Numeracy and Understanding of Quantitative Aspects of Predictive Models: A Pilot Study

Gary E. Weissman[1,2,3,4] Kuldeep N. Yadav[2,3,4] Vanessa Madden[2,3,4] Katherine R. Courtright[1,2,3,4] Joanna L. Hart[1,2,3,4] David A. Asch[1,4,5,6] Marilyn M. Schapira[1,6] Scott D. Halpern[1,2,3,4]

[1] Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States
[2] Department of Medicine, Palliative and Advanced Illness Research Center, University of Pennsylvania, Philadelphia, Pennsylvania, United States
[3] Fostering Improvement in End-of-Life Decision Science Program, University of Pennsylvania, Philadelphia, Pennsylvania, United States
[4] Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, Pennsylvania, United States
[5] Center for Health Care Innovation, University of Pennsylvania, Philadelphia, Pennsylvania, United States
[6] The Center for Health Equity Research and Promotion, Philadelphia VA Medical Center, Philadelphia, Pennsylvania, United States

Address for correspondence Gary E. Weissman, MD, MSHP, Palliative and Advanced Illness Research Center, Perelman School of Medicine, University of Pennsylvania, 306 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104, United States
(e-mail: gary.weissman@uphs.upenn.edu).

Appl Clin Inform 2018;9:683–692.

Abstract

Background The assessment of user preferences for performance characteristics of patient-oriented clinical prediction models is lacking. It is unknown if complex statistical aspects of prediction models are readily understandable by a general audience.

Objective A pilot study was conducted among nonclinical audiences to determine the feasibility of interpreting statistical concepts that describe the performance of prediction models.

Methods We conducted a cross-sectional electronic survey using the Amazon Mechanical Turk platform. The survey instrument included educational modules about predictive models, sensitivity, specificity, and confidence intervals (CIs). Follow-up questions tested participants' abilities to interpret these characteristics with both verbatim and gist knowledge. Objective and subjective numeracy were assessed using previously validated instruments. We also tested understanding of these concepts when embedded in a sample discrete choice experiment task to establish feasibility for future elicitation of preferences using a discrete choice experiment design. Multivariable linear regression was used to identify factors associated with correct interpretation of statistical concepts.

Results Among 534 respondents who answered all nine questions, the mean correct responses was 95.9% (95% CI, 93.8–97.4) for sensitivity, 93.1% (95% CI, 90.5–95.0) for specificity, and 86.6% (95% CI, 83.3–89.3) for CIs. Verbatim interpretation was high for all concepts, but significantly higher than gist only for CIs ($p < 0.001$). Scores on each discrete choice experiment tasks were slightly lower in each category. Both objective and subjective numeracy were positively associated with an increased proportion of correct responses ($p < 0.001$).

Keywords
- forecasting
- health communication
- decision support techniques
- clinical decision-making
- education of patients

received
May 10, 2018
accepted after revision
July 17, 2018

© 2018 Georg Thieme Verlag KG
Stuttgart · New York

DOI https://doi.org/10.1055/s-0038-1669457.
ISSN 1869-0327.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

Conclusion These results suggest that a nonclinical audience can interpret quantitative performance measures of predictive models with very high accuracy. Future development of patient-facing clinical prediction models can feasibly incorporate patient preferences for model features into their development.

Background and Significance

The quantification of patient preferences for health care interventions is important for determining their role in care delivery and for promoting preference-concordant decisions.[1–4] Prior empiric work has described patient preferences for the tradeoffs between costs, pain severity, survival, transportation time, access to care, and place of death.[2,5–7] These studies used a discrete choice experiment (DCE) design that provides a quantitative measure of these tradeoffs. While similar studies have examined preferences for characteristics associated with diagnostic tests such as cancer screening,[8–10] none has examined patient preferences for characteristics of clinical prediction models. This distinction is significant because typically a local health system cannot directly alter the performance characteristics of a diagnostic test to suit the needs and preferences of a particular patient. However, the performance characteristics of a clinical prediction model are partly dependent on modeling and statistical methods that analysts could optimize to meet individualized needs in a specific clinical scenario.

Highly customized clinical prediction models based on rich data from the electronic health record (EHR) are becoming increasingly common[11,12] in the era of learning health systems[13] and widespread EHR adoption.[14] Such models may promote "precision delivery" of health care by identifying patients at risk for a given outcome, thereby prompting targeted and timely interventions.[15] However, little is known about patient preferences for such predictive information, or how such preferences for false positive and false negative errors and for types of uncertainty might vary across clinical conditions and baseline risk estimates. For example, a person with a serious illness of otherwise uncertain prognosis may prefer a false positive to a false negative prediction of death if the response to the prediction is not too costly or is likely to be undertaken at some point anyway (e.g., advance care planning). On the other hand, a person considering prophylactic surgery to prevent future cancer occurrence may prefer a false negative to a false positive error, which might lead to an unnecessary and irreversible procedure. The validity of scientific data, time horizons of risk assessments, and presentation of statistical uncertainty are all important features of scientific knowledge important to the general public.[16]

Quantifying preferences for predictive model characteristics in such scenarios could be accomplished through a DCE. However, this approach would require at least a basic understanding of the relevant statistical concepts used to describe model performance. Although the understanding of related concepts in the context of diagnostic tests has been examined among both general audiences and clinicians,[17–21] no studies have evaluated understanding of the performance characteristics of clinical prediction models.

Objective

We conducted a pilot feasibility study to determine (1) the level of understanding of performance characteristics of clinical prediction models, and (2) whether numeracy, education, or other demographic factors are associated with understanding of these concepts. We hypothesized that using best practices in risk communication, a nonclinical audience could interpret key attributes of predictive models such as sensitivity, specificity, and confidence intervals (CIs).

Methods

Study Design

We conducted a cross-sectional electronic survey using the Web-based Qualtrics software (Qualtrics, Provo, Utah, United States) among an online population to quantify the ability to interpret the performance of clinical prediction models as described by sensitivity, specificity, and CIs.

Population

Two sequential cohorts were included in this study. The first cohort of participants was recruited via the Amazon Mechanical Turk (MTurk) platform during December 2016. Enrollment was restricted to unique participants[22] with historical task success rates of at least 95%.[23] Each participant provided an electronic informed consent and received US$4 upon completion of the survey.

Given the low median age of respondents in this first cohort (►Table 1) and the desire to recruit a cohort generalizable to patient populations likely to utilize predictive information in decisions about health care,[24,25] a second cohort was recruited via the TurkPrime platform, which allows for detailed filtering of eligible respondents by demographic features within the MTurk population.[26] During January 2017, we recruited an additional 301 nonduplicated respondents over the age of 60 years, each of whom received US$2 for completing the survey. The reimbursement amount was less in the second cohort based on the lower-than-expected median study completion time found in the first (►Table 2).

In both cohorts, respondents who failed any attention checks or completed the entire survey in less than 3 minutes were excluded from the analysis.[27]

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

Table 1 Characteristics of the study population

Characteristic, *n* (%)	All	Cohort 1	Cohort 2
Participants, *n*	534	280	254
Age (y), median (IQR)	51 (30–63)	31 (27–39)	63 (60–67)
Gender			
Male	257 (48.1)	165 (58.9)	92 (36.2)
Female	273 (51.1)	112 (40.0)	161 (63.4)
Other	4 (< 1)	3 (1.1)	1 (< 1)
Race			
White	439 (82.2)	206 (73.6)	233 (91.7)
Asian	38 (7.1)	33 (11.8)	5 (2.0)
Black	33 (6.2)	24 (8.6)	9 (3.5)
Other	24 (4.5)	17 (6.1)	7 (2.8)
Highest level of education			
High school	117 (21.9)	56 (20.0)	61 (24.0)
GED or equivalent	45 (8.4)	26 (9.3)	19 (7.5)
Associate's degree	123 (23.0)	65 (23.2)	58 (22.8)
Bachelor's degree	168 (31.5)	105 (37.5)	63 (24.8)
Master's degree	60 (11.2)	24 (8.6)	36 (14.1)
Doctoral degree	21 (3.9)	4 (1.4)	17 (6.7)
Daily meds, median (IQR)	0 (0–2)	0 (0–1)	2 (0–4)
Marital status			
Single, never married	187 (35.0)	152 (54.3)	35 (13.8)
Married	234 (43.8)	102 (36.4)	132 (52.0)
Divorced/Widowed	113 (21.2)	26 (9.3)	87 (34.3)
Numeracy, median (IQR)			
Objective (range 0–8)	7 (6–8)	7 (6–8)	7 (6–8)
Subjective (range 1–6)	4.9 (4.3–5.4)	4.8 (4.1–5.3)	4.9 (4.4–5.4)

Abbreviations: GED, general equivalency degree; IQR, interquartile range.

Presentation of Statistical Concepts

The survey instrument included didactic modules to explain what is a predictive model and describe the relevant statistical concepts. First, a single-page visual and text description of a predictive model was displayed. Next, using best practices in risk communication, explanatory modules with text exemplars, icon arrays, integer annotations, summary explanations, and simple sentences[28,29] were included in three separate modules to describe the concepts of sensitivity, specificity, and CIs as they relate to the performance of predictive models. We depicted CIs using a variation of a "blurred" icon array.[19] Sample explanatory icon arrays are presented in ►Fig. 1. Each module presented the statistical concept in the context of a weather prediction. Weather examples were chosen to remove any potential cognitive or affective influences common in medical decision making[30,31] and to isolate ascertainment of participant interpretation of these concepts. The explanatory text of these modules were written at a 10th-grade Flesch–Kincaid reading level for clarity. The instrument was iteratively piloted with five experienced research coordinators not involved with the study to improve clarity. A copy of the final survey instrument with explanatory modules is available in the ►Supplementary Material (available in the online version).

Knowledge Testing

Each module was followed by two questions, and each question separately tested verbatim and gist knowledge. Both types of knowledge are associated with identifying optimal medical treatments in a comparison task[29] and with understanding of numeric concepts.[32] There were three verbatim and three gist knowledge questions in total. Finally, to assess participants' abilities to compare two models given a complex presentation of information, participants were presented with three DCE tasks each comparing performance and other characteristics of two different prediction models. A DCE task was chosen because DCEs are commonly used to determine the relative utilities of time, cost, and health states

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

Table 2 Characteristics of participant responses, and performance on questions by content area and type of knowledge

Measure	All	Cohort 1	Cohort 2
Study duration (min), median (IQR)	14.3 (10.8–18.5)	12.5 (9.3–15.6)	16.5 (13.3–19.6)
Overall knowledge score	91.9 (89.2–94.0)	89.9 (85.6–93.1)	94.1 (90.3–96.6)
Sensitivity			
All	95.9 (93.8–97.4)	93.8 (90.1–96.2)	98.3 (95.6–99.4)
Verbatim	97.0 (95.1–98.2)	94.6 (91.3–96.9)	99.6 (97.5–99.9)
Gist	98.5 (96.9–99.3)	97.1 (94.2–98.7)	100.0 (98.1–100.0)
DCE task	92.3 (89.6–94.4)	89.6 (85.3–92.8)	95.3 (91.7–97.4)
Specificity			
All	93.1 (90.5–95.0)	91.3 (87.2–94.2)	95.0 (91.4–97.2)
Verbatim	93.8 (91.3–95.6)	91.4 (87.4–94.3)	96.5 (93.2–98.3)
Gist	95.3 (93.1–96.9)	92.1 (88.2–94.9)	98.8 (96.3–99.7)
DCE task	90.1 (87.1–92.4)	90.4 (86.1–93.4)	89.8 (85.2–93.1)
Confidence interval			
All	86.6 (83.3–89.3)	84.3 (79.4–88.3)	89.1 (84.5–92.5)
Verbatim	94.4 (92.0–96.1)	93.6 (90.0–96.0)	95.3 (91.7–97.4)
Gist	83.1 (79.6–86.2)	79.3 (74.0–83.9)	87.4 (82.5–91.1)
DCE task	82.1 (78.4–85.3)	79.4 (73.6–84.2)	84.6 (79.5–88.7)

Abbreviations: DCE, discrete choice experiment; IQR, interquartile range.
Note: All scores are reported as mean percentages with 95% binomial confidence intervals.

among patients.[2,5–7] Thus, they represent an ideal study design in future work for assessing preferences for the performance characteristics of clinical prediction models and allow researchers to quantify tradeoffs between measures such as sensitivity and specificity, as are commonly encountered in predictive model development. Only one of six model characteristics (i.e., DCE attributes) varied at a time for each of these sample tasks and participants were asked to choose the better of the two models. The following model attributes were presented for each sample task: outcome, time horizon, sensitivity, specificity, sensitivity CI, and specificity CI. The response was scored as correct if the model with larger sensitivity or specificity, or smaller CI was chosen. Since the goal was to use these sample tasks to assess feasibility of interpreting complex presentations of information, rather than to elicit preferences for model characteristics themselves, we did not conduct a DCE in this study.

We reported the primary outcome as the overall knowledge score, which is the mean percentage of correct responses to all nine conceptual questions (two questions each for understanding of three different statistical concepts plus three sample DCE tasks). We also reported mean scores broken down by type of knowledge and concept.

Numeracy Measures and Demographics

Numeracy influences the interpretation of quantitative risk information.[17,18,33] Following the education modules and knowledge questions, we tested both objective and subjective numeracy because they each gauge distinct types of understanding and preferences.[34] We used the short Numeracy Understanding in Medicine Instrument (S-NUMi; range, 0–8)[35] and the Subjective Numeracy Scale (SNS; range, 1–6).[36] The order of these two instruments was randomized for each participant. The last section of the instrument asked for the participants' age, gender, race, ethnicity, level of education, marital status, and number of prescription medications taken each day, a crude measure of comorbidity and health.[37,38]

Statistical Analysis

The overall knowledge score, the mean correct responses to nine questions, is reported with 95% binomial CIs. We evaluated differences in scores and participant characteristics between cohorts using chi-square and two-sample, unpaired, two-sided *t*-tests with = 0.05 for categorical and continuous variables, respectively. We developed a multivariable linear regression model to examine the relationship between numeracy measures and the overall knowledge score. The following covariates were selected based on likely relevance and included in the multivariable model: age, gender, race, education level, and minutes spent on the survey. Pearson's correlation coefficients between continuous model inputs were reported in a correlation matrix. All analyses were conducted using the R language for statistical computing (version 3.3.1).[39] The deidentified data and source code used for these analyses are available online (https://github.com/gweissman/numeracy_pilot_study).

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

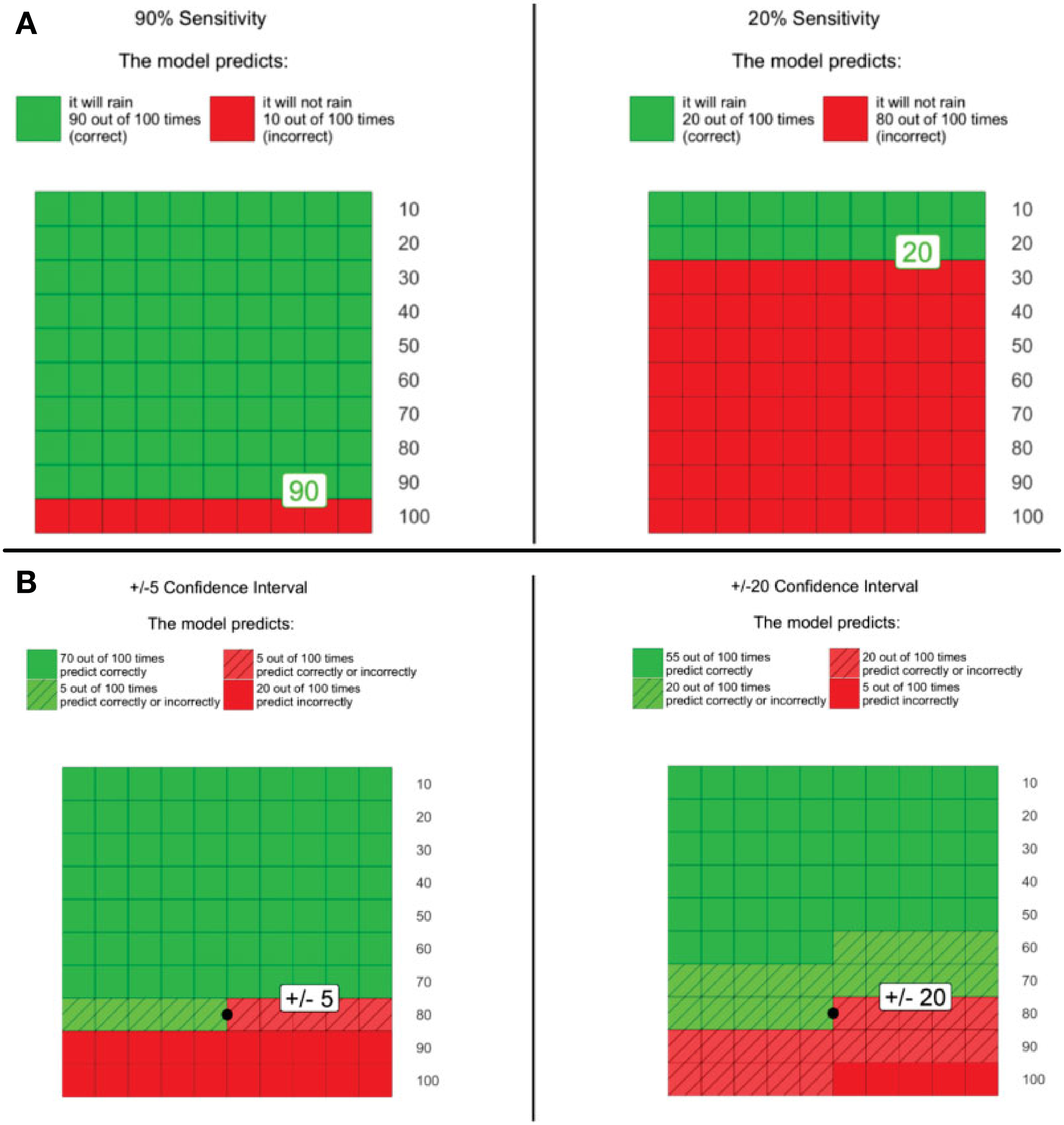

Fig. 1 Sample visual explanatory tools used in the didactic modules to convey sensitivity (A) and confidence intervals (B) using icon arrays, integer annotations, and juxtaposed comparisons.

Results

A total of 608 participants completed the survey, and 69 were excluded for failing one or more attention checks, and another five for completing the survey in less than 3 minutes (►Fig. 2). Among the remaining 534 respondents, 273 (51.1%) were women and 439 (82.2%) self-identified as white. The median age was 51 years (interquartile range [IQR], 30–63; range, 18–81) and 249 (46.6%) participants had completed at least a 4-year college degree (►Table 1).

The second cohort ($n = 254$) was older (64 vs. 34 years, $p < 0.001$), took longer to complete the survey (17.2 vs. 13.6 minutes, $p < 0.001$), had higher mean subjective numeracy (4.8 vs. 4.6, $p = 0.006$), and included more women (63.6% vs. 40.4%, $p < 0.001$). The cohorts had similar mean objective numeracy (6.8 vs. 6.7, $p = 0.186$) and similar rates of bachelor or higher degrees (47.5% vs. 45.7%, $p = 0.736$).

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

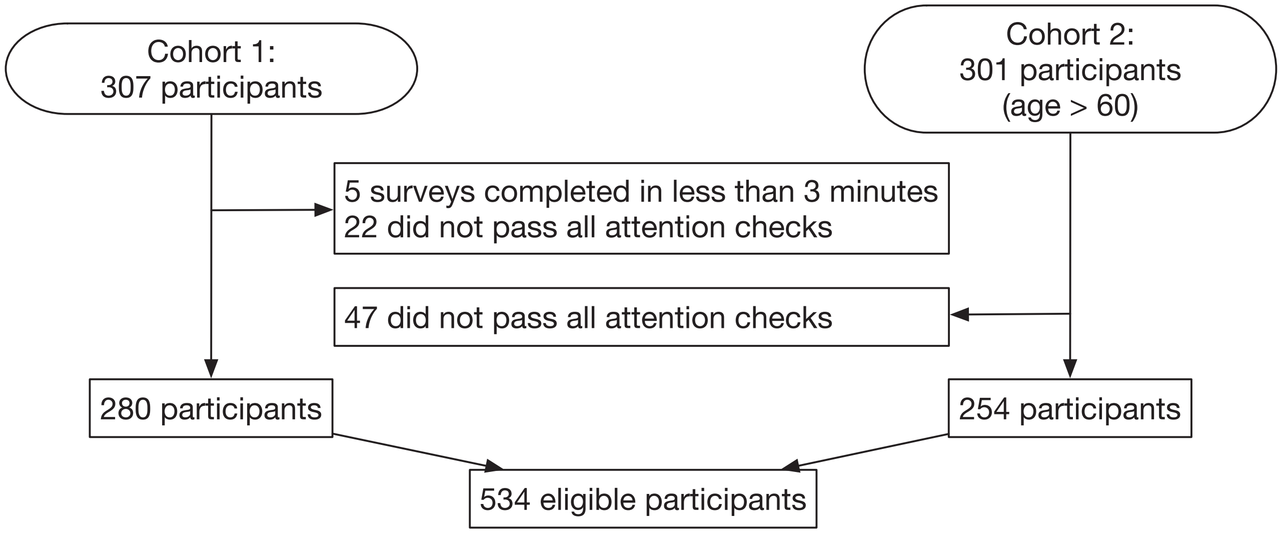
Cohort 1:
307 participants
Cohort 2:
301 participants
(age > 60)
5 surveys completed in less than 3 minutes
22 did not pass all attention checks
47 did not pass all attention checks
280 participants
254 participants
534 eligible participants

Fig. 2 Patient enrollment and exclusions.

Given these similarities and differences, we reported all results both overall and separately by cohort.

The overall knowledge score ranged from 11.1 to 100.0% (median number of correct responses 9, IQR, 8–9). Verbatim knowledge was similar to gist knowledge for sensitivity and specificity, but significantly exceeded gist knowledge for CIs by 11.2% ($p < 0.001$; ►**Table 2**). The mean score for the subset of questions embedded in a DCE task was 88.5% (95% CI, 85.4–91.0). The second cohort scored higher than the first by overall knowledge score (94.1% vs. 89.9%, $p < 0.001$). In the adjusted multivariable analysis, a one-point increase in the S-NUMi or SNS was associated with a 4.7% (95% CI, 3.9–5.6) or 2.2% (95% CI, 0.9–3.5) increase, respectively, in the overall knowledge score (►**Table 3**). Age was weakly correlated with both measures of numeracy and the test duration (►**Table 4**). The first 46 responses to the DCE-embedded question testing understanding of CIs (Question 9 in the ►**Supplementary Material**, available in the online version) were excluded from all analyses due to an error found in the survey instrument which was corrected for all subsequent participants. This resulted in the exclusion of 42 participant responses that would otherwise have been included in the final analytic sample for that question only.

Discussion

These data provide preliminary evidence of the feasibility of interpreting statistical concepts underlying the performance characteristics of a prediction model among a nonclinical audience. The findings from this pilot study support the

Table 3 Multivariable linear regression results

Variable	All			Cohort 1			Cohort 2		
	Coefficient	95% CI	*p*-Value	Coefficient	95% CI	*p*-Value	Coefficient	95% CI	*p*-Value
S-NUMi score	4.7	3.9–5.6	< 0.001	5.7	4.4–6.9	< 0.001	1.9	0.8–3.1	0.001
SNS score	2.2	0.9–3.5	0.001	3.1	1.0–5.2	0.004	1.0	−0.5 to 2.4	0.200
Male gender	-1.3	−3.5 to 1.0	0.215	−3.93	−7.3 to −0.5	0.024	2.1	−0.1 to 4.4	0.066
White race	5.8	2.9–8.7	< 0.001	6.2	2.1–10.2	0.003	2.3	−1.8 to 6.3	0.267
Bachelor's degree or higher	-0.9	−3.0 to 1.2	0.397	−2.7	−6.1 to 0.7	0.111	2.4	0.1–4.7	0.038
Age (y)	0.0	−0.01 to 0.1	0.292	−0.01	−0.1 to 0.1	0.772	−0.1	−0.3 to 0.1	0.539
Study duration (min)	0.1	−0.1 to 0.1	0.127	0.23	−0.1 to 0.4	0.077	−0.1	−0.1 to 0.2	0.837

Abbreviations: CI, confidence interval; S-NUMi, Short Numeracy Understanding in Medicine Instrument; SNS, Subjective Numeracy Scale.
Note: Coefficient estimates represent the percent change in overall knowledge score associated with a one-unit change in each variable.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

Table 4 Pearson's correlations (*p*-value) between continuous variables used in the multivariable prediction model

	S-NUMi	SNS	Age (y)	Duration (min)
S-NUMi	1.0	0.275 (< 0.001)	0.108 (0.013)	0.028 (0.51)
SNS		1.0	0.133 (0.002)	0.068 (0.12)
Age (y)			1.0	0.293 (< 0.001)
Duration (min)				1.0

Abbreviations: S-NUMi, Short Numeracy Understanding in Medicine Instrument; SNS, Subjective Numeracy Scale.

possibility of using DCEs or other methods to elicit quantitatively expressed preferences for aspects of clinical prediction models. Such an approach may increase the relevance of future prediction models in real clinical decision-making scenarios. There are several ways in which these findings suggest future directions for study in a wide range of populations for whom clinical prediction models are likely to be of use.

First, the inability to understand risk information limits the potential for widespread deployment of patient-centered complex risk models. Objective numeracy as measured by the S-NUMi was strongly associated with performance on the survey in both cohorts, although the association was stronger in the younger cohort. Subjective numeracy, on the other hand, was significantly associated with the knowledge score only in the younger cohort, while the level of education was significant only in the older cohort. This finding is consistent with prior work demonstrating that numeracy is positively associated with understanding statistical risk information.[27,40] Adaptation of risk models based on patient preferences will require different approaches in low-numeracy populations.

Second, the role of educational models in explaining quantitative performance measures to a nonclinical audience remains unknown. Observed knowledge scores in this study compared favorably to understanding of statistical concepts in other populations in both stand-alone questions and when concepts were embedded in complex DCE tasks. For example, in a review of six studies that assessed the ability of health care professionals to identify accuracy measures of diagnostic tests based on multiple choice definitions or written vignettes, sensitivity and specificity were correctly identified 76 to 88% and 80 to 88% of the time, respectively.[21] Similarly, in a sample of medicine residents, only 56.7% correctly determined which of two example tests had higher specificity.[41] Although the presentation and testing of statistical knowledge varied between these[21,41] and the present study, we speculate that our participants scored higher because of the inclusion of didactic modules prior to testing. These modules presented icon arrays, integer annotations, and plain language explanations.[28,29] Although the performance scores are not directly comparable across the studies, the differences do suggest a potential role for such risk communication techniques in a nonclinical audience.

Third, further work is needed to explain age- and education-dependent influences on knowledge, and to describe potential interactions with the level of numeracy on knowledge of statistical concepts. Additionally, white race was also associated with higher scores in the younger cohort, which may represent residual confounding due to factors not fully assessed in this survey.[42–44]

Fourth, this is the first study to demonstrate a difference between the interpretation of CIs as measured by verbatim and gist knowledge. Prior studies have demonstrated, using both quantitative and qualitative methods, that interpretation of CIs is difficult for the general population.[45,46] The use of CIs to convey statistical uncertainty may even worsen understanding of risk information compared with the presentation of a point estimate alone.[27,47] To date, the exact features of CIs or the mechanisms of their interpretation that confuse have not been elucidated. Although our survey instrument did provide helpful heuristics to guide interpretation of statistical information, our study design did not test heuristics explicitly. Given that verbatim knowledge of CIs was high in all groups while gist knowledge was markedly lower, we hypothesize that respondents may have been able to understand the numeric description of CIs, but misinterpreted their qualitative comparison due to a "bigger is better" heuristic.[48] This heuristic is appropriate for sensitivity and specificity, but is reversed for the correct interpretation of CIs where "smaller is better." Gist interpretation typically relies on a qualitative assessment of what a numeric estimate means to the reader.[32] Without an explicit assessment of how each participant views uncertainty with respect to weather information, the categories provided in the multiple choice questions may not have encoded a standard meaning. Further work in the use and prompting of heuristics to understanding quantitative features of prediction models and interpretation of CIs is warranted.

The strengths of this study include the testing of both gist and verbatim knowledge, the adjustment for subjective and objective numeracy and other demographic factors, and the representativeness with respect to age range, number of medications used in an ambulatory population, and gender.[49] Additionally, this study is the first to test the interpretation of statistical concepts as they describe prediction models in both stand-alone examples and when embedded in a complex DCE sample task.

However, the results of this study should be interpreted in light of some limitations. First, this study conveyed predictive information in the context of weather examples, which may not elicit the same cognitive and affective decision-making mechanisms as those relating to health states.[30,31]

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

Second, the generalizability of these results is limited by the analytic sample, which was primarily white and of very high numeracy compared with the general population.[35,36,50] People with lower numeracy may be especially vulnerable to misinterpretations of these statistical concepts,[17,18,33] and thus are an important population in which further validation is warranted. Third, our study tested knowledge of these statistical concepts immediately following provision of education modules, but we did not administer a pretest knowledge assessment. Therefore, we cannot draw conclusions about the efficacy of the education modules themselves in improving baseline knowledge. Future work should include a pretest baseline assessment to better characterize effective strategies for describing statistical concepts related to prediction models. Similarly, future testing should characterize the temporal duration of an intervention's effect on knowledge, which may decay with time,[51] and which may better distinguish between true knowledge and immediate recall. Fourth, the multiple-choice format limits more robust assessments of the ability to apply these statistical concepts, and may result in overly optimistic performance scores if participants employed other test-taking strategies.[52] Fifth, this study did not measure knowledge of false positive and false negative concepts directly, which may be more directly relevant to the development of clinical prediction models than sensitivity and specificity—indirect measures of these error rates—which were tested in this study. Sixth, we did not perform reliability testing in the development of this pilot instrument which may threaten the validity of the findings. Finally, because subjects were not actively screened and approached for recruitment, our study design cannot account for the self-selection of potential participants from the online platform who saw but chose not to complete the survey.

Conclusion

In conclusion, this study demonstrates that a nonclinical audience can interpret predictive model features such as sensitivity and specificity with high accuracy using both gist and verbatim knowledge. Such understanding was high even when interpreted within a complex DCE task. These findings highlight the feasibility of future DCEs to quantify preferences for tradeoffs between performance characteristics of predictive models, and suggest the need for validating these results in more generalizable patient populations.

Clinical Relevance Statement

The rapidly growing interest in and use of prediction models in health care settings warrant increased focus on patient preferences for information. In order for clinical prediction models to achieve significant impact in informing decisions about care, they must incorporate different preferences for tradeoffs between false positive and false negative errors, bias and variance, and performance across varying predictive time horizons. Although preferences for these particular tradeoffs will likely exhibit significant variation depending on the clinical scenario, this study demonstrates the feasibility of assessing such preferences for model characteristics in a nonclinical population. The incorporation of patient preferences into predictive model development would better align with practices for cancer treatments, medical devices, and organ transplant protocols, all of which are informed by research into patient preferences for tradeoffs between their features.

Multiple Choice Questions

1. In this study, interpretation of performance characteristics as measured by verbatim knowledge was high for which of the following:
   a. Sensitivity, confidence intervals, and time horizons.
   b. Specificity.
   c. Sensitivity and specificity.
   d. Sensitivity, specificity, and confidence intervals.

   Correct Answer: The correct answer is option d. The proportion of correctly answered questions was greater than 90% when interpreting verbatim knowledge of sensitivity, specificity, and confidence intervals. This suggests participants were able to interpret the specific numbers associated with the model performance characteristics described in the question stems.

2. In this study, interpretation of performance characteristics as measured by gist knowledge was high for which of the following:
   a. Sensitivity, confidence intervals, and time horizons.
   b. Specificity.
   c. Sensitivity and specificity.
   d. Sensitivity, specificity, and confidence intervals.

   Correct Answer: The correct answer is option c. The proportion of correctly answered questions was greater than 90% when interpreting gist knowledge of sensitivity and specificity, but was considerably lower when interpreting confidence intervals. This suggests participants were less frequently able to interpret the "gist" meaning of the values (e.g., "good" or "bad" performance) of confidence intervals. We hypothesized this might be due to a "bigger is better" heuristic that works well for sensitivity and specificity, but fails for confidence intervals, where smaller is better.

Authors' Contributions

All authors have materially participated in the research and/or article preparation. All authors have made substantial contributions to all of the following: (1) the conception and design of the study, or acquisition of data, or analysis and interpretation of data, (2) drafting the article or revising it critically for important intellectual content, (3) final approval of the version to be submitted.

Protection of Human and Animal Subjects

This study was considered exempt by the Institutional Review Board of the University of Pennsylvania.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

Funding

This work was supported by the National Institutes of Health/NHLBI (T32-HL098054), the Masters of Science in Health Policy Research Program, and the Center for Health Incentives and Behavioral Economics/Penn Roybal Center at the Leonard Davis Institute through a grant from the National Institutes of Health/NIA (P30-AG034546) at the Perelman School of Medicine, University of Pennsylvania. The views expressed here are those of the authors and not of the funders.

Conflict of Interest

None.

References

1 Ryan M. Discrete choice experiments in health care. BMJ 2004; 328(7436):360–361

2 Bridges JF, Hauber AB, Marshall D, et al. Conjoint analysis applications in health–a checklist: a report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. Value Health 2011;14 (04):403–413

3 Ibrahim SA. Decision aids and elective joint replacement — how knowledge affects utilization. N Engl J Med 2017;376(26): 2509–2511

4 Elwyn G, Cochran N, Pignone M. Shared decision making—the importance of diagnosing preferences. JAMA Intern Med 2017; 177(09):1239–1240

5 Malhotra C, Farooqui MA, Kanesvaran R, Bilger M, Finkelstein E. Comparison of preferences for end-of-life care among patients with advanced cancer and their caregivers: a discrete choice experiment. Palliat Med 2015;29(09):842–850

6 Halpern SD, Berns JS, Israni AK. Willingness of patients to switch from conventional to daily hemodialysis: looking before we leap. Am J Med 2004;116(09):606–612

7 Armstrong K, Putt M, Halbert CH, et al. The influence of health care policies and health care system distrust on willingness to undergo genetic testing. Med Care 2012;50(05):381–387

8 Wordsworth S, Ryan M, Skåtun D, Waugh N. Women's preferences for cervical cancer screening: a study using a discrete choice experiment. Int J Technol Assess Health Care 2006;22(03): 344–350

9 Hol L, de Bekker-Grob EW, van Dam L, et al. Preferences for colorectal cancer screening strategies: a discrete choice experiment. Br J Cancer 2010;102(06):972–980

10 de Bekker-Grob EW, Rose JM, Donkers B, Essink-Bot ML, Bangma CH, Steyerberg EW. Men's preferences for prostate cancer screening: a discrete choice experiment. Br J Cancer 2013;108(03): 533–541

11 Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017;24(01):198–208

12 Amarasingham R, Audet A-MJ, Bates DW, et al. Consensus statement on electronic health predictive analytics: a guiding framework to address challenges. EGEMS (Wash DC) 2016;4(01):1163

13 Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. Health Aff (Millwood) 2014;33(07):1163–1170

14 Charles D, Gabriel M, Searcy T. Adoption of electronic health record systems among U.S. non-federal acute care hospitals: 2008 - 2014. ONC Data Brief 2015;23:1–10

15 Darcy AM, Louie AK, Roberts LW. Machine learning and the profession of medicine. JAMA 2016;315(06):551–552

16 Schapira MM, Imbert D, Oh E, Byhoff E, Shea JA. Public engagement with scientific evidence in health: a qualitative study among primary-care patients in an urban population. Public Underst Sci 2016;25(05):612–626

17 Peters E, Hart PS, Fraenkel L. Informing patients: the influence of numeracy, framing, and format of side effect information on risk perceptions. Med Decis Making 2011;31(03):432–436

18 Bodemer N, Meder B, Gigerenzer G. Communicating relative risk changes with baseline risk: presentation format and numeracy matter. Med Decis Making 2014;34(05):615–626

19 Schapira MM, Aggarwal C, Akers S, et al. How patients view lung cancer screening. The role of uncertainty in medical decision making. Ann Am Thorac Soc 2016;13(11):1969–1976

20 Hoffrage U, Gigerenzer G. Using natural frequencies to improve diagnostic inferences. Acad Med 1998;73(05):538–540

21 Whiting PF, Davenport C, Jameson C, et al. How well do health professionals interpret diagnostic information? A systematic review. BMJ Open 2015;5(07):e008155

22 Ott M. 2017. Available at: https://uniqueturker.myleott.com/. Accessed August 5, 2018

23 Amazon Web Services. Mechanical Turk Concepts; 2017. Available at: http://docs.aws.amazon.com/AWSMechTurk/latest/RequesterUI/mechanical-turk-concepts.html. Accessed August 5, 2018

24 Yourman LC, Lee SJ, Schonberg MA, Widera EW, Smith AK. Prognostic indices for older adults: a systematic review. JAMA 2012;307(02):182–192

25 Cardona-Morrell M, Hillman K. Development of a tool for defining and identifying the dying patient in hospital: Criteria for Screening and Triaging to Appropriate aLternative care (CriSTAL). BMJ Support Palliat Care 2015;5(01):78–90

26 Litman L, Robinson J, Abberbock T. TurkPrime.com: a versatile crowdsourcing data acquisition platform for the behavioral sciences. Behav Res Methods 2017;49(02):433–444

27 Sladakovic J, Jansen J, Hersch J, Turner R, McCaffery K. The differential effects of presenting uncertainty around benefits and harms on treatment decision making. Patient Educ Couns 2016;99(06):974–980

28 Fagerlin A, Zikmund-Fisher BJ, Ubel PA. Helping patients decide: ten steps to better risk communication. J Natl Cancer Inst 2011; 103(19):1436–1443

29 Hawley ST, Zikmund-Fisher B, Ubel P, Jancovic A, Lucas T, Fagerlin A. The impact of the format of graphical presentation on health-related knowledge and treatment choices. Patient Educ Couns 2008;73(03):448–455

30 Halpern J, Arnold RM. Affective forecasting: an unrecognized challenge in making serious health decisions. J Gen Intern Med 2008;23(10):1708–1712

31 Fagerlin A, Peters E, Schwartz A, et al. Cognitive and affective influences on health decisions. In: Suls JM, Davidson KW, Kaplan RM, eds. Handbook of Health Psychology and Behavioral Medicine. New York, NY: Guilford Press; 2010:49–63

32 Reyna VF. A theory of medical decision making and health: fuzzy trace theory. Med Decis Making 2008;28(06):850–865

33 Reyna VF, Nelson WL, Han PK, Dieckmann NF. How numeracy influences risk comprehension and medical decision making. Psychol Bull 2009;135(06):943–973

34 Dolan JG, Cherkasky OA, Li Q, Chin N, Veazie PJ. Should health numeracy be assessed objectively or subjectively? Med Decis Making 2016;36(07):868–875

35 Schapira MM, Walker CM, Miller T, et al. Development and validation of the numeracy understanding in Medicine Instrument short form. J Health Commun 2014;19(Suppl (Suppl 2):240–253

36 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(05):672–680

37 Wallace E, McDowell R, Bennett K, Fahey T, Smith SM. Comparison of count-based multimorbidity measures in predicting emergency admission and functional decline in older community-dwelling adults: a prospective cohort study. BMJ Open 2016;6(09):e013089

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

38 Brilleman SL, Salisbury C. Comparing measures of multimorbidity to predict outcomes in primary care: a cross sectional study. Fam Pract 2013;30(02):172–178

39 R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2015

40 Brewer NT, Richman AR, DeFrank JT, Reyna VF, Carey LA. Improving communication of breast cancer recurrence risk. Breast Cancer Res Treat 2012;133(02):553–561

41 Windish DM, Huot SJ, Green ML. Medicine residents' understanding of the biostatistics and results in the medical literature. JAMA 2007;298(09):1010–1022

42 Waldrop-Valverde D, Osborn CY, Rodriguez A, Rothman RL, Kumar M, Jones DL. Numeracy skills explain racial differences in HIV medication management. AIDS Behav 2010;14(04):799–806

43 Hall CW, Davis NB, Bolen LM, Chia R. Gender and racial differences in mathematical performance. J Soc Psychol 1999;139(06):677–689

44 Langford AT, Resnicow K, Roberts JS, Zikmund-Fisher BJ. Racial and ethnic differences in direct-to-consumer genetic tests awareness in HINTS 2007: sociodemographic and numeracy correlates. J Genet Couns 2012;21(03):440–447

45 Schapira MM, Fletcher KE, Gilligan MA, et al. A framework for health numeracy: how patients use quantitative skills in health care. J Health Commun 2008;13(05):501–517

46 Brundage MD, Smith KC, Little EA, Bantug ET, Snyder CF; PRO Data Presentation Stakeholder Advisory Board. Communicating patient-reported outcome scores using graphic formats: results from a mixed-methods evaluation. Qual Life Res 2015;24(10):2457–2472

47 Brundage M, Feldman-Stewart D, Leis A, et al. Communicating quality of life information to cancer patients: a study of six presentation formats. J Clin Oncol 2005;23(28):6949–6956

48 Bromgard GD, Trafimow D, Silvera DH. The influence of presentation format on the "bigger is better" (BIB) effect. Psychol Rep 2013;112(02):458–468

49 NCHS, National Health and Nutrition Examination Survey. See Appendix I, National Health and Nutrition Examination Survey (NHANES). Available at: https://www.cdc.gov/nchs/data/hus/hus16.pdf#079. Accessed June 15, 2018

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

51 Korfage IJ, Fuhrel-Forbis A, Ubel PA, et al. Informed choice about breast cancer prevention: randomized controlled trial of an online decision aid intervention. Breast Cancer Res 2013;15(05):R74

52 Daneman M, Hannon B. Using working memory theory to investigate the construct validity of multiple-choice reading comprehension tests such as the SAT. J Exp Psychol Gen 2001;130(02):208–223

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.