Principles of survey development for telemedicine applications

George Demiris
Department of Health Management and Informatics, University of Missouri-Columbia, USA

Summary
Surveys can be used in the evaluation of telemedicine applications but they must be properly designed, consistent and accurate. The purpose of the survey and the resources available will determine the extent of testing that a survey instrument should undergo prior to its use. The validity of an instrument is the correspondence between what is being measured and what was intended to be measured. The reliability of an instrument describes the ‘consistency’ or ‘repeatability’ of the measurements made with it. Survey instruments should be designed and tested following basic principles of survey development. The actual survey administration also requires consideration, for example data collection and processing, as well as the interpretation of the findings. Surveys are of two different types. Either they are self-administered, or they are administered by interview. In the latter case, they may be administered by telephone or in a face-to-face meeting. It is important to design a survey instrument based on a detailed definition of what it intends to measure and to test it before administering it to the larger sample.

Introduction
A survey is a tool for collecting data from respondents from a given population, and may be composed of closed structure or open-ended questions (items). Surveys are widely used in telemedicine to assess the opinions, attitudes and perceptions of users. How users perceive and understand a telemedicine application will influence its level of acceptability and consequently its rate of adoption. The importance of patients’ perceptions and attitudes has been emphasized in the literature. Donabedian\(^1\) suggested that patients are in a unique position to assess important aspects of the quality of the care they receive. Delbanco\(^2\) argued that by using survey instruments designed to solicit focused reports from patients that address each dimension of care, health-care providers can obtain feedback about the mode of health-care delivery. Such reports can be used to identify areas that require improvement.

The success of telemedicine applications depends not only on patients’ but on care providers’ acceptance as well. Thus, health-care providers can also be the respondents of surveys in telemedicine.

What are we measuring?
The most important step in survey development is a detailed definition of the concept that is to be measured. What information or what kind of feedback do we need from our respondents? Often, the terms ‘satisfaction’, ‘attitude’ and ‘usage’ are used interchangeably in the context of surveys, although in fact the concepts are quite distinct. Patients may have an overall positive attitude towards telemedicine applications but be dissatisfied with a specific system due to its hardware or due to scheduling difficulties. Users may like a system, but choose not to use it in future. When defining the concept that is to be measured, one has to determine what the underlying constructs are, before phrasing the specific items of the survey. This process helps one determine whether a survey already exists that measures this concept or whether a new one needs to be developed. What we measure helps us also decide how often we need to measure it.
Types of survey

Surveys are of two different types. Either they are self-administered, or they are administered by interview. In the latter case, they may be administered by telephone or in a face-to-face meeting.

Self-completion questionnaire

A self-completion (or self-report) questionnaire is an instrument that the respondent can complete by reading the questions and providing answers without an interviewer or other person taking part. It can be used for single individuals or with groups. The strengths of a self-completion questionnaire are that it ensures that all the subjects being studied are exposed to the same instrument and it is a convenient method of data collection that can be used in more than one study. A self-completion questionnaire is usually a low-cost technique, which allows data to be collected from many people in a short period of time. The questionnaire is particularly useful when the phenomenon being studied is ‘amenable to self-observation and well-defined data about it can be elicited in simple straightforward questions, and respondents can read and write’.3

One of the weaknesses of surveys’ whether they are self-administered, conducted by telephone or face-to-face, is that the instrument itself may promote change. Respondents can easily fall into role selection when answering questionnaires (i.e. respondents are likely to anticipate what the interviewer expects of them in an interpersonal situation and act accordingly),4,5 since no one is present to observe the role they are taking and challenge it. Specifically for self-administered surveys, a problem may arise if respondents misunderstand the questions or wish for clarification of some items.

Interviewing

Face-to-face

The interpersonal interview is preferable to the self-completion questionnaire when the questions being investigated are long and complex, or the content area is not well defined. An interview will also be preferable if the respondents cannot read or write. The major strength of the face-to-face interview is that by using a well trained interviewer, respondents can be queried intensively, and answers can be clarified and further investigated.4 The interview provides the most flexible method for the use of descriptive cues.4 ‘Descriptive cues’ refer to keywords that an interviewer can use to stimulate a reaction and further the discussion. A variety of questions can be asked and judgements can be made about the respondent’s state.

The face-to-face interview is, on the other hand, susceptible to a variety of biases. Role selection can occur. Interviewers become more proficient over time and more subtle in asking questions, so that later interviews may be conducted differently from earlier ones (which affects the consistency of the research instrument).4,5 The interview may fail to obtain accurate information on highly sensitive subjects and is generally an expensive method of data collection.

Telephone

Telephone interviewing is an alternative to face-to-face interviewing and the information is often cheaper to collect per interview. Telephone interviewing is susceptible to additional biases relative to face-to-face interviewing. A primary concern is population restrictions (e.g. some people do not have a telephone, there are people with unlisted phone numbers, some people may have moved and the public listing of their number has not changed yet). A telephone survey cannot usually be maintained for the same length of time as a face-to-face interview.6

Practical issues of survey administration

It is important that surveys are designed so that they do not become burdensome to the respondents. For example, a long survey with numerous questions may appear likely to produce more information than a short survey. However, a long survey is also more likely to become burdensome to the respondent, thus increasing the likelihood of incomplete or inaccurate responses. Often the timing for surveys administered to patients can be a sensitive issue, if the patient is frail or emotional.

The frequency with which a survey is administered should be based on the information needs of the organization. If, for example, one is interested in measuring how patients change their perceptions of health-care delivery via telemedicine, a questionnaire should be administered more than once, perhaps prior to patients’ experience with the system, again after a few teleconsultations and then periodically after their experience continues to grow. For purposes of quality improvement one may be interested in administering surveys that reflect on patients’ or providers’ experience periodically. Other settings and interventions may require a one-time administration of the survey.
The anonymity of subjects is important, since the responses to survey items may be affected by fear of identification. For example, patients may be reluctant to state that they dislike a telemedicine system if they know that their health-care provider is a great supporter of the system. In such cases it is important to ensure anonymity of responses to the survey items and also to inform the respondents that their responses will be processed and presented only in aggregate form, without identifying individual participants.

If periodic administration of the same survey to the same subjects is required, then there needs to be a means of linking the surveys to an individual respondent. Obviously, the privacy and confidentiality of subjects’ responses need to be protected in accordance with the relevant regulations and any guidelines from institutional review boards or local ethics committees.

A further factor that needs to be considered is data entry. A clear procedure needs to be determined to describe when and how the collected survey data will be entered into a computer for further analysis and how and where the original survey forms will be stored. In order to reduce errors, double data-entry is preferred: i.e. two people enter the survey data into the computer independently, allowing a comparison. Data storage is likely to be governed by regulations designed to protect the privacy and confidentiality of data.

### Designing a survey

In designing a survey, the items have to be specific and a range of semantically and logically distinct response options, including a non-response (‘no opinion’), must be provided. When assessing respondents’ agreement with statements, it is good practice to phrase some items positively (e.g. the telemedicine system saves me time) and some negatively (e.g. the telemedicine system is difficult to use). This can minimize the ‘halo’ effect, where respondents may have an overall attitude towards the system and respond to all items consistently without reading the individual statements carefully. When respondents realize that statements are not always positive, they tend to rate each statement individually and express the level of agreement or disagreement more carefully.

Questionnaires often use a five-point Likert scale to cover several degrees of agreement (1 for ‘strongly disagree’, 2 for ‘disagree’, 3 for ‘neutral’, 4 for ‘agree’ and 5 for ‘strongly agree’). For the overall sum of the items to make sense in a summed scale, it is necessary to ensure that they are all scored in the same direction. That is, items worded in a negative manner need to have a reversed scoring (1 for ‘strongly agree’, 2 for ‘agree’, etc.).

Items should be specific, to avoid generalities or ambiguity. One should also avoid using loaded terms. For example, the question ‘Is a telemedicine visit easier to schedule than the time-consuming, complicated scheduling of a face-to-face visit?’ obviously carries a bias as it labels one position more favourably than the other. Furthermore, it is better to avoid recall times (e.g. ‘In the last 10 years how many times have you seen any health-care provider?’), as people’s ability to recall the past can be limited. Items with compound clauses may introduce undue complexity. For instance, the item, ‘Do you have or have you ever had a physical, mental, or other health condition which has lasted over one year and for which you had to travel more than once to an urban medical centre’ is better broken into two items: ‘Do you have or have you ever had a physical, mental, or other health condition which has lasted over one year?’ and the follow-up item, ‘If you answered yes to the previous question, did this condition require that you travel more than once to an urban medical centre?’

### Readability

Once the first draft of the survey has been created, its readability should be assessed using the Flesch Reading Ease Index and the Flesch–Kincaid Grade Level. The Reading Ease Index is based on the number of words in the instrument, the average sentence length and the number of syllables. It can range from 0 to 100 (the higher the index the easier it is to read the instrument). The Flesch–Kincaid Grade Level is based on the average sentence length and the length of words, and indicates the grade level required to understand the document. A score of 8.0 means, for example, that an eighth-grade student would understand the document. Standard writing approximately equates to the seventh- to eighth-grade level. The Flesch–Kincaid Grade Level can range from 1.0 to 50. A high readability index means that the writing is complex and difficult to read.

Readability statistics can be calculated with commercial word processing software. Readability is important when respondents may have difficulty understanding the items of the survey.

### Validity and reliability

Validity is the correspondence between what is being measured with what was intended to be measured (in
other words, am I really measuring what I think I am measuring?). Several types of validity can be investigated.

Face validity is the easiest to address. It pertains to the question ‘Does my survey look right?’ and can be addressed by obtaining feedback about the instrument from a number of researchers with experience both in instrument design and in telemedicine applications, as well as representatives of the respondent group. Content validity, which describes the extent to which the items in the instrument cover all the relevant issues, can be addressed by including all related issues that need to be covered by the instrument as identified in an extensive literature review as well as during focus groups. Criterion-related validity investigates agreement between the responses to one instrument and some other measure obtained independently with a different instrument. This can be difficult if one is attempting to develop a survey measuring a specific concept or a common concept but within a unique context that does not already exist. Finally, construct validity can be investigated by examining the results of this testing procedure and comparing these to published findings. For example, suppose a survey is developed to assess patients’ perceptions of a home-based telemedicine system. Assume that the findings resulting from this survey indicate high overall patient satisfaction, some concerns about the privacy and confidentiality of medical data, concerns about the use and reliability of the required equipment and a generally positive attitude towards telemedicine in home care. These would be findings that are in agreement with those in the literature. Thus, the findings support the construct validity of the questionnaire.

Reliability of an instrument describes the ‘consistency’ or ‘repeatability’ of the measurements made with an instrument. Parallel forms reliability is used to assess the consistency of the results of two sets constructed to measure the same parameters. Test–retest reliability is used to assess the consistency of a measure from one occasion to another and, therefore, requires that the same subjects are exposed a second time to the same instrument. If one administers a survey assessing patients’ attitudes towards telemedicine, one may expect that these attitudes could change over time (i.e. after exposure to such a system for a few months) but the overall attitude should remain the same if the survey is administered again half an hour after the first assessment. If there is a wide discrepancy of findings across all subjects, this is an indication that the instrument itself may be unreliable.

Finally, internal consistency reliability indicates the consistency of results across items within a test. In effect, the reliability of the instrument is judged by estimating how well the items that reflect the same construct yield similar results. For example, if three items of a survey instrument constitute a subscale that covers the issue of access to care, results for these three items should be consistent in reflecting the overall attitude towards this concept. An index of internal consistency that is widely used is Cronbach’s alpha. Its value, which ranges from zero to one, depends on the variance of the total score and the sum of the variances for each item. A higher value indicates a higher reliability.

There are several other techniques that can be used to test a survey. For example, in order to determine if items are redundant or not correlated with the underlying construct, an item-analysis can be performed for every item by calculating the mean and variance of the total score and then Cronbach’s alpha if the item were to be deleted. Such an item-analysis can identify an item as inconsistent or redundant. A more complicated method used to investigate the covariance structure of the questionnaire items and provide an explanation of the relationships among those variables in terms of a smaller number of unobserved latent variables is the so-called factor analysis. This is a statistical method in which variations in scores on a number of variables are expressed in a smaller number of dimensions (factors). This results in the construction of a few hypothetical variables (factors), which are supposed to contain the essential information in a larger set of observed variables or objects. Thus, the overall complexity of the data is reduced.

Conclusions

Surveys are often used in the evaluation of telemedicine applications. However, if they are not designed and tested properly, the quantity and quality of the resulting information may be questionable and lead to erroneous conclusions. Depending on the intent of the survey, some or all types of reliability and validity need to be addressed. Often, administrators are interested in findings that indicate an overall trend of patient acceptance in their specific setting, for example, and not in developing a survey that produces findings that can be generalized to other settings or an instrument that can be made available to other researchers. In these cases, addressing only basic forms of reliability and validity is sufficient. It is important, however, to design a survey based on a detailed definition of what it intends to measure and to test it before administering it to the larger sample. Surveys can be powerful tools when designed and applied properly, and may be useful in the evaluation of telemedicine applications.
References

1 Donabedian A. Evaluating the quality of medical care. Milbank Q 1966;44 (Suppl.):166
8 Flesch R. Art of Readable Writing. Indianapolis, IN: Macmillan Publishing Company, 1976
11 Krzanowski WJ, Kline P. Cross-validation for choosing the number of important components in principal component analysis. Multivariate Behav Res 1995;30:149–65