Contents lists available at ScienceDirect

Patient Education and Counseling

journal homepage: www.elsevier.com/locate/pateducou

Patient Perception, Preference and Participation

Assessing risk communication in breast cancer: Are continuous measures of patient knowledge better than categorical?

Jeffrey Belkora [a,*], Dan H. Moore [a], David W. Hutton [b]

[a] University of California, San Francisco, United States
[b] Stanford University, United States

ARTICLE INFO

Article history:
Received 29 March 2008
Received in revised form 29 October 2008
Accepted 12 November 2008

Keywords:
Risk communication
Patient knowledge
Realistic expectations
Shared decision making
Physician–patient relations
Decision aids
Adjuvant!
Breast cancer
Software
Patient education
Measurement
Information theory

ABSTRACT

Objective: To compare the performance of categorical and continuous measures of patient knowledge in the context of risk communication about breast cancer, in terms of statistical and clinical significance as well as efficiency.

Methods: Twenty breast cancer patients provided estimates of 10-year mortality risk before and after their oncology visit. The oncologist reviewed risk estimates from Adjuvant!, a well-validated and commonly used prognostic model. Using the Adjuvant! estimates as a gold standard, we calculated how accurate the patient estimates were before and after the visit. We used three novel continuous measures of patient accuracy, the absolute bias, Brier, and Kullback–Leibler scores, and compared them to a categorical measure in terms of sensitivity to intervention effects. We also calculated the sample size required to replicate the primary study using the categorical and continuous measures, as a means of comparing efficiency.

Results: In this sample, the Kullback–Leibler measure was most sensitive to the intervention effects ($p = 0.004$), followed by Brier and absolute bias (both $p = 0.011$), and finally the categorical measure (0.125). The sample size required to replicate the primary study was 18 for the Kullback–Leibler measure, 23 for absolute bias and Brier, and 37 for the categorical measure.

Conclusions: The continuous measures led to more efficient sample sizes and to rejection of the null hypothesis of no intervention effect. However, the difference in sensitivity of the continuous measures was not statistically significant, and the performance of the categorical measure depends on the researcher's categorical cutoff for accuracy. Continuous measures of patient accuracy may be more sensitive and efficient, while categorical measures may be more clinically relevant.

Practice implications: Researchers and others interested in assessing the accuracy of patient knowledge should weigh the trade-offs between clinical relevance and statistical significance while designing or evaluating risk communication studies.

1. Introduction

Many researchers are currently using categorical measures to evaluate whether educational interventions for patients, including those focused on specific decisions known as decision aids, help patients absorb and recall risk information. For example, a recent Cochrane review of decision aids included five studies that quantitatively assessed the accuracy of patient knowledge using categorical measures. See Table 1 for a summary. The categorical measures in these five studies assessed the number of patients whose risk estimates fell within a certain margin of error of a gold standard estimate derived from outcome data.

Continuous measures are generally thought to be more sensitive and powerful than categorical measures [1]. Yet none of the studies included in the Cochrane review of decision aids used continuous measures of patient accuracy, even though continuous measures are available.

Several continuous measures of accuracy have been used in diverse fields from education to weather forecasting [2]. A continuous measure is a weighted or transformed difference between the patient estimate and the gold standard, on a continuous scale. Three continuous measures used to evaluate probabilistic forecasts are: the absolute bias (absolute value of the difference or bias between an estimate and a gold standard); the Brier score or squared difference; and the Kullback–Leibler logarithmic score, which is a logarithmic function of the estimate and the gold

[*] Corresponding author at: University of California, 3333 California Street, Suite 265, San Francisco, CA 94118, United States. Tel.: +1 650 533 6965; fax: +1 415 651 8574.
E-mail address: jeff.belkora@ucsfmedctr.org (J. Belkora).

0738-3991/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.pec.2008.11.012

Table 1
Properties of studies of quantitative patient risk estimates included in Cochrane Review of Decision Aids[a] [14].

Study	Health concern	Sample size	Intervention comparison	Measure	Gold standard	Standard of accuracy[b]
Dodin and Légaré [15]	Hormone replacement therapy	101	Decision aid versus usual care	Lifetime probability of heart disease, hip fracture, and breast cancer	Data calculated from Grady et al. [16]	Correct quartile
O'Connor et al. [17]	Hormone replacement therapy	165	Decision aid versus usual care	Lifetime probability of heart disease, hip fracture, and breast cancer	Data calculated from Grady et al. [16]	Correct quartile
Rostom et al. [18]	Hormone replacement therapy	51	Decision aid versus usual care	Lifetime probability of heart disease, hip fracture, and breast cancer	Data calculated from Grady et al. [16]	Within ±15%
McBride et al. [19]	Hormone replacement therapy	539	Decision aid versus usual care	Ten year probability of breast cancer	Gail Model Probability Risk Score [20]	Within ±10%
Man-Son-Hing et al. [21]	Antithrombotic therapy	287	Decision aid versus usual care	Chance of stroke and bleeding	Data from SPAF III trial and [22]	Within 1–3%

[a] A decision aid is a presentation of printed or audio-visual material that educates patients about specific choices and their potential risks and benefits.
[b] None of these studies provided justification for the clinical relevance of the standard of accuracy that they used to evaluate patient estimates.

standard. The Brier score was initially developed for use in weather forecasting, but has been used for calibration of professional or model-based probabilistic estimates, such as ICU mortality estimates [2,3]. The Kullback–Leibler logarithmic measure is a measure developed and used in information theory, but has been used in imaging to identify disease such as breast cancer microcalcifications [4,5]. However, we have found no studies of patient knowledge that use these continuous measures. As a result, researchers in the field of risk communication do not know the feasibility of using continuous measures to measure accuracy of patient knowledge, nor do we know the advantages or disadvantages of continuous measures relative to the existing categorical measures.

We hypothesize that widely used categorical measures of patient knowledge are potentially insensitive to clinically and statistically significant changes, and may result in inappropriate interpretations of study results, compared to potentially more sensitive continuous measures. In addition, the categorical measures may contribute to inflated sample sizes and therefore inefficient use of resources.

We explored our hypotheses in a secondary analysis of empirical data from a risk communication study. Our research questions were: how sensitive and efficient were each of the categorical and continuous measures; and were there significant differences in sensitivity across these measures?

2. Methods

2.1. Study design

We conducted a secondary analysis of data originally collected to establish the feasibility and efficacy of using a novel risk communication aid during breast oncology consultations. The study provides an appropriate and interesting dataset in which to conduct a preliminary exploration of categorical and continuous accuracy measures. The dataset is small enough to be included in the first three columns of Table 2, along with key results, so that readers can easily reproduce and verify calculations. The intervention was designed to improve accuracy, as measured by a categorical standard with clinical relevance, but the improvement was not statistically significant ($p = 0.125$). This allowed us to explore whether continuous measures would make a material difference in hypothesis testing (i.e. achieve statistical significance), and if so whether the continuous measures were as clinically relevant as the categorical one. We now briefly describe aspects of the primary study design that were relevant to the secondary analysis. We provide a detailed description of the primary study in another report [6].

2.2. Population, setting, and study site

The primary study took place at the breast care center in an academic medical center in San Francisco. The center treats over 500 newly diagnosed breast cancer patients per year. The population of new breast cancer patients at this center is mostly White, college educated, affluent, and insured.

2.3. Subjects, recruitment, consent

Researchers recruited a convenience sample of 20 patients consecutively referred for oncology consultations with either of two senior oncologists. Patients were eligible to participate in the study if they could speak and read English, if they had completed surgery for stage I, II, or IIIa breast cancer, if they had not initiated any form of adjuvant therapy, and if their medical charts included tumor size, tumor grade, hormone receptor status, node status, and age. Patients were not eligible to participate in the study if they had metastatic disease, if they needed further surgery to complete staging, or if they were unable to provide informed consent. The institutional Committee on Human Research and the funding agency's Institutional Review Board approved the study protocol. Patients were enrolled between October 2001 and February 2002.

2.4. Outcomes and instruments

Subjects filled out a brief survey asking them to estimate their 10-year mortality risk with and without adjuvant therapy, before and after an educational presentation of gold standard estimates by their oncologist. The pre- and post-visit surveys took the form: "The chance that I will die from my breast cancer within the next 10 years after having [therapy] is (circle one): [response]." The response format was a list of potential responses ranging from 0% to 100% in increments of 5%. Patients were prompted to respond for local therapy (surgery and local radiation) and adjuvant therapy (systemic chemotherapy or hormone therapy) scenarios. This secondary analysis examined the accuracy of patient estimates as to the risk of dying within 10 years after local therapy only, i.e. with no adjuvant therapy. We focused our analysis on this topic because local therapy prognosis is a baseline prognosis that patients should understand before they consider adding therapy.

2.5. Intervention

The intervention consisted of an oncologist reviewing a printout of the graphs from the Adjuvant! software program, which presents estimates of the patient prognosis based on

Table 2

Patient and gold standard estimates of 10-year local therapy mortality risk, with categorical and continuous measures of accuracy. Highlighted numbers are undefined boundary values for Kullback–Leibler. For Kullback–Leibler calculation, the boundary value 0 was replaced with the minimum (0.05) and the boundary value 1 with the maximum (0.95) allowable response on the study survey. *Note*: Values in this table were log-transformed to make them normally distributed for *t*-tests and *d*-values discussed in the text.

	Patient Estimate		Categorical (within ± 5%)		Absolute Bias			Brier			Kullback-Leibler Divergence		
Adjuvant!	Before	After	Before	After	Before	After	Change	Before	After	Change	Before	After	Change
4%	10%	**0%**	0	1	0.060	0.040	0.020	0.004	0.002	0.002	0.025	0.0011	0.024
6%	10%	10%	1	1	0.040	0.040	0.000	0.002	0.002	0.000	0.010	0.0102	0.000
8%	10%	5%	1	1	0.020	0.030	-0.010	0.000	0.001	-0.001	0.002	0.0081	-0.006
8%	20%	**0%**	0	0	0.120	0.080	0.040	0.014	0.006	0.008	0.055	0.0081	0.047
9%	30%	5%	0	1	0.210	0.040	0.170	0.044	0.002	0.043	0.130	0.0138	0.117
9%	90%	**0%**	0	0	0.810	0.090	0.720	0.656	0.008	0.648	1.802	0.0138	1.789
12%	10%	20%	1	0	0.020	0.080	-0.060	0.000	0.006	-0.006	0.002	0.0226	-0.020
13%	10%	10%	1	1	0.030	0.030	0.000	0.001	0.001	0.000	0.005	0.0046	0.000
14%	30%	10%	0	1	0.160	0.040	0.120	0.026	0.002	0.024	0.070	0.008	0.062
15%	10%	20%	1	1	0.050	0.050	0.000	0.003	0.003	0.000	0.012	0.0084	0.004
17%	10%	10%	0	0	0.070	0.070	0.000	0.005	0.005	0.000	0.023	0.023	0.000
22%	5%	20%	0	1	0.170	0.020	0.150	0.029	0.000	0.029	0.172	0.0012	0.171
23%	5%	10%	0	0	0.180	0.130	0.050	0.032	0.017	0.016	0.189	0.0714	0.118
24%	90%	75%	0	0	0.660	0.510	0.150	0.436	0.260	0.176	1.224	0.5715	0.653
27%	**0%**	25%	0	1	0.270	0.020	0.250	0.073	0.000	0.073	0.263	0.001	0.262
28%	10%	10%	0	0	0.180	0.180	0.000	0.032	0.032	0.000	0.128	0.1276	0.000
28%	10%	40%	0	0	0.180	0.120	0.060	0.032	0.014	0.018	0.128	0.0314	0.096
39%	5%	50%	0	0	0.340	0.110	0.230	0.116	0.012	0.104	0.531	0.0244	0.506
41%	70%	40%	0	1	0.290	0.010	0.280	0.084	0.000	0.084	0.180	0.0002	0.180
89%	**100%**	80%	0	0	0.110	0.090	0.020	0.012	0.008	0.004	0.029	0.0291	0.000
Averages			5	10	0.199	0.089	0.110	0.080	0.019	0.061	0.249	0.049	0.200
St. Dev.					0.207	0.108	0.174	0.166	0.057	0.146	0.460	0.127	0.414

patient-specific inputs consisting of age, tumor size and grade, estrogen receptor status, node status, and number of comorbidities [7]. Adjuvant! is a validated, widely used prognostic model [8,9]. Its estimates functioned as a gold standard for patient 10-year local therapy mortality risk in this study. Oncologists presented several estimates; our analysis focuses on local therapy only.

2.6. Data collection and management procedures

A research assistant transcribed the survey responses to an Excel workbook, and entered the corresponding gold standard estimates from Adjuvant! for comparison with patient estimates.

2.7. Measures of patient accuracy

2.7.1. Categorical

An indicator of whether each patient estimate was within ±5% of the gold standard. This was a binary variable with 1 indicating an estimate within 5% and 0 indicating an estimate outside the 5% threshold. The oncologists recruiting patients to the original study felt that 5% was a clinically meaningful threshold for accuracy in this patient population, and it corresponds to a conservative estimate for the Adjuvant! model's margin of error [8].

2.7.2. Continuous

- Absolute bias, defined as the magnitude of the difference between the patient and Adjuvant! estimate.
- Brier score, defined as the square of the difference between the patient and gold standard estimates.
- Kullback–Leibler divergence score, defined by

$$D_{KL}(P|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

where P and Q represent the gold standard and patient estimates, respectively. In this case the summation is over the two possible outcomes, survival or death and $P(2) = 1 - P(1)$, $Q(2) = 1 - Q(1)$ with $P(1)$ and $Q(1)$ representing the 10-year mortality estimates. The Kullback–Leibler divergence score is undefined for estimates of 0 or 1. We substituted 5% for 0 and 95% for 1 in the computation of the Kullback–Leibler score, since these were the closest allowable responses to 0 and 1 on the patient survey.

All four measures are used in information theory. A thorough discussion of these and other measures can be found in Grunwald and Dawid [10]. We selected these measures because they are easy to compute and are widely used in information theory.

All of the continuous measures are lowest (zero) when the patient estimate equals the gold standard and increase as the patient estimate moves away from the gold standard. Since we were interested in the measuring the effect of the oncologists' risk communication, we used the difference in these statistics, *before* minus *after* scores, as a measure of the information gained by the patient as a result of exposure to the oncologist presentation. For the continuous measures, positive values in this difference were associated with improvements in patient accuracy, since that means the *before* score was bigger (worse) than the *after* score.

2.8. Analysis plan

The study questions were: "How sensitive and efficient were each of the categorical and continuous measures; and were there significant differences in sensitivity across these measures?".

We began by calculating the sensitivity and efficiency for each measure. As a first measure of sensitivity, we tested the null hypothesis of no change in patient accuracy after compared to before the oncology visit to see if the measures detected a statistically significant change. For the categorical measure, because our data were paired we used McNemar's test to compare the number of patients that improved to within ±5% versus the number that started within but ended up outside that margin of error. We ran additional scenarios to explore the effect on statistical sensitivity, as measured by *p*-values, of relaxing our categorical standard of accuracy from ±5% to ±10% and 15%.

For the continuous measures, we applied the Shapiro–Wilks test for normality, transformed the distributions as needed, and then used a paired *t*-test to determine whether the observed differences were statistically significant. Since it was possible that the risk presentation could have led to decreased patient accuracy, through patient overload, both tests were two-sided at 5% level of significance.

As an index of sensitivity, we calculated the *p*-value for the paired comparisons defined by the categorical and three continuous measures.

For each measure we then calculated the sample size needed to have 90% power (at 5% significance level) to replicate the observed effect in a separate, independent study. See Appendix 1 for sample size calculation details. These calculations were performed using Stata 10 [11] and provided a measure of the efficiency of each of the categorical and continuous measures considered separately for this sample of patients.

In order to statistically test whether the continuous measures differed significantly in their sensitivity, we tested for differences in standardized effect sizes, denoted *d*. We had to take into account that our data consists of a single set of patient scores. The differences in measures are due to different calculation methods rather than different samples of scores. The measures are therefore correlated, which complicates statistical comparisons. In particular, we could not use standard meta-analytic techniques for testing the homogeneity of effect sizes across different samples [12]. Instead we used a bootstrap method to calculate 95% confidence intervals for the differences between *d*'s across the continuous measures [13].

Specifically, we wrote a bootstrap function in the statistical programming language R to resample, with replacement, scores from our 20 patients. For each bootstrap sample we calculated the difference between the log-transformed Kullback–Leibler and Brier estimated *d*'s (the *d* for the absolute bias is identical to that for the Brier). This was repeated 1000 times and the 2.5 and 97.5 percentile values were used to estimate the 95% confidence interval for the difference in *d*'s. A confidence interval spanning zero would indicate no statistically significant difference in the effect size across the continuous measures.

3. Results

3.1. How sensitive and efficient were each of the categorical and continuous measures?

For the categorical measure, 5 of 20 patients were accurate within ±5% before, one of whom became inaccurate. Of the 15 that were inaccurate before the intervention, 6 improved to within ±5%. Therefore we applied McNemar's test to the 7 who changed (representing 35% of the sample of 20), of whom 6 (86% of the patients who changed) improved to within ±5% ($p = 0.125$). Using the categorical measure, we failed to reject the null hypothesis of no change in accuracy of patient knowledge.

Regarding the continuous measures, absolute bias was reduced from an average of 0.199 to 0.089; the Brier score went from 0.080 to 0.019; and Kullback–Leibler divergence went from 0.249 to 0.049 based on resetting boundary values to 5% or 95%. However, all three distributions were skewed in this sample so a transformation was necessary in order to apply a paired *t*-test to the differences.

Upon transforming to logarithms, the Brier score is equal to twice the absolute bias score ($\log(\text{abs}(x-y)) = (1/2)\log((x-y)^2)$). All transformed measures passed the Shapiro–Wilks test for normality. Thus, we used a paired *t*-statistic to test whether the improvement in log-transformed units was statistically significant. For the pilot data $t = 2.84$ for the absolute bias and the Brier and $t = 3.24$ for the Kullback–Leibler Divergence, each statistic having 19 degrees of freedom.

The associated two-sided *p*-values for the continuous measures were $p = 0.011$ for the absolute bias and Brier and $p = 0.004$ for the Kullback–Leibler Divergence. See Table 3. Using any of the continuous measures, we rejected the null hypothesis of no change in accuracy of patient knowledge.

Regarding the efficiency of the continuous measures compared to the categorical one, Table 3 also displays the sample sizes required for 90% power to detect our observed effect sizes in an independent sample. The sample size of 17 for Kullback–Leibler, the most sensitive, powerful, and efficient measure, is less than half of the sample size of 38 calculated for the categorical measure. As we increased the categorical threshold for accuracy from within 5% to within 10% and 15%, we observed decreasing *p*-values, as illustrated in Table 4. A categorical measure with a threshold of 15% led to an observed *p*-value (0.004) matching the most sensitive continuous measure.

3.2. Were there significant differences in sensitivity across these measures?

3.2.1. Confidence interval for difference in d's

The bootstrap method gave a 95% confidence interval for the difference, Kullback–Liebler minus Brier, *d* from −0.02 to 0.24. We failed to reject the hypothesis of no difference in *d*'s.

Table 3
Summary statistics for categorical and continuous measures.

	Categorical (*N* within ±5%)	Absolute bias[a]	Brier[*]	Kullback–Leibler[b]
Observed sample size	20	20	20	20
p-Value	0.125	0.011	0.011	0.004
d		0.63	0.63	0.72
Required sample size (1)	38	22	22	17
Difference in *d*'s [95% confidence interval] as a test of whether measures detect different effect sizes				[−0.02, 0.24]

Notes: (1) Sample size is for 90% power for one-sided replication of observed effects at the 5% level of significance. See Appendix 1 for calculation details.

[a] Absolute bias and Brier are identical upon log transformation because ($\log(\text{abs}(x-y)) = (1/2)\log((x-y)^2)$).

[b] Five imputed (closest) responses replaced five patient responses for which Kullback–Leibler Divergence is undefined.

Table 4
Sensitivity of categorical measure to threshold.

Categorical threshold	Within threshold before	Within threshold after	Number changing	Improvements	*p*-Value
±5%	5	10	7	6	0.125
±10%	7	15	8	8	0.008
±15%	9	18	9	9	0.004

4. Discussion and conclusion

4.1. Discussion

4.1.1. Analysis and interpretation

Any of the continuous measures led to rejection of the null hypothesis of no change in accuracy of patient knowledge, while the categorical measure did not. One implication is that the intervention might have been retained for further investigation using continuous measures, but discarded using categorical measures. In addition, studies of this nature might be less than half as costly or lengthy using continuous compared to categorical measures.

For this sample, the Kullback–Leibler performed best compared to the Brier, absolute bias, and a categorical measure with an accuracy threshold of ±5%. However, the effect sizes for the continuous measures were too similar for us to rule out the role of chance in creating an apparent advantage, in this sample, of the Kullback–Leibler over the absolute bias and Brier measures. Likewise, we cannot generalize about the performance of continuous measures compared to the categorical measures beyond this sample, since at other accuracy thresholds, the categorical measure might be more sensitive and efficient.

However, our study does illuminates the possible trade-offs between statistical significance and clinical relevance that can be encountered in general cases. The performance of the continuous and categorical measures must be weighed against their clinical relevance, interpretability, and ease of use. For example, the Kullback–Leibler Divergence was not only a sensitive and powerful measure in our sample, but also was the least interpretable, and hardest to use. It measures distance on a complex log scale; it is not symmetric; and it is not defined for patient estimates of 0% or 100%. On the other hand, the Kullback–Leibler Divergence has advantageous information-theoretic properties. It penalizes inaccuracies much more heavily if the gold standard probability is close to 0% or 100%, where an error of 5% may be more consequential than when the gold standard estimate is close to 50%. For example, among breast cancer patients, a patient at very low mortality risk after surgery may not need chemotherapy, which would be recommended for a higher risk patient. So a patient at very low mortality risk who overestimates their risk may change their treatment decision, whereas a patient at intermediate risk can be inaccurate without it affecting the care plan.

We found absolute bias and Brier scores to have the advantage of being much more familiar and intuitive than the Kullback–Leibler Divergence. Absolute bias is simply the magnitude of the difference between the patient and gold standard estimates. Brier is the square of the difference. Because these scores are non-negative, bounded by 0 and 1, and likely to be skewed, they will often need to be log-transformed, which renders them identical.

Finally, while categorical measures of accuracy may be least sensitive and powerful, they can be designed to be most clinically relevant. Researchers, clinicians and patients may be able to identify treatment thresholds, or points at which their preferences for care change from one treatment to another. From these treatment thresholds, researchers can derive appropriate standards for accuracy of patient estimates. For example, in breast cancer, there is a treatment threshold where recommendations change from local therapy only to chemotherapy. This threshold will be dependent on patient biology and preferences. However, for many early-stage patients the 10-year local therapy mortality risk may be under 10%, and the absolute benefit of chemotherapy may be under 5%. Our study team reasoned therefore that the maximum margin of error for informed decision making for early-stage patients should be ±5%, which also corresponds to a conservative margin of error for Adjuvant! estimates.

Researchers, clinicians, and patients can easily interpret the meaning of study results when expressed in categorical terms. For example, in our study, 5 patients estimated their risk within ±5% of the gold standard before their oncologist reviewed their risk, whereas 10 estimated their risk accurately after the intervention. One study oncologist indicated that this was clinically unsatisfactory and that a higher proportion, perhaps 80% or more of patients, should be within ±5% of the gold standard risk estimate. Conversely, the fact that the average Brier score went from 0.080 before to 0.019 after the intervention is more difficult to interpret.

These continuous and categorical measures have different strengths that may be suited for different purposes: continuous measures may be better for researchers determining if an intervention has an effect, and categorical measures may be better for determining the clinical significance of an intervention. If researchers wish to use categorical measures, they should determine the accuracy threshold a priori based on clinical reasoning about treatment thresholds, as we did. Otherwise, our results suggest that a categorical measure can be configured post hoc to match the sensitivity of continuous measures, but without providing the usual discrimination of the categorical approach. In our analysis, an accuracy threshold of ±15% did match the best continuous measure in terms of statistical sensitivity. However, a 15% margin of accuracy would lead to situations where a patient with a 15% local therapy mortality risk could estimate their risk as anywhere in the range from 0% to 30% and be considered accurate. This is problematic since at the lower end of the spectrum, adjuvant treatment would not be necessary, while at the upper end, most oncologists would recommend it. Researchers should choose categorical measures with thresholds that are not so broad that they blur the lines among patients with materially different risks and materially different treatment benefits.

4.1.2. Limitations

Our study has strengths and limitations. Regarding strengths, we used empirical data to explore the performance of four candidate measures of patient understanding in the increasingly important area of risk communication. As a secondary analysis, our study extracted additional insight out of an existing dataset, which conserved resources and leveraged the efforts of prior investigators and the patients who participated in the original study. Our data illustrate how a categorical measure that is clinically relevant may not be statistically significant, while continuous measures that may not be as clinically relevant are statistically significant. The small size of the study allows us in this methodological paper to present all the data and results of calculations for readers to verify their understanding of the techniques presented.

As for limitations, we had not planned our comparison prior to data collection as this was a secondary analysis. Therefore we did not anticipate the issue of five patient responses at boundary values (0% and 100%) for which the Kullback–Leibler Divergence measure is undefined.

4.2. Conclusions

In this study, we found continuous measures to be more powerful and resource-efficient than a categorical measure. The categorical measure we examined was used with an accuracy standard of ±5%, which is relevant to breast cancer treatment decision making. Categorical measures with appropriate standards of accuracy may be worth using due to their clinical relevance, even if they are less sensitive and powerful, and therefore more expensive to use. Conversely, categorical measures with an overly broad standard of accuracy may be insensitive to intervention effects, and also clinically irrelevant.

For continuous measures, in this study we also found tensions between power and interpretability. The Kullback–Leibler measure was powerful, but more complex than the absolute bias or Brier. Among continuous measures, we believe that the Brier offers a good combination of ease of use and interpretability, statistical power, and desirable information-theoretic properties.

Overall, we conclude that researchers and other interested parties in the field of risk communication must strike a balance between statistical significance and clinical relevance.

4.3. *Practice implications*

Our target audience includes researchers who are designing risk communication studies; funders who are paying for such studies; and patients, physicians, payers, and other stakeholders who may be interpreting risk communication studies with an eye to adopting or rejecting proposed interventions.

For researchers, the implications of this study are that categorical measures of patient knowledge should be designed with clinical treatment thresholds in mind, to assure that the standards for evaluating a patient estimate are clinically relevant. The maximum margin of error must be identified prior to studies, otherwise researchers may be tempted to select an accuracy threshold after data analysis that maximizes the statistical significance but is less clinically relevant.

Funders or payers should be wary of adopting interventions based on categorical measures with overly broad accuracy thresholds, even if these were associated with statistically significant effects. These interventions may not contribute to clinically relevant improvements in patient knowledge.

Physicians and payers or policymakers may find it feasible to establish categorical standards for patient knowledge such as “80% of my patients should be able to identify their risk within 5% of the Adjuvant! gold standard”. They may wish to survey patients to monitor and continuously improve patient education practices.

Patients should seek out physicians whose educational efforts are proven to result in improved patient knowledge, as a path to making informed treatment decisions.

Acknowledgements

The authors thank the patients who participated in the primary study, as well as other members of the primary study team including Laura Esserman, Hope Rugo, Debu Tripathy, Fern Hassin, and Daniel Chen. We thank Deborah Grady and Anita Stewart for guidance on measurement issues, Brian Jersky and Helena Kraemer for statistical advice, Pam Derish for instruction on scientific writing, Martha Daschbach for assistance with regulatory compliance, and the staff at the H.M. Fishbon Memorial Library at the University of California, San Francisco, for their assistance with background research and literature searches.

Role of funding: The primary study was funded by the Department of Defense, grant DAMD17-96-1-6260. Funding sources for this secondary analysis included the Department of Defense, DAMD17-96-1-6260, as well the National Institutes of Health (NIH) Office of Research on Women’s Health (ORWH) and National Institute of Child Health and Human Development (Grant 9K12HD052163-06). David Hutton was supported by a Stanford Graduate Fellowship. None of the funders had any involvement in the design, conduct, or reporting of this research.

Appendix A. Sample size calculations

We calculated sample sizes using standard formulas for normally distributed measurements using the function sampsi in STATA version 10 [11]. For the categorical measure, the sample size is that for testing the proportion of patients who improved their accuracy with denominator equal to the number whose accuracy changed. In our study, six patients improved their accuracies to with ±5% while one patient became inaccurate with respect to the ±5% criterion. Thus, six of seven (86%) improved. The sample size is that required to test 86% improvement against a null hypothesis of 50% (chance) improvement. The calculated sample size then was inflated to an enrollment size, by dividing by 35%, equal to the percentage of patients whose accuracy changed.

Sample sizes were calculated for one-sided tests, since the data from this pilot study indicated that changes in accuracy, as a result of the intervention, would be in the positive direction. All tests are based on paired responses so that, for sample size calculations, one-sample tests are appropriate.

The following is a listing of the STATA commands to calculate sample sizes:

1. For the categorical accuracy assessment:

```
Input line> sampsi .5 .86, onesample onesided pow(.9)

STATA output>
Estimated sample size for one-sample comparison of proportion
  to hypothesized value

Test Ho: p = 0.5000, where p is the proportion in the population

Assumptions:

         alpha =   0.0500  (one-sided)
         power =   0.9000
 alternative p =   0.8600

Estimated required sample size:

             n =       13

. display 13/0.35
37.142857
```

2. For the log-transformed absolute bias and Brier scores the d (average of differences divided by standard deviation of differences) is 0.63.

```
Input> sampsi 0 .63, sd(1) onesided onesample
STATA output>
Estimated sample size for one-sample comparison of mean
  to hypothesized value

Test Ho: m =      0, where m is the mean in the population

Assumptions:

         alpha =   0.0500  (one-sided)
         power =   0.9000
 alternative m =      .63
            sd =        1

Estimated required sample size:
n =        22
```

3. For the Kulback–Leibler with $d = 0.72$.

```
Input> sampsi 0 .72, sd(1) onesided onesample
STATA output>
Estimated sample size for one-sample comparison of mean
  to hypothesized value

Test Ho: m =      0, where m is the mean in the population

Assumptions:

         alpha =   0.0500  (one-sided)
         power =   0.9000
 alternative m =      .72
            sd =        1

Estimated required sample size:

             n =       17
```

Conflict of interest

The authors declare no conflicts of interest.

References

[1] Hulley SB, Cummings SR, Browner WS, Grady DG, Newman T. Designing clinical research: an epidemiologic approach. Philadelphia, PA: Lippincott Williams & Wilkins; 2007.

[2] Brier G. Verification of forecasts expressed in terms of probability. Monthly Weather Review 1950;78:1–3.

[3] McClish DK, Powell SH. How well can physicians estimate mortality in a medical intensive care unit? Med Decis Making 1989;9:125–32.

[4] García JA, Fernández-Valdivia J, Rodriguez-Sánchez R, Fernández-Vidal XR. Performance of the Kullback–Leibler information gain for predicting image fidelity. In: Proceedings of the 16th International Conference on Pattern Recognition (ICPR'02), vol. 3; 2002. p. 843–8.

[5] Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat 1951;22:79–86.

[6] Belkora J, Rugo H, Moore D, Hutton D, Chen D, Esserman L. Using the Adjuvant! model during oncologist–patient consultations: changes in patient knowledge of prognosis. 2007:San Francisco, submitted for publication.

[7] Ravdin PM, Siminoff LA, Davis GJ, Mercer MB, Hewlett J, Gerson N, Parker HL. Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. J Clin Oncol 2001;19:980–91.

[8] Olivotto IA, Bajdik CD, Ravdin PM, Speers CH, Coldman AJ, Norris BD, Davis GJ, Chia SK, Gelmon KA. Population-based validation of the prognostic model ADJUVANT! for early breast cancer. J Clin Oncol 2005;23:2716–25.

[9] Love N. Management of breast cancer in the adjuvant and metastatic settings. Patterns Care 2005;2:10–24.

[10] Grünwald PD, Dawid AP. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. Ann Stat 2004;32:1367–433.

[11] Stata Statistical Software: Release 10. College Station, TX: StataCorp LP; 2007.

[12] Hedges LV, Olkin I. Statistical methods for meta-analysis. Orlando: Academic Press; 1985.

[13] Efron B, Tibshirani R. An introduction to the bootstrap. New York: Chapman & Hall; 1993.

[14] O'Connor AM, Stacey D, Entwistle V, Llewellyn-Thomas H, Rovner D, Holmes-Rovner M, Tait V, Tetroe J, Fiset V, Barry M, Jones J. Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev 2003. CD001431.

[15] Dodin S, Légaré F. Prise de decision en matière d'hormonothérapie de remplacement. Can Fam Physician 2001;47:1586–93.

[16] Grady D, Rubin SM, Petitti DB, Fox CS, Black D, Ettinger B, Ernster VL, Cummings SR. Hormone therapy to prevent disease and prolong life in postmenopausal women. Ann Intern Med 1992;117:1016–37.

[17] O'Connor AM, Tugwell P, Wells GA, Elmslie T, Jolly E, Hollingworth G, McPherson R, Drake E, Hopman W, Mackenzie T. Randomized trial of a portable, self-administered decision aid for postmenopausal women considering long-term preventive hormone therapy. Med Decis Making 1998;18:295–303.

[18] Rostom A, O'Connor A, Tugwell P, Wells G. A randomized trial of a computerized versus an audio-booklet decision aid for women considering post-menopausal hormone replacement therapy. Patient Educ Couns 2002;46: 67–74.

[19] McBride CM, Bastian LA, Halabi S, Fish L, Lipkus IM, Bosworth HB, Rimer BK, Siegler IC. A tailored intervention to aid decision-making about hormone replacement therapy. Am J Public Health 2002;92:1112–4.

[20] Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 1989;81:1879–86.

[21] Man-Son-Hing M, Laupacis A, O'Connor AM, Biggs J, Drake E, Yetisir E, Hart RG. A patient decision aid regarding antithrombotic therapy for stroke prevention in atrial fibrillation: a randomized controlled trial. JAMA 1999;282:737–43.

[22] Risk factors for stroke and efficacy of antithrombotic therapy in atrial fibrillation. Analysis of pooled data from five randomized controlled trials. Arch Intern Med 1994;154:1449–57.