proportion of Bayesian responses, absolute deviation, and the 50%-Split—consistently lead to the conclusion that natural frequencies foster understanding of medical test results.

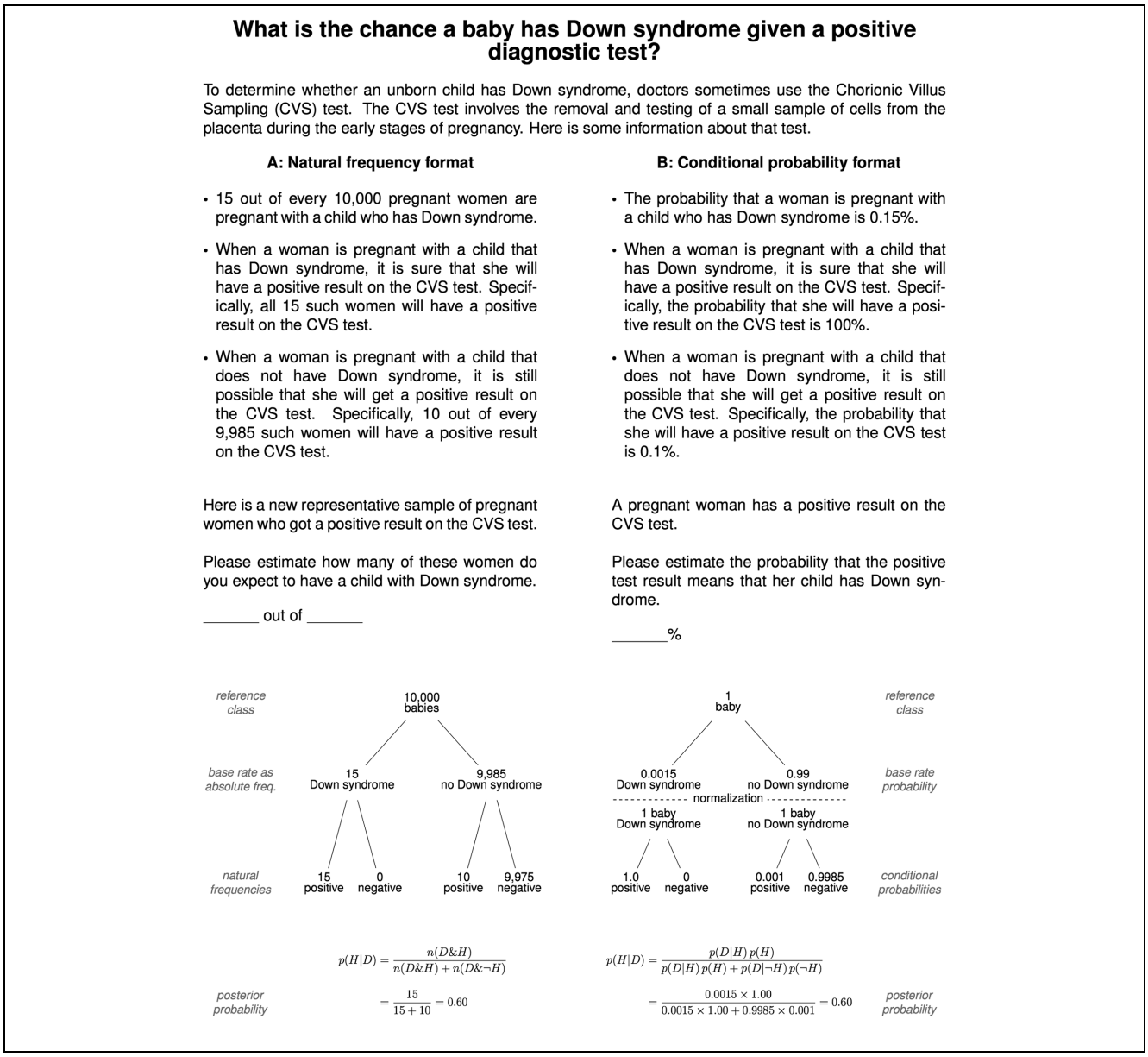

Figure 1 Natural frequencies and conditional probabilities: Two formats for presenting the information necessary for estimating the positive predictive value (Bayesian posterior probability) that a baby has Down syndrome given a positive chronic villus sampling (CVS) test result, as well as the respective computations to calculate the answer. Figure 1A presents natural frequencies where n(D&H) refers to the joint frequency of women who receive a positive test result (D for Data) and have a Down syndrome baby (H for Hypothesis) and n(D&¬H) refers to the joint frequency of women who receive a positive test result but do not have a Down syndrome baby. Figure 1B presents conditional probabilities where p(H) refers to the hypothesis that a child has Down syndrome (prevalence or base rate) and p(D) refers to a positive test result from the data (with p(¬H) and p(¬D) representing their negations). Conditional probabilities (1B) fix the marginal probabilities a priori; that is, the information is normalized, meaning that the base rates need to be reintroduced into the calculation of p(H|D), as shown in the equation at the bottom right of the figure. Natural frequencies do not normalize the information, meaning that information about the naturally occurring base rates are retained in the joint frequencies and base rates can therefore be ignored, thus facilitating computation (see equation on bottom left). The numerical values in Figure 1 are taken from Pighin and others.[11] Note: Natural frequencies are an alternative to "conditional probabilities," not to "single-event statements in terms of percentages" (p 686) as stated by Pighin and others. The numbers in 1B could represent a single-event, as shown here (the probability that a baby has Down syndrome) or a relative frequency among multiple events (a percentage of babies) and could be displayed numerically in percentages, relative frequencies, or probabilities. The crucial difference is that the information in 1B is normalized.

Table 1 Percentage of Bayesian Responses Across Studies Where At Least One Medical Problem Was Used

Study	Sample	Natural Frequencies		Conditional Probabilities[a]	
		n	%	*n*	%
Hoffrage and Gigerenzer[19]	Physicians	24	46.0	24	10.0
Krauss and others[27]	College students	41	53.7	41	12.2
Mellers and McGraw[28]	College students	46	28.0	42	7.0
Evans and others[29]	College students	28	29.0	26	21.0
Lindsey and others[24]	Law students	127	40.0	127	1.0
	Professional jurists	27	74.0	27	10.5
Bramwell and others[30]	Pregnant women	21	14.0	22	5.0
	Companions	20	15.0	20	15.0
	Midwives	20	0.0	22	0.0
	Obstetricians	20	65.0	21	5.0
Chapman and Liu[31]					
Low numeracy	College students	65	7.7	87	0.0
High numeracy	College students	92	28.3	92	1.1
Misuraca and others[32]	College students	120	22.8	30	0.0
Siegrist and Keller[9]	General population	132	11.4	134	0.7
	General population	77	11.7	71	1.4
Tsai and others 33	College students	12	54.0	12	31.0
Ferguson and Starmer[34]	College students	35	31.0	34	3.0
	Medical students	12	50.0	12	8.0
Lesage and others[35]	College students	41	43.0	42	0.0
	College students	43	40.0	43	4.0
	Secondary school students	44	42.0	44	12.0
Friederichs and others[36]	Medical students	29	37.9	34	14.7
Sirota and others[25]	College students	151	57.0	151	19.9

[a]Studies may have used conditional probability or other normalized formats (see Figure 1 caption). For simplicity, we use the term conditional probabilities to refer to normalized formats.

A Survey of the Evidence that Natural Frequencies Foster Understanding of Medical Test Results

First, we address the conclusion by Pighin and others[11] that natural frequencies do not foster an understanding of medical test results, a claim that contradicts an extensive literature in support of the opposite conclusion.[13-15] Natural frequencies have been recommended for communicating information about medical test results because they offer improvement on the conditional probability formats that have previously been used. That is, natural frequencies foster understanding relative to conditional or normalized formats. In their initial study on natural frequencies, Gigerenzer and Hoffrage[5] found across 15 different problems, of which 6 were medical problems, that more participants could solve problems presented using natural frequencies than when using conditional probabilities. As shown in Table 1, numerous studies have replicated this general facilitative effect using health-related problems with physicians and patients, lay samples, and even high school students. Natural frequencies facilitate understanding not only of hypothetical but also real world medical test results, such as the Down syndrome, HIV or mammography screening tests, a finding that has direct implications for informed decision making in health, as natural frequency formats are recommended for communicating test results to patients.[3] Studies on natural frequencies have also been conducted across various problems in other domains (e.g., management, law, social problems) and have even been studied with children.[10] In each study, medical or not, natural frequency formats were better than conditional probability formats (Table 1).[13-15] Even though effect sizes vary, a recent meta-analysis has documented the robustness of the effect across studies and in light of several different individual, methodological, and problem representation moderators.[15]

Natural frequencies not only improve the percentage of people who can calculate a correct estimate of a positive predictive value (or posterior probability in Bayesian inference terminology), but the representation can also be trained to even further bolster performance. In sessions that lasted less than 2 h, Sedlmeier and

Gigerenzer[16] demonstrated that participants trained to convert conditional probabilities into natural frequencies improved from a median performance of around 10% to over 80%, an effect that was maintained over periods of 5 wk and 3 mo. Participants who received rule-based training (e.g., learning to apply Bayes theorem, see equation in Figure 1B) improved to a smaller degree post-training (e.g., around 60% correct solutions) but performance declined substantially over the longer-term (down to 43% at the 5-wk follow-up). Similar effects have been demonstrated with medical students and health professionals.[17,18] This review of the evidence shows that: 1) contrary to the statement made by Pighin and others[11], the study by Galesic and others[7] is not the only one to have studied the effect of natural frequencies on the understanding of medical test results, and 2) in each study, using medical or other topics, natural frequencies improve the number of participants who can provide a correct solution, relative to conditional probabilities.

The 50%-Split

Natural frequencies facilitate performance on Bayesian inference problems, such as the Down syndrome diagnosis problem illustrated in Figure 1, because the computations required to calculate an estimate are reduced compared to conditional probability formats. The representation does part of the computation: the equation in Figure 1A with natural frequencies requires fewer computations than in 1B given conditional probabilities.[5] Most studies have measured performance by the percentage of correct positive predictive value estimates.[5,19,20]

We agree with the suggestion by Pighin and others[11] that alternative scoring metrics should be explored to evaluate people's understanding of medical test results, yet disagree with the metric used in their study. The authors proposed a scoring metric we refer to as the 50%-Split: if the correct positive predictive value is above 50%, they classified any estimate above 50% as correct ("in line with the correct hypothesis") and vice versa.[ii] In their study, they used the problem described in Figure 1 where the task was to infer the probability that a baby had Down syndrome given that the mother tested positive on the chorionic villus sampling (CVS) test. The base rate of Down syndrome was 0.15% (15 in 10,000 babies have Down syndrome), the hit rate was 100% (15 out of 15 babies with Down syndrome test positive), and the false-positive rate was 0.1% (10 out 9,985 babies without Down syndrome test positive). The correct positive predictive value estimate was 60% (or 15 out of 25), meaning that the hypothesis that the baby did have Down syndrome was slightly more probable.

We evaluate the scoring approach of Pighin and others[11] using three analyses. First, we examine the results from Pighin and others and show that the 50%-Split misclassifies known errors in Bayesian reasoning as Bayesian. Second, we reanalyze 21 additional Bayesian inference problems using the 50%-Split and show that the conclusions from the single study by Pighin and others do not hold across most problems. Third, we employ a common scoring metric to evaluate how far participants' estimates deviated from the correct solution.[iii] In each case, we find that natural frequency formats improve performance relative to conditional probabilities.

In our first analysis, we show how the use of 50%-Split scoring misclassifies common errors as supporting Bayesian reasoning. There are 4 common errors found in Bayesian reasoning tasks, 2 of which account for the high percentage of participants in the conditional probability condition who were classified as supporting the correct hypothesis in the study by Pighin and others[11]: 38% of participants gave estimates of 99.9%, which is consistent with the likelihood subtraction error (subtracting the false-positive rate from the hit rate; $p(\mathrm{D}|\mathrm{H})-p(\mathrm{D}|\neg\mathrm{H})=99.9\%$). Another 6% gave estimates of 100%, which is consistent with the hit rate error ($p(\mathrm{D}|\mathrm{H})=100\%$). Altogether, 55% of participants gave an estimate between 99% and 100%. None of the participants who received natural frequencies committed these errors. Using the 50%-Split, Pighin et al. classified 71% of participants who received conditional probabilities as showing understanding of the correct Bayesian hypothesis, of which all but 16% (71% to 55%) clearly show no understanding of the positive predictive value. Thus, this reanalysis shows that the 50%-Split misclassifies non-Bayesian estimates as supporting Bayesian reasoning.

The real and interesting result of their study is that, in the conditional probability condition, more than half of the participants thought the test result was almost certain, or certain. In the natural frequency condition, not a single participant thought so. This illusory certainty has been documented with other medical test results. Many HIV counsellors believe that a positive HIV test result in a low-risk client means that the client definitely has HIV, despite the fact that around 1 in 25 HIV test results are false positives, even in the best tests.[22,23]

Second, we reanalyzed the 2 problems from Galesic and others[7] along with 19 problems from Hafenbrädl and Hoffrage[12] to determine how well the 50%-Split classifies estimates in a larger set of problems. In contrast to the single-problem results for Pighin and others,[11] Table

Table 2 Numerical Values for Problems, and Proportion of Participants Who Provided the Correct Bayesian Solution, Gave an Estimate in line with a 50%-Split, and the Absolute Deviation of Responses from the Correct Solution[a]

					Percentage Bayes'		50%-Split		Mean Absolute Deviation	
Problem	**Base Rate**	**Hit Rate**	**False Alarm**	**Bayes**	**NF**	**Prob**	**NF**	**Prob**	**NF**	**Prob**
Hafenbrädl and Hoffrage[12]										
Pimp	0.005	80	0.05	7.41	33.3	17.9	89.7	82.1	12.2	18.5
HIV infection	0.01	100	0.10	9.09	41.4	10.3	89.7	24.1	11.7	67.7
Heroin addict	0.01	100	0.19	5	66.7	17.9	93.3	29.1	8.6	61.3
Committing suicide	0.024	15	12	0.03	13.8	7.1	96.6	96.4	6.3	7.8
Prenatal damage in child	0.21	47.6	0.50	16.7	27.6	10.3	100.0	88.9	12.3	20.3
Car accident	1	55	5.10	9.91	33.3	16.7	89.7	76.7	9.5	19.8
Breast cancer	1	80	9.60	7.77	37.9	17.2	69.0	41.4	23.9	43.2
Pregnant[b]	2	95	0.51	79.71	45.5	3.4	91.0	79.3	11.9	26.8
Accident on way to school	3	90	40	6.51	40.0	17.2	96.7	93.1	6.4	8.4
Active feminist	5	0.4	2.11	0.99	43.3	17.2	100.0	100.0	3.1	3.7
Bad posture in child	5	40	20	9.52	33.3	16.7	86.2	86.7	16.6	17.4
Blue cab	15	80	20	41.38	23.3	13.8	66.7	39.3	23.8	25.1
Incorrect tax return	20	30	10	42.86	45.5	11.9	89.1	84.8	16.7	24.7
Supplier A	30	15	10	39.13	51.9	36.5	92.6	90.2	11.4	15.0
Choosing economics course	30	70	50	37.5	40.0	17.9	58.6	55.6	11.9	16.5
Admission to school	36	75	20	67.8	43.3	13.8	76.7	89.3	15.3	10.4
Produced in Ohio	60	5	10	42.86	44.1	27.5	95.0	91.0	18.4	20.5
Get contract	60	70	50	67.74	25.5	9.3	69.0	61.7	14.2	17.0
Red ball	80	75	25	92.31	62.1	24.1	96.6	76.9	10.6	29.5
Total					**39.6**	**16.1**	**86.6**	**73.0**	**12.9**	**23.9**
Galesic and others[7]										
Diabetes	0.5	95	50	0.95	31.3	11.0	71.6	26.6	25.0	59.7
Trisomy 21	0.15	80	8	1.47	16.0	3.6	60.6	28.6	39.8	56.4
Pighin and others[11]										
Trisomy 21 – CVS test	0.15	100	0.10	60.0	16.7	6.4	18.5	74.5	47.6	40.0

NF, natural frequency format; Prob, conditional probability format.

[a]50%-Split refers to the proportion of participants whose positive predictive value estimates supported or did not support the most likely hypothesis (e.g., if the positive predictive value was 5%, participants who gave an estimate of $<50\%$ were deemed as supporting the correct hypothesis). Mean absolute deviation refers to the mean absolute difference between a participant's estimate and the Bayesian solution (i.e., the absolute distance irrespective of the direction of the error). Participants who did not provide an answer were excluded (fewer than 5% across problems) and those who gave an estimate of exactly 50% were excluded from the calculation of the 50%-Split (35 instances in Hafenbrädl & Hoffrage; 43 instances in Galesic et al.; and 1 instance in Pighin et al). Problems varied according to the size of the probability/reference class (e.g., the reference class ranged from 100 to 1,000,000 in NF problems).

[b]One participant was excluded from analysis as their problem data was missing.

2 (column 50%-Split) shows that the percentage of participants classified as Bayesian according to this criterion was higher for natural frequencies than for conditional probability formats in 19 out of 21 cases, even in cases where the positive predictive value was over 50% as in the study by Pighin and others (e.g., see problems: pregnant, get contract, and red ball; see also the results of Hafenbrädl and Hoffrage,[10] who found that participant's response strategies were not influenced by the numerical value of the Bayesian estimate.) Contrary to their claims, the alternative scoring criterion used by Pighin and others[11], despite its limits shown before, actually supports natural frequency formats across a variety of different problem scenarios with different numerical values (e.g., base rate, hit rate statistical values).

In a third analysis, we used a common measure of fit, the mean absolute deviation of estimates from the Bayesian positive predictive value (i.e., the absolute distance irrespective of the direction of the error, see Fiedler and others[21]). We applied this metric to the 21 problems plus Pighin and others[11] Down syndrome problem. Using this metric, the last 2 columns of Table 2, show that participant estimates were closer to the correct solution for natural frequencies than conditional probabilities across all but 2 of the 22 problems we reanalyzed.

(One of these 2 problems was the Down syndrome problem used in Pighin and others.)

A final point regarding Pighin and others[11] approach is that a more direct metric for evaluating understanding would be to measure inference or choice, not as categorized by the authors based on the participant's estimates but by asking the participant to make the inference or choice him or herself. In a study examining how jurors and law students understood the implications of a positive DNA test result on the probability that an individual had committed a crime, Lindsey and others[24] examined understanding in terms of a correct estimate and also in terms of a verdict. Natural frequencies improved performance (from 1% with conditional probabilities to 40% with natural frequencies), as determined by a higher proportion of correct estimates (the correct probability estimate was 9.1%), and reduced the number of guilty verdicts from 54.5% to 32.5% for the law student sample and from 44.5% to 32.0% for the juror sample for conditional probability and natural frequency formats, respectively. The advantage of asking the participant to make the choice or inference rather than determining the inference based on their estimates is that one can tease apart calculation errors from errors in understanding. We briefly discuss this point below.

Understanding Errors and Non-Bayesian Strategies

Table 3 provides a detailed analysis of how frequently the 50%-Split misclassifies the 4 common errors as consistent with Bayesian responses: base rate, hit rate, joint occurrence, or the likelihood subtraction error. The first 2 columns show the percentage of Bayesian answers from Table 2, for natural frequencies and conditional probabilities, respectively. The next 2 columns show the percentage of responses that are identical to the base rate, for each format, and the following columns for the other 3 common non-Bayesian strategies. The shaded areas show the cases where the errors would be classified as correct given the 50%-Split scoring method, again showing that this scoring procedure misclassifies known errors as Bayesian (see also the final column in Table 3). Contrary to the claim by Pighin and others, certain error types are not necessarily more common in one format compared to another. For example, in some problems, more participants commit the hit rate error in natural frequency formats compared to conditional probabilities (e.g., see heroin addict, supplier A, produced in Ohio, and Trisomy 21). It is also evident that many of the estimates participants gave across problems could not be categorized as Bayesian or as one of these 4 errors, and that more work is needed to understand the types of errors people make.

A key argument underlying Pighin and others[11] study 2 is that numerical values (e.g., numerical value of the hit rate) can influence numerical estimates, and this can account for errors across formats. This is correct. Hafenbrädl and Hoffrage[12] found that numerical values did influence numerical estimates. Higher base rates and hit rates were associated with more participants finding the correct Bayesian solution and also resulted in higher numerical estimates from participants, whereas higher false-positive rates were associated with lower Bayesian solution rates. Higher base rates were also associated with more participants finding the Bayesian solution and, for those that did not, this led to greater absolute deviations from the correct solution. We conducted similar analyses of Hafenbrädl and Hoffrage's data for the 50%-Split scoring and found similar results. Higher base rates and hit rates increased the odds that a participants' estimate was classified as correct according to the 50%-Split, whereas higher false-positive rates decreased the odds. Further, some of these effects were stronger when problems were presented in conditional probabilities. Participants who received problems in conditional probabilities were more strongly affected by the different numerical values (e.g., higher base rates and hit rates were associated with greater absolute deviations from the correct Bayesian estimate and higher hit rates were associated with higher numerical estimates) than when participants received problems presented in natural frequencies.[12]

Exploring the relationship between numerical values and errors in numerical estimates can help identify potential misunderstandings; however, this approach is limited by its focus on inferring strategies from outcomes. Thus, it does not allow us to determine where potential errors in participant's estimates lie. Gigerenzer and Hoffrage,[5] along with others,[25,26] have studied the written protocols of participants to disentangle calculation errors (e.g., basic arithmetic errors) from errors in understanding. Rather than making inferences about errors in relation to specific numerical estimates, the inclusion of written protocols allows for an assessment of both outcome and process. Written protocols also allowed Gigerenzer and Hoffrage to explore pictorial analogs (e.g., the participant draws a pictorial "beam" to represent the class of alternative outcomes) and non-Bayesian strategies participants used that approximated Bayesian solutions given different numerical values. The aim of this analysis was ecological: to examine whether different cognitive

Table 3 Percentage of Participants Using Different Solution Strategies for Each Problem in Haffenbrädl and Hoffrage,[12] Galesic and Others,[7] and Pighin and Others[11]

	Percentage Bayes'		Base Rate p(H) [BR]		Joint Occurrence p(H)p(D|H) or p(H&D) [JO]		Hit Rate p(D|H) [HR]		Likelihood Subtraction p(D|H) – p(D|¬H) [LS]		
Problem	**NF**	**Prob**	**NF**	**Prob**	**NF**	**Prob**	**NF**	**Prob**	**NF**	**Prob**	**Error Favors 50%-Split**
Hafenbrädl and Hoffrage[12]											
Pimp	33.3	17.9	3.3	7.1	-	-	6.7	7.1	-	7.1	BR, JO
HIV infection	41.4	10.3	5.2[a]	-	5.2[a]	-	3.4	3.4	-	44.8	BR, JO
Heroin addict	66.7	17.9	3.4[a]	-	3.4[a]	-	6.7	-	-	39.3	BR, JO
Committing suicide	13.8	7.1	-	3.6	-	-	10.3	25.0	-	3.6	BR, JO, HR, LS
Prenatal damage in child	27.6	10.3	6.9	-	3.4	20.7	-	-	-	-	BR, JO, HR, LS
Car accident	33.3	16.7	-	3.3	-	6.7	10.0	16.7	-	-	BR, JO, LS
Breast cancer	37.9	17.2	6.9	-	-	3.4	10.3	10.3	-	-	BR, JO
Pregnant	45.5	3.4	-	-	-	-	13.6	17.2	-	-	HR, LS
Accident on way to school	40.0	17.2	-	3.4	10.0	34.5	3.3	-	-	-	BR, JO
Active feminist	43.3	17.2	3.3	3.4	3.3	3.4	10.0	20.7	-	-	BR, JO, HR
Bad posture in child	33.3	16.7	-	-	3.3	10.0	10.0	16.7	-	-	BR, JO, HR, LS
Blue cab	23.3	13.8	-	6.9	30.0	13.8	16.7	34.5	-	13.8	BR, JO
Incorrect tax return	45.5	11.9	1.0[b]	5.0[b]	7.9	12.9	5.9	3.0	1.0[b]	5.0[b]	BR, JO, HR, LS
Supplier A	51.9	36.5	-	10.6	9.3	11.5	15.7	3.8	-	2.9	BR, JO, HR, LS
Choosing economics course	40.0	17.9	-	3.6	-	7.1	3.3	14.3	-	-	BR, JO, LS
Admission to school	43.3	13.8	-	-	-	3.4	6.7	24.1	-	3.4	HR, LS
Produced in Ohio	44.1	27.5	-	2.9	6.9	14.7	23.5	10.8	-	-	JO, HR
Get contract	25.5	9.3	6.7	26.8	5.6	15.5	12.2	5.2	1.1	1.0	BR, HR
Red ball	62.1	24.1	-	-	10.3	6.9	20.7	24.1	-	10.3	BR, JO, HR
Galesic and others[7]											
Diabetes	31.3	11.0	1.3	2.4	-	-	-	1.2	-	1.2	BR, JO
Trisomy 21	16.0	3.6	-	-	-	-	5.0	3.7	-	2.5	BR, JO
Pighin and others[11]											
Trisomy 21 – CVS test	16.7	6.4	22.2[a]	5.3[a]	22.2[a]	5.3[a]	-	6.4	-	38.3	HR, LS

NF, natural frequency format; Prob, conditional probability format. Shaded cells indicate the error would favor the 50%-Split scoring criteria.
[a]The joint occurrence and base rate strategies are equivalent given the problem characteristics and proportions were halved between the 2 errors.
[b]The base rate and likelihood subtraction strategies are equivalent given the problem characteristics and proportions were halved between the 2 errors. Errors that would result in an estimate of 50% are considered as not favoring the 50%-Split as they cannot be classified for this strategy. Participants who did not provide an answer were excluded (less than 5% across problems).

strategies could lead to approximate estimates for probability and natural frequency formats and under which conditions these were appropriate. Gigerenzer and Hoffrage found that, depending on the numerical values of the problem, non-Bayesian strategies could approximate correct solutions. For example, when the base rate is very small, one can take a cognitive short-cut by replacing $p(D|\neg H)p(\neg H)$ by $p(D|\neg H)$ in the calculation for conditional probabilities (see Equation Figure 1B).

If we are to make strong conclusions about how people understand medical test statistics, a more comprehensive analysis of the errors is necessary. By understanding the errors, we can potentially target such errors in future representations; for example, by providing visual aids that illustrate the relationship between the hypothesis (e.g., Down syndrome) and the outcomes of the test (e.g., positive or negative result) to overcome errors where a perfect hit rate is provided. To our knowledge, few studies have sought to design representations to address specific or common errors, and we welcome further work in this direction.

Conclusions

Over 20 y of research has supported the finding that natural frequencies foster understanding relative to conditional or normalized formats. Pighin and others[11] argue against the use of natural frequency formats based on a single reasoning task, with a perfect hit rate, and a crude scoring criterion of a 50%-Split. We have responded to the study by Pighin and others using 3 analyses. First, we emphasize how the 50%-Split systematically misclassifies known errors in Bayesian reasoning as supporting an understanding of medical test results. Second, in a

reanalysis of numerous Bayesian inference tasks, we show how their crude criterion does not support their results in 19 out of 21 cases. Third, we apply a scoring metric, the mean deviation score, to show that estimates from natural frequency formats are closer to the correct estimate than those from conditional probability formats. In each analysis, natural frequencies improve performance relative to conditional probabilities, in medical problems as well as in others. Nevertheless, we welcome the discussion of alternative performance metrics that can provide additional insights into understanding, such as deviation from the correct estimate, and methodologies, such as write-aloud protocols, that can identify non-Bayesian strategies and their unknown causes.

Acknowledgments

We would like to thank Stefania Pighin and Sebastian Hafenbrädl for providing data from their studies.

Notes

i. In the Bayesian reasoning literature this is called the posterior probability.
ii. A more direct metric would be to measure inference by asking the participant to select the hypothesis him or herself.
iii. As yet, the scoring metric is not frequently applied to score Bayesian inference problems; for an exception, see Fiedler et al.[21])

References

1. NHS. NHS breast screening: helping you decide. 2017 [updated 4 May 2017; cited 2017 10 September]; Available from: https://www.gov.uk/government/publications/breast-screening-helping-women-decide.
2. IQWiG. Einladungsschreiben und Entscheidungshilfe zum Mammographie-Screening. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen; 2017 [updated 01 August 2017. Available from: URL: https://www.iqwig.de/download/P14-03_Abschlussbericht_Einladungsschreiben-und-Entscheidungshilfe-zum-M....pdf. [Accessed 10 September 2017].
3. Akl EA, Oxman AD, Herrin J, et al. Using alternative statistical formats for presenting risks and risk reductions. *Cochrane Database Syst Rev*. 2011;16(3):CD006776.
4. Trevena L, Zikmund-Fisher B, Edwards A, et al. Presenting quantitative information about decision outcomes: a risk communication primer for patient decision aid developers. *BMC Med Inform Decis Mak*. 2013;13(Suppl 2):S7.
5. Gigerenzer G, Hoffrage U. How to improve Bayesian reasoning without instruction: Frequency formats. *Psychol Rev*. 1995;102:684–704.
6. Kleiter GD. Natural sampling: Rationality without base rates. In: Fischer GH, Laming D, editors. Contributions to Mathematical Psychology, Psychometrics, and Methodology. New York: Springer-Verlag New York; 1994. p 375–88.
7. Galesic M, Gigerenzer G, Straubinger N. Natural frequencies help older adults and people with low numeracy to evaluate medical screening tests. *Med Decis Making*. 2009;29:368–71.
8. Garcia-Retamero R, Hoffrage U. Visual representation of statistical information improves diagnostic inferences in doctors and their patients. *Soc Sci Med*. 2013;83:27–33.
9. Siegrist M, Keller C. Natural frequencies and Bayesian reasoning: the impact of formal education and problem context. *J Risk Res*. 2011;14:1039–55.
10. Zhu LQ, Gigerenzer G. Children can solve Bayesian problems: the role of representation in mental computation. *Cognition*. 2006;98:287–308.
11. Pighin S, Gonzalez M, Savadori L, Girotto V. Natural frequencies do not foster public understanding of medical test results. *Med Decis Making*. 2016;36:686–91.
12. Hafenbradl S, Hoffrage U. Toward an ecological analysis of Bayesian inferences: how task characteristics influence responses. *Front Psychol*. 2015;6:1–15.
13. Brase GL, Hill WT. Good fences make for good neighbors but bad science: a review of what improves Bayesian reasoning and why. *Front Psychol*. 2015;6:1–9.
14. Johnson ED, Tubau E. Comprehension and computation in Bayesian problem solving. *Front Psychol*. 2015;6:1–19.
15. McDowell M, Jacobs P. Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychol Bull*. 2017 Dec;143:1273–312.
16. Sedlmeier P, Gigerenzer G. Teaching Bayesian reasoning in less than two hours. *J Exp Psychol Gen*. 2001;130:380–400.
17. Hoffrage U, Lindsey S, Hertwig R, Gigerenzer G. Communicating statistical information. *Science*. 2000;290:2261–2.
18. Feufel MA, Keller N, Kendel F, Spies CD, Gigerenzer G. A 1-hour training to improve skills for interpreting medical tests. (manuscript under review). 2017.
19. Hoffrage U, Gigerenzer G. Using natural frequencies to improve diagnostic inferences. *Acad Med*. 1998;73:538–40.
20. Hoffrage U, Krauss S, Martignon L, Gigerenzer G. Natural frequencies improve Bayesian reasoning in simple and complex inference tasks. *Front Psychol*. 2015;6:1–14.
21. Fiedler K, Brinkmann B, Betsch T, Wild B. A sampling approach to biases in conditional probability judgments: Beyond base rate neglect and statistical format. *J Exp Psychol Gen*. 2000;129:399–418.
22. Prinz R, Feufel M, Gigerenzer G, Wegwarth O. What counselors tell low-risk clients about HIV test performance. *Curr HIV Res*. 2015;13:369–80.
23. Gigerenzer G, Hoffrage U, Ebert A. AIDS counselling for low-risk clients. *AIDS Care*. 1998;10:197–211.
24. Lindsey S, Hertwig R, Gigerenzer G. Communicating Statistical DNA Evidence. *Jurimetrics J*. 2003;43:147–63.
25. Sirota M, Juanchich M, Hagmayer Y. Ecological rationality or nested sets? Individual differences in cognitive processing predict Bayesian reasoning. *Psychon Bull Rev*. 2014;21:198–204.

26. Hoffrage U, Hafenbradl S, Bouquet C. Natural frequencies facilitate diagnostic inferences of managers. *Front Psychol*. 2015;6:642.
27. Krauss S, Martignon L, Hoffrage U. Simplifying Bayesian inference: The general case. In: Magnani L, Nersessian NJ, Thagard P, editors. Model-based Reasoning in Scientific Discovery. New York: Kluwer Academic/Plenum Publishers; 1999. p 165–79.
28. Mellers BA, McGraw AP. How to improve Bayesian reasoning: Comment on Gigerenzer and Hoffrage (1995). *Psychol Rev*. 1999;106:417–24.
29. Evans JSBT, Handley SJ, Perham N, Over DE, Thompson VA. Frequency versus probability formats in statistical word problems. *Cognition*. 2000;77:197–213.
30. Bramwell R, West H, Salmon P. Health professionals' and service users' interpretation of screening test results: experimental study. *BMJ*. 2006;333:284-6.
31. Chapman GB, Liu J. Numeracy, frequency, and bayesian reasoning. *Judgm Decis Mak*. 2009;4:34–40.
32. Misuraca R, Carmeci FA, Pravettoni G, Cardaci M. Facilitating effect of natural frequencies: Size does not matter. *Percept Motor Skills*. 2009;108:422–30.
33. Tsai J, Miller S, Kirlik A. Interactive visualizations to improve bayesian reasoning. *Proc Hum Factor Ergon Soc Annu Meet*. 2011;55:385–9.
34. Ferguson E, Starmer C. Incentives, expertise, and medical decisions: Testing the robustness of natural frequency framing. *Health Psychol*. 2013;32:967–77.
35. Lesage E, Navarrete G, De Neys W. Evolutionary modules and Bayesian facilitation: The role of general cognitive resources. *Think Reason*. 2013;19:27–53.
36. Friederichs H, Ligges S, Weissenstein A. Using tree diagrams without numerical values in addition to relative numbers improves students' numeracy skills: A randomized study in medical education. *Med Decis Making*. 2014;34:253–7.