Simple and Multistate Survival Curves: Can People Learn to Use Them?

Tim Rakow, Rebecca J. Wright, Catherine Bull and David J. Spiegelhalter

Med Decis Making published online 29 June 2012
DOI: 10.1177/0272989X12451057

Simple and Multistate Survival Curves: Can People Learn to Use Them?

Tim Rakow, PhD, Rebecca J. Wright, MSc, Catherine Bull, MD, David J. Spiegelhalter, PhD

Objective and Sample: *This investigation assessed the comprehension of survival curves in a community sample of 88 young and middle-aged adults when several aspects of good practice for graphical communication were implemented, and it compared comprehension for alternative presentation formats.* ***Design, Method, and Measurements:*** *After reading worked examples of using survival curves that provided explanation and answers, participants answered questions on survival data for pairs of treatments. Study 1 compared presenting survival curves for both treatments on the same figure against presentation via 2 separate figures. Study 2 compared presenting data for 3 possible outcome states via a single "multistate" figure for each treatment against presenting each outcome on a separate figure (with both treatments on the same figure). Both studies compared alternative forms of questioning (e.g., "number alive" versus "number dead"). Numeracy levels (self-rated and objective measures) were also assessed.* ***Results:*** *Comprehension was generally good—exceeding 90% correct answers on half the questions—and was similar across alternative graphical formats. Lower accuracy was observed for questions requiring a calculation but was significantly lower only when the requirement for calculation was not explicit (13%–28% decrements in performance). In study 1, this effect was most acute for those with lower levels of numeracy. Subjective (self-rated) numeracy and objective (measured) numeracy were both moderate positive predictors of overall task accuracy (r ≈ 0.3).* ***Conclusions:*** *A high degree of accuracy in extracting information from survival curves is possible, as long as any calculations that are required are made explicit (e.g., finding differences between 2 survival rates). Therefore, practitioners need not avoid using survival curves in discussions with patients, although clear and explicit explanations are important* ***Key words:*** *Risk communication; visual displays; survival curves; numeracy.* ***(Med Decis Making XXXX;XX:XXX–XXX)***

Risk communication prior to major medical intervention often focuses on the immediate risk of adverse outcomes. However, to make or share in a truly informed decision, a patient will often need to consider several risks and benefits—the profiles of which may change over time. Additionally, these profiles may vary across treatment options, which can be an important consideration in the choice among treatment alternatives. Survival curves are used routinely in medical journals to chart how the probability of survival for a given patient group changes over time. They are not used widely with patients, although some assessment of their suitability for communicating risk has been undertaken.[1-11] We extend these investigations by examining how accurately people can extract information from survival curves when several aspects of agreed good practice in risk communication are in place. We also extend this research to the case of multistate survival curves, which simultaneously show the time-dependent probabilities for alternative health states (which may differ in attractiveness to the patient) in addition to the probability of survival. These graphics increase the information available to a patient—in principle, allowing consideration of alternative possible outcomes from treatment that may differentially affect his or her daily functioning

Received 13 January 2011 from University of Essex, Colchester, United Kingdom (TR, RJW); the Hospital for Sick Children, Great Ormond Street, London, United Kingdom (KB); and the University of Cambridge, Cambridge, United Kingdom (DJS). We thank the Children's Heart Federation for discussions that informed this work and our participants who gave generously of their time. We also thank Lukasz Walasek for collecting the data that were used to evaluate our task instructions. All those who contributed to the scientific content of this investigation have been included as authors for this article. Revision accepted for publication 31 March 2012.

Address correspondence to Dr. Tim Rakow, Department of Psychology, University of Essex, Colchester, CO4 3SQ, UK; e-mail: timrakow@essex.ac.uk.

DOI: 10.1177/0272989X12451057

and quality of life. However, a search of the literature yielded no studies examining the suitability of multistate survival curves as a tool for communicating such risks and benefits that are liable to become more or less likely to occur over time.

Line graphs are recommended for showing trends over time,[12] and so survival curves should be a good way of presenting survival data to patients.[13] The efficacy of using survival curves with patients has been explored in 2 ways: 1) testing how accurately people extract information from them and 2) considering patients' choices using survival curves—examining preferences for consistency, coherence, and sensitivity to the shape of the curve.

Armstrong and others[2] found that only around 50% of a broad-based community sample correctly answered all of 3 straightforward questions about survival probabilities shown via 2 survival curves. Armstrong and others[1] found that accuracy on such questions improved from 74% to 83% accuracy with practice—unless people were asked about *changes* in survival over time, where accuracy remained around 55%.

Patients' choices imply that survival curves allow them to understand the tradeoffs among different treatment outcomes.[5] Moreover, increasing the degree of explanation that is given when survival curves are shown increases sensitivity to the information represented by the curves.[7] The proportion of patients preferring a given treatment can differ according to whether survival rates are shown as ''proportion alive'' or ''proportion dead'' (i.e., a mortality curve). However, it is not clear that such ''loss-gain'' framing effects are any larger for survival curves than for nongraphical presentation.[14] Intriguingly, although people are less accurate at reading mortality curves than survival curves,[2,11] among those who can extract valid information from them, the consistency of repeated choices is greater for mortality than for survival curves.[11] Relative to nongraphical information, survival curves seem to reduce the impact of immediate outcomes and increase reliance on long-term outcomes.[8] However, few patients (though a much larger proportion of physicians) report paying attention to the middle section of survival curves (i.e., the intermediate time frame).[4,6] Nonetheless, data from our lab show that preferences between 2 options do alter when the middle section of one curve is changed even if the endpoints and crossing points of the curves are unaltered.[15]

Thus, although survival curves offer the potential for reducing the habitual tendency for myopic decision making,[16,17] research on their use is far from uniformly favorable. Seemingly, people often fail to extract correct information from survival curves[2]; can be influenced by irrelevant procedural, semantic, or visual features[9,10]; and may not attend to all the information that they contain.[4,6] However, these studies often did not provide detailed explanations of, or the opportunity to learn how to use, survival curves—which ought to represent good practice.[12,18] Therefore, participants in the studies reported here were given a ''tutorial'' on reading survival curves—our intent being to explore how accurate people can be under favorable circumstances. To inform best practice, we manipulated the complexity of the information presented: contrasting more ''informative'' displays with simpler but less ''efficient'' ones (e.g., 2 options shown together v. options shown separately).

Our investigation also evaluated the use of multistate survival curves (see Methods), which simultaneously show probabilities over time for 3 or more outcomes states—for instance, 1) death, 2) survival with poor heart function, and 3) survival with good heart function[19]; or 1) death, 2) survival with further surgical intervention, or 3) event-free survival. Multistate survival curves offer the potential for a fuller consideration of time-dependent outcomes that patients involved in treatment decisions may wish to take into account—particularly those outcomes that affect quality of life as opposed to, or in addition to, survival per se. However, whether patients are able to use these efficient yet more complex graphics is not yet known.

METHODS

Participants

The participants were 88 volunteers with a mean (standard deviation, s) age of 38.5 (11.1) years and a range of 19 to 68 years and an interquartile range of 30 to 46 years. Most were parents or other adult family members of a child undergoing treatment/investigation for congenital heart disease, recruited in a hospital (n = 44) or via family support groups (n = 9). The remaining participants were (nonclinical) administrative/clerical staff in a hospital (n = 33) or university (n = 2). Fifty-eight percent indicated that they were female, 72% were a parent, and 92% had English as their first language. The sample varied widely by education: 22% had no formal educational qualification beyond those obtained by age 16 years; 26% had education to age 17 or 18 years (but not to degree level); and 30% had an undergraduate degree and 19% a postgraduate degree as their highest qualification (education level was not known for the remaining 3% of participants).

Table 1 Percentage Correct: Study 1—Simple Survival Curves

	Combined Across Conditions	**Language**		**Presentation Format**	
		Congruent	**Incongruent**	**One Figure, Each With 2 Lines**	**Two Figures, Each With 1 Line**
Exercise 1		*n* = 45	*n* = 41	*n* = 42	*n* = 44
1. Number out of 100 **ALIVE/DEAD** 5 years after the start of treatment	95	93	98	95	95
2. Number out of 100 **ALIVE/DEAD** 6 years after the start of treatment	97	98	95	98	95
3. Number out of 100 who died between 4 and 9 years after the start of treatment	77[a]	*73*	*80*	79	75
4. Number out of 100 who died between 2 and 7 years after the start of treatment	74[b]	*76*	*73*	69	80
5. When exactly 50 out of 100 **ALIVE/DEAD**?	97	98	95	98	95
6. Which treatment has more **ALIVE/DEAD** 2 years after the start of treatment	97	96	98	95	98
7. Which treatment has more **ALIVE/DEAD** 7 years after the start of treatment	98	98	98	98	98
Mean (*s*)	91.2 (14.7)	90.1 (15.5)	90.9 (16.8)	90.1 (15.6)	90.9 (16.0)
Exercise 2		*n* = 45	*n* = 41	*n* = 43	*n* = 43
1. Number out of 100 who **ARE/(NOT)** alive & well 3 yrs after the start of treatment	95	98	93	93	98
2. Number out of 100 who **ARE/(NOT)** alive & well 5 yrs after the start of treatment	91	96	85	88	96
3. Number out of 100 **MORE/FEWER** alive & well at 7 years compared with 9 years	87	87	88	84	91
4. Number out of 100 **MORE/FEWER** alive & well at 1 year compared with 8 years	90[ab]	89	90	84	95
5. When 50 out of 100 **ARE/(NOT)** alive and well?	93	93	93	93	93
6. Which treatment has **MORE/FEWER** alive & well 2 yrs after the start of treatment	95	98	93	98	93
7. Which treatment has **MORE/FEWER** alive & well 9 yrs after the start of treatment	98	100	95	98	98
Mean (*s*)	92.7 (10.7)	94.3 (8.8)	90.9 (12.3)	91.0 (11.3)	94.4 (9.9)
Overall mean (*s*)	91.9 (9.6)	92.2 (8.7)	91.6 (10.6)	90.4 (10.4)	93.4 (8.6)

Note: The order of the wording for options shown in bold-type capitals corresponds to the congruent-incongruent (left-right) order. Where percentages are shown in italics, there was no language manipulation. Therefore, the column headings refer to variation in *other* questions. Any differences between conditions for these questions merely reflect sampling variability associated with the random allocation of participants to condition. For accuracy comparisons of equivalent questions *between exercises*, pairs of percentages (in the *same column*) that share a common superscript differ significantly ($P < 0.05$).

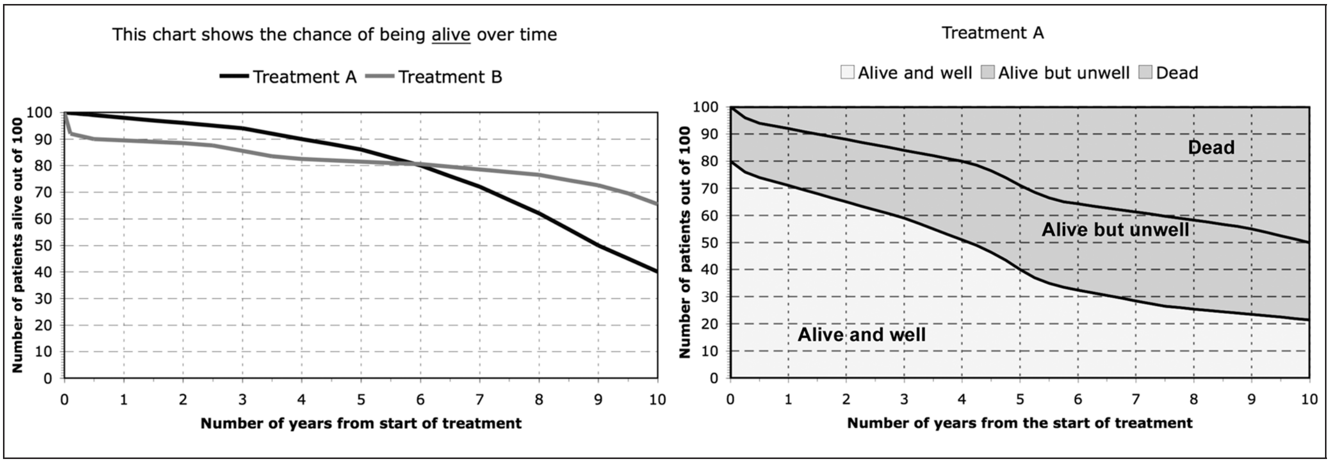


Figure 1 *Sample graphics from the study materials: a pair of simple survival curves (study 1, left) and a multistate survival curve (study 2, right).*

Tasks, Apparatus, and Design

Study 1: Simple survival curves. Two exercises were developed, each comprising survival curves for 2 treatments (''A'' represented by a blue line and ''B'' by an orange line). Vertical axes were labeled ''number of patients alive out of 100'' (0–100 scale), and horizontal axes were labeled ''number of years from start of treatment'' (0–10). Figure 1 shows an example from the study materials. The survival data were hypothetical but realistic: exercise 1 presented survival rates (see Appendix 1), and exercise 2 showed the chance of being alive and well. Seven questions with objective answers were posed for each exercise (see Table 1); plus, participants indicated which treatment they thought was best. The order of the exercises was counterbalanced.

Two factors were manipulated in a 2 × 2 between-subjects design: presentation format and language. For format, the curves were presented as 2 separate figures or on 1 figure showing both curves. In exercise 1, language was manipulated to be congruent or incongruent with the survival frame of the curves for 5 of the 7 questions (''number alive'' versus ''number dead''). In exercise 2, the 2 language formats differed according to the description of the target category (''alive and well'' v. its complement ''**not** alive and well'') or the quantifier used (''more'' versus ''fewer'') as appropriate to the question.

Study 2: Multistate survival curves. The task was equivalent to study 1, except that 3 possible outcomes were represented for each of the 2 treatments: ''alive,'' ''alive and well,'' and ''alive but unwell'' (Figure 1). Six questions were asked for each of the 2 exercises (order counterbalanced); plus, participants stated which treatment was preferred (see Appendix 2).

Two factors were manipulated in a 2 × 2 between-subjects design: figure type and language. For figure type, participants saw either 3 simple survival curve figures (1 for each outcome) each showing both treatments or 2 multistate survival figures for treatments A and B separately. The multistate graphs consisted of 2 lines showing probabilities for survival and for the outcome ''alive and well,'' which partitioned the graph into 3 color-coded labeled regions (''alive and well'' in yellow, ''alive but unwell'' in green, and ''dead'' in blue). The language was varied on 4 questions from each exercise: ''alive'' versus ''dead'' or ''more'' versus ''fewer,'' as appropriate to the question.

Numeracy measures. Participants completed 2 previously validated measures: a Subjective Numeracy Scale (SNS) comprising 8 questions (e.g., ''How good are you at working with fractions?) answered on a 6-point scale labeled at each endpoint (''not at all good'' to ''extremely good'')[20] and an Objective Numeracy Scale (ONS) with 11 questions (e.g., ''Which of the following numbers represents the biggest risk of getting a disease? 1%, 10%, 5%'').[21] Minor changes to the original wording of some SNS and ONS questions were made to reflect UK English idiom. SNS scores were the mean rating for each participant (after reverse-coding 1 item), and ONS scores were the number correct (out of 11).

Procedure

Potential participants were approached by an experimenter and given written information about the study. Volunteers then signed a consent form that reiterated their right to withdraw and to anonymity. Participants were asked not to confer with anyone else, and they were given a pack of materials to complete in the following order:

1. a 550-word explanation of survival curves, using 3 graphs and 8 examples of extracting information (available at http://understandinguncertainty.org/survival-paper-materials);
2. the study 1 exercises (participants being randomly assigned by language and presentation format);
3. a 300-word explanation, with 3 examples of how to extract information for multiple-outcome states—either from multistate survival curves or from several simple curves (appropriate to their assigned condition);
4. the study 2 exercises (participants were randomly assigned to 1 of the 2 language formats independently for each exercise and to 1 of the 2 figure types; random assignment to conditions in study 2 was independent of the assignment to conditions in study 1);
5. a series of exercises for another risk communication study, unrelated to survival curves (not reported here); and
6. the SNS and ONS questions, plus a short demographic questionnaire.

Participants either completed the questionnaire then and there or took it away to be returned in person or by post a few hours or days later. Typical completion times were 45 to 55 minutes. Participants were thanked and offered UK£5 (approximately US$8) for participating. Most participants elected to give their fee to a medical charity. The figure of 88 participants represents approximately two-thirds of those who expressed sufficient interest in participating to request the study materials. Only a small proportion (i.e., <5%) of those approached to participate had declined to take a study pack. Given informal feedback from potential and actual participants, we conclude that the relatively time-consuming nature of our study was a hindrance to completing the study.

Statistical Analysis

Data analysis was performed using Excel and SPSS software. Inferential statistical tests were conducted and are reported 2-tailed with an α level of 0.05 without correction for multiple tests. Mean differences for the percentage of correct responses over a set of questions were examined using analysis of variance (ANOVA). For these analyses, we report partial η^2 as a measure of effect size: this gives the proportion of variance in the dependent variable that is accounted for by the effect under examination. Between-group or paired differences in the proportion of correct responses for individual questions were examined with χ^2 tests. The relationships between individual difference measures (e.g., level of numeracy) and overall accuracy in reading survival curves were assessed using correlation. The sample size provides standard errors of 5%, 3%, and 2% for observed proportions of 75%, 90%, and 95%, respectively. A sample size of 44 per group also provides 80% power (for a single question) to detect a true difference in the proportion of correct responses between groups having 60% and 88% correct responses, at the 5% level.

Funding Source and Research Governance

The research was funded by a research grant from the British Academy, an academic society (part-funded by government) that funds research in the humanities and social sciences. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. The research received ethical approval from the Joint UCL/UCLH Committees on the Ethics of Human Research (Committee A), which is a member of the UK National Health Service National Research Ethics Service.

RESULTS

Each response was scored for accuracy allowing a margin of ±3% (absolute difference) on questions with a percentage answer. For questions regarding time, the margin allowed was ±0.5 years. Four participants completed only 1 of the 2 exercises in study 1. (As they went on to complete further questions, we assume that they inadvertently skipped a page of the materials). Their responses to individual questions were analyzed alongside other participants—however, no overall accuracy score was imputed for these participants.

Study 1: Simple Survival Curves

Table 1 shows the accuracy of response for each question, for each exercise, and overall (combined across all conditions and separately by presentation

Table 2 Percentage Correct: Study 2—Multistate Survival Curves

		Language		Figure Type	
	Combined Across Conditions	**Form A**	**Form B**	**Simple**	**Multistate**
Exercise 1		$n = 46$	$n = 42$	$n = 43$	$n = 45$
1. Number out of 100 **ALIVE/DEAD** 3 years after the start of treatment	89	87	90	93	84
2. Number out of 100 alive but unwell 4 years after the start of treatment	72	*78*	*64*	86^{d}	58^{d}
3. Number out of 100 alive and well 4 years after the start of treatment	89	*83*	*95*	91	87
4. Which treatment has **MORE/FEWER** deaths in the first 3 years	94	93	95	95	93
5. Which treatment has **MORE/FEWER** alive and well 7 years after the start of treatment	86	93	79	88	84
6. Which treatment has **MORE/FEWER** alive 9 years after the start of treatment	83	87	79	88	78
Mean (s)	85.4 (19.7)	87.0 (18.9)	83.7 (20.7)	90.3^{c} (18.6)	80.7^{c} (85.4)
Exercise 2		$n = 46$	$n = 42$	$n = 43$	$n = 45$
1. Number out of 100 **ALIVE/DEAD** 7 years after the start of treatment	90	85	95	93	87
2. Number out of 100 alive but unwell 9 years after the start of treatment	75	*78*	*71*	79	71
3. Number out of 100 alive and well 2 years after the start of treatment	95	*100*	*90*	93	98
4. Which treatment had **MORE/FEWER** deaths between 4 & 6 yrs after starting treatment	85	87	83	81	89
5. Which treatment had **MORE/FEWER** alive but unwell 7 yrs after the start of treatment	84	87	81	86	82
6. Which treatment had **MORE/FEWER** alive 3 years after the start of treatment	84	89	79	93^{e}	76^{e}
Mean (s)	85.6 (17.7)	87.0 (17.5)	84.1 (18.0)	87.6 (17.9)	83.7 (17.6)
Overall mean (s)	85.5 (17.1)			89.0 (16.8)	82.2 (16.8)

Note: Participants completed both exercises with the same type of figure, but language format could vary between exercises. The order of the wording for options shown in bold-type capitals corresponds to the Form A-B (left-right) order. Where percentages are shown in italics, there was no language manipulation. Therefore, the column headings refer to variation in *other* questions. Any differences between conditions for these questions merely reflect sampling variability associated with the random allocation of participants to condition. For accuracy comparisons *between conditions*, pairs of percentages (in the *same row*) that share a common superscript differ significantly ($P < 0.05$).

format and by language). A 2-way ANOVA (presentation format × language) with percentage correct as the dependent variable indicated a small but nonsignificant effect of presentation format (small advantage for 2 figures over 1), $F(1, 80) = 1.91$, $P = 0.17$, partial $\eta^2 = 0.023$, no significant effect of language and no significant 2-way interaction (both $F < 1$). Overall, participants were highly accurate, and this differed little between conditions.

Accuracy exceeded 90% in 11 of the 14 questions. Notably, accuracy fell below 80% for questions 3 and 4 of exercise 1, both of which involved finding differences between 2 points. However, this calculation is also required in questions 3 and 4 of exercise 2, where accuracy was more similar to the questions that involved reading of a single point. McNemar χ^2 tests confirmed that participants were significantly more accurate on question 4 of exercise 2 than they were on questions 3 or 4 of exercise 1 ($P = 0.04$ and $P = 0.02$, respectively). For question 3 of exercise 2, the 11% to 14% advantage over questions 3 and 4 of exercise 1 was not significant ($P = 0.15$ and $P = 0.06$, respectively). We note that the term ‘‘died’’ (exercise 1) may not make the need for subtraction explicit in the same way as the terms ‘‘more’’ or ‘‘fewer’’ (exercise 2)—see discussion. Consistent with the general

Table 3 Correlation (P Value) Between Overall Accuracy and Ability Measures (Numeracy and Education Level)

	Measure of Ability		
Task Accuracy Measure: % Correct	**Objective Numeracy**	**Subjective Numeracy**	**Education Level**
Study 1: Simple survival curves	0.28 (0.01)	0.27 (0.02)	–0.12 (0.30)
Study 2: Multiple outcome states	0.28 (0.01)	0.35 (0.001)	0.23 (0.03)

pattern, accuracy on individual questions differed little across the 2 language formats or presentation formats—and never differed by more than 11% between conditions.

Study 2: Multistate Survival Curves

Table 2 shows the accuracy of response: overall, for each exercise, and for each question (combined across conditions, by language, and by figure type). Accuracy was slightly lower than in study 1. The mean percentage correct was analyzed separately for each exercise using 2-way ANOVA (language × figure type).

For exercise 1, neither the main effect of language nor the 2-way interaction was significant (both $F < 1$). However, the main effect of figure type was significant, $F(1, 84) = 5.57$, $P = 0.02$, partial $\eta^2 = 0.062$, with accuracy being lower for multistate curves than for simple figures. Notably, approximately half the 10% gap between the 2 figure types is attributable to the markedly inferior performance with multistate figures on question 2 (concerning ''alive but unwell'' patients). There was a significant difference between simple and multistate figures for the proportion correct on this question, $\chi^2(1, N = 88) = 8.64$, $P = 0.003$. None of the other questions in exercise 1 had a significant difference between the conditions. We note that reading the ''alive but unwell'' category on a multistate figure required identifying a ''gap'' between 2 curves—but this was not required in the simple figures condition, as probabilities for ''alive but unwell'' were represented by a single curve (see Discussion).

For exercise 2, neither the main effect of language nor the 2-way interaction were significant (both $F < 1.02$), and also, the small accuracy advantage for simple figures was nonsignificant, $F(1, 84) = 1.18$, $P = 0.28$, partial $\eta^2 = 0.014$, The largest individual effect of figure type was for question 6 (''more/fewer alive''), significant $\chi^2(1, N = 88) = 5.01$, $P = 0.03$, which alone accounted for most of the 4% accuracy difference between the simple and multistate figures. When exercises 1 and 2 are taken together, there is a 7% accuracy advantage for simple figures over multistate figures—a difference that approached significance, $t(86) = 1.88$, $P = 0.06$.

There is some evidence that asking which treatment has more patients in a given state yields more accurate responses than asking which treatment has fewer patients (7% difference in percentage correct). However, this difference was nonsignificant in both exercises (P of 0.12 and 0.15, respectively).

Numeracy

SNS scores were calculated for each participant ($\bar{x} = 4.14$, $s = 1.01$, range of 1.38–6.00, interquartile range of 3.63–4.88), and this scale was found to have very good internal consistency (Cronbach $\alpha = 0.84$). In keeping with other studies using this scale,[22] participants' ONS scores were often close to the maximum score of 11 ($\bar{x} = 9.24$, $s = 2.08$, range of 1–11, interquartile range of 9–11), although this scale also had good internal consistency ($\alpha = 0.78$). Consistent with prior research,[20] higher SNS predicted better ONS, $r = 0.48$ ($N = 85$), $P < 0.001$. SNS was not significantly related to educational level (5-point scale), $r = 0.16$ ($N = 84$), $P = 0.15$, but ONS and educational level were significantly related, $r = 0.29$ ($N = 83$), $P = 0.007$.

We used correlation to examine the relationship between task accuracy (percentage correct) and each ability measure (ONS, SNS, and educational level). Table 3 shows that the subjective and objective numeracy measures predict task accuracy to a similar degree and that these brief specific measures are at least as good linear predictors of accuracy as a participant's level of general education.* Examination of scatter plots indicated that none of these correlations were substantially affected by outliers. Post hoc, following a reviewer's suggestion, we further inspected the data to determine whether these relationships between

*Post hoc, we identified "How good are you with percentages?" as the SNS item with the strongest item-scale correlation. Interestingly, this single item was a significant predictor of task accuracy ($r = 0.25$ and $r = 0.28$ for studies 1 and 2) and was superior to education level as a linear predictor of accuracy.

Table 4 Mean (s) Percentage Correct for Separate Levels of the Ability Measures

	Objective Numeracy			
	Low ($\leq$ 7)	Low-Average (8/9)	High-Average (10)	High (11)
Study 1: Simple survival curves	85.1 (9.4)	91.0 (11.6)	93.7 (7.1)	95.0 (9.3)
Study 2: Multiple outcome states	79.2 (21.9)	86.0 (14.2)	84.6 (18.0)	90.7 (13.5)
	Subjective Numeracy (SNS)			
	Low ($\leq$ 3.5)	Low-Average ($3.5 < \text{SNS} \leq 4.3$)	High-Average ($4.3 < \text{SNS} < 4.9$)	High ($\geq$ 4.9)
Study 1: Simple survival curves	88.2 (8.8)	90.6 (12.2)	94.9 (6.8)	94.0 (9.9)
Study 2: Multiple outcome states	78.6 (23.2)	82.1 (20.6)	89.4 (8.4)	90.4 (11.5)
	Education Level			
	To Age 16/17 Years	To Age 18 Years	Undergraduate Degree	Postgraduate Degree
Study 1: Simple survival curves	91.9 (9.4)	94.4 (6.3)	93.1 (9.1)	88.0 (13.2)
Study 2: Multiple outcome states	81.6 (19.5)	80.1 (17.4)	93.0 (8.4)	89.2 (11.3)

ability and task accuracy were in any way nonlinear (e.g., better described by quadratic or step functions). To this end, Table 4 reports mean accuracy for different levels of ONS, SNS, and educational level (levels collapsed to create 4 groups, with groups as equally sized as possible). For ONS, the decrement in accuracy with decreasing numeracy is greatest at the lowest level of ONS: the difference between the ''low'' and ''low average'' ONS groups being slightly greater than that between the ''low average'' and ''high'' SNS groups. Inspection of a (nonlinear) Loess regression plot concurred with this analysis, indicating a distinct ''step down'' in accuracy on the simple survival curves between ONS scores of 7 and 8. For SNS, the ''high'' and ''high average'' groups perform similarly, with performance falling steadily across the lower SNS groups. For educational level, performance on the simple survival curves does not vary systematically by level of education (consistent with the r of –0.12 in Table 3). However, for study 2, those with a university degree performed somewhat better than those without one.

As a further, more detailed analysis of the possible impact of low numeracy, we examined accuracy on individual items in relation to ONS and SNS. Our analysis of study 1 had identified 2 ''hard'' items, where a subtraction is required but perhaps not explicitly indicated by the question wording (questions 3 and 4 of exercise 1). Interestingly, of the 14 items in study 1, these 2 items had the strongest positive correlations with SNS and were the first and fourth most strongly correlated items with ONS. Specifically, participants in the ''low'' ONS or SNS groups achieved slightly less than 50% accuracy on these 2 questions, while those in the ''low average'' and ''high average'' ONS or ''low average'' SNS groups were approximately 75% accurate for these items. Importantly, these same participants performed better on other items, with accuracy for most questions being close to 90%. In contrast, participants in the remaining higher numeracy groups did not perform worse on these 2 questions: they maintained close to 90% accuracy irrespective of the question. Thus, it would seem that the difficulties posed by this item were not experienced uniformly by all participants—only those of average to low levels of numeracy (relative to our sample) experienced difficulties. However, no equivalent pattern could be discerned in the data for the multistate curves in study 2, where question 2 of exercise 1 and questions 2 and 5 of exercise 2 each required subtractions that were not explicitly cued by the language of question that was posed.

DISCUSSION

Our findings encourage greater optimism regarding people's ability to interpret survival curves correctly, compared to what some previous studies have done (e.g., Armstrong et al.[2]). Accuracy was typically in excess of 90% when people were asked about 1 of 2 possible outcomes (e.g., ''alive'' or ''dead'') represented via a simple survival curve. Even for the more complex task of considering 3 outcome states (study 2), accuracy was generally well in excess of 80% irrespective of the particular form of words used. We acknowledge the limitation that our sample was more highly educated than the general population—which could have contributed to the higher levels of performance in our studies. Certainly, in future investigations, we would wish to recruit samples where more than 50% of participants have no college-level education (reflecting UK norms). However, educational level was not a particularly good predictor of task accuracy, which, given that there was reasonable variability in educational attainment, would have been expected if general educational level were an important factor in performance. Specific measures of numeracy—self-reported or objectively assessed—were more reliable predictors of accuracy than general educational level, and it is important to note that these specific measures were not strongly related to levels of general education. We assume that providing a detailed tutorial on how to interpret survival curves (which drew on prior research) explains some of the enhanced performance of our participants relative to those in other studies.[†] In the absence of formally manipulating this within this investigation, we cannot be certain of this—nonetheless, it would be odd if it did not contribute to the accuracy of our participants given the evidence for the efficacy of worked examples across a range of domains[23,24] and the effects of practice that have been noted with interpreting survival curves.[1]

A further limitation of our study sample was that we surveyed neither older adults (over 70 years of age) nor adolescents. Individuals from these groups can be expected to take an active role in the treatment decisions that affect them, but—important for the kinds of information-processing tasks that we examined—the distribution of cognitive ability within such groups may not be closely matched to that of our study participants.[25] Similarly, we had too few participants with English as an additional language to undertake meaningful analysis of the comprehension of individuals having to follow medical information provided to them in their second or third language. This limits the generalization our findings to these important subgroups of patients. Therefore, future research on graphical risk communication would benefit from a specific focus on these subgroups. These further investigations could be enhanced by taking measurements of numeracy (both self-reported and directly assessed, as we did) as well as perhaps more specific measures, such as health literacy.[26,27] This would allow one to disentangle the separate effects of factors that might be (somewhat) related (e.g., age, numeracy, and literacy). This would permit more widely applicable recommendations for best practice than those that we are able to make on the basis of the studies reported here.

These limitations aside, our findings do have valuable implications for best practice in communicating complex risk information. Participants often performed less well on questions that required them to find the difference between 2 points (e.g., changes in survival over time or the difference between 2 curves on a multistate graph). This has been observed before with simple survival curves[1] and should be expected as the opportunities for error increase as the number of mental operations increases. However, our data provide some additional insight into the boundary conditions of this finding. First, the difficulty may not primarily reside in calculating a difference but in knowing *when* this calculation is appropriate. In 2 questions of exercise 1 in study 1, participants were required to say how many people died between 2 time points (which requires finding a difference)—and performed less well on these questions than on other questions in that study. However, crucially, participants performed well when asked how many fewer (or more) participants were alive and well at one point in time in comparison to another (which also requires finding a difference). It may seem obvious to those familiar with survival data that the number of people dying is simply the difference between 2 values for the number of people alive. However, it does strike us that the calculation of a difference is more explicitly signaled by the words ''more'' or ''fewer,'' as these words orientate one toward the outcome that is explicitly labeled on

[†] Some evidence in support of this assumption comes from a small-scale evaluation of the "tutorial" on survival curves used in study 1, conducted subsequent to the studies that we report. This tested the efficacy of providing instructions on interpreting survival curves against providing general instructions on the use of graphs. As a filler task in an unrelated study, 100 participants from a university population (mainly students) were randomly assigned to read either the "specific" 550-word explanation of survival curves used in study 1 or a 620-word "generic" description (including figures) of using and constructing simple graphs (taken almost verbatim from the online resource Wikipedia). All participants then completed the same pair of study 1 exercises (single figures showing 1 line, with incongruent language). Participants given specific instructions made 20% fewer errors than those given generic ones: mean (standard deviation) percentage correct, 86.4 (14.3) with specific instructions versus 83.0 (17.5) with generic instructions. An independent-samples t test found no significant difference in accuracy between these 2 participant groups, $t(98) = 1.07$, $P = 0.29$, $d = 0.21$. However, a more powerful "items analysis," which compared task accuracy for specific and generic instructions across the 14 questions that were answered, detected a significant advantage for the specific instructions, $t(14) = 2.28$, $P = 0.04$.

a survival curve (e.g., ''alive'' or ''alive and well''). In study 2, where participants were asked about ''more/fewer'' deaths, number alive, or number alive and well, performance was generally better. It would seem that the more explicit the signaling of a calculation, the better. A second possible boundary condition is that, seemingly, the difficulties with such calculations (or the language associated with them) arose primarily among those with lower levels of numeracy. Our investigation lacked the statistical power to confidently identify complex interactions between numeracy levels and language variations on *individual* items. However, our unanticipated findings in this regard suggest that such interactions warrant further investigation.

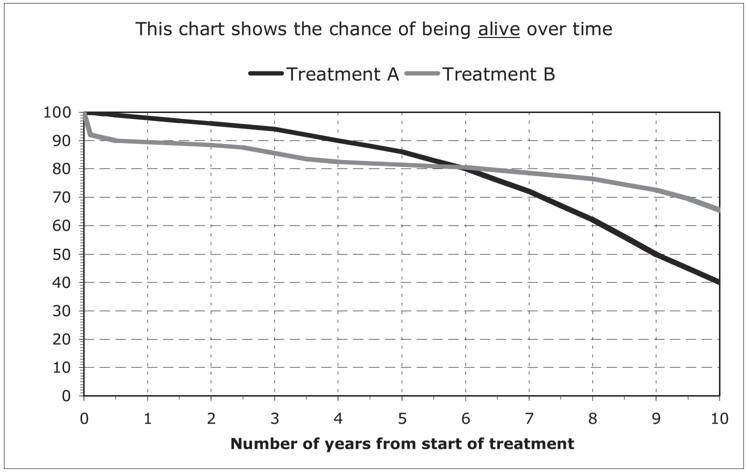
Problems with ''nonsignaled calculations'' may also have hindered performance on the multistate curves. Here, a probability for the event ''alive but unwell'' is not denoted by a point on a curve but by the difference between points on 2 separate curves. This was indicated in the tutorial that participants read but probably needs emphasizing more clearly, as it is not necessarily obvious that the proportion of patients who are alive but unwell is the difference between the proportion who are alive and the proportion who are alive and well. Conversely, while there is a single line on our multistate curves that denotes the probability of survival (the uppermost curve), the region below is divided into two subregions (''alive and well'' plus ''alive but unwell''). This may have prompted participants to perform additions that were not necessary—thereby increasing the chances of error. Consistent with this proposal, question 6 of each exercise in study 2 was answered less well for multistate figures than for simple figures. For this reason, we recommend labeling curves (e.g., ''number alive'') as well as regions above or below the line (e.g., ''alive and well'' and ''alive but unwell''). Thus, although there was some evidence that the added complexity (alongside improved efficiency) of a multistate curve was, to some degree, harmful to accuracy, the improvements to the presentation and explanation of these graphics suggested above would be worth testing to see if accuracy levels for multistate curves can match those for simple curves.

Our studies focused primarily on the initial *extraction* or *comprehension* of information represented via survival curves that were necessary but perhaps not sufficient for using that information appropriately. Future research should examine the *interpretation* and *use* of survival data more directly. For instance, while most participants were able to extract valid information when asked to, we do not know whether they had a clear conception of what it meant for 2 survival curves (for different treatments) to be seen to cross over.

We conclude that, with appropriate explanation, survival curves representing a range of outcome states can be read with a high degree of accuracy. Patients with lower levels of numeracy may experience some difficulties. However, even in our hardest task (study 2), participants below the lower quartile for ONS or SNS maintained close to 80% accuracy. High levels of accuracy were maintained over a range of different forms of language, though if a calculation is not signaled explicitly, this increases the chance of error or misunderstanding—perhaps particularly so for less numerate individuals. Improved ''signaling'' for extracting information from survival curves is likely to be achieved through careful attention to the language used when discussing or examining curves and, particularly in the case of multistate curves, through improved labeling of the graphic.

The high levels of accuracy attained by our participants point to the importance of good explanations for graphical risk information. Our findings regarding the signaling of calculations illustrate the potential danger of ''hidden knowledge'': features that are obvious to those who are familiar with particular kinds of information but which need to be made explicit to the nonexpert. It may be wise to use multiple forms of language (e.g., ''mortality'' and ''survival'' rates) to ensure that concepts, or categories and their complements, are well understood. Regarding the tailoring of numeric information, our studies suggest that general educational background may be a poor predictor of what can be grasped. Specific measures of numeracy are preferred. From a practical standpoint, it is promising that a subjective numeracy measure (i.e., simply asking someone how skilled he or she is) was at least as effective as an objective test in identifying who was more or less able to cope with the graphical/numeric information that we provided. In sum, there is no reason to avoid using survival curves as long as resources are given to assist patients to understand and use them.

Please look at the two survival curves below then answer the questions at the bottom of the page. It may sometimes be difficult to read the scale on the graph accurately. If this is the case, please do not worry. Please just answer as accurately as you are able to.



1) **For Treatment A**, how many patients are dead **5 years** after the start of treatment? _________ patients out of 100

2) **For Treatment B**, how many patients are dead **6 years** after the start of treatment? _________ patients out of 100

3) **For Treatment A**, how many patients have died **between 4 years and 9 years** after the start of treatment? _________ patients out of 100

4) **For Treatment B**, how many patients have died **between 2 years and 7 years** after the start of treatment? _________ patients out of 100

5) **For Treatment A**, when are exactly **50 out of 100 patients** dead? _________ years after the start of treatment

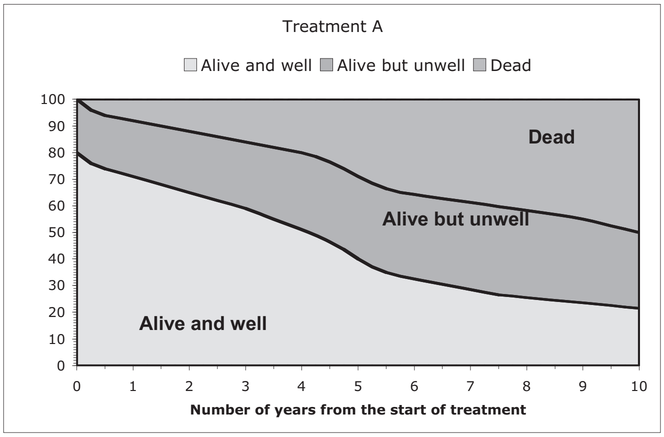
6) For which treatment are more patients dead **after 2 years**? (PLEASE CIRCLE ONE OPTION) Treatment A Both treatments have the same number dead Treatment B

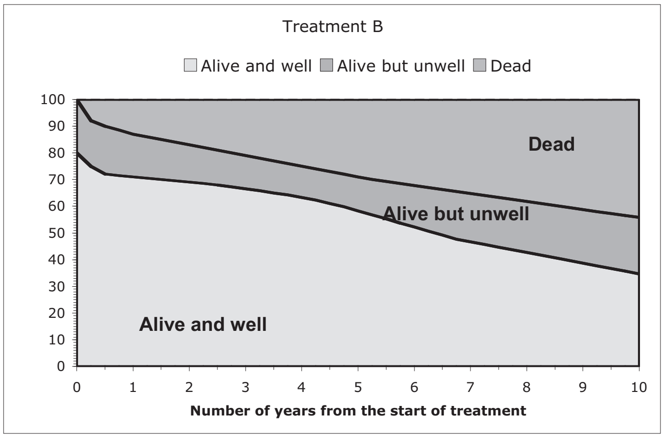
7) For which treatment are more patients dead **after 7 years**? (PLEASE CIRCLE ONE OPTION) Treatment A Both treatments have the same number dead Treatment B

8) **In your opinion**, which is the best treatment? (PLEASE CIRCLE ONE OPTION) Treatment A Treatment B No preference (the two options are equal) Too difficult to make this choice

Appendix 1 Study 1: Simple Survival Curves (Example of One Exercise)

Please take a look at the charts below, and then answer the questions on the right-hand side of the page.





1) **For treatment A,** how many patients are *alive* **7 years** after the start of treatment?

________ out of 100

2) **For Treatment B,** how many patients are *alive but unwell* **9 years** after the start of treatment?

________ out of 100

3) **For Treatment A,** how many patients are *alive and well* **2 years** after the start of treatment?

________ out of 100

4) For which treatment did more patients die **between 4 years and 6 years** after the start of treatment?

Treatment ________

5) For which treatment are more patients *alive but unwell* **7 years** after the start of treatment?

Treatment ________

6) For which treatment are more patients *alive* **3 years** after the start of treatment?

Treatment ________

7) **In your opinion**, which is the best treatment? (PLEASE TICK ONE OPTION)

☐ Treatment A
☐ Treatment B
☐ No preference (both options equal)
☐ Too difficult to choose

If you are able and willing to give a reason for your answer to Question 7, please write here:

Note to appendix, if viewing in gray-scale: Original colors are light blue for dead, light green for 'alive but unwell', and yellow for 'alive and well'. Horizontal and vertical gridlines were visible on the original images.

Appendix 2 Study 2: Multistate Survival Curves (Example of One Exercise)

REFERENCES

1. Armstrong K, Fitzgerald G, Schwartz JS, Ubel PA. Using survival curve comparisons to inform patient decision making: can a practice exercise improve understanding? J Gen Intern Med. 2001;16:482–5.

2. Armstrong K, Schwartz JS, Fitzgerald G, Putt M, Ubel PA. Effect of framing as gain versus loss on understanding and hypothetical treatment choices: survival and mortality curves. Med Decis Making. 2002;22:76–83.

3. Armstrong K, Weber B, Ubel PA, Peters N, Holmes J, Schwartz JS. Individualized survival curves improve satisfaction with cancer risk management decisions in women with BRAC1/2 mutations. J Clin Oncol. 2005;23:9319–28.

4. Mazur DJ, Hickam DH. Interpretation of graphic data by patients in a general medicine clinic. J Gen Intern Med. 1990;5:402–5.

5. Mazur DJ, Hickam DH. Patients' preferences: survival versus quality of life considerations. J Gen Intern Med. 1993;8:374–7.

6. Mazur DJ, Hickam DH. Patients' and physicians' interpretations of graphic data displays. Med Decis Making. 1993;13:59–63.

7. Mazur DJ, Hickam DH. The effect of physicians' explanations on patients' treatment preferences: five-year survival data. Med Decis Making. 1994;14:255–8.

8. Mazur DJ, Hickam DH. Five-year survival curves: how much data are enough for patient-physician decision making in general surgery. Eur J Surg. 1996;162:101–4.

9. Mazur DJ, Merz JF. Five-year survival data in surgical decision making: what aspects of graphical data influence patients' preferences? Theoretical Surg. 1994;9:76–81.

10. Zikmund-Fisher BJ, Fagerlin A, Ubel PA. What's time got to do with it? Inattention to duration in interpretation of survival graphs. Risk Anal. 2005;25:589–95.

11. Zikmund-Fisher BJ, Fagerlin A, Ubel PA. Mortality versus survival graphs: improving temporal consistency in perceptions of treatment effectiveness. Patient Educ Couns. 2007;66:100–7.

12. Lipkus IM. Numeric, verbal and visual formats of conveying health risks: suggested best practices and future recommendations. Med Decis Making. 2007;27:696–713.

13. Lipkus IM, Hollands JG. The visual communication of risk. JNCI Monographs. 1999;25:149–63.

14. McNeil BJ, Pauker SG, Sox HC, Tversky A. On the elicitation of preferences for alternative therapies. N Eng J Med. 1982;306: 1259–62.

15. Rakow T, Balcombe L, Bhadal P, Edwards J. What do people look for in a survival curve—and what don't they look for? Paper presented at: European Society for Medical Decision Making; June 2004; Rotterdam, Netherlands.

16. Frederick S, Loewenstein G, O'Donoghue T. Time discounting and time preference: a critical review. J Econ Lit 2002;40:351–401.

17. Shiv B, Loewenstein G, Bechara A, Damasio H, Damasio AR. Investment behaviour and the negative side of emotion. Psych Sci. 2005;16:435–9.

18. Ancker JS, Kaufman D. Rethinking health numeracy: a multidisciplinary literature review. J Am Med Informatics Assoc. 2007;14:713–21.

19. Rakow T, Bull C. Same patient, different advice: a study into why doctors vary. Arch Dis Childhood. 2003;88:497–502.

20. Fagerlin A, Zikmund-Fisher PA, Jankovic A, Derry HA, Smith DM. Measuring numeracy without a math test: development of the Subjective Numeracy Scale. Med Decis Making. 2007;27:672–80.

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

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

23. Sweller J, van Merrienboer JJG, Paas FGWC. Cognitive architecture and instructional design. Educational Psych Rev. 1998; 10:251–96.

24. Ward M, Sweller J. Structuring effective worked examples. Cogn Instr. 1990;7:1–39.

25. Craik FIM, Bialystock E. Cognition through the lifespan: mechanisms of change. Trends Cog Sci. 2006;10:131–8.

26. Baker DW, Williams MV, Parker RM, Gazmararian JA, Nurss J. Development of a brief test to measure functional health literacy. Patient Educ Couns. 1999;38:33–42.

27. Schwartz LM, Woloshin S, Welch HG. Can patients interpret health information? An assessment of the medical data interpretation test. Med Decis Making. 2006;25:290–300.