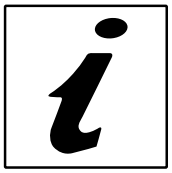


Influence of data display formats on physician investigators' decisions to stop clinical trials: prospective trial with repeated measures

Linda S Elting, Charles G Martin, Scott B Cantor, Edward B Rubenstein

Abstract

Objective To examine the effect of the method of data display on physician investigators' decisions to stop hypothetical clinical trials for an unplanned statistical analysis.
Design Prospective, mixed model design with variables between subjects and within subjects (repeated measures).
Setting Comprehensive cancer centre.
Participants 34 physicians, stratified by academic rank, who were conducting clinical trials.
Interventions Participants were shown tables, pie charts, bar graphs, and icon displays containing hypothetical data from a clinical trial and were asked to decide whether to continue the trial or stop for an unplanned statistical analysis.
Main outcome measure Percentage of accurate decisions with each type of display.
Results Accuracy of decisions was affected by the type of data display and positive or negative framing of the data. More correct decisions were made with icon displays than with tables, pie charts, and bar graphs (82% *v* 68%, 56%, and 43%, respectively; P = 0.03) and when data were negatively framed rather than positively framed in tables (93% *v* 47%; P = 0.004).
Conclusions Clinical investigators' decisions can be affected by factors unrelated to the actual data. In the design of clinical trials information systems, careful consideration should be given to the method by which data are framed and displayed in order to reduce the impact of these extraneous factors.

Introduction

Monitoring interim results of clinical trials is a complex task. Formal interim monitoring points, at which statistical tests are conducted, are designated a priori, but investigators also conduct informal interim safety monitoring. No statistical tests accompany such monitoring in order to avoid the statistical difficulties associated with sequential comparisons. However, an implicit component of informal monitoring is the decision whether to continue the trial or to stop for an unplanned statistical analysis when interim results suggest either dramatic benefit or harmful effects of treatment. When a clear benefit is demonstrated by interim results it is usually considered unethical to continue to expose patients to the inferior treatment.[1]

We hypothesised that in informal safety monitoring the decision to stop the trial for an unplanned statistical analysis could be influenced not only by the actual interim results from the trial but also by the method of displaying those results. Thus, we conducted a prospective study of the effect of the method of displaying results on decisions to conduct unplanned analyses of hypothetical clinical trials.

Participants and methods

Thirty four full time faculty members at the University of Texas MD Anderson Cancer Center volunteered to participate. All 34 participants were physicians, certified by their specialty boards, who were involved in conducting clinical trials in medical oncology. The sample comprised 17 (50%) assistant professors, 13 (38%) associate professors, and four full professors. Five (15%) of the participants were women.

Design

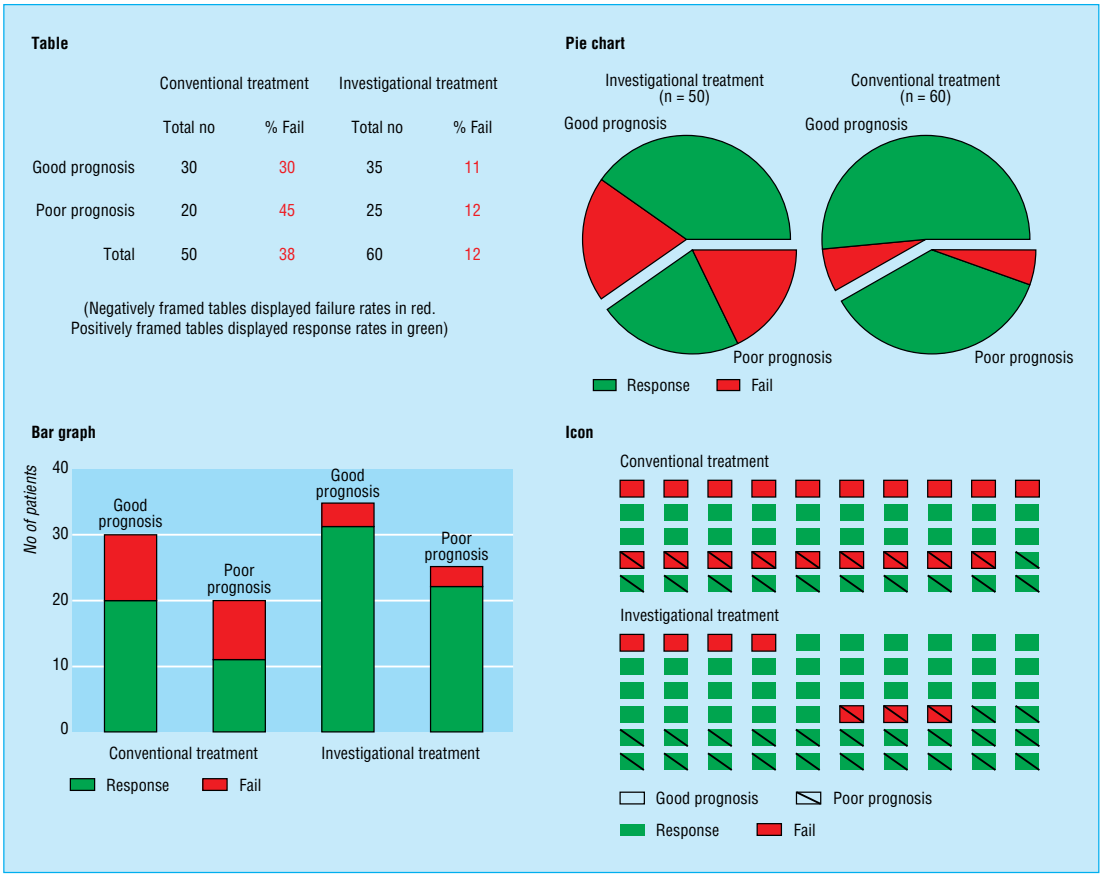
The participants viewed each of four displays of preliminary results from hypothetical clinical trials of a generic "conventional treatment" compared with a generic "investigational treatment." With the exception of the generic treatment names, the experiment mimicked the task of interim monitoring of a clinical trial. A mixed model design was used with comparisons both between participants and within participants (repeated measures). The primary hypotheses concerned the time taken to make decisions, the percentage of correct decisions, and preferences among displays as functions of academic rank, method of display, and framing used. Because of the small numbers of participants at the instructor and professor levels, we divided academic rank into two groups: assistant professor + instructor and associate professor + professor.

We read standard instructions to each participant before the experiment (see appendix). Five points were stressed: four different displays would be evaluated, the data were hypothetical, the four trials were separate and unrelated, the decision to stop required conducting an unplanned statistical test, and the decision to collect more data required the entry of additional patients to the trial. Each participant then evaluated

Editorial by Wyatt

Department of Medical Specialties, University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard—Box 40, Houston, TX 77030-4095, USA
Linda S Elting, *associate professor of epidemiology*
Charles G Martin, *assistant professor of biomathematics*
Scott B Cantor, *assistant professor of medicine*
Edward B Rubenstein, *associate professor of medicine*
Correspondence and reprint requests to: Dr Elting lelting@mdanderson.org

BMJ 1999;318:1527–31



Examples of the four types of display of hypothetical clinical trial data

the four displays of hypothetical data and, for each display, chose whether to stop the trial, to conduct an unplanned statistical analysis, or to collect more data.

The displays

The four types of display used were a table (the most commonly used display format), a stacked bar graph, a pie chart, and an icon display (see figure). Although the use of stacked bar graphs and pie charts has been questioned by some authorities,[2] these were tested because they were the graphical displays requested most often by physician investigators in our institution. Each display showed the results of a clinical trial comparing two treatments identified only as conventional or investigational; their outcomes were categorised as either response or failure. Within each treatment group patients were categorised as having either a good prognosis or a poor prognosis.

An assumption basic to this study was that in each case there was a correct decision. Although the participants were told that the four trials were unrelated, the underlying data used to construct the displays were identical. From a clinical perspective, one of the treatments was clearly superior to the other in response rate (88% *v* 62%). The superiority of the hypothetical treatment persisted across both prognostic groups but was larger in the group with poor prognosis (88% *v* 55%). We used a sample size typical of those used in oncology studies with adequate power (50 patients in one treatment group and 60 in the other), and the difference in response rate was significant (P = 0.002), although P values were not included in the displays. Because of the large difference in efficacy, it would be clinically appropriate in each case to stop the trial for an unplanned analysis.

Control of bias

We stratified the participants by academic rank to control bias due to previous experience of decision making. In these strata we randomly varied the order in which the displays were presented to avoid bias due to learning effect. We hypothesised that stronger evidence would be required to stop trials when investigational treatment was superior than when conventional treatment was better. Thus, we randomly varied the superior treatment from one display to the next for each participant. The graphical displays showed both responses and failures to treatment. Since that is not typically the case in a tabular display, we randomly varied the format of tables among participants to avoid bias due to negative or positive framing. Thus, half of the participants viewed a table with response rates, and half saw a table with failure rates.

Statistical analysis

For each display, we recorded the decision taken, time required for the decision, and each participant's preference, academic rank, sex, and comments. Differences in decision times, a continuous variable, were tested with a mixed model analysis of variance of means, with academic rank being a variable between participants and type of display being a variable within participants. For the discrete variables, correct decisions and preferences, we used Cochran's Q statistic to

test differences in repeated measures among displays and between treatments.[3] For two group comparisons, Cochran's Q test reduces to the McNemar test.[4] We used Pearson's χ^2 statistic for comparisons between participants (independent group) in the proportions of correct decisions, that is, by academic rank and positive or negative framing of tables. Statistical tests were computed with BMDP-Dynamic (BMD Statistical Software, 1993).

Results

All 34 participants viewed the four displays, resulting in 136 decisions. The mean times to make decisions were remarkably similar for each display: 35 seconds for the table, 36 seconds for the pie chart, 34 seconds for the bar graph, and 37 seconds for the icon display ($P = 0.81$). Likewise, there was no difference between academic ranks in the time to make decisions ($P = 0.22$) and no interaction between rank and display ($P = 0.31$). No interactions between display type and other variables, continuous or discrete, were significant. Although the displays were constructed from identical data, none of the participants commented on the similarities, and six volunteered that the data were so different that comparisons of the displays were meaningless. When viewing the table, pie chart, and bar graph displays, some participants requested additional information: five requested P values, and one asked for standard deviations. When viewing icon displays, 11 participants commented on the large, impressive differences between the treatments, seven in terms of response rates and four in terms of failure rates.

Twenty one of the participants preferred the table display, eight preferred the bar graph, and five preferred the pie chart. Despite the superior accuracy of the icon display, none of the participants preferred that method, and eight voiced considerable contempt for the display. Cochran's Q statistic for preferences among the four displays was $\chi^2_3 = 28.4$, $P < 0.0001$.

Display effect

The relation between display format and likelihood of a correct decision was significant (Cochran's Q test = 8.8; $P = 0.0326$). Correct decisions were significantly more common with the icon displays (82%) than with either pie charts or bar graphs, both 56% (McNemar test = 4.8, $P = 0.03$) (table 1). The table display gave intermediate results (68%) not significantly different from those with the icon display (McNemar test = 1.9, $P = 0.17$).

Sources of bias

There was no consistent relation between the order in which the displays were presented and the number of erroneous decisions (table 1). However, there was a slight learning effect in that early displays had more errors overall, although the differences were not significant ($P = 0.77$).

Although all participants were managing clinical trials, their experience varied. Academic rank was used as a surrogate measure of decision making experience. Professors and associate professors made more accurate decisions than assistant professors (71% *v* 60%), but this difference was not significant. This pattern was true for all the displays except for the pie chart (table 2). Use of pie and bar charts resulted in many inaccurate decisions regardless of academic rank.

The likelihood of erroneous decisions was not related to which treatment (conventional or investigational) was superior (table 2). However, the way in which table results were framed made significant differences in the number of erroneous decisions (table 3). Negatively framed tables (those reporting failure rates) resulted in significantly more decisions to stop the trial than positive ones (93% *v* 47%, $P = 0.004$).

Table 1 Evaluation of learning effect: proportion of correct decisions related to type of display and order in which the displays were presented to participants

	% (95% CI) of correct decisions				
Test order	Table	Pie chart	Bar graph	Icon	Overall value for test order*
1	40 (12 to 74)	56 (21 to 86)	62 (24 to 91)	86 (42 to 99)	59 (41 to 75)
2	86 (42 to 99)	40 (12 to 74)	67 (30 to 93)	75 (35 to 97)	65 (47 to 80)
3	62 (24 to 91)	71 (29 to 96)	50 (19 to 81)	89 (52 to 99)	68 (49 to 83)
4	89 (52 to 99)	62 (24 to 91)	43 (10 to 82)	80 (44 to 97)	70 (53 to 85)
Overall value for display†	68 (49 to 83)	56 (38 to 73)	56 (38 to 73)	82 (65 to 93)	

*χ^2 statistic for overall value for test order: χ^2_3=1.14; P=0.7679.
†Q statistic for overall value for display: χ^2_3=8.76; P=0.0326 (Cochran's Q statistic). Icon *v* table: P=0.1655. Icon *v* pie chart or bar graph: P=0.0290 (McNemar test).

Table 2 Proportion of correct decisions related to type of display and to academic rank and which treatment was superior

	% (95% CI) of correct decisions				
	Table	Pie chart	Bar graph	Icon	Overall value
Academic rank:					
Assistant professor (n=17)	59 (33 to 82)	59 (33 to 82)	47 (23 to 72)	76 (50 to 93)	60 (48 to 72)
Associate or full professor (n=17)	76 (50 to 93)	53 (28 to 77)	65 (38 to 86)	88 (64 to 99)	71 (58 to 81)
Superior treatment:					
Investigational	67 (38 to 88)	60 (32 to 84)	42 (20 to 67)	84 (29 to 76)	63 (42 to 67)
Conventional	68 (43 to 87)	53 (29 to 76)	73 (45 to 92)	80 (52 to 96)	68 (55 to 78)
Overall value	68 (49 to 83)	56 (38 to 73)	56 (38 to 73)	82 (65 to 93)	

Discussion

Our data suggest that various factors influence decisions to stop clinical trials for unplanned statistical analyses. These include the method of displaying data and the way in which results are framed. Pie charts and bar graphs seemed to be inferior to table and icon displays, although they were preferred by 15% and 23% of participants respectively. Icon displays led to superior decisions by participants at all levels of experience, but they were not liked by the participants.

Table 3 Impact of how results in tables were framed: proportion of correct decisions with table display related to negative and positive framing*

	% (95% CI) of correct decisions	
	Negative framing	Positive framing
Superior treatment:		
Investigational	86 (42 to 99)	50 (16 to 84)
Conventional	100 (63 to 100)	45 (17 to 77)
Total	93 (68 to 99)	47 (24 to 71)

*Negatively framed tables displayed failure rates, and positively framed tabled displayed response rates. χ^2=8.092; P=0.0044.

Methodological issues

To ensure that observed differences were due to the displays rather than to other issues that might affect decision making, we used a repeated measures experiment with simulated clinical trial data rather than a randomised controlled clinical trial. This artificial setting is a limitation of the study; participants may have made very different decisions in real life situations or when they were not being observed and "graded." Since the experiment was conducted in only one centre, our results may not be generalisable: research practice may evolve locally as clinical practice does, particularly with respect to informal monitoring tasks. In the absence of confirmatory studies from other centres, these results should be interpreted with caution.

As well as the impact of the method of displaying results on the decision to stop a trial for an unplanned analysis, we explored the effect of other factors—prior experience in decision making, loss aversion, framing effect, and learning effect. This was possible because we used a repeated measures design with a separate randomisation for each of three factors: which treatment was superior (conventional or investigational), the order in which displays were presented, and the way in which table results were framed (negative or positive).

Errors in decision making

Despite initial concern about the influence of learning effect on the time to make decisions and their accuracy, the participants performed similarly regardless of the order in which the displays were presented. Likewise, experience in clinical trial decision making, measured here by academic rank, conferred only a slight, non-significant advantage in accuracy (71% *v* 60%). These somewhat surprising findings may reflect the extensive clinical trial experience of the faculty members at this large comprehensive cancer centre, in which over 500 clinical trials are conducted annually. The monitoring task simulated in our experiment is a familiar activity for clinical investigators at a research institution. This hypothesis is supported by the extremely short time taken by participants to make decisions.

Given their extensive experience and resistance to learning effect, why should the participants have made any errors at all? Firstly, the errors could reflect the short time used for decision making (30 seconds on average). However, these times are similar to those recorded in other trials comparing display methods: physicians took an average of 50 seconds to compare two displays and answer a question about the data,[5] and respiratory therapists required an average of 14 seconds to view a complex flow sheet or icon display of seven days' data from a ventilator and identify changes in a patient's status.[6 7] Despite the brief decision times in these studies, the accuracy rates were high, particularly with icon displays. However, the hypothetical nature of our study may have reduced the time or care with which the participants analysed the data; more time would undoubtedly be used to make decisions in actual clinical research practice. While extra time might improve accuracy in general, there are no data suggesting that its effect would vary among the display methods.

Secondly, our results may provide an example of status quo bias—the tendency to maintain one's current position despite explicit evidence supporting change.[8] There are several explanations for this seemingly irrational behaviour,[8 9] but the most likely is that our findings illustrate a form of status quo bias termed endowment effect—the tendency to require more to give up a possession than one is willing to pay to acquire it.[9] It is possible that our participants quite rationally obeyed the axiom never to make a clinical trial decision based on a single observation. Given data from only one monitoring event in the experiment, they might require more compelling evidence than a benefit of 26% to stop a trial for an unplanned test. With this reasoning, erroneous decisions in this experiment might not be considered errors by some investigators.

Impact of type of data display

In our study icon displays produced significantly more accurate decisions than the other displays. Icon displays have been shown to be an effective method for acquiring information in complex medical situations, often resulting in more accurate responses to questions[5] and patient assessments.[6 7] To our knowledge, ours is the first study to explore the use of icon displays for decision making in clinical trials. Our results suggest that they may be as useful for this task as they have been for communicating complex medical information.

Showing every failure (and response), the icon display provides a provocative visual illustration of the differences between treatments' effects that may make it easier to reach correct decisions than the comparison of two numerical rates in a table. This explanation is supported by research in graphical displays of the progress of labour: altering the scale of the *x* and *y* axes on partograms changed the frequency of medical interventions.[10]

Alternatively, the superiority of icon displays may be an illustration of the influence of focusing decision makers' attention on the probability of loss, or in this case failure of treatment. We found that participants were more likely to make correct decisions when the failure rate was displayed (all graphical displays and negatively framed tables) than when only the response rate was displayed in positively framed tables (68% *v* 47%). In our artificial scenario capable clinical investigators made decisions that seem to depart from a rational choice model. Although one treatment provided a benefit of 26%, researchers decided to continue to randomise patients to the other treatment in 53% of cases in which positively framed tables were used. This common phenomenon, termed loss aversion, occurs because decision makers tend to place greater weight on losses than on gains.[11 12] When faced with decisions in which the status quo is an option, decision makers value the potential losses from change greater than the potential gains.

This feature of icon displays could be a problem if they encourage decision makers to overreact to clinically unimportant differences. Studies suggest that the features of graphical displays that make them visually attractive to users may also detract from proper comprehension.[13–15] In our study the correct decision

Key messages

- In clinical trials formal interim monitoring points, at which statistical tests are conducted, are designated a priori, but investigators also conduct informal interim monitoring, when statistical tests are not used
- This study investigated the effect of the method of displaying results on clinical investigators' decisions to conduct unplanned analyses of a hypothetical clinical trial
- The method of displaying results significantly influenced the accuracy of decisions, as did the framing of these results (positive or negative)
- The display formats preferred by the clinical investigators did not lead to the most accurate decisions
- Careful consideration should be given to the method of data display in information systems supporting clinical research

was to stop the trial, so this potential negative effect of icon displays could not be examined.

In view of the apparent superiority of icon displays, it is regrettable that they were so unpopular with the participants. In other studies icon displays were often preferred by nurses, students, and allied healthcare workers but were considered unacceptable by physicians.[5–7] Physicians have generally preferred table displays of data, the most common method for medical data display.[5] Physicians may require considerable persuasion to accept icon displays. Moreover, if clinicians choose display formats they may select the most familiar display rather than the one supporting the best decisions.

We conclude that careful consideration should be given to the use of graphical icon displays when designing monitoring systems for clinical trials. Because of the importance of these decisions, further studies of decision making in clinical trials should be undertaken, and tools to support such decisions should be developed.

We thank Cynthia Karl and Gwen Amos for their assistance in conducting this study. Part of this study was presented at the spring conference of the American Medical Informatics Association, Portland, Oregon, 1992.

Contributors: LSE initiated the research, designed and coordinated the study, participated in analysing and interpreting the data, and cowrote the paper. CGM conducted the statistical analysis, participated in interpreting the results, and cowrote the paper. SBC and EBR participated in interpreting the results and editing the manuscript. LSE and CGM are guarantors for the paper.

Funding: This research is based in part on work supported by the Texas Advanced Technology Program under Grant No 000015004.

Competing interest: None declared.

Appendix: Instructions to participants

The purpose of this study is to determine whether the format in which data are displayed affects the decisions made by physician investigators about clinical trials. The four formats include a standard table of data, a pie chart, a bar graph, and an icon display in which each block represents a person. You will see the four different displays of data from four separate hypothetical randomised clinical trials. The trials are completely unrelated. In fact, the direction of the difference is opposite from one trial to the next in order to avoid bias due to learning effect.

For purposes of this study, please consider the conventional and investigational treatments as "generic" treatment—that is, not antineoplastics, antibiotics, or analgesics, merely some treatment being studied.

You have only one task. View the data supplied for interim monitoring of the clinical trial. As the principal investigator, decide whether the differences are large enough to stop the trial for an unplanned statistical analysis or whether more data need to be collected. The decision to collect more data requires the entry of additional patients (not merely collection of more data on currently enrolled patients).

We will record your decision and the time required to reach that decision. Your decisions will not be compared with those of other participants.

Do you have any questions?

Here is the first data display. Would you stop and analyse the trial or collect more data?

1 Friedman LM, Furberg CD, DeMets DL. *Fundamentals of clinical trials*. 2nd ed. Littleton, MA: PSG Publishing, 1985.
2 Cleveland WS, McGill R. Graphical perception: the visual decoding of quantitative material when on graphical displays of data. *J R Stat Soc Ser A* 1987;150:192-229.
3 Cochran WG. The comparison of percentages in matched samples. *Biometrika* 1950;37:256-66.
4 McNemar Q. Note on the sampling error of the differences between correlated proportions or percentages. *Psychometrika* 1947;12:153-7.
5 Elting LS, Bodey GP. Is a picture worth a thousand medical words? A randomized trial of reporting formats for medical research data. *Methods Inf Med* 1991;30:145-50.
6 Cole WG, Stewart JG. Metaphor graphics to support integrated decision making with respiratory data. *Int J Clin Monit Comput* 1993;10:91-100.
7 Cole WG, Stewart JG. Human performance evaluation of a metaphor graphic display for respiratory data. *Methods Inf Med* 1994;33:390-6.
8 Cartmill RSV, Thornton JG. Effect of presentation of partogram information on obstetric decision-making. *Lancet* 1992;339:1520-2.
9 Dwyer FM. The effect of questions on visual learning. *Percept Motor Skills* 1970;30:51-4.
10 Morgan RL. The effects of color in textbook illustrations on the recall and retention of information by students of varying socio-economic status [doctoral dissertation]. *Diss Abstr Intern* 1971;32(3-B):1892-3.
11 Spaulding S. Communication potential of pictorial illustrations. *AV Commun Rev* 1956;4:31-41.
12 Samuelson W, Zeckhauser R. Status quo bias in decision making. *J Risk Uncertain* 1988;1:7-59.
13 Thaler R. Toward a positive theory of consumer choice. *J Econ Behav Organ* 1980;1:39-60.
14 Kahneman D, Tversky A. Choices, values and frames. *Am Psychol* 1984;39:341-50.
15 Kahneman D, Knetsch JL, Thaler RH. The endowment effect, loss aversion and status quo bias. *J Econ Perspect* 1991;5:193-206.

(Accepted 9 February 1999)

Endpiece

Ideas must be comforting

It is important to realise that ideas are much easier to believe if they are comforting, and that many clinical notions are accepted because they are comforting rather than because there is any evidence to support them. Just as we swallow food because we like it not because of its nutritional content, so do we swallow ideas because we like them and not because of their rational content.

Richard Asher, *Talking Sense,* Pitman Medical Publishing Co Ltd, 1972