Natural frequencies do not foster public understanding of medical test results

Stefania Pighin[1], Michel Gonzalez[2], Lucia Savadori[3], Vittorio Girotto[1*]

Affiliations:

[1]Center for Experimental Research in Management and Economics, DCP, University IUAV of Venice, Italy

[2]Laboratory of Cognitive Psychology, CNRS and Aix-Marseille University, France

[3]Department of Economics and Management, University of Trento, Italy

[*]Corresponding author

Keywords:

Natural frequencies; Single-event probability; Test result understanding; Diagnostic reasoning; Numeracy.

Acknowledgments

Financial support for this study was provided in part by grants of the Swiss & Global Ca` Foscari Foundation, and of the Italian Ministry of Research (PRIN2010-RP5RNM). The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. The study was approved by the Ethics Committee of CERME (University Ca' Foscari of Venice). We thank Mirta Galesic for providing us with the text of the Trisomy 21 problem (Study 1).

Abstract

Major organizations recommend presenting medical test results in terms of natural frequencies, rather than single-event probabilities. The evidence, however, is that natural frequency presentations benefit at most one-fifth of samples of health-service users and patients. Only one study reported a substantial benefit of these presentations. Here, we replicate that study, testing online survey respondents. Study 1 attributed the previously reported benefit of natural frequencies to a scoring artifact. Study 2 showed that natural frequencies may elicit evaluations that conflict with the normatively correct one, potentially hindering informed decision-making. Ironically, these evaluations occurred less often when respondents reasoned about single-event probabilities. These results suggest caution in promoting natural frequencies as the best way to communicate medical test data to health-service users and patients.

"Did you know", I ask, "that the rate of mortality from snake bites is only three to ten percent?" [...]

"You and your statistics!" she says. "If I had a hundred daughters, each of them bitten by a viper, then yes! Then I would lose only three to ten daughters. Surprisingly few! [But] I have only a single child!"

Max Frisch, *Homo Faber*, 1957

Patients, and to a lesser extent doctors, often err in evaluating the probability that a person with a positive test result has a given disease (1, 2). A common view is that their errors depend on the format in which information is provided: Respondents fail when the prevalence of the disease and the properties of the test (i.e., the true positive and false positive rates) are expressed by single-event statements in terms of percentages (e.g., respectively, "The probability that a tested person has the disease is 0.15%"; "If she is diseased, the probability that she has a positive result is 80%"; "If she is not diseased, the probability that she has a positive result is 8%").

Respondents perform better when the same data are expressed by natural frequency statements (i.e., respectively, "15 out of the 10,000 people tested were diseased"; "12 out of the 15 diseased people had a positive result"; "799 out of the 9,985 non-diseased people had a positive result"), and they have to predict how many members of a new sample of individuals with a positive test result will actually be diseased (3). Accordingly, a common recommendation is to use natural frequencies, rather than single-event probabilities, for communicating test results to patients (4, 5). Despite its popularity, this recommendation is questionable. In fact, only about half of doctors and other samples of educated respondents (3, 6-9), and at most one-fifth of samples of health-service users (8) and patients (9) reason correctly about test results expressed as natural frequencies.

One study, however, reported a different response pattern (10). Respondents sampled from the general public, including elderly and low-numeracy individuals, reasoned about the results concerning the neck-fold skin test for trisomy 21, and genetic testing for diabetes (see Figure 1).

When data were expressed as single-case probabilities, respondents failed. When data were expressed as natural frequencies, about half respondents made correct estimations. Does their success prove that natural frequencies actually benefit heath-service users?

A closer examination of the way in which their answers have been coded suggests a different interpretation. In both problems, the correct solution was about "1%".[1] The authors, however, coded all answers that were below "5%" as accurate estimations, because their "… focus was on whether participants could give estimates that were functional for health-related decisions" (10, p. 369). The choice of such a loose criterion conflicts with the recommendation made by the advocates of the natural frequencies, that is, of following strict coding criteria to avoid classification errors (3). Indeed, applying the loose criterion may lead to classifying wrong answers as accurate evaluations.

Consider the frequent error of using the prevalence rate as the positive predictive value of the test (11). If respondents make this error in completing the natural frequency problems of Figure 1, they will answer "15 out of 10,000" in the Trisomy 21 problem, and "50 out of 10,000" in the Diabetes one. Under the loose criterion used in (10), both answers will be scored as correct. Likewise, if respondents make the other frequent error of using the true positive frequency (7), they will answer "12 out of 10,000" in the Trisomy 21 problem, and "48 out of 10,000" in the Diabetes one. Under the loose criterion, these answers too will be scored as correct. Thus, our analysis suggests that the success rate reported by (10) reflects an inaccurate coding of respondents' evaluations, and does not prove any benefit of natural frequencies. To test our analysis, we replicated the study reported in (10).

[1] Following an estimate rate criterion, the correct solution is close to the estimated relative frequency of individuals with a pathological condition among those with a positive test result, namely, 12/(12+799) = 1.47%, in the Trisomy 21 problem, and 48/(48+4975) = 0.95%, in the Diabetes problem.

STUDY 1

METHODS

The details of procedure and sampling are available as Supplementary Material (see the online Appendix). Respondents were 160 US residents (mean age: 36 y; age range 20-67 y; 70 women) recruited using Amazon Mechanical Turk (AMT) platform. AMT allows researchers to conduct studies with samples of the US population, whose results are similar to those obtained with laboratory samples (12). Respondents were paid $1.50 to complete a scenario (Figure 1), and an 11-point numeracy scale (13) as in (10). They were randomly assigned to one of four groups (N = 40), in a 2 (Scenario content: Trisomy 21 vs. Diabetes) x 2 (Information type: Percentage vs. Natural Frequency) between-participants design.

RESULTS

We analyzed the answers using the same criterion as in (10), that is, we scored all answers smaller than "5%" as correct evaluations. The results (Table 1) replicated those reported in (10): About 60% of respondents solved the natural frequency version of the two problems, whereas only about 15% did so in the percentage version. The difference was significant (Trisomy 21 problem: difference, 47.5 percentage points [CI 26-63], $P < 0.001$; Diabetes problem: difference, 42.5 percentage points [CI 21-59], $P < 0.001$). We then analyzed the answers using the strict criterion (similar to 3), according to which correct answers have to be numerically the same as the normatively correct evaluation. In the natural frequency version, the correct answer was "12 out of 811" in the Trisomy 21 problem, and "48 out of 5023" in the Diabetes one. In the percentage version, the correct answer was "1.47%" (i.e., 1.2%/[1.2% + 79.8%]) in the Trisomy 21 problem, and "0.95%" (i.e., 0.48%/[0.48% + 49.8%]) in the Diabetes one. Under this criterion, virtually no respondent solved the natural frequency version nor the percentage version of these problems. This pattern of results did not change when we used an approximate criterion, according to which all answers within ±10% of the normatively correct evaluation were scored as correct. This liberal criterion treats some non-normative answers as correct. For example, the response "12 out of 799"

(Trisomy 21 problem) is non-normative because it consists of dividing the true positive frequency by the false positive frequency, but it is approximatively correct (the correct evaluation is "12 out of 811"). Yet, even under the approximate criterion, very few respondents (less than 8%) solved the natural frequency or the percentage version of the problems.

Studies testing samples of patients (9) reported that the most frequent errors consisted of using the prevalence rate in the natural frequency versions (21%), and the true positive rate in the percentage ones (35%). We obtained similar results (28% and 31%, respectively). We suspect that many of the correct responses on the natural frequency versions reported in (10) were actually erroneous answers of this sort.

Could our results be attributed to low numeracy or education levels? In fact, 58% of our respondents had a university degree, and their median score on the 11-point numeracy scale was 10 (mean: 9.3). These numeracy scores are similar to those reported in (10), as well as in studies testing more educated samples (13).

STUDY 2

Advocates of natural frequencies might argue that patients' typical concern is understanding test results, rather than making precise probability estimates. Therefore, the finding that natural frequencies do not elicit normatively correct evaluations is clinically irrelevant. We posit, however, that natural frequencies may elicit answers that not only differ from but also conflict with the normatively correct response, potentially hindering informed decision-making.

In the Trisomy 21 scenario in Study 1, the normative answer ("12 out of 811") favored the hypothesis that the child did not have Down syndrome, consistent with the erroneous answers frequently observed using natural frequencies (e.g., "15 out of 10,000"). Consider the scenario in Figure 2. It is the same as the Trisomy 21 scenario of Study 1, except that it describes the CVS test (14). Unlike Study 1, in the natural frequency version, the normative answer ("15 out of 25") favors

the hypothesis that the child has Down syndrome, whereas the typical erroneous answers (e.g., the prevalence rate: "15 out of 10,000") favor the hypothesis that the child does not have it. Respondents who make these errors will endorse the less likely hypothesis, potentially making misinformed choices. But do respondents actually make these errors? Or do natural frequencies protect them from the misunderstandings often attributed to single-event probabilities? Study 2 addressed these questions.

METHODS

A new sample of 104 US residents (mean age: 34 y; age range 19-72 y; 45 women) was recruited using AMT. They were paid $1 to complete either the natural frequency (N = 55) or the percentage version (N = 49) of the scenario in Figure 2.

RESULTS AND DISCUSSION

The correct answer was "15 out of 25" in the natural frequency version, and "60%" in the percentage one. Under the strict criterion used in Study 1, only 16% of respondents correctly answered the natural frequency version, and only 6% of respondents did so in the percentage version. The proportion of correct answers trended towards being higher in the natural frequency than in the percentage version. The difference, however, was not significant (P = 0.103). In the natural frequency version, the most frequent errors (52% of all errors) consisted of reporting the value of the prevalence or true positive rate. In this version, only 20% of respondents correctly assigned a greater than 50% chance to the Down syndrome hypothesis, whereas 71% of respondents did so in the percentage version. The difference was significant (51.4 percentage points [CI 33-65], P < 0.001). In sum, in the natural frequency version, respondents endorsed the less likely hypothesis more often than in the percentage one. This result suggests that, in some cases, natural frequencies result in more non-normative than normative responses.

Differing from Study 1, in the percentage version the most common error (37%) consisted of subtracting the false positive rate from the true positive rate. One possibility is that respondents did so because they assumed that they could not use the true positive rate (100%) as the positive predictive value. This tendency suggests that respondents' evaluations may depend on the specific numerical values provided in the scenarios (1, 7).

GENERAL DISCUSSION

We replicated the only study reporting that natural frequencies foster public understanding of medical test results (10). Study 1 showed that the benefit was an artifact of the scoring procedures: It occurred when judgments were scored according to the loose criterion used in the original study. When they were scored according to the strict criterion recommended by the proponents of natural frequencies (3), virtually no respondents reasoned correctly. Study 2 showed that natural frequencies may elicit evaluations potentially hampering informed decision-making. Ironically, these evaluations occurred less often when respondents reasoned about single-event probabilities.

Our studies tested AMT population samples. AMT respondents appear to be more literate and knowledgeable than the average US population (15). The finding that they make erroneous frequency evaluations suggests that respondents sampled from the general population may be even less likely to benefit from natural frequency presentations.

Along with previous results (7-9), our studies call into question the recommendation that health-care professionals use natural frequencies, in preference to single-event probabilities, to communicate test data (4-5). Some trials have documented the shortcomings of using natural frequencies in risk communication (16). These trials have been criticized on the ground that they did not use proper natural frequency information (17). Such a dismissal, however, does not concern the present results because they have been obtained using the same scenarios and tasks used by the advocates of natural frequencies (10).

Alternative procedures could lead health-service users and patients to make accurate frequency predictions (7, 9). The clinical utility of frequency predictions, however, is questionable (11). As the passage quoted at the start of the paper indicates, individuals are generally interested in evaluating their personal case (e.g., "Is my daughter actually diseased?"), rather than in making predictions about a sample of cases (e.g., "How many individuals similar to my daughter will actually be diseased?") Accordingly, research should try to improve inferences about the test results of a single individual patient.

REFERENCES

1. Casscells W, Schoenberger A, Graboys T. Interpretation by physicians of clinical laboratory results. N Engl J Med. 1978;299:999–1000.
2. Eddy DM. Probabilistic reasoning in clinical medicine: Problems and opportunities. In: Kahneman D, Slovic P, & Tversky A, eds. Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press. 1998. pp. 249-267.
3. Gigerenzer G, Hoffrage U. How to improve Bayesian reasoning without instruction: Frequency formats. Psychol Rev. 1995; 102:684-704.
4. Akl EA, Oxman AD, Herrin J, et al. Using alternative statistical formats for presenting risks and risk reductions. Cochrane Database Syst Rev. 2011;(4):CD006776.
5. Elwyn G, O'Connor A, Stacey D, et al. International patient decision aids standards (IPDAS) collaboration. Developing a quality criteria framework for patient decision aids: online international Delphi consensus process. BMJ. 2006;333:417.
6. Hoffrage U, Gigerenzer G. Using natural frequencies to improve diagnostic inferences. Acad Med. 1998;73:538–540.
7. Girotto V, Gonzalez M. Solving probabilistic and statistical problems: a matter of information structure and question form. Cognition. 2001;78:247–276.

8. Bramwell R, West H, Salmon P. Health professionals' and service users' interpretation of screening test results: experimental study. BMJ. 2006;333:284–286.

9. 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.

10. Galesic M, Gigerenzer G, Straubinger N. Natural frequencies help older adult and people with low numeracy to evaluate medical screening tests. Med Decis Making. 2009;29:368-371.

11. Pighin S, Gonzalez M, Savadori L, Girotto V. Improving public interpretation of probabilistic test results: Distributive evaluations. Med Decis Making. 2015;35:12-15.

12. Paolacci G, Chandler J, Ipeirotis PG. Running experiments on Amazon Mechanical Turk. Judgm Decis Making. 2010;5:411–419.

13. Lipkus IM, Samsa G, Rimer BK. General performance on a numeracy scale among highly educated samples. Med Decis Making. 2001;21:37–44.

14. Hahnemann JM, Vejerslev LO. Accuracy of cytogenetic findings of Chorionic Villus Sampling (CVS) – Diagnostic consequences of CVS mosaicism and non-mosaic discrepancy in centres contributing to Eucromic* 1986-1992. Prenat Diagn. 1997;17: 801-820.

15. Cooper EA, Farid H. Does the Sun revolve around the Earth? A comparison between the general public and online survey respondents in basic scientific knowledge. Public Underst Sci 2014; doi: 10.1177/0963662514554354.

16. Woloshin S, Schwartz LM. Communicating data about the benefits and harms of treatment: a randomized trial. Ann Intern Med. 2011;155:87–97.

17. Gigerenzer G. What are natural frequencies? BMJ 2011;343:d6386.

Table 1. Percentage of Suitable Evaluations for Scenario, Information Type and Coding Criterion

(a)

	Scenario			
	Trisomy 21		Diabetes	
Criterion	Natural Frequency	Percentage	Natural Frequency	Percentage
Loose (b)	62	15	60	17
Strict	2	0	2	0
Approximate	7	2	5	5

(a) N = 40 per condition (b) criterion used by Galesic et al. (2009)

Figure 1. Scenarios used in Study 1

Trisomy 21

To determine whether an unborn child has Down syndrome, doctors sometimes measure the thickness of the fetus' neck skin fold. Here is some information about that 'neck-fold' test.

(Natural frequencies version)

- 15 out of every 10,000 pregnant women are pregnant with a child who has Down syndrome.
- When a woman is pregnant with a child that has Down syndrome, it is not sure that she will have a positive result on the 'neck-fold' test. Specifically, 12 of every 15 such women will have a positive result on the 'neck-fold' test.
- When a woman is pregnant with a child that does not have Down syndrome, it is still possible that she will get a positive result on the 'neck-fold' test. Specifically, 799 out of every 9,985 of such women will have a positive result on the 'neck-fold' test.

Here is a new representative sample of pregnant women who got a positive result on the 'neck-fold' test.
Please estimate how many of these women do you expect to have a child with Down Syndrome.
___ out of ____

(Percentages version)

- The probability that a woman is pregnant with a child who has Down syndrome is 0.15%.
- When a woman is pregnant with a child that has Down syndrome, it is not sure that she will have a positive result on the 'neck-fold' test. Specifically, the probability that she will have a positive result on the 'neck-fold' test is 80%.
- When a woman is pregnant with a child that does not have Down syndrome, it is still possible that she will get a positive result on the 'neck-fold' test. Specifically, the probability that she will have a positive result on the 'neck-fold' test is 8%.

A pregnant woman has a positive result on the 'neck-fold' test.
Please estimate the probability that the positive test result means that her child has Down syndrome:
_____%

Diabetes

To determine whether a person is at risk of insulin-dependent diabetes, doctors sometimes conduct genetic testing. If a person tests positive for a certain gene, he or she might have insulin-dependent diabetes. Here is some information about that genetic test.

(Natural frequencies version)

- 50 out of every 10,000 people have insulin-dependent diabetes.
- If a person has insulin-dependent diabetes, it is not sure that he or she will have a positive result on the genetic test. More precisely, 48 of every 50 of such people will have a positive result on the genetic test.
- If a person does not have insulin-dependent diabetes, it is still possible that he or she will have a positive result on the genetic test. More precisely, 4,975 out of every 9,950 such people will have a positive result on the genetic test.

Here is a new representative sample of people who got a positive result on the genetic test.

Please estimate how many of these people actually have insulin-dependent diabetes.
_____ out of _____

(Percentages version)

- The probability that a person has insulin-dependent diabetes is 0.5%.
- If a person has insulin-dependent diabetes, it is not sure that he or she will have a positive result on the genetic test. More precisely, he or she has a 95% probability of having a positive result on the genetic test.
- If a person does not have insulin-dependent diabetes, it is still possible that he or she will have a positive result on the genetic test. More precisely, he or she has a 50% probability of having a positive result on the genetic test.

Please estimate the probability that a person has insulin-dependent diabetes if he or she has a positive genetic test:
_____%

Figure 2. Scenarios used in Study 2

Trisomy 21 – CVS test

To determine whether an unborn child has Down syndrome, doctors sometimes use the Chorionic Villus Sampling (CVS) test. Here is some information about that test.

(Natural frequencies version)

- 15 out of every 10,000 pregnant women are pregnant with a child who has Down syndrome.
- When a woman is pregnant with a child that has Down syndrome, it is sure that she will have a positive result on the CVS test. Specifically, all 15 such women will have a positive result on the CVS test.
- When a woman is pregnant with a child that does not have Down syndrome, it is still possible that she will get a positive result on the CVS test. Specifically, 10 out of every 9,985 such women will have a positive result on the CVS test.

Here is a new representative sample of pregnant women who got a positive result on the 'neck-fold' test.

Please estimate how many of these women do you expect to have a child with Down Syndrome.

___ out of ____

(Percentages version)

- The probability that a woman is pregnant with a child who has Down syndrome is 0.15%.
- When a woman is pregnant with a child that has Down syndrome, it is sure that she will have a positive result on the CVS test. Specifically, the probability that she will have a positive result on the CVS test is 100%.
- When a woman is pregnant with a child that does not have Down syndrome, it is still possible that she will get a positive result on the CVS test. Specifically, the probability that she will have a positive result on the CVS test is 0.1%.

A pregnant woman has a positive result on the CVS test.

Please estimate the probability that the positive test result means that her child has Down syndrome:

_____%

Supplementary Material of the paper

Natural frequencies do not foster public understanding of medical test results

Stefania Pighin, Michel Gonzalez, Lucia Savadori, and Vittorio Girotto

Table of contents

1. Details about the procedure and participants

In all studies, to compare performance, we estimated two-sided P-values based on the Pearson chi-square statistics applied to between-respondent data, and we determined the confidence intervals [CI], at the P = .95 level, for the differences between proportions of correct responses.

Our aim was to replicate Galesic et al.'s (2009) study, whose group size ranged between about 24 (older adults) and 58 (younger adults) respondents. Accordingly, we have tested 40 respondents per group in Study 1, and about 50 respondents per group in Study 2. Moreover, a power analysis based on the results reported by Galesic et al. (2009) showed that at least 27 respondents in each group were needed to detect a difference between groups with 90% power at the confidence level of 95%.

In both studies, we used Qualtrics, a subscription online survey tool. Its design features place a cookie on a respondent's browser to prevent them from completing the survey more than once. This measure could be bypassed, if an individual uses different browsers or different computers. However, in order to receive their payment, M-Turk workers are asked to provide their M-Turk ID number. This measure reduces the possibility of multiple completions by a single individual.

Participation was voluntary and anonymous, and respondents were free to withdraw from the study at any time.

2. Tables

Table 1SI. Percentage of common responses elicited in the four conditions of Study 1.

	Scenario			
	Trisomy 21		Diabetes	
Response	Natural Frequency	Percentage	Natural Frequency	Percentage
Correct evaluation	2	0	2	0
Prevalence rate	28	0	28	5
True positive rate	5	38	2	23
False positive rate	2	8	5	20
True positive rate - False positive rate	0	17	0	8
Other(a)	63	37	63	44

1.Responses that could not be easily classified under any standard category (see 3) were scored as “Other”

Table 2SI. Percentage of common responses elicited in the four conditions of Study 2.

Response	Condition	
	Natural Frequency	Percentage
Correct evaluation	16	6
Prevalence rate	44	10
True positive rate	0	6
False positive rate	2	6
False positive rate- True positive rate	0	37
Other (a)	38	35

(a) Responses that could not be easily classified under any standard category (see 3) were scored as "Other"

3. On-line invitation

We invite you to participate in a scientific experiment.

Participation requires that you give your informed consent. Before proceeding, please consider the following information:

The study task consists of reading a text and answering a few questions.

The survey will take on average about 10-15 minutes to complete.

There are no risks involved in this study.

You will be paid for your participation at the posted rate (provided that you complete the whole study, including demographic questions).

Your individual privacy will be maintained in all published and written data resulting from the study.

Participation in this research study is voluntary. You may choose not to participate.

By ticking the box below, and proceeding to the study task you certify that you have read this form, and agreed to participate in accordance with the above conditions.

□ I give my informed consent to participate in this study

4. Numeracy Scale and Percentage of Correct Answers

Question		Percentage of correct answer
1.	Imagine that we roll a fair, six-sided die 1000 times. Out of 1000 roles, how many times do you think the die would come up even (2, 4, or 6)?	78
2.	In a specific lottery, the chances of winning a $10 prize are 1%. What is your best guess about how many people would win a $10 prize if 1000 people each buy a single ticket from the lottery?	82
3.	In the scratch lottery, the chance of winning a car is 1 in 1000. What percent of tickets of the scratch lottery win a car?	66
4.	Which of the following numbers represents the biggest risk of getting a disease? [1 in 100; 1 in 1000; 1 in 10]	98
5.	Which of the following represents the biggest risk of getting a disease? [1%; 10%; 5%]	99
6.	If Person A's risk of getting a disease is 1% in ten years, and Person B's risk is double that of A's, what is B's risk?	91
7.	If Person A's chance of getting a disease is 1 in 100 in ten years, and person B's risk is double that of A, what is B's risk?	72
8.	If the chance of getting a disease is 10%, how many people would be expected to get the disease out of 100?	96
9.	If the chance of getting a disease is 10%, how many people would be expected to get the disease out of 1000?	87

10.	If the chance of getting a disease is 20 out of 100, this would be the same as having a ____% chance of getting the disease.	95
11.	The chance of getting a viral infection is .0005. Out of 10,000 people, about how many of them are expected to get infected?	68

5. Demographic questions

5.1. Questionnaire used in both Studies

Please answer the following questions concerning your personal data:

Gender:

☐ Male

☐ Female

Age:

Education:

☐ Less than High School

☐ High School

☐ University Degree