BMJ Open

Simulated patient cases versus numerical summaries: An experimental study of clinician judgments about diagnostic test results

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-019241
Article Type:	Research
Date Submitted by the Author:	21-Aug-2017
Complete List of Authors:	Armstrong, Bonnie; Ryerson University, Psychology Spaniol, Julia; Ryerson University, Psychology Persaud, Nav; St Michael`s Hospital, Keenan Research Centre in the Li Ka Shing Knowledge Institute
Keywords:	Diagnostic inference, Experience-based learning, PPV, NPV, Judgment accuracy, Subjective probability

Simulated patient cases versus numerical summaries:

An experimental study of clinician judgments about diagnostic test results

Bonnie A. Armstrong, MA (PhD Candidate)[1]

Julia Spaniol, PhD (Associate Professor)[1]

Nav Persaud, MSc, MD (Associate Scientist/Staff Physician)[2, 3]

1. Department of Psychology, Ryerson University, 350 Victoria Street, Toronto, Canada, M5B 2K3
2. Li Ka Shing Knowledge Institute, St Michael's Hospital, 209 Victoria Street, Toronto, Canada, M5B 1T8
3. Department of Family and Community Medicine, University of Toronto, 27 King's College Circle, Toronto, Canada, M5S 1A1

Author Note

Correspondence can be addressed to: Bonnie Armstrong, 350 Victoria Street, Ryerson University, Toronto, Canada, M5B 2K3. Email: bonnie.armstrong@psych.ryerson.ca.

Word Count: 1410

Abstract

Objective: Investigate whether clinicians' interpretations of diagnostic test results are sensitive to the format – experiential or numerical – in which relevant statistical information is acquired.

Design: Decision making study involving a 2 (format: experiential vs. numerical) x 3 (diagnostic test: gold standard vs. low sensitivity vs. low specificity) within-subjects design.

Setting: Online experimental study completed by clinicians varying in expertise from one medical center.

Participants: 50 physicians (12 clinicians and 38 residents) from the Department of Family and Community Medicine at St. Michael's Hospital in Toronto, Canada. All participants with at least one year of residency complete were selected for the study.

Main Outcome Measures: Accuracy of clinicians' numerical estimates of the positive and negative posterior values (PPV, NPV) of each diagnostic test.

Results: Clinicians judged the PPV more accurately in the experience condition (95% CI: 12.5 to 19.3) compared to the numerical condition (95% CI: 28.2 to 36.9; $d = 0.697$; P=.001), and the NPV more accurately in the experience condition (95% CI: 6.8 to 15.2) compared to the numerical condition (95% CI: 14.3 to 34.44; $d = 0.303$; P=.015).

Conclusions: Compared to numerical information, experiential exposure to representative simulated patient cases significantly improves clinicians' PPV and NPV judgments without training and in a short period of time.

Strengths:

- The effect of format was tested within-subjects, that is, any existing effect can be observed within each individual.
- The experience-based approach to judging posterior probabilities presented in this paper is an intuitive, quick-and-dirty technique used to improve understanding of diagnostic test results, that does not require statistical training.
- Because fictitious diseases and diagnostic tests were used in the current study, the results will reveal whether clinicians were able to estimate the PPV accurately for a common disease in a prior study [8] due to their familiarity with the disease and test screening or whether it was due to experiencing a simulation of patient cases.

Limitations:

- Participants were recruited from one medical center (i.e., the Department of Community and Family Medicine at St. Michael's Hospital in Toronto, Canada), and any existing effect that surfaces from this study should be replicated across varying specialists in future research.

Clinicians often misinterpret diagnostic test results. Common errors include overestimation of the probability of disease following a positive test result (positive predictive value, PPV) and underestimation of the probability of no disease given a negative test result (negative predictive value, NPV) [1,2]. These errors have negative effects on patient care. Overestimation of the PPV, for example, may lead to costly and harmful overtreatment [3-5]. Typically, clinicians interpret test results on the basis of relevant statistics such as disease prevalence, test sensitivity and test specificity. These details are frequently communicated via numerical summary (e.g., leaflets). The current study examined whether an experiential format – i.e., exposure to representative simulated patient cases – would enhance the accuracy of physicians' PPV and NPV estimates.

Clinicians' inferences about diagnostic test results have long been shown to be sensitive to format effects, with frequency formats producing more accurate judgments than probabilities [1,6,7]. For example, one study showed that when relevant statistics were presented as probabilities clinicians correctly identified the PPV out of four options in only 10% of cases, whereas 46% of clinicians identified the PPV when statistics were presented in natural frequencies [6]. Similarly, retrospectively thinking about past patients in terms of natural frequencies led OB-GYNs to make more accurate PPV judgments for a Down syndrome screening test compared to thinking about these experiences in terms of probabilities [8]. These findings suggest that the format in which judgment-relevant information is presented has a powerful effect on clinicians' diagnostic inferences. In a recent study with non-clinicians (i.e., younger and older adults), PPV and NPV judgments for fictitious diseases and tests were significantly more accurate following sequential exposure to a small (N=100) set of fictitious patient cases than following exposure to numerical summaries of relevant statistics [9]. This

finding suggests that simulated patient cases are superior to numerical summaries, at least for laypeople without formal training in medical diagnosis or statistics.

In the current study, we sought to test whether the advantage of an experiential, "numbers-free" format would extend to clinicians. In the *experience* format, participants viewed samples of representative fictitious patient scenarios that illustrated the frequency-of-occurrence of correct and incorrect positive and negative diagnoses in a population. In the numerical format, participants were presented with explicit statistical summaries, framed in natural frequencies. We predicted that, similar to laypeople, clinicians would provide more accurate estimates of the PPV and NPV in the experience format than in the numerical format.

Methods

Fifty clinicians (Table 1) affiliated with the Department of Community and Family Medicine from St. Michael's Hospital in Toronto, Canada, provided informed consent before completing a 1-hour online experiment in which they received information about a fictitious disease and three separate fictitious diagnostic tests. Tests included a "gold standard" test (high sensitivity, high specificity) as well as two lower-quality tests (one with low sensitivity, another with low specificity).

Information about each test was provided in a numerical format and an experience format. Each participant viewed all of the 6 resulting versions (2 formats x 3 tests), with presentation order counterbalanced across participants. In the numerical format, participants read a summary about each diagnostic test, explicitly describing the disease prevalence (constant throughout the experiment), as well as test sensitivity and the false-positive rate. In the experience format, disease prevalence and test characteristics for each test were communicated via a slideshow of 100 representative patients. Within each of the three slideshows, patients were

presented one at a time, along with information indicating their disease status (has disease/does not have disease) and test result (positive/negative). No numerical statistics were provided and participants were instructed not to take written notes.

In both the numerical and experience formats, relevant information was presented for 3 minutes before participants were prompted for judgments. Participants were not told that the three diagnostic tests were identical in both formats. Subsequent to reading the numerical summary or experiencing patient information, participants judged the PPV and NPV, using a natural frequency response format (e.g., "6 out of 98"). PPV and NPV judgment errors, defined as the absolute difference between true and judged values (Figure 1), were submitted to separate 2 (format: numerical vs. experience) x 3 (test: gold standard vs. low sensitivity vs. low specificity) repeated-measures analyses of variance. Given the sample size (N=50) and the repeated-measures design, the statistical power to detect medium-sized effects (10), with an alpha of .05, was .93 for the "format" factor, and .98 for the "test" factor (11). Statistical analysis was performed using SPSS.

Results

Thirty-one females and nineteen males completed the online study (Table 1), comprising thirty-eight residents and twelve practicing clinicians. On average residents completed 1.4 years of residency, with clinicians averaging 4.3 years of practice.

For PPV judgments, the mean absolute error rate (MAE) was higher in the numerical format condition (*MAE*=32.6%; 95% CI: 28.2 to 36.9) than in the experience format condition (*MAE*=15.9%; 95% CI: 12.5 to 19.3; $d = 0.697$, $P < .001$). As seen in Figure 1, the extent to which PPV judgments were overestimated was reduced dramatically when information was experienced, whereas the classic overestimation of the PPV was replicated when information

was described numerically. For NPV judgments, the numerical format also produced more errors (*MAE*=24.4%; 95% CI: 14.3 to 34.4) compared with the experience format (*MAE*=11.0%; 95% CI: 6.8 to 15.2; $d = 0.303$; $P=.015$), with less underestimation when information was experienced. The effect of format was stable across the three tests ($P=0.54$). There was also no effect of presentation order of format ($P=0.48$) and no statistically significant difference between residents' and qualified clinicians' PPV ($P= 0.35$) or NPV ($P=0.80$) judgment accuracy.

Discussion

Compared to a numerical format in which disease prevalence and test characteristics were provided in a statistical summary, sequential exposure to representative patient scenarios reduced overestimation of the PPV and underestimation of the NPV in clinicians. This effect was replicated across three information scenarios that described separate diagnostic tests with different test characteristics. As predicted, simulated patient cases were superior to numerical summaries in supporting accurate estimation of posterior probabilities, even among medical experts.

For decades it has been known that natural frequency information is more digestible than probabilistic information expressed as conditionals [1,6], but even with frequencies, clinicians' PPV and NPV judgment accuracy is worryingly low. Results of the current study show, for the first time, that experiencing representative cases is an effective strategy through which clinicians can achieve more accurate judgments. Critically, the "experience advantage" was demonstrated under experimental conditions that controlled for disease-specific prior knowledge, both through the use of fictitious diseases and by holding "experience" constant across participants.

Trainees and fully licensed clinicians struggle and commonly commit errors when making Bayesian inferences in medicine such as estimating the PPV [1,2,5-7], which can lead to

a variety of negative consequences [2,3]. Presenting frequency information removes the requirement of inference. Experiencing event frequencies allow the subject to instead directly infer a posterior probability from the cell counts, or in this case each of the four possible combinations of a patient case. That is, those inexperienced with statistics are able to accurately judge posterior probabilities without having to make Bayesian inferences. Indeed presenting frequency events makes the problem at hand easier, and more intuitive, and based on the results of the current study, judgment errors and biases commonly made are reduced dramatically. The approach of experiential learning through frequency information is also quite flexible, as the statistical distribution underlying patient cases such as the base rate (i.e., prevalence of disease) and test characteristics (i.e., sensitivity and specificity) can be defined without restriction. That is, the technique of experience-based learning of medical information introduced in the current paper can be applied to a variety of scenarios. Critically, further studies are required to determine whether these findings from a single center on judgments about hypothetical tests translate to improvements in actual clinical decision making.

Exposure to repeated frequency events (analogous to a fast-paced version of encountering one's patients over time) is the first technique to enhance experts' posterior probability judgments to this extent without any training and in a short timespan. While the numerical format is commonly used in medical education and in real patient cases, it has not been proven to be more effective than other approaches. These results suggest the typical method for describing test characteristics to clinicians may be inferior to more intuitive approaches. It is worth determining in future research whether this effect holds across a variety of medical specialists, and whether patients benefit from learning medical information this way.

Required Manuscript Submission Statements

a) Details of contributors:
1. Bonnie Armstrong (study guarantor): study programming, participant recruitment, data collection, data analysis, manuscript writing
2. Julia Spaniol: manuscript revisions
3. Nav Persaud: study funder, participant recruitment, manuscript writing

All authors had full access to all of the data (including statistical reports and tables) in the study and can take responsibility for the integrity of the data and the accuracy of the data analysis.

b) Competing interests:
All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: all authors had financial support from the Department of Family and Community Medicine of St Michael's Hospital, the Department of Family and Community Medicine of the University of Toronto, the Physicians Services Incorporated Graham Farquharson Knowledge Translation Fellowship, as well as an Early Researcher Award from the Ministry of Research and Innovation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

c) Funding Statement:
This work was supported by the Department of Family and Community Medicine of St Michael's Hospital, the Department of Family and Community Medicine of the University of Toronto, the Physicians Services Incorporated Graham Farquharson Knowledge Translation Fellowship and an Early Researcher Award from the Ministry of Research and Innovation.

d) Data Sharing Statement:
The full dataset is available from the Dryad repository, DOI: [will include DOI once available from Dryad repository]. Consent was not obtained by participants for data sharing but the presented data are anonymized and risk of identification is low.

Ethics Approval Statement:
Ethics approval to conduct the current study was obtained from both St. Michael's Hospital Research Ethics Board (REB number: 16-282), as well as the Ryerson Ethics Board (REB number: 2014-129). All participants gave informed consent before participating in the study.

References

1. Gigerenzer G, Gaissmaier W, Kurz-Milcke E, et al. Helping doctors and patients make sense of health statistics. *Psychol Sci Public Interest*. 2007;8(2):53-96.

2. Whiting PF, Davenport C, Jameson C, et al. How well do health professionals interpret diagnostic information? A systematic review. *BMJ Open.* 2015;5(7):e008155. doi:10.1136/bmjopen-2015- 008155.

3. Wegwarth O, Gigerenzer G. Overdiagnosis and overtreatment: evaluation of what physicians tell their patients about screening harms. *JAMA Intern Med.* 2013;173(22):2086-2087.

4. Bhatt JR, Klotz Laurence. Overtreatment in cancer – is it a problem? *Expert Opin Pharmacother.* 2016;17(1):1-5. doi: 10.1517/14656566.2016.1115481

5. Wegwarth O, Gigerenzer G. Statistical illiteracy in doctors. In: Gigerenzer G, Gray JA, ed. *Better doctors, better patients, better decisions: Envisioning health care 2020*. Cambridge: MIT Press; 2011:137-151.

6. Hoffrage U, Gigerenzer G. Using natural frequencies to improve diagnostic inferences. *Acad Med.* 1998;73(5):538-40.

7. Anderson BL, Gigerenzer G, Parker S, et al. Statistical literacy in obstetricians and gynecologists. *J Healthc Qual.* 2012;36(1):5-17.

8. Obrecht NA, Anderson B, Schulkin J, et al. Retrospective frequency formats promote consistent experience-based Bayesian judgments. *Appl Cognit Psychol.* 2012;26:436-440. doi: 10.1002/acp.2816.

9. Armstrong BA, Spaniol J. Experienced probabilities increase understanding of diagnostic test results in younger and older adults [published online February 15, 2017]. *Med Decis Making.* doi: 10.1177/0272989X17691954.

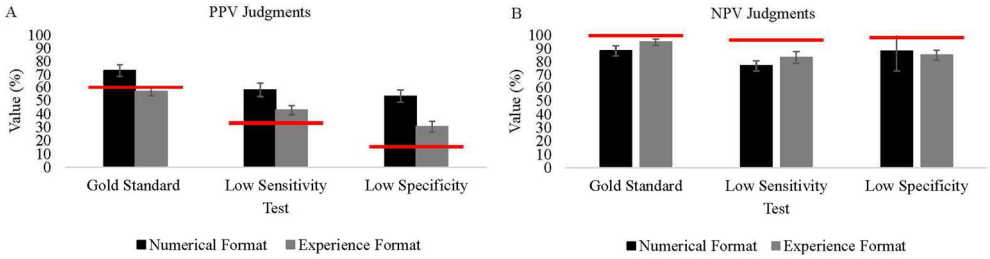
10. Cohen J. A power primer. *Psychol Bull.* 1992;112(1):155.

11. Erdfelder E, Faul, F, Buchner A. GPOWER: A general power analysis program. *Behav Res Methods Instrum Comput.* 1996;28(1):1-11.

Table 1. Demographics of Study Participants

	Age (years)	Gender Female : Male	Practice (years)
Residents (n=38)	27.7 (.25)	44% : 32%	1.4 (.08)
Clinicians (n=12)	35.2 (1.8)	18% : 6%	4.3 (1.6)

Note. Values represent means (standard errors in parentheses) for age and years of medical practice, and the proportion of female and male participants.



84x23mm (600 x 600 DPI)

BMJ Open

Does exposure to simulated patient cases improve accuracy of clinicians' predictive value estimates of diagnostic test results? A within-subjects experiment

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-019241.R1
Article Type:	Research
Date Submitted by the Author:	22-Nov-2017
Complete List of Authors:	Armstrong, Bonnie; Ryerson University, Psychology Spaniol, Julia; Ryerson University, Psychology Persaud, Nav; St Michael`s Hospital, Keenan Research Centre in the Li Ka Shing Knowledge Institute
Primary Subject Heading:	Medical education and training
Secondary Subject Heading:	Evidence based practice
Keywords:	Diagnostic inference, Experience-based learning, PPV, NPV, Estimate accuracy

Does exposure to simulated patient cases improve accuracy of clinicians' predictive value estimates of diagnostic test results? A within-subjects experiment

Bonnie A. Armstrong, MA (PhD Candidate)[1]

Julia Spaniol, PhD (Associate Professor)[1]

Nav Persaud, MSc, MD (Associate Scientist/Staff Physician)[2, 3]

1. Department of Psychology, Ryerson University, 350 Victoria Street, Toronto, Canada, M5B 2K3
2. Li Ka Shing Knowledge Institute, St Michael's Hospital, 209 Victoria Street, Toronto, Canada, M5B 1T8
3. Department of Family and Community Medicine, University of Toronto, 27 King's College Circle, Toronto, Canada, M5S 1A1

Author Note

Correspondence can be addressed to: Bonnie Armstrong, 350 Victoria Street, Ryerson University, Toronto, Canada, M5B 2K3. Email: bonnie.armstrong@psych.ryerson.ca.

Word Count: 1667

Abstract

Objective: Clinicians often overestimate the positive and negative predictive values (PPV, NPV) of diagnostic tests. The purpose of this study was to investigate whether experiencing simulated patient cases (i.e., an "experience format") would promote more accurate PPV and NPV estimates compared with a numerical format.

Design: Participants were presented with information about three diagnostic tests for the same fictitious disease, and were asked to estimate the PPV and NPV of each test. Tests varied with respect to sensitivity and specificity. Information about each test was presented once in the numerical format and once in the experience format. The study used a 2 (format: numerical vs. experience) x 3 (diagnostic test: gold standard vs. low sensitivity vs. low specificity) within-subjects design.

Setting: The study was completed online, via Qualtrics (Provo, UT).

Participants: 50 physicians (12 clinicians and 38 residents) from the Department of Family and Community Medicine at St. Michael's Hospital in Toronto, Canada completed the study. All participants had completed at least one year of residency.

Results: Estimation accuracy was quantified by the mean absolute error (MAE; absolute difference between estimate and true predictive value). PPV estimation errors were higher in the numerical format (*MAE*=32.6%; 95% CI: 26.8 to 38.4) compared to the experience format (*MAE* =15.9%; 95% CI: 11.8 to 20.0; $d = 0.697$, $P < .001$). Likewise, NPV estimation errors were higher in the numerical format (*MAE*=24.4%; 95% CI: 14.5 to 34.3) than in the experience format (*MAE*=11.0%; 95% CI: 6.5 to 15.5; $d = 0.303$; $P=.015$).

Conclusions: Exposure to simulated patient cases promotes accurate estimation of predictive values in clinicians. This finding carries implications for diagnostic training and practice.

Strength:

- The use of fictitious diseases and diagnostic tests provided information about performance that was not biased by respondents' prior knowledge about real diseases and tests.

Limitation:

- All participants were recruited from the Department of Community and Family Medicine at St. Michael's Hospital in Toronto, Canada. Future studies should replicate this research in other settings and with other populations.

Probabilistic reasoning is central to medical diagnosis [1–4]. Calculating or estimating the probability of a disease given a positive test result (positive predictive value; PPV), or the probability of no disease given a negative test result (negative predictive value, NPV) is notoriously difficult for clinicians, although commonly required for diagnostic inference [5-7]. Specifically, clinicians have difficulty understanding and applying test accuracy evidence to pre-test odds of disease [5-10]. Systematic errors include overestimation of the PPV and the NPV [5-10], which may have negative effects on patient care. Overestimation of the PPV, for example, increases the risk of overdiagnosis, which may lead to costly and harmful overtreatment such as unnecessary surgery or chemotherapy [11,12].

The accuracy of probabilistic inference has been shown to be sensitive to the format in which relevant statistics are presented [13-20]. The distinction between numerical and experience formats is most critical in the current context. In numerical formats, PPV and NPV estimates are based on numerical summaries of disease prevalence, test sensitivity, and test specificity or false-positive rates [5-8,14-20]. In so-called experience formats, in contrast, decision makers accrue information about the prevalence of disease and test reliability through exposure to representative patient cases whose true disease status and test outcome are revealed [21-25]. Thus, rather than manipulating statistical information to arrive at PPV and NPV estimates, decision makers must rely on their memory for previously-experienced patient scenarios (i.e., true and false positives and negatives) when estimating predictive values.

A series of studies suggests that experience formats may be superior to numerical formats in medical non-experts. An experience format led to greater sensitivity to the prevalence of genetic disease in unborn children, as well as a decreased subjective sense of worry about the

disease [21]. In another study, an experience format increased respondents' understanding of patients' knowledge of the risks and benefits of lung cancer screening [22].

We recently showed that both younger and older adults, regardless of numeracy skills, were more successful at estimating PPVs and NPVs for fictitious diagnostic tests when information was presented in an experience format, compared with when it was presented in a numerical format [23]. Similar findings were reported in a study comparing PPV estimates for a Down Syndrome screening [24].

In summary, there is significant evidence suggesting an advantage of experience over numerical formats in the context of diagnostic inference. However, all studies to date have been conducted with medical non-clinicians. In the current study, we sought to test whether the experience advantage would extend to clinicians. We predicted that, similar to laypeople, clinicians would provide more accurate estimates of the PPV and NPV after being exposed to relevant information in an experience format, compared to a numerical format. To test the robustness of the format effect across different types of diagnostic tests, participants provided estimates of PPV and NPV for 3 different fictitious diagnostic tests that differed in sensitivity and specificity.

Methods

Fifty clinicians affiliated with the Department of Community and Family Medicine from St. Michael's Hospital in Toronto, Canada, provided informed consent before completing a 1-hour online experiment via Qualtrics (Provo, UT), in which they received information about a fictitious disease and three separate fictitious diagnostic tests.

Information about each of the three tests was provided in a numerical format and an experience format. The numerical format was based on prior literature [5-8,14-20], and involved

reading a verbal passage describing the prevalence of a disease, as well as the sensitivity and the false-positive rate (i.e., 1 – specificity) of the diagnostic test. Numerical information was expressed in normalized frequencies, in which the base rate frequency was normalized to 100 (see Appendix for an example). In the experience format (see Figure 1), participants were presented with a slideshow of 100 representative patient cases. Each patient was characterized by a combination of disease status (does vs. does not have the disease) and diagnosis (positive vs. negative). The words "Has Disease" and "Positive Test Result" appeared in red, and the words "Does Not Have Disease" and "Negative Test Result" appeared in blue. Therefore, same-colour patient cases indicated a true test result (e.g., Has Disease and Positive Test Result), whereas different-colour patient cases indicated false test results (e.g., Has Disease and Negative Test Result). Each slide presented a single patient case for 3 seconds. Participants were instructed not to take notes.

In order to test the robustness of the format effect (numerical vs. experience) on the accuracy of PPV and NPV estimates, three separate diagnostic tests with varying test characteristics were used. The Gold Standard test had high sensitivity and high specificity; the Low Sensitivity test had low sensitivity but high specificity, and the Low Specificity test had high sensitivity but low specificity (see Table 1 for details). Each participant completed testing for all 6 combinations of format (numerical vs. experience) and test (gold standard vs. low sensitivity vs. low specificity), with presentation order counterbalanced across participants.

In both the numerical and experience formats, information for each test was presented for a total of 3 minutes before participants were prompted for estimates. Participants were not told that the three diagnostic tests were identical in both formats. PPV and NPV estimates were solicited using a frequency response format in which participants had to fill in both the

numerator and the denominator (e.g., "6 out of 98"). PPV and NPV estimate errors, defined as the absolute difference between true and estimated values, were submitted to separate 2 (format: numerical vs. experience) x 3 (test: gold standard vs. low sensitivity vs. low specificity) repeated-measures analyses of variance. Given the sample size (N=50) and the repeated-measures design, the statistical power to detect medium-sized effects [26], with an alpha of .05, was .93 for the "format" factor, and .98 for the "test" factor [27]. Statistical analysis was performed using SPSS, with alpha set to .05.

Results

Thirty-one female and 19 male clinicians completed the online study. The sample included 38 residents and 12 practicing clinicians. On average residents had completed 1.4 years of residency, and practicing clinicians had completed 4.3 years of practice.

Figure 2 presents mean estimates of the positive and NPVs for each format and test, along with the true predictive values. For PPV estimates, the mean absolute error rate (MAE) was higher in the numerical format (*MAE*=32.6%; 95% CI: 26.8 to 38.4) than in the experience format (*MAE*=15.9%; 95% CI: 11.8 to 20.0; $d = 0.697$, $P < .001$). As seen in Figure 2, the classic overestimation of the PPV was replicated when information was described numerically. In contrast, the extent to which PPVs were overestimated was reduced dramatically when information was experienced. For NPV estimates, the numerical format also produced larger errors (*MAE*=24.4%; 95% CI: 14.5 to 34.3) compared with the experience format (*MAE*=11.0%; 95% CI: 6.5 to 15.5; $d = 0.303$; $P=.015$), with less underestimation and reduced variability in estimates when information was experienced. For PPV and NPV estimates, the effect of format was stable across the three tests ($P=0.54$). There was also no effect of presentation order of

format (*P*=0.48) and no statistically significant difference between residents' and qualified clinicians' accuracy for either the PPV (*P*= 0.35) or the NPV (*P*=0.80).

Discussion

Compared to a numerical format, an experience format in which simulated patient cases were viewed over time produced more accurate PPV and NPV estimates in clinicians. The format effect was replicated across three separate diagnostic tests, demonstrating the robustness of the effect across variations of the problem. Critically, the experience format reduced overestimation of the PPV. Trainees and fully licensed clinicians commonly commit errors when making Bayesian inferences. Most notably, overestimating the PPV [5-10] can lead to a variety of negative consequences [11,12]. The current study thus adds to a growing literature demonstrating that the format in which decision-relevant information is presented influences predictive value estimates [13-20]. More specifically, the current data lends further support to the finding that experience formats boost diagnostic inference relative to numerical formats [21-25], and it extends this finding to a clinician population.

Why does the "experience advantage" occur? While the current study was not designed to address this question, there are several possible explanations. First, the experience format promotes an intuitive estimation strategy, requiring little in the way of statistical knowledge or active manipulation of numerical information. Second, the experience format presented participants with "natural frequencies" of the four possible diagnostic scenarios (i.e., the absolute number of true positives, false positives, true negatives, and false negatives). This is in contrast to the "normalized frequencies" presented in the numerical format. For example, in the numerical format, participants learned that the sensitivity of one of the tests was 83.33%. This number represents the relative frequency of true positive findings *among those with the disease*. In

contrast, in the experience format, participants encountered five true positives and one true negative in the slideshow of 100 patients. While both formats convey the same statistical information, the experience format may produce superior predictive value estimates because of its use of natural frequencies [5,13,16-20,28-30]. To what extent the strength of the experience format is due to the "slideshow" method that encourages intuitive responses, or from the use of natural as opposed to normalized frequencies, remains to be addressed in future work.

The current study shows that exposure to simulated patient cases is an effective technique for enhancing experts' predictive probability estimates without need for statistical training. Importantly, the experience format significantly reduced the common error of overestimating the PPV relative to the numerical format. Of note, the latter is commonly used in medical education and in real patient cases [1-4]. As discussed, more research is needed to shed light on the mechanisms underlying the experience advantage. In particular, it would be important to contrast the experience format with a numerical format in which decision-relevant information is presented in natural, rather than in normalized, frequencies [28-30]. Additional avenues for future research include studying the impact of experience formats on clinicians' treatment decisions and other clinical outcomes, and examining the viability of these formats for communicating test results to patients.

Required Manuscript Submission Statements

a) Details of contributors:

1. Bonnie Armstrong (study guarantor): study design, study programming, participant recruitment, data collection, data analysis, manuscript writing

2. Julia Spaniol: study design, manuscript writing

3. Nav Persaud: study design, study funder, participant recruitment, manuscript writing

All authors had full access to all of the data (including statistical reports and tables) in the study and can take responsibility for the integrity of the data and the accuracy of the data analysis.

b) Competing interests:

All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no support from any organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

c) Funding Statement:

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. NP was supported by the Department of Family and Community Medicine of St Michael's Hospital, the Department of Family and Community Medicine of the University of Toronto, an Early Researcher Award from the Ministry of Research and Innovation, and the

Physicians Services Incorporated Graham Farquharson Knowledge Translation Fellowship. The funders had no role in the study.

d) Data Sharing Statement:

The full dataset is available from the Dryad repository, DOI: [will include DOI once available from Dryad repository].

e) Ethics Approval Statement:

Ethics approval to conduct the current study was obtained from both St. Michael's Hospital Research Ethics Board (REB number: 16-282), as well as the Ryerson Ethics Board (REB number: 2014-129). All participants gave informed consent before participating in the study.

Acknowledgments

We thank Ryan Marinacci for his help with programming the online study, as well as Taehoon Lee and Anjli Bali for their help recruiting participants.

References

1. Kostopoulou O, Oudhoff J, Nath R, et al. Predictors of diagnostic accuracy and safe management in difficult diagnostic problems in family medicine. *Med Decis Making.* 2008;28:668-80.

2. Heneghan C, Glasziou P, Thompson M, et al. Diagnostic strategies used in primary care. *BMJ.* 2009;338-b946.

3. Dowie J, Elstein A, editors. *Professional judgment: a reader in clinical decision making.* Cambridge: Cambridge University Press;1988.

4. Falk G, Fahey T. Clinical prediction rules. *BMJ.* 2009;339:b2899.

5. Gigerenzer G, Gaissmaier W, Kurz-Milcke E, et al. Helping doctors and patients make sense of health statistics. *Psychol Sci Public Interest*. 2007;8:53-96.

6. Wegwarth O, Gigerenzer G. Statistical illiteracy in doctors. In: Gigerenzer G, Gray JA. (eds.) *Better doctors, better patients, better decisions: Envisioning health care 2020.* Cambridge: MIT Press;2011.p.137-51.

7. Anderson BL, Gigerenzer G, Parker S, et al. Statistical literacy in obstetricians and gynecologists. *J Healthc Qual.* 2012;36:5-17.

8. Lyman GH, Balducchi L. The effect of changing disease risk on clinical reasoning. *J Gen Intern Med.* 1994;9:488-95.

9. Whiting PF, Davenport C, Jameson C, et al. How well do health professionals interpret diagnostic information? A systematic review. *BMJ Open.* 2015;5:e008155.

10. Steurer J, Fischer JE, Bachmann LM, et al. Communicating accuracy of tests to general practitioners: a controlled study. *BMJ*. 2002;324:824-26.

11. Wegwarth O, Gigerenzer G. Overdiagnosis and overtreatment: evaluation of what physicians tell their patients about screening harms. *JAMA Intern Med.* 2013;173:2086-87.

12. Bhatt JR, Klotz L. Overtreatment in cancer – is it a problem? *Expert Opin Pharmacother.* 2016;17:1-5.

13. Gigerenzer G, Hoffrage U. How to improve Bayesian reasoning without instruction: frequency formats. *Psychol Rev.* 1995;102:684-704.

14. Akobeng AK. Understanding diagnostic tests 2: likelihood ratios, pre- and post-test probabilities and their use in clinical practice. *Acta Paediatr.* 2007;96:487-91.

15. Galesic M, Garcia-Retamero R, Gigerenzer G. Using icon arrays to communicate medical risks: overcoming low numeracy. *Health Psychol.* 2009;28:210-16.

16. Hoffrage U, Gigerenzer G. How to improve the diagnostic inferences of medical experts. In. Kurz-Milcke E, Gigerenzer G, eds. *Experts in Science and Society.* New York: Kluwer Academic/Plenum;2004;249-68.

17. Gigerenzer G. *Adaptive thinking: rationality in the real world.* New York: Oxford University Press;2000.

18. Gigerenzer G. What are natural frequencies? Doctors need to find better ways to communicate risk to patients. *BMJ.* 2011;343:d6386.

19. 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.

20. Hoffrage U, Gigerenzer G. Using natural frequencies to improve diagnostic inferences. *Acad Med.* 1998;73:538-40.

21. Tyszka T, Sawicki P. Affective and cognitive factors influencing sensitivity to probabilistic information. *Risk Anal.* 2011;31:1832-45.

22. Fraenkel L, Peters E, Tyra S, et al. Shared medical decision making in lung cancer screening: experienced versus descriptive risk formats. *Med Decis Making.* 2015;36:518-25.

23. Armstrong BA, Spaniol J. Experienced probabilities increase understanding of diagnostic test results in younger and older adults. *Med Decis Making.* 2017;37:670-79.

24. Wegier P, Shaffer VA. Aiding risk information learning through simulated experience (ARISE): using simulated outcomes to improve understanding of conditional probabilities in prenatal Down syndrome screening. *Patient Educ Couns.* 2017;100:1882-89.

25. Obrecht NA, Anderson B, Schulkin J, et al. Retrospective frequency formats promote consistent experience-based Bayesian estimates. *Appl Cognit Psychol.* 2012;26:436-40.

26. Cohen J. A power primer. *Psychol Bull.* 1992;112:155.

27. Erdfelder E, Faul, F, Buchner A. GPOWER: a general power analysis program. *Behav Res Methods Instrum Comput.* 1996;28:1-11.

28. Johnson ED, Tubau, E. Comprehension and computation in Bayesian problem solving. *Front Psychol.* 2015;6:1-19.

29. Gigerenzer G, Hoffrage U. The role of representation in Bayesian reasoning: correcting common misconceptions. *Behav. Brain Sci.* 2007;30:264-67.

30. Gigerenzer G, Hoffrage U. Overcoming difficulties in Bayesian reasoning: a reply to Lewis and Keren (1999) and Mellers and McGraw (1999). *Psychol. Rev.* 1999;106:425-30.

Table 1.

Test Characteristics

	Test Type		
Test Characteristics	Gold Standard	Low Sensitivity	Low Specificity
Prevalence	6%	6%	6%
Sensitivity	100%	50%	83.33%
Specificity	95.74%	93.62%	71.28%
False-Positive Rate	4.26%	6.38%	28.72%
PPV	60%	33.33%	15.63%
NPV	100%	96.70%	98.53%

Note. The prevalence of disease and all test characteristics are presented as percentages (i.e., normalized by a base-rate frequency of 100) reflecting what was presented in the numerical format. The joint event combinations (has vs. does not have disease and positive vs. negative test result) underlying the percentages were presented in the experience format.

Figure 1. A schematic of the experience format. 100 representative patient cases were viewed in the slideshow for each of the 3 tests. Each slide was presented for 3 seconds, and describes each patient in terms of disease status (i.e., has disease or does not have disease) and test result (negative or positive). “Has Disease” and “Positive Test Result” were shown in red font, and “Does Not Have Disease” and “Negative Test Result” were shown in blue font.

Figure 2. Mean PPV and NPV estimates for each format and test type. The x-axis displays the experimental factors (format x test) and y-axis displays mean estimate values. The grey bars represent mean estimates in the experience format. The black bars represent mean estimates in the numerical format. The red lines indicate the true PPVs and NPVs. Error bars for each mean represent standard errors.

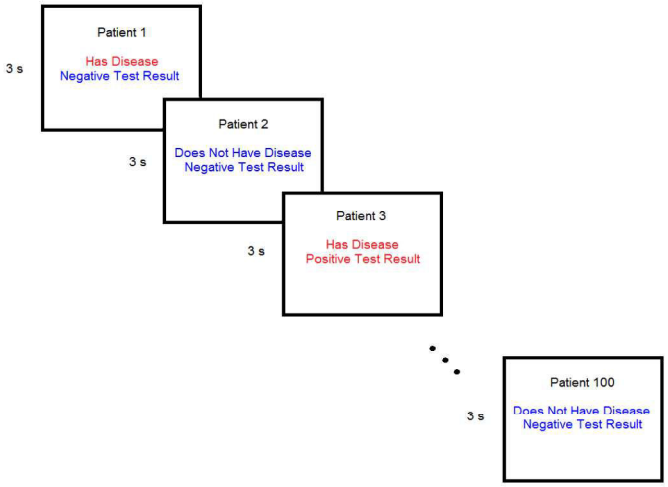


Figure 1. A schematic of the experience format.

254x190mm (300 x 300 DPI)

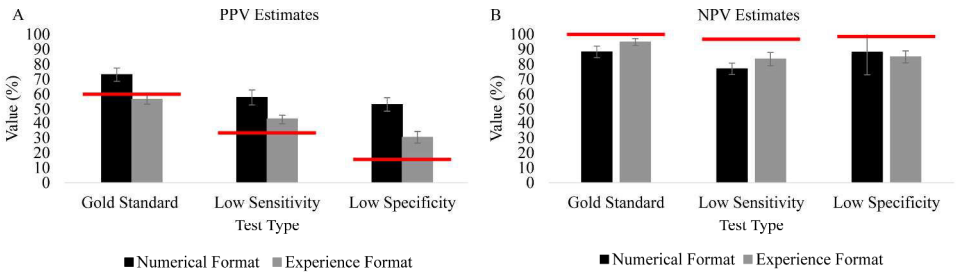


Figure 2. Mean PPV and NPV estimates for each format and test type.

254x190mm (300 x 300 DPI)

Appendix

Medical Screening Test
Disease X

To determine whether a person is at risk of Disease X, doctors sometimes conduct genetic testing. The passage below displays how common the disease is and how accurate the diagnostic test is.

6 out of every 100 people have Disease X.

If a person has Disease X, it is not certain whether he or she will have a positive result on the genetic test. More precisely, only 83.33 of every 100 such people will have a positive result on the genetic test.

If a person does not have Disease X, it is still possible that he or she will have a positive result on the genetic test. More precisely, 28.72 out of every 100 such people will have a positive result on the genetic test.

BMJ Open

Does exposure to simulated patient cases improve accuracy of clinicians' predictive value estimates of diagnostic test results? A within-subjects experiment at St. Michael's Hospital, Toronto, Canada

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-019241.R2
Article Type:	Research
Date Submitted by the Author:	01-Dec-2017
Complete List of Authors:	Armstrong, Bonnie; Ryerson University, Psychology Spaniol, Julia; Ryerson University, Psychology Persaud, Nav; St Michael`s Hospital, Keenan Research Centre in the Li Ka Shing Knowledge Institute
Primary Subject Heading:	Medical education and training
Secondary Subject Heading:	Evidence based practice
Keywords:	Diagnostic inference, Experience-based learning, PPV, NPV, Estimate accuracy

Does exposure to simulated patient cases improve accuracy of clinicians' predictive value estimates of diagnostic test results? A within-subjects experiment at St. Michael's Hospital, Toronto, Canada

Bonnie A. Armstrong, MA (PhD Candidate)[1]

Julia Spaniol, PhD (Associate Professor)[1]

Nav Persaud, MSc, MD (Associate Scientist/Staff Physician)[2, 3]

1. Department of Psychology, Ryerson University, 350 Victoria Street, Toronto, Canada, M5B 2K3
2. Li Ka Shing Knowledge Institute, St Michael's Hospital, 209 Victoria Street, Toronto, Canada, M5B 1T8
3. Department of Family and Community Medicine, University of Toronto, 27 King's College Circle, Toronto, Canada, M5S 1A1

Author Note

Correspondence can be addressed to: Bonnie Armstrong, 350 Victoria Street, Ryerson University, Toronto, Canada, M5B 2K3. Email: bonnie.armstrong@psych.ryerson.ca.

Word Count: 2530

Abstract

Objective: Clinicians often overestimate the probability of a disease given a positive test result (positive predictive value; PPV) and the probability of no disease given a negative test result (negative predictive value; NPV). The purpose of this study was to investigate whether experiencing simulated patient cases (i.e., an "experience format") would promote more accurate PPV and NPV estimates compared with a numerical format.

Design: Participants were presented with information about three diagnostic tests for the same fictitious disease, and were asked to estimate the PPV and NPV of each test. Tests varied with respect to sensitivity and specificity. Information about each test was presented once in the numerical format and once in the experience format. The study used a 2 (format: numerical vs. experience) x 3 (diagnostic test: gold standard vs. low sensitivity vs. low specificity) within-subjects design.

Setting: The study was completed online, via Qualtrics (Provo, UT).

Participants: 50 physicians (12 clinicians and 38 residents) from the Department of Family and Community Medicine at St. Michael's Hospital in Toronto, Canada completed the study. All participants had completed at least one year of residency.

Results: Estimation accuracy was quantified by the mean absolute error (*MAE*; absolute difference between estimate and true predictive value). PPV estimation errors were larger in the numerical format (*MAE*=32.6%; 95% CI: 26.8 to 38.4) compared to the experience format (*MAE* =15.9%; 95% CI: 11.8 to 20.0; $d = 0.697$, $P < .001$). Likewise, NPV estimation errors were larger in the numerical format (*MAE*=24.4%; 95% CI: 14.5 to 34.3) than in the experience format (*MAE*=11.0%; 95% CI: 6.5 to 15.5; $d = 0.303$; $P = .015$).

Conclusions: Exposure to simulated patient cases promotes accurate estimation of predictive values in clinicians. This finding carries implications for diagnostic training and practice.

Strength:

- The use of fictitious diseases and diagnostic tests provided information about performance that was not biased by respondents' prior knowledge about real diseases and tests.

Limitation:

- All participants were recruited from the Department of Community and Family Medicine at St. Michael's Hospital in Toronto, Canada. Future studies should replicate this research in other settings and with other populations.

Probabilistic reasoning is central to medical diagnosis [1–4]. Calculating or estimating the probability of a disease given a positive test result (positive predictive value; PPV), or the probability of no disease given a negative test result (negative predictive value, NPV) is notoriously difficult for clinicians, although commonly required for diagnostic inference [5-7]. Specifically, clinicians have difficulty understanding and applying test accuracy evidence to pre-test odds of disease [5-10]. Systematic errors include overestimation of the PPV and the NPV [5-10], which may have negative effects on patient care. Overestimation of the PPV, for example, may increase the risk of overtreatment such as unnecessary surgery or chemotherapy [11,12].

The accuracy of probabilistic inference has been shown to be sensitive to the format in which relevant statistics are presented [13-20]. The distinction between numerical and experience formats is most critical in the current context. In numerical formats, PPV and NPV estimates are based on numerical summaries of disease prevalence, test sensitivity (i.e., the proportion of patients with the disease who receive a positive test result [9]), and test specificity (i.e., the proportion of patients without the disease who receive a negative test result [9]) or false-positive rates [5-8,14-20]. In so-called experience formats, in contrast, decision makers accrue information about the prevalence of disease and test reliability through exposure to representative patient cases whose true disease status and test outcome are revealed [21-25]. Thus, rather than manipulating statistical information to arrive at PPV and NPV estimates, decision makers must rely on their memory for previously-experienced patient scenarios (i.e., true and false positives and negatives) when estimating predictive values.

A series of studies suggests that experience formats may be superior to numerical formats in non-experts. An experience format led to greater sensitivity to the prevalence of genetic disease in unborn children, as well as a decreased subjective sense of worry about the disease

[21]. In another study, an experience format increased patients' knowledge of the risks and benefits of lung cancer screening [22]. We recently showed that both younger and older adults, regardless of numeracy skills, were more successful at estimating PPVs and NPVs for fictitious diagnostic tests when information was presented in an experience format, compared with when it was presented in a numerical format [23]. Similar findings were reported in a study comparing PPV estimates for a Down Syndrome screening [24].

In summary, there is strong evidence suggesting an advantage of experience over numerical formats in the context of diagnostic inference. However, no study to date has tested this effect in clinicians. In the current study, we sought to test whether the experience advantage would extend to clinicians. We predicted that, similar to laypeople, clinicians would provide more accurate estimates of the PPV and NPV after being exposed to relevant information in an experience format, compared to a numerical format. To test the robustness of the format effect across different types of diagnostic tests, participants provided estimates of PPV and NPV for 3 different fictitious diagnostic tests that differed in sensitivity and specificity.

Methods

Fifty clinicians affiliated with the Department of Community and Family Medicine from St. Michael's Hospital in Toronto, Canada, provided informed consent before completing a 1-hour online experiment via Qualtrics (Provo, UT), in which they received information about a fictitious disease and three separate fictitious diagnostic tests.

Information about each of the three tests was provided in a numerical format and an experience format. The numerical format was based on prior literature [5-8,14-20], and involved reading a verbal passage describing the prevalence of a disease, as well as the sensitivity and the false-positive rate (i.e., 1 – specificity) of the diagnostic test. Numerical information was

expressed in normalized frequencies, in which the base rate frequency was normalized to 100 (see Figure 1, Panel A). In the experience format (see Figure 1, Panel B), participants were presented with a slideshow of 100 representative patient cases. Each patient was characterized by a combination of disease status (does vs. does not have the disease) and diagnosis (positive vs. negative). The words "Has Disease" and "Positive Test Result" appeared in red, and the words "Does Not Have Disease" and "Negative Test Result" appeared in blue. Therefore, same-colour patient cases indicated a true test result (e.g., Has Disease and Positive Test Result), whereas different-colour patient cases indicated false test results (e.g., Has Disease and Negative Test Result). Each slide presented a single patient case for 3 seconds. Participants were instructed not to take notes.

In order to test the robustness of the format effect (numerical vs. experience) on the accuracy of PPV and NPV estimates, three separate diagnostic tests with varying test characteristics were used. The Gold Standard test had high sensitivity and high specificity; the Low Sensitivity test had low sensitivity but high specificity, and the Low Specificity test had high sensitivity but low specificity (see Table 1 for details). Each participant completed testing for all 6 combinations of format (numerical vs. experience) and test (gold standard vs. low sensitivity vs. low specificity). Presentation order was counterbalanced, such that half of the participants completed the scenarios in the numerical format first (with test order counterbalanced across participants), followed by the scenarios in the experience format (with test order once again counterbalanced). The other half of participants received the reverse order (experience then numerical). Participants were not told that the three diagnostic tests were identical in both formats.

In both the numerical and experience formats, information for each test was presented for a total of 3 minutes before participants were prompted for estimates, specifically "how many patients had the disease, out of all patients who received a positive test result" (PPV) and "how many patients did not have the disease, out of all patients who received a negative test result" (NPV).

PPV and NPV estimates were solicited using a frequency response format in which participants had to fill in both the numerator and the denominator (e.g., "6 out of 98"). PPV and NPV estimate errors, defined as the absolute difference between true and estimated values, were submitted to separate 2 (format: numerical vs. experience) x 3 (test: gold standard vs. low sensitivity vs. low specificity) repeated-measures analyses of variance. Given the sample size (N=50) and the repeated-measures design, the statistical power to detect medium-sized effects [26], with an alpha of .05, was .93 for the "format" factor, and .98 for the "test" factor [27]. Statistical analysis was performed using SPSS, with alpha set to .05.

Results

Thirty-one female and 19 male clinicians completed the online study. The sample included 38 residents and 12 practicing clinicians. On average residents had completed 1.4 years of residency, and practicing clinicians had completed 4.3 years of practice.

As a measure of task performance, mean absolute estimation errors (*MAE*) are reported. Low *MAE* values indicate more accurate estimates [23]. We chose *MAE* over alternative performance measures (e.g., % respondents with responses close to the true value) because the *MAE* provides fine-grained information about the distance between estimates and true values. Because *MAE* does not distinguish between under- and overestimation, Figure 2 additionally shows the mean raw PPV and NPV estimates for each experimental condition, as well as the true

values. For PPV estimates, errors were larger in the numerical format (*MAE*=32.6%; 95% CI: 26.8 to 38.4) than in the experience format (*MAE*=15.9%; 95% CI: 11.8 to 20.0; $d = 0.697$, $P<.001$). As seen in Figure 2, the classic overestimation of the PPV was replicated when information was described numerically. In contrast, the extent to which PPVs were overestimated was reduced dramatically when information was experienced. For NPV estimates, the numerical format also produced larger errors (*MAE*=24.4%; 95% CI: 14.5 to 34.3) compared with the experience format (*MAE*=11.0%; 95% CI: 6.5 to 15.5; $d = 0.303$; $P=.015$), with less underestimation and reduced variability in estimates when information was experienced. For PPV and NPV estimates, the effect of format was stable across the three tests ($P=0.54$). There was also no effect of presentation order of format ($P=0.48$) and no statistically significant difference between residents' and qualified clinicians' accuracy for either the PPV ($P= 0.35$) or the NPV ($P=0.80$).

Discussion

Compared to a numerical format, an experience format in which simulated patient cases were viewed over time produced more accurate PPV and NPV estimates in clinicians. The format effect was replicated across three separate diagnostic tests, demonstrating the robustness of the effect across variations of the problem. Critically, the experience format reduced overestimation of the PPV. Trainees and fully licensed clinicians commonly commit errors when making Bayesian inferences. Most notably, overestimating the PPV [5-10] can lead to a variety of negative consequences [11,12]. The current study thus adds to a growing literature demonstrating that the format in which decision-relevant information is presented influences predictive value estimates [13-20]. More specifically, the current data lends further support to the

finding that experience formats boost diagnostic inference relative to numerical formats [21-25], and it extends this finding to a clinician population.

Why does the "experience advantage" occur? While the current study was not designed to address this question, there are several possible explanations. First, the experience format promotes an intuitive estimation strategy, requiring little in the way of statistical knowledge or active manipulation of numerical information. Second, the experience format presented participants with naturally occurring frequencies of the four possible diagnostic scenarios (i.e., the absolute number of true positives, false positives, true negatives, and false negatives). This is in contrast to the "normalized frequencies" presented in the numerical format. For example, in the numerical format, participants learned that the sensitivity of one of the tests was 83.33%. This number represents the relative frequency of true positive findings *among those with the disease*. In contrast, in the experience format, participants encountered five true positives and one true negative in the slideshow of 100 patients, and could subsequently derive subjective natural frequency values based on memory of the patient cases. While both formats convey the same statistical information, the experience format may produce superior predictive value estimates because of its use of naturally occurring frequencies [5,13,16-20,28-30]. To what extent the strength of the experience format is due to the "slideshow" method that encourages intuitive responses, or from the use of natural as opposed to normalized frequencies, remains to be addressed in future work.

The current study shows that exposure to simulated patient cases is an effective technique for enhancing experts' predictive probability estimates without need for statistical training. Importantly, the experience format significantly reduced the common error of overestimating the PPV relative to the numerical format. Of note, the latter is commonly used in medical education

and in real patient cases [1-4]. As discussed, more research is needed to shed light on the mechanisms underlying the experience advantage. In particular, it would be important to contrast the experience format with a numerical format in which decision-relevant information is presented in natural, rather than in normalized, frequencies [28-30]. Additional avenues for future research include studying the impact of experience formats on clinicians' treatment decisions and other clinical outcomes, and examining the viability of these formats for communicating test results to patients.

Required Manuscript Submission Statements

a) Details of contributors:

1. Bonnie Armstrong (study guarantor): study design, study programming, participant recruitment, data collection, data analysis, manuscript writing

2. Julia Spaniol: study design, manuscript writing

3. Nav Persaud: study design, study funder, participant recruitment, manuscript writing

All authors had full access to all of the data (including statistical reports and tables) in the study and can take responsibility for the integrity of the data and the accuracy of the data analysis.

b) Competing interests:

All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no support from any organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

c) Funding Statement:

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. NP was supported by the Department of Family and Community Medicine of St Michael's Hospital, the Department of Family and Community Medicine of the University of Toronto, an Early Researcher Award from the Ministry of Research and Innovation, and the

Physicians Services Incorporated Graham Farquharson Knowledge Translation Fellowship. The funders had no role in the study.

d) Data Sharing Statement:

The full dataset is available from the Dryad repository, DOI: [will include DOI once available from Dryad repository].

e) Ethics Approval Statement:

Ethics approval to conduct the current study was obtained from both St. Michael's Hospital Research Ethics Board (REB number: 16-282), as well as the Ryerson Ethics Board (REB number: 2014-129). All participants gave informed consent before participating in the study.

Acknowledgments

We thank Ryan Marinacci for his help with programming the online study, as well as Taehoon Lee and Anjli Bali for their help recruiting participants.

References

1. Kostopoulou O, Oudhoff J, Nath R, et al. Predictors of diagnostic accuracy and safe management in difficult diagnostic problems in family medicine. *Med Decis Making* 2008;28:668-80.

2. Heneghan C, Glasziou P, Thompson M, et al. Diagnostic strategies used in primary care. *BMJ* 2009;338:b946.

3. Dowie J, Elstein A, editors. *Professional judgment: a reader in clinical decision making.* Cambridge: Cambridge University Press;1988.

4. Falk G, Fahey T. Clinical prediction rules. *BMJ* 2009;339:b2899.

5. Gigerenzer G, Gaissmaier W, Kurz-Milcke E, et al. Helping doctors and patients make sense of health statistics. *Psychol Sci Public Interest* 2007;8:53-96.

6. Wegwarth O, Gigerenzer G. Statistical illiteracy in doctors. In: Gigerenzer G, Gray JA. (eds.) *Better doctors, better patients, better decisions: Envisioning health care 2020.* Cambridge: MIT Press;2011.p.137-51.

7. Anderson BL, Gigerenzer G, Parker S, et al. Statistical literacy in obstetricians and gynecologists. *J Healthc Qual* 2012;36:5-17.

8. Lyman GH, Balducchi L. The effect of changing disease risk on clinical reasoning. *J Gen Intern Med* 1994;9:488-95.

9. Whiting PF, Davenport C, Jameson C, et al. How well do health professionals interpret diagnostic information? A systematic review. *BMJ Open* 2015;5:e008155.

10. Steurer J, Fischer JE, Bachmann LM, et al. Communicating accuracy of tests to general practitioners: a controlled study. *BMJ* 2002;324:824-26.

11. Wegwarth O, Gigerenzer G. Overdiagnosis and overtreatment: evaluation of what physicians tell their patients about screening harms. *JAMA Intern Med* 2013;173:2086-87.

12. Bhatt JR, Klotz L. Overtreatment in cancer – is it a problem? *Expert Opin Pharmacother* 2016;17:1-5.

13. Gigerenzer G, Hoffrage U. How to improve Bayesian reasoning without instruction: frequency formats. *Psychol Rev* 1995;102:684-704.

14. Akobeng AK. Understanding diagnostic tests 2: likelihood ratios, pre- and post-test probabilities and their use in clinical practice. *Acta Paediatr* 2007;96:487-91.

15. Galesic M, Garcia-Retamero R, Gigerenzer G. Using icon arrays to communicate medical risks: overcoming low numeracy. *Health Psychol* 2009;28:210-16.

16. Hoffrage U, Gigerenzer G. How to improve the diagnostic inferences of medical experts. In. Kurz-Milcke E, Gigerenzer G, eds. *Experts in Science and Society.* New York: Kluwer Academic/Plenum;2004;249-68.

17. Gigerenzer G. *Adaptive thinking: rationality in the real world.* New York: Oxford University Press;2000.

18. Gigerenzer G. What are natural frequencies? Doctors need to find better ways to communicate risk to patients. *BMJ* 2011;343:d6386.

19. 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.

20. Hoffrage U, Gigerenzer G. Using natural frequencies to improve diagnostic inferences. *Acad Med* 1998;73:538-40.

21. Tyszka T, Sawicki P. Affective and cognitive factors influencing sensitivity to probabilistic information. *Risk Anal* 2011;31:1832-45.

22. Fraenkel L, Peters E, Tyra S, et al. Shared medical decision making in lung cancer screening: experienced versus descriptive risk formats. *Med Decis Making* 2015;36:518-25.

23. Armstrong BA, Spaniol J. Experienced probabilities increase understanding of diagnostic test results in younger and older adults. *Med Decis Making* 2017;37:670-79.

24. Wegier P, Shaffer VA. Aiding risk information learning through simulated experience (ARISE): using simulated outcomes to improve understanding of conditional probabilities in prenatal Down syndrome screening. *Patient Educ Couns* 2017;100:1882-89.

25. Obrecht NA, Anderson B, Schulkin J, et al. Retrospective frequency formats promote consistent experience-based Bayesian estimates. *Appl Cognit Psychol* 2012;26:436-40.

26. Cohen J. A power primer. *Psychol Bull* 1992;112:155.

27. Erdfelder E, Faul, F, Buchner A. GPOWER: a general power analysis program. *Behav Res Methods Instrum Comput* 1996;28:1-11.

28. Johnson ED, Tubau, E. Comprehension and computation in Bayesian problem solving. *Front Psychol* 2015;6:1-19.

29. Gigerenzer G, Hoffrage U. The role of representation in Bayesian reasoning: correcting common misconceptions. *Behav. Brain Sci* 2007;30:264-67.

30. Gigerenzer G, Hoffrage U. Overcoming difficulties in Bayesian reasoning: a reply to Lewis and Keren (1999) and Mellers and McGraw (1999). *Psychol. Rev* 1999;106:425-30.

Table 1.

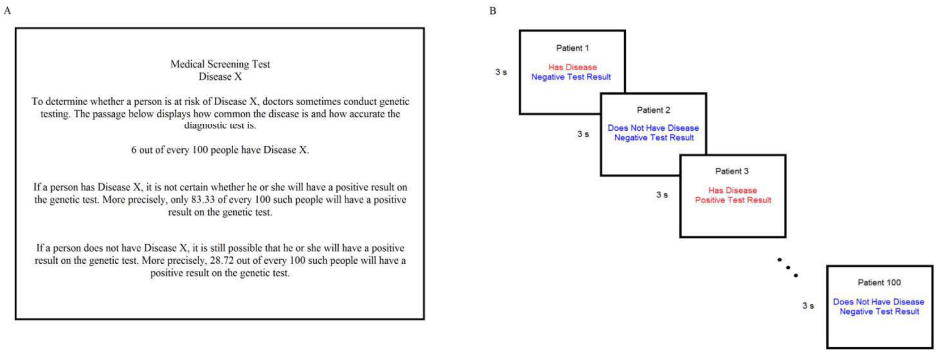
Test Characteristics

	Test Type		
Test Characteristics	Gold Standard	Low Sensitivity	Low Specificity
Prevalence	6%	6%	6%
Sensitivity	100%	50%	83.33%
Specificity	95.74%	93.62%	71.28%
False-Positive Rate	4.26%	6.38%	28.72%
PPV	60%	33.33%	15.63%
NPV	100%	96.70%	98.53%

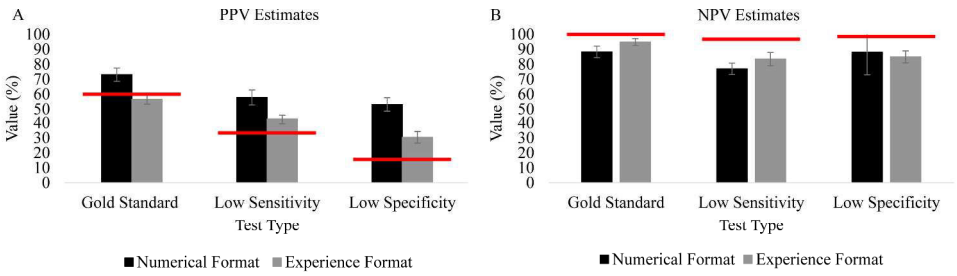
Note. The prevalence of disease and all test characteristics are presented as percentages (i.e., normalized by a base-rate frequency of 100) reflecting what was presented in the numerical format. The joint event combinations (has vs. does not have disease and positive vs. negative test result) underlying the percentages were presented in the experience format.

Figure 1. Panel A is an example of the numerical format, and Panel B is an example of the experience format. The numerical format provides the prevalence of disease, as well as the sensitivity and the false-positive rate of the diagnostic test. In the experience format, 100 representative patient cases were viewed in the slideshow for each of the 3 tests. Each slide was presented for 3 seconds, and describes each patient in terms of disease status (i.e., has disease or does not have disease) and test result (negative or positive). “Has Disease” and “Positive Test Result” were shown in red font, and “Does Not Have Disease” and “Negative Test Result” were shown in blue font.

Figure 2. Mean PPV and NPV estimates for each format and test type. The x-axis displays the experimental factors (format x test) and y-axis displays mean estimate values. The grey bars represent mean estimates in the experience format. The black bars represent mean estimates in the numerical format. The red lines indicate the true PPVs and NPVs. Error bars for each mean represent standard errors.



254x190mm (300 x 300 DPI)



254x190mm (300 x 300 DPI)

BMJ Open

Does exposure to simulated patient cases improve accuracy of clinicians' predictive value estimates of diagnostic test results? A within-subjects experiment at St. Michael's Hospital, Toronto, Canada

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-019241.R3
Article Type:	Research
Date Submitted by the Author:	13-Dec-2017
Complete List of Authors:	Armstrong, Bonnie; Ryerson University, Psychology Spaniol, Julia; Ryerson University, Psychology Persaud, Nav; St Michael`s Hospital, Keenan Research Centre in the Li Ka Shing Knowledge Institute
Primary Subject Heading:	Medical education and training
Secondary Subject Heading:	Evidence based practice
Keywords:	Diagnostic inference, Experience-based learning, PPV, NPV, Estimate accuracy

Does exposure to simulated patient cases improve accuracy of clinicians' predictive value estimates of diagnostic test results? A within-subjects experiment at St. Michael's Hospital, Toronto, Canada

Bonnie A. Armstrong, MA (PhD Candidate)[1]

Julia Spaniol, PhD (Associate Professor)[1]

Nav Persaud, MSc, MD (Associate Scientist/Staff Physician)[2, 3]

1. Department of Psychology, Ryerson University, 350 Victoria Street, Toronto, Canada, M5B 2K3
2. Li Ka Shing Knowledge Institute, St Michael's Hospital, 209 Victoria Street, Toronto, Canada, M5B 1T8
3. Department of Family and Community Medicine, University of Toronto, 27 King's College Circle, Toronto, Canada, M5S 1A1

Author Note

Correspondence can be addressed to: Bonnie Armstrong, 350 Victoria Street, Ryerson University, Toronto, Canada, M5B 2K3. Email: bonnie.armstrong@psych.ryerson.ca.

Word Count: 2109

Abstract

Objective: Clinicians often overestimate the probability of a disease given a positive test result (positive predictive value; PPV) and the probability of no disease given a negative test result (negative predictive value; NPV). The purpose of this study was to investigate whether experiencing simulated patient cases (i.e., an "experience format") would promote more accurate PPV and NPV estimates compared with a numerical format.

Design: Participants were presented with information about three diagnostic tests for the same fictitious disease, and were asked to estimate the PPV and NPV of each test. Tests varied with respect to sensitivity and specificity. Information about each test was presented once in the numerical format and once in the experience format. The study used a 2 (format: numerical vs. experience) x 3 (diagnostic test: gold standard vs. low sensitivity vs. low specificity) within-subjects design.

Setting: The study was completed online, via Qualtrics (Provo, UT).

Participants: 50 physicians (12 clinicians and 38 residents) from the Department of Family and Community Medicine at St. Michael's Hospital in Toronto, Canada completed the study. All participants had completed at least one year of residency.

Results: Estimation accuracy was quantified by the mean absolute error (*MAE*; absolute difference between estimate and true predictive value). PPV estimation errors were larger in the numerical format (*MAE*=32.6%; 95% CI: 26.8 to 38.4) compared to the experience format (*MAE* =15.9%; 95% CI: 11.8 to 20.0; $d = 0.697$, $P < .001$). Likewise, NPV estimation errors were larger in the numerical format (*MAE*=24.4%; 95% CI: 14.5 to 34.3) than in the experience format (*MAE*=11.0%; 95% CI: 6.5 to 15.5; $d = 0.303$; $P = .015$).

Conclusions: Exposure to simulated patient cases promotes accurate estimation of predictive values in clinicians. This finding carries implications for diagnostic training and practice.

Strengths:

- The use of fictitious diseases and diagnostic tests provided information about performance that was not biased by participants' prior knowledge about real diseases and tests.
- Three separate diagnostic tests that varied in sensitivity and specificity were presented in each format, within-subjects, in order to show the robustness of the format effect.

Limitations:

- All participants were recruited from the Department of Community and Family Medicine at St. Michael's Hospital in Toronto, Canada. Future studies should replicate this research in other settings and with other populations.
- The study was conducted online, which may affect the ecological validity of the results.

Probabilistic reasoning is central to medical diagnosis [1–4]. Calculating or estimating the probability of a disease given a positive test result (positive predictive value; PPV), or the probability of no disease given a negative test result (negative predictive value, NPV) is notoriously difficult for clinicians, although commonly required for diagnostic inference [5-7]. Specifically, clinicians have difficulty understanding and applying test accuracy evidence to pre-test odds of disease [5-10]. Systematic errors include overestimation of the PPV and the NPV [5-10], which may have negative effects on patient care. Overestimation of the PPV, for example, may increase the risk of overtreatment such as unnecessary surgery or chemotherapy [11,12].

The accuracy of probabilistic inference has been shown to be sensitive to the format in which relevant statistics are presented [13-20]. The distinction between numerical and experience formats is most critical in the current context. In numerical formats, PPV and NPV estimates are based on numerical summaries of disease prevalence, test sensitivity (i.e., the proportion of patients with the disease who receive a positive test result [9]), and test specificity (i.e., the proportion of patients without the disease who receive a negative test result [9]) or false-positive rates [5-8,14-20]. In so-called experience formats, in contrast, decision makers accrue information about the prevalence of disease and test reliability through exposure to representative patient cases whose true disease status and test outcome are revealed [21-25]. Thus, rather than manipulating statistical information to arrive at PPV and NPV estimates, decision makers must rely on their memory for previously-experienced patient scenarios (i.e., true and false positives and negatives) when estimating predictive values.

A series of studies suggests that experience formats may be superior to numerical formats in non-experts. An experience format led to greater sensitivity to the prevalence of genetic disease in unborn children, as well as a decreased subjective sense of worry about the disease

[21]. In another study, an experience format increased patients' knowledge of the risks and benefits of lung cancer screening [22]. We recently showed that both younger and older adults, regardless of numeracy skills, were more successful at estimating PPVs and NPVs for fictitious diagnostic tests when information was presented in an experience format, compared with when it was presented in a numerical format [23]. Similar findings were reported in a study comparing PPV estimates for a Down Syndrome screening [24].

In summary, there is strong evidence suggesting an advantage of experience over numerical formats in the context of diagnostic inference. However, no study to date has tested this effect in clinicians. In the current study, we sought to test whether the experience advantage would extend to clinicians. We predicted that, similar to laypeople, clinicians would provide more accurate estimates of the PPV and NPV after being exposed to relevant information in an experience format, compared to a numerical format. To test the robustness of the format effect, participants provided estimates of the PPV and NPV for 3 different fictitious diagnostic tests that differed in sensitivity and specificity.

Methods

Fifty clinicians affiliated with the Department of Community and Family Medicine from St. Michael's Hospital in Toronto, Canada, provided informed consent before completing a 1-hour online experiment via Qualtrics (Provo, UT), in which they received information about a fictitious disease and three separate fictitious diagnostic tests.

Information about each of the three tests was provided in a numerical format and an experience format. The numerical format was based on prior literature [5-8,14-20], and involved reading a verbal passage describing the prevalence of a disease, as well as the sensitivity and the false-positive rate (i.e., 1 – specificity) of the diagnostic test. Numerical information was

expressed in normalized frequencies, in which the base rate frequency was normalized to 100 (see Figure 1, Panel A). In the experience format (see Figure 1, Panel B), participants were presented with a slideshow of 100 representative patient cases. Each patient was characterized by a combination of disease status (does vs. does not have the disease) and diagnosis (positive vs. negative). The words "Has Disease" and "Positive Test Result" appeared in red, and the words "Does Not Have Disease" and "Negative Test Result" appeared in blue. Therefore, same-colour patient cases indicated a true test result (e.g., Has Disease and Positive Test Result), whereas different-colour patient cases indicated false test results (e.g., Has Disease and Negative Test Result). Each slide presented a single patient case for 3 seconds. Participants were instructed not to take notes.

In order to test the robustness of the format effect (numerical vs. experience) on the accuracy of PPV and NPV estimates, three separate diagnostic tests with varying test characteristics were used. The Gold Standard test had high sensitivity and high specificity; the Low Sensitivity test had low sensitivity but high specificity, and the Low Specificity test had high sensitivity but low specificity (see Table 1 for details). Each participant completed testing for all 6 combinations of format (numerical vs. experience) and test (gold standard vs. low sensitivity vs. low specificity). Presentation order was counterbalanced, such that half of the participants completed the scenarios in the numerical format first (with test order counterbalanced across participants), followed by the scenarios in the experience format (with test order once again counterbalanced). The other half of participants received the reverse order (experience then numerical). Participants were not told that the three diagnostic tests were identical in both formats.

In both the numerical and experience formats, information for each test was presented for a total of 3 minutes before participants were prompted for estimates, specifically "how many patients had the disease, out of all patients who received a positive test result" (PPV) and "how many patients did not have the disease, out of all patients who received a negative test result" (NPV).

PPV and NPV estimates were solicited using a frequency response format in which participants had to fill in both the numerator and the denominator (e.g., "6 out of 98"). PPV and NPV estimate errors, defined as the absolute difference between true and estimated values, were submitted to separate 2 (format: numerical vs. experience) x 3 (test: gold standard vs. low sensitivity vs. low specificity) repeated-measures analyses of variance. Given the sample size (N=50) and the repeated-measures design, the statistical power to detect medium-sized effects [26], with an alpha of .05, was .93 for the "format" factor, and .98 for the "test" factor [27]. Statistical analysis was performed using SPSS, with alpha set to .05.

Results

Thirty-one female and 19 male clinicians completed the online study. The sample included 38 residents and 12 practicing clinicians. On average residents had completed 1.4 years of residency, and practicing clinicians had completed 4.3 years of practice.

As a measure of task performance, mean absolute estimation errors (*MAE*) are reported. Low *MAE* values indicate more accurate estimates [23]. We chose *MAE* over alternative performance measures (e.g., % of participants with responses close to the true value) because the *MAE* provides fine-grained information about the distance between estimates and true values. Because *MAE* does not distinguish between under- and overestimation, Figure 2 additionally shows the mean raw PPV (Panel A) and NPV (Panel B) estimates for each experimental

condition, as well as the true values. For PPV estimates, errors were larger in the numerical format (*MAE*=32.6%; 95% CI: 26.8 to 38.4) than in the experience format (*MAE*=15.9%; 95% CI: 11.8 to 20.0; $d = 0.697$, $P < .001$). As seen in Figure 2 (Panel A), the classic overestimation of the PPV was replicated when information was described numerically. In contrast, the extent to which PPVs were overestimated was reduced dramatically when information was experienced. For NPV estimates, the numerical format also produced larger errors (*MAE*=24.4%; 95% CI: 14.5 to 34.3) compared with the experience format (*MAE*=11.0%; 95% CI: 6.5 to 15.5; $d = 0.303$; $P = .015$), with less underestimation and reduced variability in estimates when information was experienced (see Panel B). For PPV and NPV estimates, the effect of format was stable across the three tests ($P = 0.54$). There was also no effect of presentation order of format ($P = 0.48$) and no statistically significant difference between residents' and qualified clinicians' accuracy for either the PPV ($P = 0.35$) or the NPV ($P = 0.80$).

Discussion

Compared to a numerical format, an experience format in which simulated patient cases were viewed over time produced more accurate PPV and NPV estimates in clinicians. The format effect was replicated across three separate diagnostic tests, demonstrating the robustness of the effect across variations of the problem. Critically, the experience format reduced overestimation of the PPV. Trainees and fully licensed clinicians commonly commit errors when making Bayesian inferences. Most notably, overestimating the PPV [5-10] can lead to a variety of negative consequences [11,12]. The current study thus adds to a growing literature demonstrating that the format in which decision-relevant information is presented influences predictive value estimates [13-20]. More specifically, the current data lends further support to the

finding that experience formats boost diagnostic inference relative to numerical formats [21-25], and it extends this finding to a clinician population.

Why does the "experience advantage" occur? While the current study was not designed to address this question, there are several possible explanations. First, the experience format promotes an intuitive estimation strategy, requiring little in the way of statistical knowledge or active manipulation of numerical information. Second, the experience format presented participants with naturally occurring frequencies of the four possible diagnostic scenarios (i.e., the absolute number of true positives, false positives, true negatives, and false negatives). This is in contrast to the "normalized frequencies" presented in the numerical format. For example, in the numerical format, participants learned that the sensitivity of one of the tests was 83.33%. This number represents the relative frequency of true positive findings *among those with the disease*. In contrast, in the experience format, participants encountered five true positives and one true negative in the slideshow of 100 patients, and could subsequently derive subjective natural frequency values based on memory of the patient cases. While both formats convey the same statistical information, the experience format may produce superior predictive value estimates because of its use of naturally occurring frequencies [5,13,16-20,28-30]. To what extent the strength of the experience format is due to the "slideshow" method that encourages intuitive responses, or from the use of natural as opposed to normalized frequencies, remains to be addressed in future work.

There are both strengths and weaknesses of the current study. A main strength is that we controlled for the potential confound of prior knowledge through the use of fictitious information. Previous research has investigated clinicians' probability estimates for real diseases and tests [5,7,10]. However, knowledge of medical statistics, such as disease prevalence or test

sensitivity and specificity, may have influenced clinicians' estimates. Results presented here demonstrate the effect of format on clinicians' estimate accuracy more cleanly. Another important strength of the study is that participants were shown information for three separate diagnostic tests, varying in sensitivity and specificity, presented in both formats within-subjects. The purpose of this design was to demonstrate the stability of the format effect across individuals, as well as different versions of the problem (i.e., for reliable and unreliable diagnostic tests that are subject to different types of errors such as false alarms or misses). The findings of the study illustrate the robustness of the format effect. An important limitation of the study is that the sample includes clinicians from one discipline (Family and Community Medicine) from the same hospital, restricting the generalizability of the results. A second limitation is that the study was conducted online, which may affect the ecological validity of the study findings because the experimental setting cannot be fully controlled by experimenters. For example, participants may have had different browser experiences, or distractions in the physical environment. Future studies should test the effect of format on medical experts' probability estimates in more controlled settings (e.g., an in-lab environment).

The current study shows that exposure to simulated patient cases is an effective technique for enhancing experts' predictive probability estimates without need for statistical training. Importantly, the experience format significantly reduced the common error of overestimating the PPV relative to the numerical format. Of note, the latter is commonly used in medical education and in real patient cases [1-4]. As discussed, more research is needed to shed light on the mechanisms underlying the experience advantage. In particular, it would be important to contrast the experience format with a numerical format in which decision-relevant information is presented in natural, rather than in normalized, frequencies [28-30]. Additional avenues for

future research include studying the impact of experience formats on clinicians' treatment decisions and other clinical outcomes across a variety of medical disciplines, and examining the viability of these formats for communicating test results to patients.

Required Manuscript Submission Statements

a) Details of contributors:

1. Bonnie Armstrong (study guarantor): study design, study programming, participant recruitment, data collection, data analysis, manuscript writing

2. Julia Spaniol: study design, manuscript writing

3. Nav Persaud: study design, study funder, participant recruitment, manuscript writing

All authors had full access to all of the data (including statistical reports and tables) in the study and can take responsibility for the integrity of the data and the accuracy of the data analysis.

b) Competing interests:

All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no support from any organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

c) Funding Statement:

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. NP was supported by the Department of Family and Community Medicine of St Michael's Hospital, the Department of Family and Community Medicine of the University of Toronto, an Early Researcher Award from the Ministry of Research and Innovation, and the

Physicians Services Incorporated Graham Farquharson Knowledge Translation Fellowship. The funders had no role in the study.

d) Data Sharing Statement:

The full dataset is available from the Dryad repository, DOI: [will include DOI once available from Dryad repository].

e) Ethics Approval Statement:

Ethics approval to conduct the current study was obtained from both St. Michael's Hospital Research Ethics Board (REB number: 16-282), as well as the Ryerson Ethics Board (REB number: 2014-129). All participants gave informed consent before participating in the study.

Acknowledgments

We thank Ryan Marinacci for his help with programming the online study, as well as Taehoon Lee and Anjli Bali for their help recruiting participants.

References

1. Kostopoulou O, Oudhoff J, Nath R, et al. Predictors of diagnostic accuracy and safe management in difficult diagnostic problems in family medicine. *Med Decis Making* 2008;28:668-80.

2. Heneghan C, Glasziou P, Thompson M, et al. Diagnostic strategies used in primary care. *BMJ* 2009;338:b946.

3. Dowie J, Elstein A, editors. *Professional judgment: a reader in clinical decision making.* Cambridge: Cambridge University Press;1988.

4. Falk G, Fahey T. Clinical prediction rules. *BMJ* 2009;339:b2899.

5. Gigerenzer G, Gaissmaier W, Kurz-Milcke E, et al. Helping doctors and patients make sense of health statistics. *Psychol Sci Public Interest* 2007;8:53-96.

6. Wegwarth O, Gigerenzer G. Statistical illiteracy in doctors. In: Gigerenzer G, Gray JA. (eds.) *Better doctors, better patients, better decisions: Envisioning health care 2020.* Cambridge: MIT Press;2011.p.137-51.

7. Anderson BL, Gigerenzer G, Parker S, et al. Statistical literacy in obstetricians and gynecologists. *J Healthc Qual* 2012;36:5-17.

8. Lyman GH, Balducchi L. The effect of changing disease risk on clinical reasoning. *J Gen Intern Med* 1994;9:488-95.

9. Whiting PF, Davenport C, Jameson C, et al. How well do health professionals interpret diagnostic information? A systematic review. *BMJ Open* 2015;5:e008155.

10. Steurer J, Fischer JE, Bachmann LM, et al. Communicating accuracy of tests to general practitioners: a controlled study. *BMJ* 2002;324:824-26.

11. Wegwarth O, Gigerenzer G. Overdiagnosis and overtreatment: evaluation of what physicians tell their patients about screening harms. *JAMA Intern Med* 2013;173:2086-87.

12. Bhatt JR, Klotz L. Overtreatment in cancer – is it a problem? *Expert Opin Pharmacother* 2016;17:1-5.

13. Gigerenzer G, Hoffrage U. How to improve Bayesian reasoning without instruction: frequency formats. *Psychol Rev* 1995;102:684-704.

14. Akobeng AK. Understanding diagnostic tests 2: likelihood ratios, pre- and post-test probabilities and their use in clinical practice. *Acta Paediatr* 2007;96:487-91.

15. Galesic M, Garcia-Retamero R, Gigerenzer G. Using icon arrays to communicate medical risks: overcoming low numeracy. *Health Psychol* 2009;28:210-16.

16. Hoffrage U, Gigerenzer G. How to improve the diagnostic inferences of medical experts. In. Kurz-Milcke E, Gigerenzer G, eds. *Experts in Science and Society.* New York: Kluwer Academic/Plenum;2004;249-68.

17. Gigerenzer G. *Adaptive thinking: rationality in the real world.* New York: Oxford University Press;2000.

18. Gigerenzer G. What are natural frequencies? Doctors need to find better ways to communicate risk to patients. *BMJ* 2011;343:d6386.

19. 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.

20. Hoffrage U, Gigerenzer G. Using natural frequencies to improve diagnostic inferences. *Acad Med* 1998;73:538-40.

21. Tyszka T, Sawicki P. Affective and cognitive factors influencing sensitivity to probabilistic information. *Risk Anal* 2011;31:1832-45.

22. Fraenkel L, Peters E, Tyra S, et al. Shared medical decision making in lung cancer screening: experienced versus descriptive risk formats. *Med Decis Making* 2015;36:518-25.

23. Armstrong BA, Spaniol J. Experienced probabilities increase understanding of diagnostic test results in younger and older adults. *Med Decis Making* 2017;37:670-79.

24. Wegier P, Shaffer VA. Aiding risk information learning through simulated experience (ARISE): using simulated outcomes to improve understanding of conditional probabilities in prenatal Down syndrome screening. *Patient Educ Couns* 2017;100:1882-89.

25. Obrecht NA, Anderson B, Schulkin J, et al. Retrospective frequency formats promote consistent experience-based Bayesian estimates. *Appl Cognit Psychol* 2012;26:436-40.

26. Cohen J. A power primer. *Psychol Bull* 1992;112:155.

27. Erdfelder E, Faul, F, Buchner A. GPOWER: a general power analysis program. *Behav Res Methods Instrum Comput* 1996;28:1-11.

28. Johnson ED, Tubau, E. Comprehension and computation in Bayesian problem solving. *Front Psychol* 2015;6:1-19.

29. Gigerenzer G, Hoffrage U. The role of representation in Bayesian reasoning: correcting common misconceptions. *Behav. Brain Sci* 2007;30:264-67.

30. Gigerenzer G, Hoffrage U. Overcoming difficulties in Bayesian reasoning: a reply to Lewis and Keren (1999) and Mellers and McGraw (1999). *Psychol. Rev* 1999;106:425-30.

Table 1.

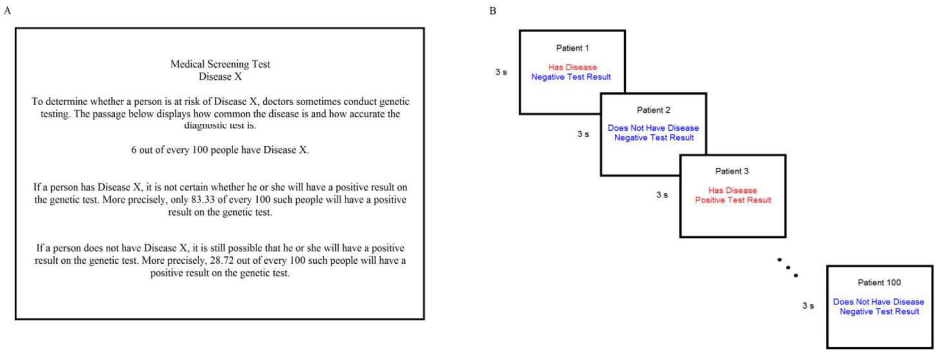
Test Characteristics

	Test Type		
Test Characteristics	Gold Standard	Low Sensitivity	Low Specificity
Prevalence	6%	6%	6%
Sensitivity	100%	50%	83.33%
Specificity	95.74%	93.62%	71.28%
False-Positive Rate	4.26%	6.38%	28.72%
PPV	60%	33.33%	15.63%
NPV	100%	96.70%	98.53%

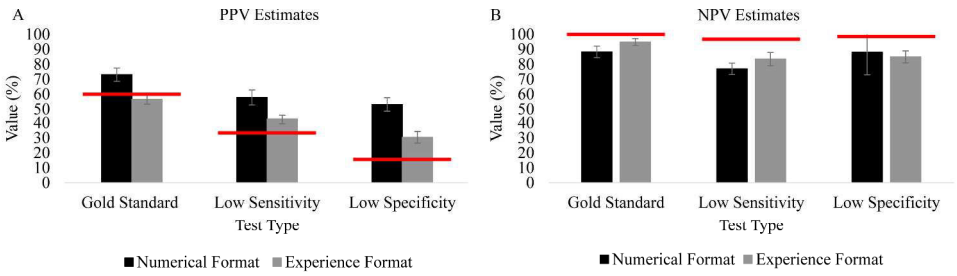
Note. The prevalence of disease and all test characteristics are presented as percentages (i.e., normalized by a base-rate frequency of 100) reflecting what was presented in the numerical format. The joint event combinations (has vs. does not have disease and positive vs. negative test result) underlying the percentages were presented in the experience format.

Figure 1. Panel A is an example of the numerical format, and Panel B is an example of the experience format. The numerical format provides the prevalence of disease, as well as the sensitivity and the false-positive rate of the diagnostic test. In the experience format, 100 representative patient cases were viewed in the slideshow for each of the 3 tests. Each slide was presented for 3 seconds, and describes each patient in terms of disease status (i.e., has disease or does not have disease) and test result (negative or positive). “Has Disease” and “Positive Test Result” were shown in red font, and “Does Not Have Disease” and “Negative Test Result” were shown in blue font.

Figure 2. Mean PPV (Panel A) and NPV (Panel B) estimates for each format and test type. The x-axis displays the experimental factors (format x test) and y-axis displays mean estimate values. The grey bars represent mean estimates in the experience format. The black bars represent mean estimates in the numerical format. The red lines indicate the true PPVs and NPVs. Error bars for each mean represent standard errors.



254x190mm (300 x 300 DPI)



254x190mm (300 x 300 DPI)