Published in final edited form as:

Med Decis Making. 2012 November ; 32(6): 840–850. doi:10.1177/0272989X12445284.

How well do commonly used data presentation formats support comparative effectiveness evaluations?

James G. Dolan, M.D. [Associate Professor],

Department of Community and Preventive Medicine, University of Rochester, Rochester, New York

Feng Qian, MD, Ph.D. [Research Assistant Professor], and

Department of Anesthesiology, University of Rochester, Rochester, New York

Peter J. Veazie, Ph.D. [Associate Professor]

Department of Community and Preventive Medicine, University of Rochester, Rochester, New York

Abstract

Background—Good decisions depend on an accurate understanding of the comparative effectiveness of decision alternatives. The best way convey data needed to support these comparisons is unknown.

Objective—To determine how well five commonly used data presentation formats convey comparative effectiveness information.

Design—Internet survey using a factorial design.

Subjects—279 members of an online survey panel.

Intervention—Study participants compared outcomes associated with three hypothetical screening test options relative to five possible outcomes with probabilities ranging from 2 per 5,000 (0.04%) to 500 per 1,000 (50%). Data presentation formats included a table, a “magnified” bar chart, a risk scale, a frequency diagram, and an icon array.

Measurements—Outcomes included the number of correct ordinal judgments regarding the more likely of two outcomes, the ratio of perceived versus actual relative likelihoods of the paired outcomes, the inter-subject consistency of responses, and perceived clarity.

Results—The mean number of correct ordinal judgments was 12 of 15 (80%), with no differences among data formats. On average, there was a 3.3-fold difference between perceived and actual likelihood ratios, 95% CI: 3.0 to 3.6. Comparative judgments based on flow charts, icon arrays, and tables were all significantly more accurate and consistent than those based on risk scales and bar charts, $p < 0.001$. The most clearly perceived formats were the table and the flow chart. Low subjective numeracy was associated with less accurate and more variable data interpretations and lower perceived clarity for icon displays, bar charts, and flow diagrams.

Conclusions—None of the data presentation formats studied can reliably provide patients, especially those with low subjective numeracy, with an accurate understanding of comparative effectiveness information.

Introduction

Provision of information to patients about the possible risks and benefits of proposed medical interventions has been an integral part of medical care since the doctrine of informed consent was adopted. In recent years it has become increasingly important due to changes in the accepted model of the doctor-patient relationship that promote more active patient involvement in decisions about their care, the rise of evidence-based medical practice, and the increasing emphasis placed on disease prevention. It is likely to become even more important as sophisticated information about future health risks becomes available through work in the “new sciences” such as medical genomics, metabolic profiling, and proteomics. (1) Richard Smith, former editor of the British Medical Journal, has called risk communication “the main work of doctors”. (2)

A realistic understanding of the differences in outcomes expected to result from alternative courses of action is an essential component of good decision making. Comparative effectiveness information can be provided to decision makers verbally, numerically, graphically, or using a combination of formats. (3–5) There is evidence that communication of quantitative data can be enhanced by using a format that conveys information about both the number of people with the outcome of interest and the size of the reference population (the “part to whole” relationship), includes a graphic representation of the data, and requires little or no additional processing by the recipient. (6–9) Current recommendations for the communication of information to support comparative assessments of decision alternatives suggest the use of graphic data displays that meet these criteria, including icon arrays, bar charts, flow diagrams, and risk scales. (3, 4) Significant knowledge gaps, however, still exist regarding how to convey information about outcomes that occur less than 1% of the time, whether graphic formats are more effective than numeric ones, whether any of the currently recommended graphic formats is superior to the others, and the impact of individual recipient characteristics. (6, 7, 10–12)

The goal of this study was to address these gaps. We used a comprehensive evaluation framework to compare the abilities of five common data presentation formats to accurately convey comparative effectiveness information across a representative range of clinically important likelihoods and consequences. We hypothesized that graphic displays meeting current formatting recommendations - bar charts and icon displays – would be the most effective formats. We also examined the relationships between communication effectiveness and recipient characteristics including age, gender, education level, literacy, and numeracy.

Methods

Study population

The study population consisted of members of an Internet survey panel who responded to a standard email invitation set by a host company (Zoomerang, MarketTools, Inc., San Francisco CA). Panel members who completed the study received points redeemable for goods and services from the company in return for their study participation.

Study intervention

The study intervention was a three part survey designed to compare the effectiveness of five data presentation formats for helping people accurately compare decision alternatives. We created a hypothetical disease screening scenario and then evaluated respondents’ abilities to compare information regarding the expected outcomes of three alternative screening tests.

The first part of the survey provided an overview of the study and introduced participants to the decision scenario. The decision scenario was described as follows: *The questions in this*

survey refer to the risks associated with a serious disease and three screening tests for the disease. The disease and tests we refer to are not real. We are using them to test different ways of communicating information about the risks and benefits of screening. However, they are similar to the risks and benefits of screening for several real diseases.

The second part of the survey asked participants to interpret a series of outcome messages created by combining one of five screening outcomes with one of five data presentation formats. The five outcomes were: 1) the lifetime chance of developing disease; 2) the chance of dying from the disease; 3) the chance of a serious screening test side effect; 4) the chance of dying from a serious screening test side effect; and 5) the chance of a false positive screening test result. The likelihoods associated with each outcome are summarized in Table 1. They ranged from 2 per 5,000 (0.04%) to 500 per 1,000 (50%). The five presentation formats consisted of: a table, a “magnified” bar chart (13), a Piling risk scale (14), a frequency diagram, and an icon array. Examples of each format are shown in Figure 1. We tested the effectiveness of the outcome messages using a factorial design that divided the study population into five groups. As illustrated in Table 1, all groups were exposed to every data format and received the same total amount of information, but varied in the format and outcome combinations evaluated.

Every outcome message provided data about outcomes associated with three different screening options, labeled as Option A, Option B, and Option C. After reviewing the message, respondents were asked to compare outcomes associated with two of the three options and indicate if they were equally likely to occur or if one was more likely to occur. If one was judged more likely to occur, they were then asked to indicate the magnitude of the relative difference in likelihoods. To avoid biasing these judgments, we used an open ended question format supplemented with several examples of possible answers including whole numbers, decimals, and fractions. The respondents then repeated this process for the other two pairs of outcomes. For all comparisons, the message being interpreted was displayed on the same page as the questions. This process resulted in a total of three comparisons per message and 15 comparisons for each respondent. An example comparison question is shown in Figure 2.

In part three of the questionnaire, respondents rated each format on a one to ten clarity scale ranging from “confusing, very unhelpful” to “clear, very helpful”, completed a demographic background questionnaire, answered the one question literacy assessment developed by Chew and colleagues (15), and completed the subjective numeracy scale. (16) These measures were all chosen to make the study questionnaire suitable for a brief, 15 to 20 minute Internet survey.

The study was approved by the University of Rochester Research Subjects Review Board.

Outcomes

We compared the effectiveness of the data presentation formats using a modified version of criteria for evaluating risk communications proposed by Weinstein and Sandman (17):

Ordinal accuracy—We assessed ordinal accuracy by tabulating the number of correct judgments regarding the more likely outcome made by each respondent. Possible scores range from 0 to 15.

Comparative accuracy—We assessed comparative accuracy using an adjusted comparative accuracy ratio. We first calculated the unadjusted accuracy ratio by dividing the perceived differences in likelihood by the actual differences for each paired comparison. We then determined the absolute values of the unadjusted ratios in order to assess errors due to

both overestimates and under-estimates in a single measure. Finally, because each outcome message assessment task involved three separate comparisons, we used the mean of each respondent's judgments for each set of comparisons to control for individual variation. Perfectly accurate judgments have an adjusted accuracy ratio equal to 1.0; inaccurate judgments have ratios greater than 1.0. To facilitate data analysis and interpretation, we transformed the data using base 10 logarithms for analysis and converted the results back into the original format for presentation when appropriate.

Consistency—We assessed how consistently respondents interpreted the outcome messages using each data format by calculating the standard deviations of the mean adjusted accuracy ratios.

Direction of communication errors—We used the relative percentage of inaccurate comparative accuracy judgments that were higher versus lower than the actual data to assess the relative frequencies of over versus under-estimates associated with each data presentation format.

User evaluation—We used the clarity scale to assess the respondents' evaluations of each data presentation format.

Data analysis

We summarized the results using standard descriptive statistical methods. Because the outcome data are not normally distributed, we determined the statistical significance of outcome differences among the presentation formats using Kruskal-Wallis analysis of variance, corrected for ties, and the related between-group difference method proposed by Conover. (18)

Because we were unable to transform the data to a normal distribution, we measured associations between respondent characteristics and both ordinal and comparative accuracy by first dividing the respondents into low, medium, and high accuracy groups as evenly as possible given the distribution of the results and then using ordinal logistic regression to determine significant multi-variable associations. Respondent characteristics examined included age, gender, race, education, literacy, and numeracy. We excluded ethnic background because there was so little variation in the study sample.

We defined statistical significance for all analyses as a p value ≤ 0.05 . Statistical calculations were performed using *Medcalc* and *JMP* statistical software. (19, 20)

Results

Study population

Two hundred seventy nine people participated in the study. As shown in Table 2, most were white, well educated, literate, and fairly numerate. Compared with the general US population, our study sample was better educated, had relatively more men, and contained smaller proportions of African-American and Hispanic respondents. (21)

Ordinal accuracy

The ordinal accuracy data are illustrated in Figure 3. The overall number of correct ordinal judgments per respondent was 12 of 15 (80%). There were no statistically significant differences in ordinal accuracy among the different data presentation formats. Rates of correct responses ranged from 78% to 83%, $p = 0.48$.

The mean rates of correct responses in the low, medium, and high accuracy groups were 5.9 (39%), 13.2 (88%), and 15 (100%) respectively. Ordinal logistic regression indicated that three of the six patient characteristics studied - female gender, white race, and formal education beyond high school - were significantly associated with higher ordinal accuracy. These data are summarized in Table 3.

Comparative accuracy

The comparative accuracy results are illustrated in Figure 4. Overall, the mean adjusted accuracy ratio was 3.3 (95% confidence interval 3.0 to 3.6) indicating a more than 3-fold difference between the perceived and actual differences in outcome likelihoods.

When the results across all five screening outcomes are combined, there is a statistically significant difference in comparative accuracy among the five formats, $p < 0.001$. Judgments based on the flow chart (mean adjusted accuracy ratio 2.8), the icon array (mean adjusted accuracy ratio 2.8), and the table (mean adjusted accuracy ratio 3.2) are all significantly more accurate than those based on the bar chart and risk scale, mean accuracy ratios 3.5 and 4.7 respectively.

There are also statistically significant differences in the data formats' comparative accuracy relative to all five screening outcomes. These results generally reflect the overall findings except for the risk of a false positive screening test where the table is the worst format and the risk scale among the best. Further analysis of this result indicated a disproportionately high number of extremely discrepant values, defined as accuracy ratios > 100 , for the table format: 47% versus 7% to 29% for the other formats. These results suggest that these respondents may have mistakenly reported absolute instead of relative differences. Repeating the comparative accuracy analysis with these extreme values removed results in findings consistent with the other four outcomes with the table among the most accurate formats and the risk scale significantly less accurate than the others (data provided in supplemental file).

There were 84 respondents in the high accuracy group, 79 in the mid-accuracy group, and 82 in the low accuracy group. Mean accuracy ratios in these three groups were 1.3, 2.3, and 12.4, respectively. Ordinal logistic regression revealed that male gender (Chi square 6.5, $p = 0.01$) and subjective numeracy higher than the median score of 4.5 (Chi square 8.9, $p = 0.003$) were significantly associated with better comparative accuracy. These results are summarized in Table 3.

Consistency

The formats with the lowest response variation were the flow diagram (mean standard deviation 5.4) and the icon array (mean standard deviation 6.1). The risk scale (mean standard deviation 71.3) was associated with the most variability in responses, $p < 0.05$. On multivariable analysis, two respondent characteristics were associated with better response consistency: older age ($p = 0.001$) and a subjective numeracy score > 4.5 ($p = 0.03$).

Direction of communication errors

Overall, there were a total of 1,005 inaccurate comparative accuracy judgments, equally divided into under-estimates (504) and over-estimates (501). There were no significant differences among the formats in the relative proportions of over and under estimates, Chi square = 8.7, $p = 0.07$. Younger respondents were more likely to overestimate the relative risk differences than older ones, Chi square = 7.2, $p = 0.03$.

Perceived clarity

Table 4 summarizes the respondents' format clarity ratings. The differences among the formats are statistically significant, $p < 0.001$. The table was the highest rated format (mean preference score 7.4) and the icon array was the lowest (mean preference score 4.6). Multivariable analysis shows that more education ($p = 0.05$) and higher subjective numeracy scores ($p = 0.002$) were associated with higher clarity ratings for the table format. Higher subjective numeracy scores were also associated with higher clarity ratings assigned to the icon display ($p = 0.008$), bar chart (0.005), and flow diagram ($p = 0.03$). Higher perceived bar chart clarity was also associated with race other than white ($p = 0.03$). None of the respondent characteristics studied were associated risk scale clarity ratings.

Discussion

The goal of this study was to address current gaps in our knowledge of how to convey comparative outcome information needed to support meaningful patient involvement in decisions about their health and healthcare. The most notable finding is that the communication outcomes associated with all formats studied were suboptimal. The overall ordinal accuracy rate was 80%, indicating that one out of every five judgments incorrectly identified which of two outcomes was more likely, and did not differ among the formats. Similarly, the comparative accuracy analysis revealed a three-fold average discrepancy between participants' perceived differences in outcome likelihoods and the actual differences in the data presented. Finally, none of the formats was judged to be particularly clear by the study participants: the average clarity ratings on the ten point clarity scale ranged from 4.6 to 7.4. These findings suggest that even the best conventional communication formats may have serious shortcomings in their abilities to convey quantitative comparative effectiveness information to members of the general population and most patients.

Despite their theoretical advantages, our results do not support the hypothesis that the bar charts and icon displays would be the most effective formats. Instead, across all outcome measures, the two most effective formats were the flow chart and the table.

Several previous studies have also demonstrated that tables are equal if not better than graphic data presentation formats for accurately conveying comparative outcome information. (22–25) Flow charts, also called natural frequency diagrams, have been shown to successfully communicate complicated quantitative information, but, to our knowledge, their effectiveness in communicating comparative effectiveness information relative to other commonly used presentation formats has not been examined previously. (9, 26)

In several previous studies, icon arrays have been found to be very effective risk communication formats over a limited range of outcome likelihoods. (27, 28) To our knowledge, this is the first study to specifically evaluate their effectiveness in supporting clinical decisions across a broad spectrum of clinically important outcome likelihoods. Consistent with previous studies, we found the icon array one of the more effective communication formats for outcome probabilities between 1% and 50%. They were much less effective in the two small risk scenarios involving likelihoods ranging from 0.04% to 0.3%. One possible explanation for this finding is the size of the array used. These small likelihoods necessitated the use of arrays consisting of 5,000 icons rather than the 1,000 icon arrays used in the other scenarios. Previous research has demonstrated that people's interpretations of icon arrays are affected by the size of the array. (29) Icon arrays also received the lowest clarity ratings from the study participants. Previous studies have similarly found icon arrays to be disliked by patients. (30) Thus our findings provide new information about both strengths and weaknesses of icon arrays and indicate that additional

work is needed to learn how to most effectively use them to support medical decision making.

Prior studies have found bar charts effective in helping people comprehend outcome information and preferred by many patients. (23, 31) Their effectiveness in this study, however, was mixed. The bar chart format was highly rated for clarity by study participants and the most consistently interpreted data format but was not as successful as several others in accurately conveying outcome information. Although we expected difficulties using the bar chart format to display the small risks (and for this reason included the magnified portion), the accuracy achieved by the bar charts was similar for low, medium, and high likelihood outcome scenarios.

Although the use of a risk scale to convey outcome information has been proposed, (4, 32–35) its usefulness for this purpose has not been thoroughly evaluated. To our knowledge, this is the first study to compare the effectiveness of risk scales in communicating comparative effectiveness information with that of other commonly used data presentation formats. Although the risk scale was ranked higher than bar charts and icon arrays in terms of clarity, it was the most inaccurately and inconsistently interpreted format. These findings probably reflect well-known difficulties with the proper interpretation of logarithmic scales. (7) (36) Alternatively, they could reflect the absence of comparative risk information that is frequently included in such scales.

Individual characteristics associated with accurate interpretations of outcome data in this study - gender, education, and numeracy – have also been found in previous studies. (10, 11, 37–39) The inconsistent gender effects we found between ordinal and comparative accuracy are difficult to reconcile and suggest that this finding may be an artifact caused by the disproportionate amount of missing data for the men with regard to comparative accuracy. The association between objectively measured numeracy skills and both data interpretation and use in decision making is becoming increasingly recognized. (10, 11, 38, 39) We found a similar relationship using a subjective numeracy measure. This result is consistent with an earlier study that examined how people interpret risk scales. (40) These findings suggest that, in addition to demonstrated skills in working with numeric information, people's preferences for verbal versus quantitative information and their confidence in working with numeric information also play a role in how well they interpret quantitative comparative effectiveness data.

This study is subject to several limitations. First, because we used a hypothetical decision scenario, we are unable to determine if the results would differ if they were based on a real decision being faced by the study participants. However the scenarios we used addressed preventive screening, a situation that affects all members of the general population.

It is also subject to the limitations of Internet surveys including sampling bias and poor data quality. (41) These limitations, however, apply equally to all data presentation format studied. They are therefore unlikely to affect the relative differences found among them.

Other limitations include the moderate amount of missing data regarding the comparative accuracies and evidence suggesting a possible misunderstanding of survey instructions with regard to the false positive question. Given the number of comparisons studied and the uniformly high amount of inaccurate interpretations found, however, it seems unlikely that these problems significantly affect the main study results.

In conclusion, the results of this study suggest that none of the five data presentation formats studied can reliably provide patients with an accurate understanding of comparative outcome information across a wide range of clinically relevant likelihoods. They also provide

additional evidence that structured numeric formats, such as a simple table or a flow chart, may work just as well as, if not better than, graphic displays for presenting small amounts of comparative outcome information and that icon displays are not well suited for presenting information about events with likelihoods less than 1%. These findings suggest that research is needed to develop more effective ways to communicate information to support meaningful patient involvement clinical decision making. Finally, our study provides new evidence that an individual's perceived numeracy affects their ability to interpret comparative effectiveness information. This finding re-enforces the already recognized need to develop communication methods that are appropriate for supporting involvement of people with less formal education and low numeracy skills, both actual and perceived, in decisions about their health and health care. (10, 38, 42)

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported by Grant number 1R21CA131760-01A1 from the National Cancer Institute.

References

1. Arab S, Gramolini AO, Ping P, Kislinger T, Stanley B, van Eyk J, et al. Cardiovascular proteomics: tools to develop novel biomarkers and potential applications. *J Am Coll Cardiol*. 2006; 48:1733–41. [PubMed: 17084242]
2. Smith R. Communicating risk: the main work of doctors. *BMJ*. 2003; 327:0-f.
3. Edwards A, Elwyn G, Mulley A. Explaining risks: turning numerical data into meaningful pictures. *BMJ*. 2002; 324:827–30. [PubMed: 11934777]
4. Paling J. Strategies to help patients understand risks. *BMJ*. 2003; 327:745–8. [PubMed: 14512489]
5. Visschers VH, Meertens RM, Passchier WW, de Vries NN. Probability information in risk communication: a review of the research literature. *Risk Anal*. 2009; 29:267–87. [PubMed: 19000070]
6. Ancker JS, Senathirajah Y, Kukafka R, Starren JB. Design Features of Graphs in Health Risk Communication: A Systematic Review. *J Med Inform Assoc*. 2006; 13:608–18.
7. Lipkus IM. Numeric, Verbal, and Visual Formats of Conveying Health Risks: Suggested Best Practices and Future Recommendations. *Med Decis Making*. 2007; 27:696–713. [PubMed: 17873259]
8. Gaissmaier W, Wegwarth O, Skopec D, Müller A-S, Broschinski S, Politi MC. Numbers can be worth a thousand pictures: Individual differences in understanding graphical and numerical representations of health-related information. *Health Psychol*. 2011
9. Kurz-Milcke E, Gigerenzer G, Martignon L. Transparency in Risk Communication. *Ann NY Acad Sci*. 2008; 1128:18–28. [PubMed: 18469211]
10. Nelson W, Reyna V, Fagerlin A, Lipkus I, Peters E. Clinical Implications of Numeracy: Theory and Practice. *Ann Behav Med*. 2008; 35:261–74. [PubMed: 18677452]
11. Reyna VF, Nelson WL, Han PK, Dieckmann NF. How numeracy influences risk comprehension and medical decision making. *Psychol Bull*. 2009; 135:943–73. [PubMed: 19883143]
12. Waldron C-A, van der Weijden T, Ludt S, Gallacher J, Elwyn G. What are effective strategies to communicate cardiovascular risk information to patients? A systematic review. *Patient Educ Counsel*. 2011; 82:169–81.
13. Dolan JG, Iadarola S. Risk communication formats for low probability events: an exploratory study of patient preferences. *BMC Med Inform Decis Mak*. 2008; 8:14.
14. Risk communication format ©John Paling. 2000.

15. Chew L, Griffin J, Partin M, Noorbaloochi S, Grill J, Snyder A, et al. Validation of Screening Questions for Limited Health Literacy in a Large VA Outpatient Population. *J Gen Intern Med*. 2008; 23:561–6. [PubMed: 18335281]
16. Fagerlin A, Zikmund-Fisher BJ, Ubel PA, Jankovic A, Derry HA, Smith DM. Measuring numeracy without a math test: development of the Subjective Numeracy Scale. *Med Decis Making*. 2007; 27:672–80. [PubMed: 17641137]
17. Weinstein N, Sandman P. Some criteria for evaluating risk messages. *Risk Anal*. 1993; 13:103–14.
18. Conover, WJ. Practical nonparametric statistics. 3. New York: Wiley; 1999.
19. JMP, Version 9. SAS Institute Inc; Cary, NC: 1989–2010.
20. MedCalc ® Version 11.6.1.0. MedCalc Software;
21. United States profile in the US Central Intelligence Agency World Factbook. 2011. Available from: <https://www.cia.gov/library/publications/the-world-factbook/geos/us.html>
22. Lipkus IM, Hollands JG. The visual communication of risk. *J Natl Cancer Inst Monogr*. 1999:149–63. [PubMed: 10854471]
23. Feldman-Stewart D, Kocovski N, McConnell BA, Brundage MD, Mackillop WJ. Perception of quantitative information for treatment decisions. *Med Decis Making*. 2000; 20:228–38. [PubMed: 10772360]
24. Stone ER, Sieck WR, Bull BE, Yates JF, Parks SC, Rush CJ. Foreground:background salience: explaining the effects of graphical displays on risk avoidance. *Organ Behav Hum Dec*. 2003; 90:19–36.
25. Lee DH, Mehta MD. Evaluation of a visual risk communication tool: effects on knowledge and perception of blood transfusion risk. *Transfusion*. 2003; 43:779–87. [PubMed: 12757530]
26. Gigerenzer G, Edwards A. Simple tools for understanding risks: from innumeracy to insight. *BMJ*. 2003; 327:741–4. [PubMed: 14512488]
27. Tait A, Voepel-Lewis T, Zikmund-Fisher B, Fagerlin A. The Effect of Format on Parents' Understanding of the Risks and Benefits of Clinical Research: A Comparison Between Text, Tables, and Graphics. *J Health Commun*. 2010; 15:487–501. [PubMed: 20677054]
28. Galesic M, Garcia-Retamero R, Gigerenzer G. Using icon arrays to communicate medical risks: overcoming low numeracy. *Health Psychol*. 2009; 28:210–6. [PubMed: 19290713]
29. Schapira MM, Nattinger AB, McHorney CA. Frequency or probability? A qualitative study of risk communication formats used in health care. *Med Decis Making*. 2001; 21:459–67. [PubMed: 11760103]
30. Elting LS, Martin CG, Cantor SB, Rubenstein EB. Influence of data display formats on physician investigators' decisions to stop clinical trials: prospective trial with repeated measures. *BMJ*. 1999; 318:1527–31. [PubMed: 10356010]
31. Fortin JM, Hirota LK, Bond BE, O'Connor AM, Col NF. Identifying patient preferences for communicating risk estimates: a descriptive pilot study. *BMC Med Inform Decis Mak*. 2001; 1:2. [PubMed: 11545684]
32. Moore RA, Derry S, McQuay HJ, Paling J. What do we know about communicating risk? A brief review and suggestion for contextualising serious, but rare, risk, and the example of cox-2 selective and non-selective NSAIDs. *Arthritis Res Ther*. 2008; 10:R20. [PubMed: 18257914]
33. Paling, J. Helping patients understand risks. Gainesville FL: The Risk Communication Institute; 2006.
34. Stallings SP, Paling JE. New Tool for Presenting Risk in Obstetrics and Gynecology. *Obstet Gynecol*. 2001; 98:345–9. [PubMed: 11510463]
35. Singh AD, Paling J. Informed consent: Putting risks into perspective. *Surv Ophthalmol*. 1997; 42:83–6.
36. Siegrist M, Orlow P, Keller C. The effect of graphical and numerical presentation of hypothetical prenatal diagnosis results on risk perception. *Med Decis Making*. 2008; 28:567–74. [PubMed: 18566487]
37. Waters EA, Weinstein ND, Colditz GA, Emmons K. Formats for improving risk communication in medical tradeoff decisions. *J Health Commun*. 2006; 11:167–82. [PubMed: 16537286]

38. Peters E. Numeracy and the Perception and Communication of Risk. *Ann NY Acad Sci.* 2008; 1128:1–7. [PubMed: 18469208]
39. Ancker JS, Kaufman D. Rethinking health numeracy: a multidisciplinary literature review. *J Med Inform Assoc.* 2007; 14:713–21.
40. Hess R, Visschers VH, Siegrist M, Keller C. How do people perceive graphical risk communication? The role of subjective numeracy. *J Risk Res.* 2011; 14:47–61.
41. Fricker RD, Schonlau M. Advantages and disadvantages of Internet research surveys: Evidence from the literature. *Field Methods.* 2002; 14:347.
42. McCaffery KJ, Smith SK, Wolf M. The Challenge of Shared Decision Making Among Patients With Lower Literacy: A Framework for Research and Development. *Med Decis Making.* 2009; 30:35–44. [PubMed: 19692709]

Figure 1.
The data presentation formats studied.

The following questions refer to the above figure.

F11. Which risk is more likely: A or B?

- ☐ A is more likely than B (now answer question F12)^{***}
- ☐ They are equally likely (Now go to the next page)^{***}
- ☐ B is more likely than A (now answer question F13)^{***}

F12. If risk A is more likely than risk B, how much more likely do you think it is?

Answer this question only if you think risk A is more likely than risk B. To answer, enter a number in the box that indicates how big a difference you think there is. You can use whole numbers (like 2 or 5), decimals (like 2.1 or 5.4), or fractions (like 1/2 or 5%). When you are done, go to the next page.

Risk A is times more likely than risk B.

F13. If risk B is more likely than risk A, how much more likely do you think it is?

Answer this question only if you think risk B is more likely than risk A. To answer, enter a number in the box that indicates how big a difference you think there is. You can use whole numbers (like 2 or 5), decimals (like 2.1 or 5.4), or fractions (like 1/2 or 5%). When you are done, go to the next page.

Risk B is times more likely than risk A.

Figure 2.
Example comparison question.

Figure 3.
Ordinal accuracy score

Figure 4.
Adjusted comparative accuracy ratio

\$watermark-text

\$watermark-text

\$watermark-text

Table 1

The study design

	Risk of disease	Risk of dying from disease	Risk of serious screening test side effect	Risk of dying from screening test side effect	Chance of a false positive screening test result
Group 1	Risk Scale	Bar	Table	Flow	Icon Array
Group 2	Flow	Icon Array	Risk Scale	Bar	Table
Group 3	Icon Array	Risk Scale	Bar	Table	Flow
Group 4	Bar	Table	Flow	Icon Array	Risk Scale
Group 5	Table	Flow	Icon Array	Risk Scale	Bar
Likelihood data *	A: 10 per 1,000 B: 22 per 1,000 C: 48 per 1,000	A: 4 per 1,000 B: 10 per 1,000 C: 22 per 1,000	A: 2 per 5,000 B: 7 per 5,000 C: 17 per 5,000	A: 2 per 5,000 B: 7 per 5,000 C: 17 per 5,000	A: 100 per 1,000 B: 300 per 1,000 C: 500 per 1,000

* The likelihood data indicate the chance that the outcome will occur following three possible screening test options, denoted as Options A, B, and C.

Table 2

The study sample

Variable	Result
Number	279
Age	Mean 45.1, range 18–85
Gender	Female: 128 (46%) Male: 151 (54%)
Ethnicity	Hispanic: 21 (8%) Non-Hispanic: 258 (92%)
Race	White: 238 (85%) African-American: 8 (3%) Hispanic: 12 (4%) Asian: 12 (4%) American Indian: 6 (2%) Other: 3 (1%)
Education, highest level completed	Less than high school: 4 (1%) High school: 43 (15%) Some college: 77 (28%) Associates degree: 27 (10%) Bachelor's degree: 68 (24%) Postgraduate level: 53 (19%) Other: 7 (3%)
Chew numeracy question: "How confident are you filling out forms by yourself?" (12)	not at all: 3 (1%) a little bit: 6 (2%) somewhat: 16 (6%) quite a bit: 82 (29%) extremely: 172 (62%)
Subjective numeracy score (13)	0–1: 0 1–1.9: 2.0% 2–2.9: 6.1% 3–3.9: 20.1% 4–4.9: 33.3% 5–5.9: 34.8% 6.0: 4.0%

Table 3

Individual characteristics associated with ordinal and comparative accuracy

Ordinal accuracy	Group *				
	Low n = 76	Medium n = 81	High, n = 122	Total	Significance †
Gender					
Female	28 (22%)	33 (26%)	67 (52%)	128	$\chi^2 = 7.3$ p = 0.007
Male	48 (32%)	48 (32%)	55 (36%)	151	
Race					
White	59 (25%)	67 (28%)	112 (47%)	238	$\chi^2 = 7.5$ p = 0.006
Non-white	17 (41%)	14 (34%)	10 (10%)	41	
Education					
High School	25 (46%)	12 (22%)	17 (31%)	54	$\chi^2 = 4.2$ p = 0.04
> High School	51 (23%)	69 (31%)	105 (47%)	225	
Comparative accuracy	Group				
	Low n = 82	Medium n = 79	High n = 84	Total	
Gender					
Female	45 (39%)	42 (36%)	29 (25%)	116	$\chi^2 = 6.5$ p = 0.01
Male	37 (29%)	37 (29%)	55 (43%)	129	
Subjective numeracy					
4.5	51 (44%)	36 (31%)	30 (26%)	117	$\chi^2 = 8.9$ p = 0.003
> 4.5	31 (24%)	43 (34%)	54 (42%)	128	

* Groups were determined by dividing the study population into thirds as equally as possible given the distribution of results.

† Chi Square and p values represent effect likelihood ratio tests derived from ordinal logistic regression with group as the dependent variable and six independent variables: age, gender, race, education, literacy, and numeracy

Table 4

Comparative accuracy

	Number	Adjusted Mean Difference ratio (95% CI)	Statistical significance*
All risks combined, all formats	1108	3.3 (3.0 to 3.6)	
All risks combined			Ht = 60.7, p < 0.001
Flow	227	2.8 (2.3 to 3.3)	Bar, scale
Icons	217	2.8 (2.2 to 3.3)	Bar, scale
Table	233	3.2 (2.6 to 4.0)	Bar, scale
Bar	217	3.5 (2.9 to 4.1)	Flow, icons, scale, table
Scale	214	4.7 (3.8 to 5.8)	Flow, icons, scale, table
Chance of disease			Ht = 9.6, p = 0.05
Icons	49	2.6 (1.7 to 3.9)	Flow, scale
Bar	44	3.0 (2.2 to 4.0)	No significant differences
Table	46	3.2 (2.1 to 5.0)	Scale
Scale	41	3.3 (2.5 to 4.5)	Icons, table
Flow	50	4.2 (2.8 to 6.3)	Icons
Chance of dying from disease			Ht = 20.8, p = 0.0003
Table	49	2.1 (1.6 to 2.7)	Icons, scale
Bar	42	2.4 (1.9 to 3.2)	Scale
Flow	45	2.9 (1.7 to 5.0)	Icons, scale
Icons	48	3.0 (2.1 to 4.3)	Flow, scale, table
Scale	45	6.3 (3.6 to 11.0)	Table, bar, flow, icons
Chance of serious screening test side effect			Ht = 34.3, p < 0.001
Table	45	2.0 (1.5 to 2.8)	Bar, scale
Flow	42	2.1 (1.6 to 2.7)	Bar, scale
Bar	47	3.8 (2.6 to 5.5)	Flow, table
Icons	45	3.8 (1.8 to 8.0)	Scale
Scale	42	8.3 (4.7 to 14.5)	Flow, icons, table
Chance of dying from test side effect			Ht = 32.8, p < 0.0001
Flow	42	1.9 (1.4 to 2.6)	Bar, scale
Icons	37	1.9 (1.4 to 2.5)	Bar, scale
Table	47	3.5 (1.8 to 6.6)	Bar, scale
Bar	41	4.1 (2.7 to 6.1)	Flow, table, icons
Scale	43	5.1 (3.0 to 9.0)	Flow, table, icons
Chance of a false positive screening test			Ht = 11.3, p = 0.023
Scale	43	2.6 (1.9 to 3.4)	No significant differences
Icons	37	2.8 (1.8 to 4.4)	Table
Flow	42	3.1 (1.9 to 5.1)	Table
Bar	41	4.3 (2.4 to 7.5)	No significant differences

	Number	Adjusted Mean Difference ratio (95% CI)	Statistical significance*
Table	47	7.6 (4.6 to 12.5)	Flow, icons

* Statistical significance of differences among the groups determined using the Kruskal-Wallis test, corrected for ties. Statistically significant inter-group differences were determined using the between-group difference method proposed by Conover. (16)

Table 5

Consistency analysis results: Standard deviations of the adjusted accuracy ratios.

Format	number	Mean (95% CI)	Median (95% CI)	Minimum	Maximum	Statistical analysis results *
All formats	1,266	19.7 (8.8 to 30.6)	1.7 (1.6 to 1.8)	0	4,269	Ht = 28.4, p < 0.0001
Bar chart	250	9.3 (5.9 to 12.6)	2.1 (1.7 to 2.5)	0	229	Different from Flow, icon, table
Flow chart	258	5.4 (3.3 to 7.6)	1.5 (1.3 to 1.7)	0	122	Different from Bar, scale
Icon Array	257	6.1 (3.5 to 8.7)	1.7 (1.4 to 1.8)	0	236	Different from Bar, scale
Risk Scale	245	71.3 (15.4 to 127)	1.8 (1.7 to 2.1)	0.1	4,269	Different from Flow, icon, table
Table	256	8.8 (5.4 to 12.1)	1.5 (1.4 to 1.7)	0	227	Different from Bar, scale

* Statistical significance of differences among the groups determined using the Kruskal-Wallis test, corrected for ties. Statistically significant inter-group differences were determined using the between-group difference method proposed by Conover. (16)

Table 6Clarity scale ratings^{*}

Format	Number	Mean preference score	95% Confidence interval	Significantly different from [*]
Table	279	7.4	7.1 to 7.7	All others
Flow chart	279	6.7	6.4 to 7.0	All others
Risk scale	279	5.9	5.6 to 6.3	All others
Bar chart	279	5.1	4.8 to 5.4	Flow, scale, table
Icon array	279	4.6	4.3 to 4.9	Flow, scale table

^{*} The differences in clarity ratings among the five data formats, determined using the Kruskal-Wallis test, corrected for ties, are statistically significant, $H = 178.6$, $p < 0.001$. Statistically significant inter-group differences were determined using the between-group difference method proposed by Conover. (16)