Communicating Relative Risk Changes with Baseline Risk: Presentation Format and Numeracy Matter

Nicolai Bodemer, PhD, Björn Meder, PhD, and Gerd Gigerenzer, PhD

Background. *Treatment benefits and harms are often communicated as relative risk reductions and increases, which are frequently misunderstood by doctors and patients. One suggestion for improving understanding of such risk information is to also communicate the baseline risk. We investigated 1) whether the presentation format of the baseline risk influences understanding of relative risk changes and 2) the mediating role of people's numeracy skills.* ***Method.*** *We presented laypeople (*N = *1234) with a hypothetical scenario about a treatment that decreased (Experiments 1a, 2a) or increased (Experiments 1b, 2b) the risk of heart disease. Baseline risk was provided as a percentage or a frequency. In a forced-choice paradigm, the participants' task was to judge the risk in the treatment group given the relative risk reduction (or increase) and the baseline risk. Numeracy was assessed using the Lipkus 11-item scale.* ***Results.*** *Communicating baseline risk in a frequency format facilitated correct understanding of a treatment's benefits and harms, whereas a percentage format often impeded understanding. For example, many participants misinterpreted a relative risk reduction as referring to an absolute risk reduction. Participants with higher numeracy generally performed better than those with lower numeracy, but all participants benefitted from a frequency format. Limitations are that we used a hypothetical medical scenario and a nonrepresentative sample.* ***Conclusions.*** *Presenting baseline risk in a frequency format improves understanding of relative risk information, whereas a percentage format is likely to lead to misunderstandings. People's numeracy skills play an important role in correctly understanding medical information. Overall, communicating treatment benefits and harms in the form of relative risk changes remains problematic, even when the baseline risk is explicitly provided.* ***Key words:*** *relative risk; absolute risk; risk communication; numeracy; presentation format; baseline risk.* ***(Med Decis Making 2014;34:615–626)***

Transparent and intuitive communication of health information is a major challenge in ensuring informed consent and shared decision making. For instance, the benefit of mammography screening for women aged 50 and older can be communicated as a risk reduction of 20% in mortality due to breast cancer.[1] But what does that number actually mean? The epidemiological data on which this benefit is based show that about 5 in every 1000 women without screening die from breast cancer within 10 years, as opposed to 4 in 1000 with screening—a relative risk reduction of 20%. Another way of conveying the same information is to say that participating in screening reduces breast cancer mortality by 0.1% (1 in 1000), namely, from 0.5% (5 in 1000) to 0.4% (4 in 1000). As this example illustrates, different formats exist for expressing treatment benefits. A *relative*

Received 25 December 2012 from Max Planck Institute for Human Development, Harding Center for Risk Literacy, Berlin, Germany (NB, GG); and Max Planck Institute for Human Development, Center for Adaptive Behavior and Cognition (ABC), Berlin, Germany (NB, BM, GG). This research was supported by the joint program ''Acciones Integradas Hispano-Alemanas'' from the Deutscher Akademischer Austauschdienst (DAAD) and the Ministerio de Ciencia y Tecnología. BM was supported by Grant ME 3717/2 from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program ''New Frameworks of Rationality'' (SPP 1516). Revision accepted for publication 8 February 2014.

DOI: 10.1177/0272989X14526305

Supplementary material for this article is available on the *Medical Decision Making* Web site at http://mdm.sagepub.com/supplemental.

Address correspondence to Nicolai Bodemer, Harding Center for Risk Literacy, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany; e-mail: bodemer@mpib-berlin.mpg.de.

Table 1 Relative and Absolute Risk Changes

Type of Change	Risk Reduction (RR)	Risk Increase (RI)
Relative (R)	$RRR = \frac{ER_{Control} - ER_{Treatment}}{ER_{Control}}$	$RRI = \frac{ER_{Treatment} - ER_{Control}}{ER_{Control}}$
	Example:	Example:
	$RRR = \frac{0.5\% - 0.4\%}{0.5\%} = 20\%$	$RRI = \frac{0.028\% - 0.014\%}{0.014\%} = 100\%$
Absolute (A)	$ARR = ER_{Control} - ER_{Treatment}$	$ARI = ER_{Treatment} - ER_{Control}$
	Example:	Example:
	$ARR = 0.5\% - 0.4\% = 0.1\%$	$ARI = 0.028\% - 0.014\% = 0.014\%$

Note: The risk reductions are based on the mammography example (breast cancer mortality reduction from 5 in 1000 to 4 in 1000 when participating in screening). The example for the risk increase measures is based on the ''pill scare'' (increase of thrombosis from 1 in 7000 to 2 in 7000 when taking the third-generation contraceptive pill). $ER_{Control}$ = event rate in the control group (baseline risk); $ER_{Treatment}$ = event rate in the treatment group.

risk reduction (RRR) is defined as the difference between event rates (risk) in the control and treatment groups, normalized by the event rate in the control group (the baseline risk). An *absolute risk reduction* (ARR), by contrast, is defined as the difference between the risk in the control group and the risk in the treatment group (Table 1).

The same holds true for communicating potential harms of treatments. The *relative risk increase* (RRI) and the *absolute risk increase* (ARI) are defined analogously to risk reductions (Table 1). For instance, in 1995 the UK Committee on Safety for Medicine used the RRI format in its statement that the third generation of the contraceptive pill increased the risk of life-threatening blood clots twofold, that is, by 100%, compared with the second-generation pill.[2,3] This RRI corresponded to an ARI of 0.014% (increase from 1 in 7000 to 2 in 7000).

Whereas relative and absolute risk formats are derived from the very same data (event rates in the control and treatment groups), they are often not psychologically equivalent. For instance, laypeople as well as health professionals evaluate treatment benefits more favorably when they are presented as an RRR rather than an ARR.[4–8] Nonetheless, RRR remains the dominant format for communicating treatment benefits in direct-to-consumer advertisements, in patient decision aids, and within the medical community.[9–14] One suggestion for improving understanding of relative reductions (or increases) is to also communicate the baseline risk.[15]

However, several important questions remain, given that neither the role of the presentation format of the baseline risk nor the influence of people's numeracy in interpreting relative risk changes with baseline risk have been systematically investigated.

Relative Risk Reduction and Baseline Risk

Schwartz and others[16] found that providing the baseline risk improved accuracy in estimating breast cancer mortality when presenting women with a treatment's benefit as an RRR or an ARR. Even when provided with the baseline risk, however, about two-thirds of the participants still gave incorrect mortality rate estimates. Similarly, most participants in Sheridan and colleagues'[17] study were unable to calculate the effect on a given baseline risk when presented with an RRR or an ARR; only about one-fifth of participants correctly estimated the effect in the treatment group based on the given information. Natter and Berry[15] showed that omitting the baseline risk led to an overestimation of both the baseline risk and the risk for the treatment group, whereas providing the baseline risk led to more accurate estimates overall.

The Role of Numeracy in Understanding Relative Risk Reduction With Baseline Risk

Both Schwartz and Sheridan and their colleagues[16,17] also investigated numeracy—the ability to comprehend and use numerical information[18]—as a moderating variable in people's understanding of RRR with baseline risk. They found that participants with lower numeracy particularly had difficulties solving the tasks. For instance, in one study, the proportion of correct answers ranged from 5.8% for those with the lowest numeracy score to 40% for those with the highest score.[16] Similarly, in another study, only 5% of participants with lower numeracy calculated the event rate in the treatment group correctly, whereas up to 50% of those with higher numeracy did.[17] These findings are consistent with related research showing that people's

numeracy mediates their understanding of statistical information, risk perception, and decision making.[18–25]

The Role of Presentation Format in Communicating Baseline Risk

What format should be used to communicate baseline risk? Whereas previous studies often used frequency formats, percentages are also frequently used. This holds both for scientific articles and when reporting results to the public. Examples include statements like ''In the United Kingdom, 35% of deaths are due to cardiovascular causes, compared with about 60% in those with type 2 diabetes and 67% of type 1 diabetic patients over 40 years old.''[26(p874)] or ''In 2004, heart disease was noted on 68% of diabetes-related death certificates among people aged 65 years and older.''[27(p8)] Here, the baseline risk is expressed as a percentage.

But what happens when information on the baseline risk is combined with RRR statements, such as ''Heart failure was reduced by 56%, strokes by 44%, and combined myocardial infarction, sudden death, stroke, and peripheral vascular disease by 34%.''[26(p874)] or ''Reducing diastolic blood pressure from 90 mmHg to 80 mmHg in people with diabetes reduces the risk of major cardiovascular events by 50%''?[27(p10)]

One important question is whether the 2 pieces of information refer to the same reference class (e.g., age group, gender, type of diabetes, etc.). But even when they do, another difficulty lies in the potential ambiguity of the words ''reduced by.'' Assume a randomized control study shows that 68% of people with diabetes have heart disease, and that a new drug reduces the risk of heart disease by 50%. The 50% risk decrease refers to an RRR, suggesting that the event rate is reduced from 68% to 34%. A conceptual misunderstanding would be to interpret the decrease as referring to an ARR. The risk reduction would then be misunderstood as referring to a decrease in *percentage points*, meaning in this case that the risk of cardiovascular events decreases by 50 percentage points from 68% to 18% when blood pressure is reduced.

Misinterpreting a relative difference as an absolute difference (i.e., difference in percentage points) might be especially likely when the baseline risk is expressed as a percentage. When the baseline risk is instead presented as a frequency (e.g., 680 of 1000 instead of 68%) in order to estimate how many people are at risk in the treatment group, it is necessary to convert the RRR into the same ''currency'' (number of people at risk), thus helping to clarify the meaning of an RRR.

Berry and others[28] provided participants with risk increase information, either in a relative or an absolute format or as number needed to treat. When no baseline risk was provided, participants strongly overestimated the side effect, particularly in the RRI condition. Provision of baseline risk improved estimates, and no differences between the 3 formats were found. Yet even with baseline risk information, the risk increase was overestimated. Because, however, participants received the baseline risk in both a percentage and a frequency format, the study does not allow for assessing potential differences between different baseline risk formats. The influence of presentation format was demonstrated in a study by Covey,[29] who examined the effects of baseline risk presentation (frequency v. percentage) on participants' understanding of RRR and ARR. Participants had to choose between 2 treatments where either both treatments were equally effective or 1 treatment was more effective than the other. When treatments differed in effectiveness, participants identified the better treatment only when the baseline risk was provided in a frequency format, regardless of whether an RRR or an ARR was used. However, the study did not investigate how participants interpret RRR when given baseline risk, under what conditions misinterpretations are more likely, or what possible misinterpretations occurred.

RESEARCH QUESTIONS

We report the findings of 2 experiments that examined laypeople's understanding of relative risk changes when the baseline risk is provided. Participants chose between different estimates for the event rate in the treatment group after being given information on the baseline risk (event rate in the control group) and the RRR (Experiments 1a and 2a) or the RRI (Experiments 1b and 2b). To pinpoint participants' understanding of relative risk changes, the available estimates corresponded to alternative interpretations of the relative risk change, such as a relative or absolute reduction (or increase) of the event rate in the treatment group.

Three questions were the focus: Does the presentation format of the baseline risk matter for correctly understanding treatment benefits and harms? To what extent is the correct understanding of relative risks mediated by people's numeracy skills? If relative risks are misunderstood, how are they actually interpreted?

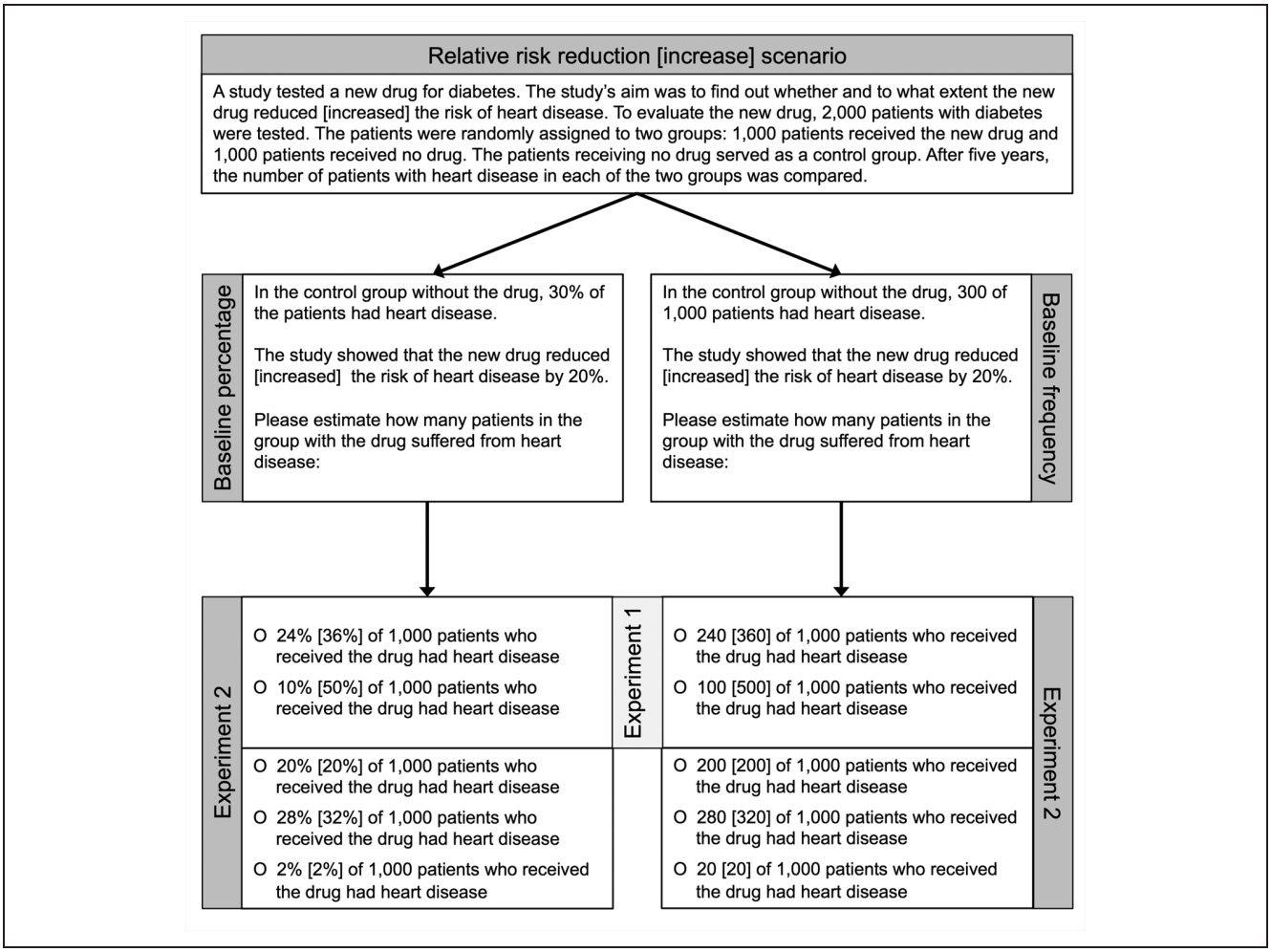


Figure 1 Hypothetical medical scenario and answers used in Experiments 1 and 2. Participants were randomly assigned to either the baseline percentage or baseline frequency condition and provided with a relative risk reduction (RRR) or relative risk increase (RRI). Words and numbers in square brackets refer to risk increase scenarios (Experiments 1b and 2b). The 5 choices correspond to the following answers (in descending order): relative reduction [increase] correctly understood, relative reduction [increase] mistaken as absolute reduction [increase]; relative reduction [increase] mistaken as event rate; Calculation Error I; Calculation Error II. See text for details.

EXPERIMENT 1

We investigated whether and under what conditions a relative risk change is erroneously interpreted as an absolute risk change. In Experiment 1a, participants had to evaluate an RRR; in Experiment 1b, participants were provided with an RRI. We hypothesized that misunderstandings are more likely when the baseline risk is presented as a percentage and for people with lower numeracy. By contrast, presenting the baseline risk in a frequency format should help people (with lower and higher numeracy) correctly understand a relative risk change.

Participants

For all experiments reported in this paper, participants were recruited through the online platform Amazon MTurk (AMT). AMT enables access to a large and diverse subject pool, often more representative of the US population than are in-person convenience samples.[30] Research suggests that findings with AMT are similar in quality and reliability to those from laboratory samples.[31–33] Participants had to be US residents and at least 18 years old.

Experiment 1a had 203 participants (55% male; $\bar{x}$ age = 33 years, SD = 20); Experiment 1b had 182 participants (52% male; $\bar{x}$ age = 33 years, SD = 12).

Participants had various occupations and educational backgrounds. Remuneration was $0.75. In each experiment participants were randomly assigned to 1 of 2 baseline risk formats (percentage v. frequency).

Materials and Procedure

In Experiment 1a, participants received a hypothetical scenario about a drug that reduces the risk of heart disease for patients with diabetes by 20% (Figure 1). Their task was to judge the event rate in the treatment group, given the baseline risk and the RRR.

All participants received the same RRR information: ''The study showed that the new drug reduced the risk of heart disease by 20%.'' Baseline risk was provided as either a percentage or a frequency. In the *percentage condition*, baseline risk was presented as 30%. The answer options were 24% of 1000 (RRR interpretation) and 10% of 1000 (ARR interpretation). In the *frequency condition*, baseline risk was presented as 300 of 1000. Possible answers were 240 of 1000 (RRR interpretation) and 100 of 1000 (ARR interpretation). Participants in Experiment 1b received the same scenario, except that the drug *increased* the risk by 20%. Here, possible answers were 36% (RRI interpretation) and 50% (ARI interpretation) in the percentage condition and 360 of 1000 (RRI interpretation) and 500 of 1000 (ARI interpretation) in the frequency condition. In all studies, the order of the answers was randomized.

We assessed participants' numeracy using Lipkus and others'[23] 11-item numeracy scale, with the items presented in random order. Numeracy is defined as ''the degree to which individuals have the capacity to access, process, interpret, communicate, and act on numerical, quantitative, graphical, biostatistical, and probabilistic health information needed to make effective health decisions.''[34(p375)] Items of the scale require participants, for instance, to transform percentages into frequencies or vice versa, or to identify the highest risk of a given set of risks (e.g., What is the highest risk of getting a disease? 1 in 10, 1 in 100, 1 in 1000). An individual's numeracy score is the total number of correct items (maximum value 11). This scale has been used in a wide range of studies, showing that people with lower numeracy have more difficulties understanding risks, are more prone to framing effects, and are more easily misled by nontransparent formats such as RRRs.[18–25]

Results

Figure 2 (left) shows that people better understood the RRR when the baseline risk was presented as a frequency rather than a percentage. In the percentage condition, 54% of participants correctly judged the event rate in the treatment group, as opposed to 85% in the frequency condition ($\chi^2 = 23.8$, $df = 1$, $P = 0.001$). Baseline risk format was also relevant for understanding a treatment's harms (Figure 2, right): 62% of participants in the percentage condition but 77% of participants in the frequency condition correctly interpreted the RRI ($\chi^2 = 4.6$, $df = 1$, $P = 0.02$).

Influence of numeracy. Aggregated across the 2 experiments, numeracy scores ranged from 4 to 11, with a median of 10 ($\bar{x} = 9.5$, skewness = -1.22). As in previous studies,[21,25] we observed a highly skewed distribution of numeracy scores. To investigate the influence of numeracy, we therefore performed a median split to divide the sample into lower ($\leq$9 items correct) and higher (>9 items correct) numeracy groups and then examined the influence of format separately for each group using χ^2 tests.* The median split facilitates interpretation and visualization of the results and enables better comparability of our findings with previous studies using the same method.[21,25]

The median split (Figure 2) shows that participants with higher numeracy generally performed better but that both lower and higher numeracy participants chose a greater number of correct answers when baseline risk was presented as a frequency (lower numeracy: $\chi^2 = 17.6$, $df = 1$, $P = 0.001$; higher numeracy: $\chi^2 = 10.6$, $df = 1$, $P = 0.001$).

Results for the RRI scenario (Experiment 1b) were similar. Participants with both lower ($\chi^2 = 3.6$, $df = 1$, $P = 0.049$) and higher ($\chi^2 = 2.5$, $df = 1$, $P = 0.08$) numeracy were more likely to correctly understand the RRI when baseline risk was presented as a frequency, although the difference was smaller than in the RRR scenario (Figure 2, right).

Discussion

The majority of participants correctly judged the benefit or harm of a treatment when the baseline risk was presented in a frequency format, whereas a percentage format led many participants to interpret the relative risk change as an absolute one. Lower numeracy

*We also conducted logistic regressions with format (categorical: percentage v. frequency) and numeracy (continuous) as predictors and the interpretation (RRR v. ARR) as dependent variable, separately for the risk reduction (Experiment 1a) and risk increase (Experiment 1b) scenario. In both experiments, format and numeracy were significant predictors for a correct understanding of a relative risk change (see web-only appendix for details). Given that regression models make stronger assumptions about the normality of the data, we here focus on the less demanding χ^2 tests. However, both types of analysis lead to the same conclusions.

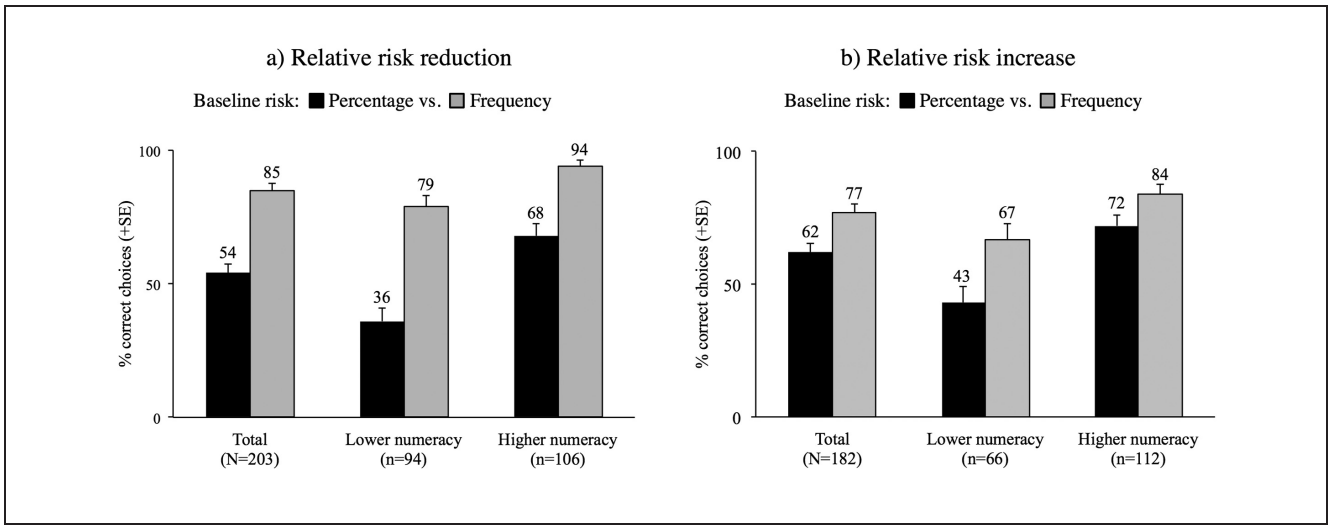


*Figure 2 Proportion of participants correctly understanding the meaning of an RRR (Experiment 1a) and an RRI (Experiment 1b). Correct choices are influenced by the baseline risk presentation format (percentage or frequency) and participants' numeracy. (a) Results of Experiment 1a. "Total" includes all participants in the percentage (*n *= 100) and frequency (*n *= 103) condition. (b) Results of Experiment 1b. "Total" includes all participants in the percentage (*n *= 100) and frequency (*n *= 82) condition. Numeracy was assessed using Lipkus and others'[23] scale; participants were categorized into lower and higher numeracy groups based on a median split. Three participants in Experiment 1a and 4 participants in Experiment 1b did not provide numeracy answers and were excluded from these analyses.*

participants more frequently interpreted the relative risk information erroneously as an absolute risk change. Yet participants with both lower and higher numeracy benefited from a frequency format.

EXPERIMENT 2

Experiment 1 showed that people often erroneously interpret a relative risk change as an absolute change. However, further misinterpretations are possible. Sheridan and colleagues[17] noted that about 20% of their participants interpreted an RRR as directly referring to the event rate in the treatment group. One explanation is that people who are unfamiliar with the basic design of randomized controlled trials might ignore the information about the baseline risk and assume instead that the reduction directly refers to the event rate in the treatment group (henceforth denoted as the ER interpretation). For instance, when the RRR in the heart disease scenario is 20%, people may mistakenly believe that 20% of all people who are treated have heart disease.

We consider these 3 interpretations (RRR, ARR, and ER) as *conceptual* interpretations of an RRR, which should be distinguished from mere calculation errors. For instance, many people have difficulties converting percentages into frequencies and vice versa.[16,18] In our example, 20% of 1000 corresponds to 200 in 1000. A false conversion could result in the answer 20 of 1000. Along with an ARR interpretation, this may lead participants to interpret the 20% risk reduction as 300 − 20 = 280 of 1000 in the frequency condition (*Calculation Error I*). People following an ER interpretation may erroneously estimate the risk in the treatment group and come up with 20 of 1000, or 2% (*Calculation Error II*). In Experiment 2 we therefore provided participants with 5 answer options (i.e., RRR, ARR, ER, Calculation Error I, Calculation Error II) for judging the event rate in the treatment group when given the baseline risk and an RRR or RRI (Figure 1, bottom).

A second goal was to examine participants' choices when the relative risk change cannot be meaningfully interpreted as an absolute risk change. For instance, given a baseline risk of 30% and an RRR of 40%, misunderstanding the reduction as a change in percentage points yields a negative number (i.e., 30% − 40% = −10%), which renders the ARR interpretation implausible. Similarly, given an RRI of 80% and a baseline risk of 30%, an ARI interpretation yields an event rate larger than 100% (i.e., 30% + 80% = 110%). Because the influence of baseline risk format should predominantly depend on whether an ARR or ARI is meaningful, we hypothesized that participants' performance in the high-benefit and high-harm condition should be influenced mainly by their numeracy skills and not by presentation format.

Participants and Design

Recruitment of participants and remuneration was identical to Experiment 1. Experiment 2a (RRR) had 443 participants (54% male, $\bar{x}$ age = 33 years, SD = 11.3); Experiment 2b (RRI) had 406 participants (51% male, $\bar{x}$ age = 30 years, SD = 9.6).

Participants in Experiment 2a were randomly assigned to 1 of 4 conditions: Format of Baseline Risk (percentage v. frequency) × Level of Risk Reduction (low benefit: 20% v. high benefit: 40%). The high-benefit condition ruled out a meaningful ARR interpretation. In Experiment 2b, the same design was used for an RRI: Baseline Risk (percentage v. frequency) × Level of Risk Increase (low harm: 20% v. high harm: 80%); here the high-harm condition ruled out a meaningful ARI interpretation. Baseline risk was 30% (300 of 1000) in all conditions.

Materials and Procedure

The materials and procedure were identical to Experiments 1a and 1b, except that 5 answer options were provided: RRR interpretation (correct answer), ARR interpretation, ER interpretation, Calculation Error I, and Calculation Error II. Figure 1 illustrates the corresponding numerical values exemplarily for the low-benefit (20%) RRR condition; corresponding numbers can be computed for the high-benefit (40%) RRR and the 2 RRI conditions (20% v. 80%). For the 40% RRR condition, the ARR interpretation was set to 0%; in the 80% RRI condition, the ARI answer was set to 100%.

Results and Discussion

Figure 3 (top row) shows the results for the 2 RRR scenarios. In the low-benefit condition, participants' judgments varied depending on baseline risk format ($\chi^2 = 27.5$, $df = 4$, $P = 0.001$) but not in the high-benefit condition ($P = 0.41$). When RRR = 20% and baseline risk was presented as a percentage, only 40% of participants interpreted the statement correctly as a relative reduction (RRR interpretation), while 40% erroneously interpreted it as an absolute difference (ARR interpretation). When baseline risk was presented as a frequency, 63% of participants gave the correct answer; only 10% erroneously interpreted the information as an ARR. Regardless of format, a substantial proportion of participants made an erroneous ER interpretation (12% and 17%, respectively) by assuming the risk reduction to refer directly to the event rate in the treatment group. Calculation Errors I and II occurred rarely.

In the high-benefit condition (RRR = 40%), in which the ARR interpretation is rendered implausible, most participants interpreted the relative risk statement correctly regardless of baseline format (65% v. 68%). Very few ARR interpretations were obtained, but a substantial proportion of participants either followed an ER interpretation or committed Calculation Error I.

Figure 4 (top row) shows participants' judgments in the 2 risk increase scenarios. In the low-harm condition, participants' judgments differed depending on the baseline risk format ($\chi^2 = 15.9$, $df = 4$, $P = 0.003$), whereas in the high-harm condition judgments were similar in pattern in both conditions ($P = 0.61$). When RRI = 20% and baseline risk was presented as a percentage, only 33% of participants correctly judged the risk in the treatment group, and 39% incorrectly interpreted the information as referring to an ARI. Performance was better when baseline risk was presented in a frequency format; 57% of participants gave the correct answer, and only 19% interpreted the risk reduction as an absolute increase. Analogous to the risk reduction scenario, a substantial proportion of participants (22% and 18%, respectively) interpreted the risk reduction in both formats as referring directly to the event rate in the treatment group (ER interpretation); Calculation Errors I and II were rare.

In the high-harm condition, 56% and 61% of participants in the percentage and frequency condition correctly interpreted the risk statement; almost none made an ARI interpretation (2%). Irrespective of format, about one-quarter of participants made an ER interpretation, and between 12% in the percentage condition and 6% in the frequency condition made Calculation Error I.

Influence of numeracy. Aggregated across the 2 experiments, numeracy scores ranged from 3 to 11, with median = 10, $\bar{x}$ = 9.4 (SD = 1.7), and skewness = −1.4. Using a median split, we compared the influence of baseline risk format separately for participants with lower and higher numeracy.[†]

[†]We also conducted logistic regressions with format (categorical: percentage v. frequency) and numeracy (continuous) as predictors and the interpretation (correct v. incorrect) as dependent variable. The regression analyses test for the influence of format and numeracy on interpreting the risk information correctly, independent of the specific misinterpretations. Consistent with the analyses based on the median split, we found that both format and numeracy were significant predictors in the low-benefit and low-harm condition; no effect of format but one of numeracy was obtained in the high-benefit and high-harm condition (see web-only appendix for details).

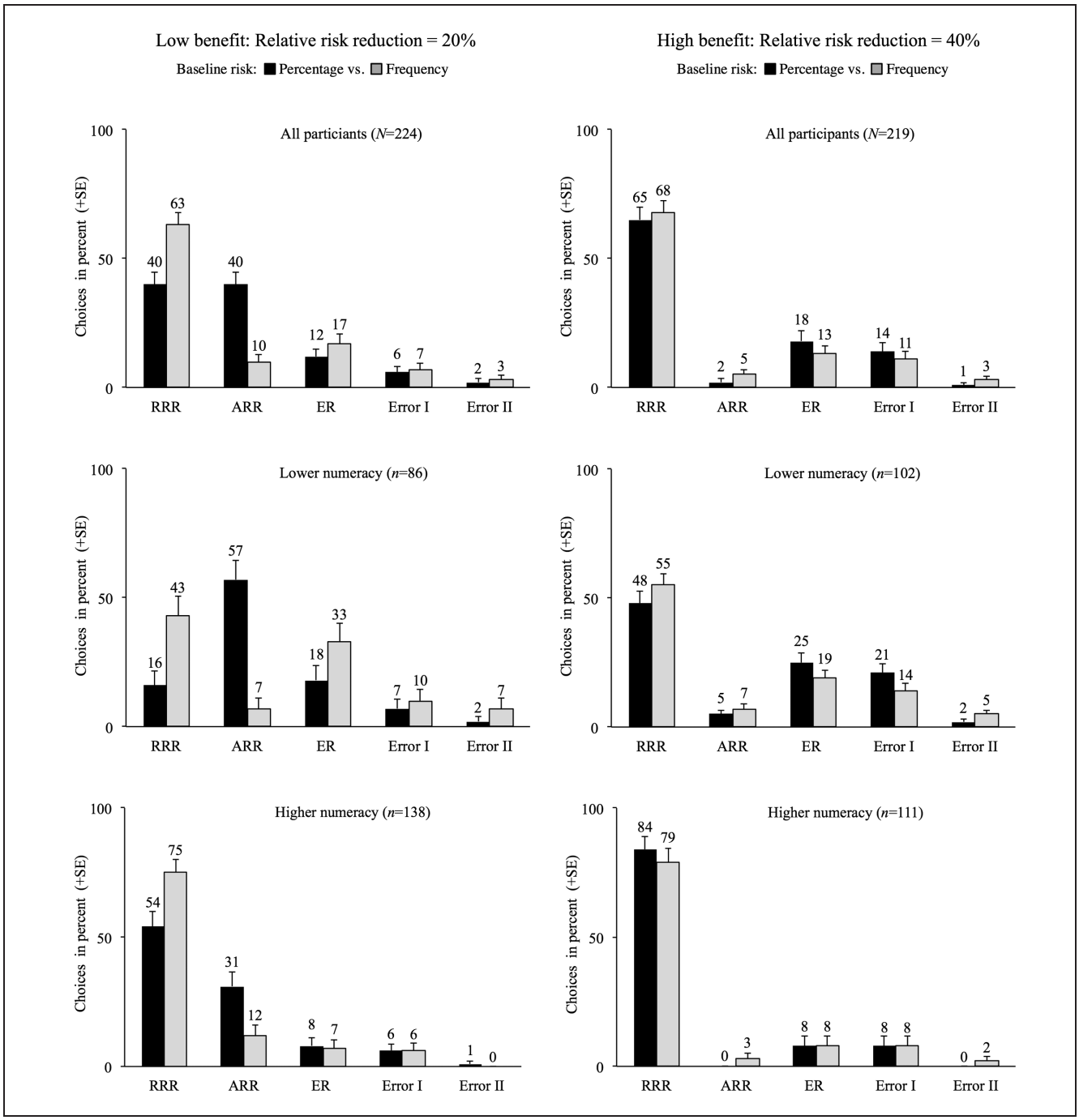


Figure 3 Results of Experiment 2a (N = 443). The interpretation of an RRR depends on the presentation format and participants' numeracy. Even when a meaningful ARR interpretation is ruled out, conceptual misunderstandings remain. The baseline risk was set to 30% (percentage condition) or 300 of 1000 (frequency condition). The left column shows the results for the condition in which the RRR was 20% (n = 224); the right column shows the results for an RRR of 40% (n = 219). The top row shows participants' answers as a function of baseline risk presentation format (percentage v. frequency). The middle and bottom rows show participants' responses separately for lower and higher numeracy participants. Six participants did not complete the numeracy questionnaire and were excluded from these analyses. ARR = absolute risk reduction interpretation; ER = event rate interpretation; Error I = Calculation Error I; Error II = Calculation Error II; RRR = relative risk reduction interpretation. See text for details.

Figure 3 (middle and bottom row) shows participants' judgments in the RRR scenarios separately for lower and higher numerates based on a median split. For participants with lower numeracy, format mattered in the low-benefit condition ($\chi^2 = 24.8$, $df = 4$, $P = 0.001$) but not in the high-benefit condition ($P = 0.73$). For participants with higher numeracy, similar results were obtained with judgments varying for the different formats in the low-benefit condition ($\chi^2 = 8.7$, $df = 4$, $P = 0.067$) but not in the high-benefit condition ($P = 0.67$). When a meaningful ARR interpretation was ruled out (high-benefit condition), participants more often interpreted the statement correctly. However, a substantial proportion of them—particularly of those with lower numeracy—misunderstood the information and based their response on an erroneous ER interpretation or Calculation Error I (Figure 3).

Analyzing lower and higher numeracy participants in the RRI scenarios separately (Figure 4, middle and bottom rows) showed that lower numeracy participants' judgments varied somewhat in both the low-harm ($\chi^2 = 8.9$, $df = 4$, $P = 0.067$) and high-harm ($\chi^2 = 7.8$, $df = 4$, $P = 0.095$) conditions. Judgments of participants with higher numeracy skills varied in the low-harm condition ($\chi^2 = 9.2$, $df = 4$, $P = 0.057$) but not in the high-harm condition ($P = 0.52$).

In sum, as in Experiment 1, more participants correctly understood relative risk changes when base rate information was presented in frequencies as opposed to percentages. In addition to confusing relative risks with absolute risks, mistaking them for event rates was a second kind of misunderstanding. Both misinterpretations tended to be more frequent among participants with lower numeracy than with higher numeracy. However, even when provided with the frequency format, many participants followed an ARI or ER interpretation in the low-harm condition or an ER interpretation along with Calculation Error I in the high-harm condition (Figure 4).

GENERAL DISCUSSION

Treatment benefits and harms are often communicated by RRRs and RRIs. One suggestion for improving the understanding of such health statistics is to include information on baseline risk. Previous studies found that inclusion of this information improved the understanding of RRRs, although many participants still gave incorrect estimates of treatment implications.[15–17] Our study extends this research by providing a more fine-grained analysis of what misinterpretations and calculation errors can occur when using relative risk changes for communicating treatments' benefits and harms.

Our findings emphasize the importance of choosing the right presentation format to communicate the baseline risk and the influence of people's numeracy skills on the perceived meaning of relative risk information. Experiment 1 showed that many participants erroneously interpreted an RRR (RRI) as referring to an absolute decrease (or increase) in risk, particularly when baseline risk was communicated as a percentage. Presenting baseline risk instead in a frequency format improved understanding. Experiment 2 demonstrated not only that people confuse relative with absolute changes but also that a substantial proportion of participants misinterpreted a relative change as referring directly to the event rate in the treatment group. Further problems in correctly understanding treatments' benefits and harms resulted from calculation errors. Notably, many participants misunderstood relative risk formats even in situations in which the numerical values of the baseline risk and the decrease (or increase) precluded misinterpreting a relative change as an absolute one.

Our studies also highlight the role of numeracy as a mediating factor in understanding relative risks. Participants with higher numeracy were more likely to judge the risk in the treatment group correctly, but they too made more accurate inferences when the baseline risk was presented in a frequency format. When an absolute risk interpretation was ruled out (high-benefit and high-harm condition, respectively, in Experiments 2a and 2b), those with lower numeracy still had more difficulties in correctly interpreting the statement than did those with higher numeracy, even when the baseline risk was provided in a frequency format. Given that they frequently opted for the ER interpretation, one possibility is that people with lower numeracy skills also tend to lack the conceptual understanding that risk changes always refer to a control group and therefore misunderstood the information as directly referring to the event rate in the treatment group.

Two important insights into people's understanding of relative risk changes stem from the present findings. First, these results may explain why people often overestimate treatment benefits when they are communicated as RRRs: Interpreting a relative reduction as an absolute reduction results in an overestimation of a treatment's effectiveness. Consider the scenarios in which the baseline risk was 30% and the RRR was 20%. This implies that the event rate in the treatment group is 24%, but misinterpreting the

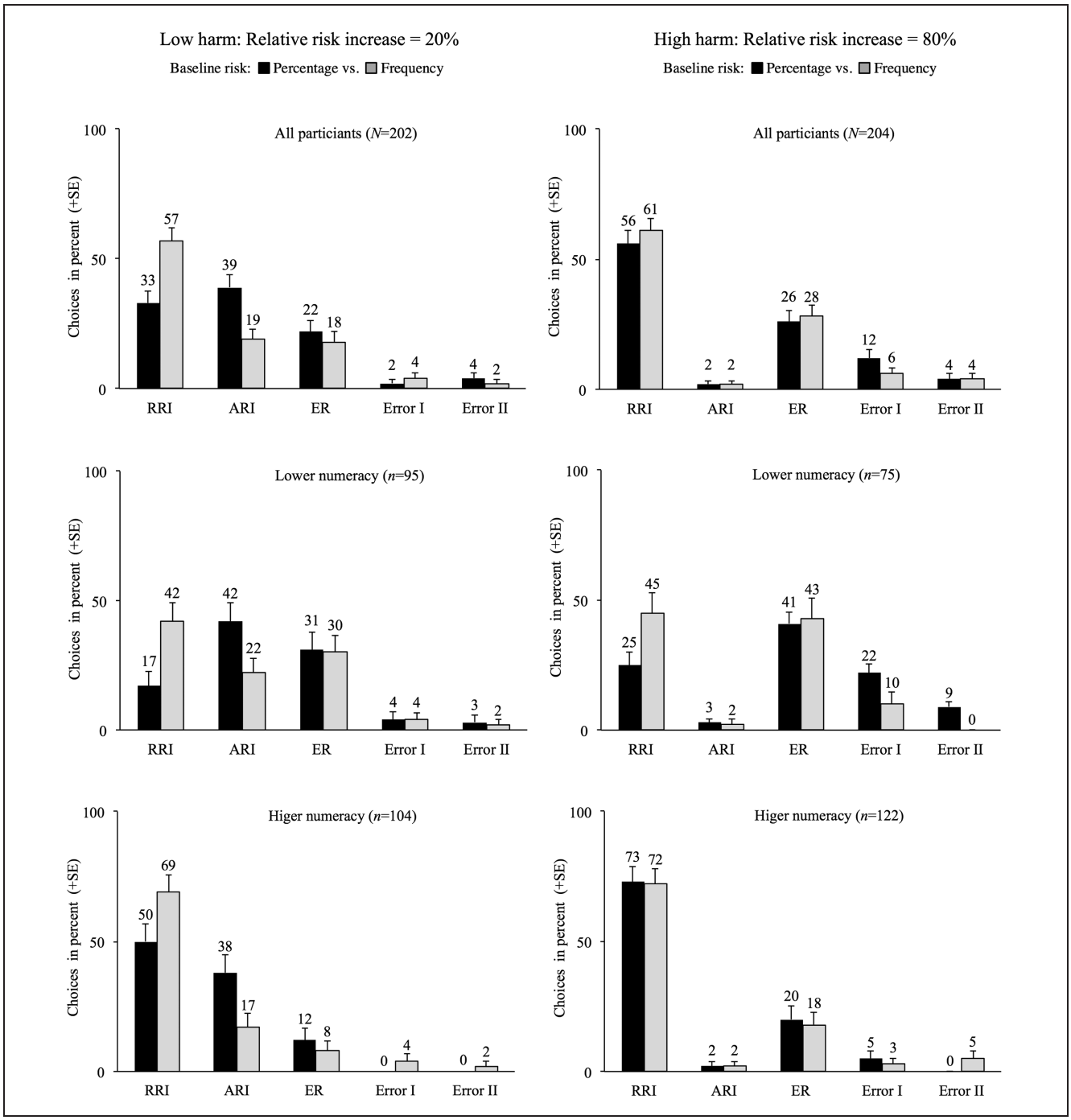
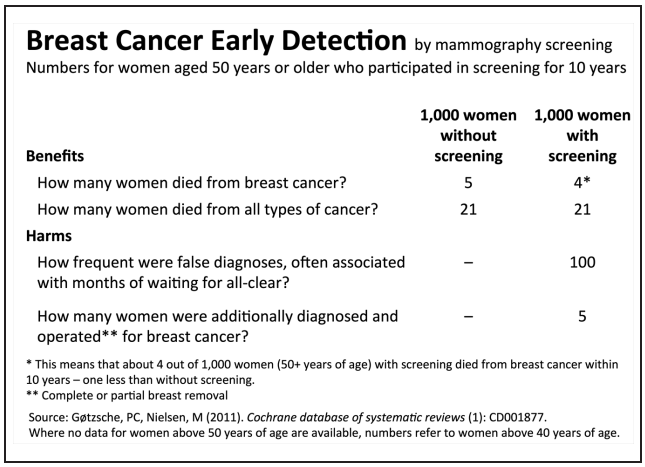


Figure 4 Results of Experiment 2b (N = 406). The interpretation of RRI depends on the presentation format and participants' numeracy. Even when a meaningful ARI interpretation is ruled out, conceptual misunderstandings remain. The baseline risk was set to 30% (percentage condition) or 300 of 1000 (frequency condition). The left column shows the results for the condition in which the RRI was 20%; the right column shows the results for an RRI of 80%. The top row shows participants' answers as a function of the baseline risk presentation format (percentage v. frequency). The middle and bottom row show participants' responses separately for lower v. higher numeracy participants. Thirteen participants did not complete the numeracy questionnaire and were excluded from these analyses. RRI = relative risk increase interpretation; ARI = absolute risk increase interpretation; ER = event rate interpretation; Error I = Calculation Error I; Error II = Calculation Error II. See text for details.



Breast Cancer Early Detection by mammography screening

Numbers for women aged 50 years or older who participated in screening for 10 years

Benefits	1,000 women without screening	1,000 women with screening
How many women died from breast cancer?	5	4*
How many women died from all types of cancer?	21	21
Harms		
How frequent were false diagnoses, often associated with months of waiting for all-clear?	–	100
How many women were additionally diagnosed and operated** for breast cancer?	–	5

* This means that about 4 out of 1,000 women (50+ years of age) with screening died from breast cancer within 10 years – one less than without screening.
** Complete or partial breast removal

Source: Gøtzsche, PC, Nielsen, M (2011). *Cochrane database of systematic reviews* (1): CD001877.
Where no data for women above 50 years of age are available, numbers refer to women above 40 years of age.

Figure 5 Example of a fact box summarizing evidence on benefits and harms of mammography screening. The fact box presents information on the event rates in the control and treatment groups in terms of frequencies and renounces any measures of relative risk reduction.[36] The authors of the Cochrane systematic review[37] distinguished between higher and lower quality studies, based on methodological criteria such as a sound randomization procedure. Across all studies, the authors estimated an absolute risk reduction of 0.5 in 1000. The fact box reports a rounded estimate of 1 in 1000. In any case, these numbers should not be interpreted as definitive but as approximate estimates of possible benefits and harms. For more details, see Gigerenzer[1] and Harding Center for Risk Literacy.[36]

information as an absolute decrease reduces the number to 10%. Second, the finding that a substantial proportion of participants interpreted the RRR (RRI) as directly referring to the event rate in the treatment group points to another problem that impedes understanding of health information: Using information on the baseline risk is beneficial only when people understand that benefits and harms of treatments are evaluated based on a comparison between a control and a treatment group. If people lack this conceptual understanding, they may simply ignore the provided baseline risk and incorrectly assume that the risk information refers directly to the event rate in the treatment group.

Given the current results and other findings from the literature,[4,15–17] we believe that the most helpful way to communicate treatment benefits and harms is to provide information on the event rate in the control group and in the treatment group—the very empirical evidence from which any measure can be derived. So-called fact boxes[35,36] aim to implement this idea. Figure 5 shows a fact box on mammography screening that provides the relevant information in a frequency format. It defines the reference class (women from the population aged 50 years and older), provides information on breast cancer mortality (5 in 1000 without screening v. 4 in 1000 with screening), compares the overall cancer mortality (21 in 1000 with and without screening), and communicates harms such as overdiagnosis (100 of 1000 of those with screening) and overtreatment (5 out of 1000 of those with screening). Thus, a fact box provides patients, doctors, and policy makers with a concise overview of the best evidence currently available (e.g., based on the Cochrane Collaboration's systematic reviews) in a transparent and intuitive manner.

Limitations and Future Research

For the experiments, we used a convenience sample with limited variability in age and educational level. Using more representative samples or specific samples (e.g., patients, doctors) might help identify further moderating variables, such as education, age, or socioeconomic status, to account for a broader range of individual differences in the interpretation of relative risk statements. We also used a forced-choice format derived from a priori specified hypotheses about alternative understandings of relative risk information. Future research should investigate how accurate people's estimates of treatment benefits and harms are when they have to perform the calculation themselves. Finally, future research should examine people's judgments across a broader range of scenarios, with varying baseline risks and treatment benefits and harms.

ACKNOWLEDGMENTS

We thank Anita Todd, Rona Unrau, and Tarlise Townsend for editing the manuscript.

REFERENCES

1. Gigerenzer G. Risk Savvy: How to Make Good Decisions. New York (NY): Viking; 2014.

2. Gigerenzer G, Gray JAM. Launching the century of the patient. In: Gigerenzer G, Gray JAM, eds. Better Doctors, Better Patients, Better Decisions: Envisioning Health Care 2020. Cambridge (MA): MIT Press; 2011. p 3–28.

3. Williams D, Kelly A, Carvalho M, Feely J. Effect of the British warning on contraceptive use in the General Medical Service in Ireland. Ir Med J. 1998;91(6):202–3.

4. Akl EA, Oxman AD, Herrin J, et al. Using alternative statistical formats for presenting risks and risk reductions. Cochrane Database Syst Rev. 2011;(3):CD006776.

5. Covey J. A meta-analysis of the effects of presenting treatment benefits in different formats. Med Decis Making. 2007;27(5): 638–54.

6. Edwards A, Elwyn G, Covey J, Matthews E, Pill R. Presenting risk information—a review of the effects of ''framing'' and other manipulations on patient outcomes. J Health Commun. 2001; 6(1):61–82.

7. Forrow L, Taylor WC, Arnold RM. Absolutely relative: how research results are summarized can affect treatment decisions. Am J Med. 1992;92(2):121–4.

8. Mühlhauser I, Kasper J, Meyer G, Federation of European Nurses in Diabetes. Understanding of diabetes prevention studies: questionnaire survey of professionals in diabetes care. Diabetologia. 2006;49(8):1742–6.

9. Jorgensen KJ, Gøtzsche PC. Presentation on websites of possible benefits and harms from screening for breast cancer: cross sectional study. BMJ. 2004;328(7432):148.

10. Moynihan R, Bero L, Ross-Degnan D, et al. Coverage by the news media of the benefits and risks of medications. N Engl J Med. 2000;342(22):1645–50.

11. Slater MD, Long M, Bettinghaus EP, Reineke JB. News coverage of cancer in the United States: a national sample of newspapers, television, and magazines. J Health Commun. 2008;13(6):523–37.

12. Gigerenzer G, Wegwarth O, Feufel M. Misleading communication of risk. BMJ. 2010;341:c4830.

13. Schwartz LM, Woloshin S, Dvorin EL, Welch HG. Ratio measures in leading medical journals: structured review of accessibility of underlying absolute risks. BMJ. 2006;333(7581):1248.

14. Sedrakyan A, Shih C. Improving depiction of benefits and harms: analyses of studies of well-known therapeutics and review of high-impact medical journals. Med Care. 2007;45(10 Suppl 2): S23–8.

15. Natter HM, Berry DC. Effects of presenting the baseline risk when communicating absolute and relative risk reductions. Psychol Health Med. 2005;10(4):326–34.

16. Schwartz LM, Woloshin S, Black WC, Welch HG. The role of numeracy in understanding the benefit of screening mammography. Ann Intern Med 1997;127(11):966–72.

17. Sheridan SL, Pignone MP, Lewis CL. A randomized comparison of patients' understanding of number needed to treat and other common risk reduction formats. J Gen Intern Med. 2003;18(11): 884–92.

18. Reyna VF, Nelson WL, Han PK, Dieckmann NF. How numeracy influences risk comprehension and medical decision making. Psychol Bull. 2009;135(6):943–73.

19. Galesic M, Garcia-Retamero R. Statistical numeracy for health: a cross-cultural comparison with probabilistic national samples. Arch Intern Med. 2010;170(5):462–8.

20. Galesic M, Garcia-Retamero R, Gigerenzer G. Using icon arrays to communicate medical risks: overcoming low numeracy. Health Psychol. 2009;28(2):210–6.

21. Garcia-Retamero R, Galesic M. Who profits from visual aids: overcoming challenges in people's understanding of risks [corrected]. Soc Sci Med. 2010;70(7):1019–25.

22. Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S. Helping doctors and patients make sense of health statistics. Psychol Sci Public Interest. 2007;8:53–96.

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

24. Lipkus IM, Peters E. Understanding the role of numeracy in health: proposed theoretical framework and practical insights. Health Educ Behav. 2009;36(6):1065–81.

25. Peters E, Västfjäll D, Slovic P, Mertz CK, Mazzocco K, Dickert S. Numeracy and decision making. Psychol Sci. 2006;17:407–13.

26. Watkins PJ. Cardiovascular disease, hypertension, and lipids. BMJ. 2003;326:874–6.

27. Centers for Disease Control and Prevention. Age-adjusted percentage of persons with diabetes aged 35 years and older reporting any cardiovascular disease condition, by sex, United States, 1997–2009. Available at: URL: http://www.cdc.gov/diabetes/statistics/cvd/fig5.htm

28. Berry DC, Knapp P, Raynor T. Expressing medicine side effects: assessing the effectiveness of absolute risk, relative risk, and number needed to harm, and the provision of baseline risk information. Patient Educ Couns. 2006;63(1–2):89–96.

29. Covey J. The effects of absolute risks, relative risks, frequencies, and probabilities on decision quality. J Health Commun. 2011;16(7):788–801.

30. Berinsky AJ, Huber GA, Lenz GS. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. Polit Anal. 2012;20(3):351–68.

31. Buhrmester M, Kwang T, Gosling SD. Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data? Perspect Psychol Sci. 2011;6(1):3–5.

32. Paolacci G, Chandler J, Ipeirotis PG. Running experiments on Amazon Mechanical Turk. Judgm Decis Making. 2010;5:411–9.

33. Sprouse J. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. Behav Res Methods. 2010;43(1):155–67.

34. Goldbeck AL, Ahlers-Schmidt CR, Paschal AM, Dismuke S. A definition and operational framework for health numeracy. Am J Prev Med. 2005;29:375–6.

35. Schwartz LM, Woloshin S, Welch HG. Using a drug facts box to communicate drug benefits and harms: two randomized trials. Ann Intern Med. 2009;150(8):516–27.

36. Harding Center for Risk Literacy. Risks and benefits of mammography screening. Available from: URL: http://harding-center.de/index.php/en/what-you-should-know/facts-boxes/mammography

37. Gøtzsche PC, Nielsen M. Screening for breast cancer with mammography. Cochrane Database Syst Rev. 2011;(1):CD001877.

Results of the Logistic Regressions for Experiments 1 and 2 with Interpretation (Correct vs. False) as Dependent Variable and Baseline Risk Format (Percentage vs. Frequency) and Numeracy Score as Predictors

Predictor	*B*	*SE B*	*p*	e^B *(odds ratio)*
Experiment 1a: Relative risk reduction = 20%				
Constant	-4.13	1.13	<.001	
Format	1.81	.37	<.001	6.11
Numeracy	0.26	.10	.002	1.29
Experiment 1b: Relative risk increase = 20%				
Constant	-3.75	1.2	.002	
Format	0.85	.36	.02	2.34
Numeracy	0.35	.11	<.001	1.41
Experiment 2a: Relative risk reduction = 20%				
Constant	-7.75	1.44	<.001	
Format	1.16	.305	<.001	3.19
Numeracy	0.62	.129	<.001	1.86
Experiment 2a: Relative risk reduction = 40%				
Constant	-2.73	1.38	<.05	
Format	0.14	.31	.65	1.15
Numeracy	0.33	.08	<.001	1.39
Experiment 2b: Relative risk increase = 20%				
Constant	-6.51	1.34	<.001	
Format	0.88	.31	<.01	2.41
Numeracy	0.52	.13	<.001	1.68
Experiment 2b: Relative risk increase = 80%				
Constant	5.97	1.65	<.001	
Format	0.31	.32	.33	1.36
Numeracy	0.55	.11	<.001	1.73