Title

Options for handling missing data in the Health Utilities Index Mark 3.

Permalink

https://escholarship.org/uc/item/9j44v9xr

Journal

Medical decision making : an international journal of the Society for Medical Decision Making, 25(2)

ISSN

0272-989X

Authors

Naeim, Arash
Keeler, Emmett B
Mangione, Carol M

Publication Date

2005-03-01

DOI

10.1177/0272989x05275153

Options for Handling Missing Data in the Health Utilities Index Mark 3

Arash Naeim, MD, PhD, Emmett B. Keeler, PhD, Carol M. Mangione, MD

Background. *The Health Utilities Index Mark 3 (HUI3) is a tool composed of 41 questions, covering 8 attributes: vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain. Responses to these questions can define more than 972,000 health situations. This tool allows respondents to answer "Don't Know," for which there is no scoring instruction, to any given question. This situation creates a break in the scoring algorithm and leads to considerable amounts of missing data. The goal of this study is to develop strategies to deal with HUI3 scores for participants who have missing data.* ***Methods***. *The authors used data from 248 individuals enrolled in the Cataract Management Trial, focusing on the HUI3 vision and ambulation attributes, which had 19% and 10% of attribute levels missing, respectively. Inspection and deduction were used to fill in values independent of the value of the missing data, then alternative analytic techniques were compared, including mean substitution, model scoring, hot deck, multiple imputation, and regression imputation.* ***Results***. *Inspection and logical deduction reduced the percentage of missing information in the HUI3 by 49% to 87%. A comparison of analytic techniques used for the remaining HUI3 vision data missing demonstrated the value of building models based on internal response patterns and that simple analytic techniques fare as well as more complicated ones when the number of missing cases is small.* ***Conclusion***. *Analyzing the pattern of responses in cases where the attribute level score is missing reduces the amount of missing data and can simplify the analytic process for the remaining missing data.* ***Key words:*** *health utilities index; missing data; nonresponse; pattern analysis; imputation.* ***(Med Decis Making 2005;25:186–198)***

The value of medical interventions has increasingly become associated with its costs and outcomes, including both life expectancy and health-related quality of life (HRQOL).[1–3] Accurately measuring quality-adjusted life years (QALYs) in clinical trials is a challenge.[4] QALY are the main metric for the denominators of cost-effectiveness (CE) analyses. CE analyses are important for policies designed to prioritize medical resources devoted to treatments such as cataract extraction. Furthermore, in clinical trials that show close calls, the CE estimate may influence which treatment is preferred.

Some have argued strongly that preference measures, such as QALYs, cannot be accurately estimated for older patients, those with low literacy, or those with little education.[5] Threats to validity of HRQOL instruments include 1) the inability to process the depth and complexity of information required in actual judgments and 2) cognitive dysfunction (especially an impaired ability to perform numerical calculations), both of which can be issues for older patients.[5] Few studies have sought to validate instruments that measure HRQOL or preferences for specific health states in the elderly.[6,7] Nevertheless, calculating QALYs is essential for the comparison of the cost-effectiveness of various treatments in any target population.

The Health Utilities Index Mark 3 (HUI3, a registered trademark of HUInc)[8] is a rigorously developed and widely used instrument to measure HRQOL. This instrument is attractive in that it provides simple descriptions of health states and should be usable even

Received 2 April 2003 from the Division of Hematology-Oncology, UCLA Department of Medicine, Los Angeles, California (AN); RAND School of Public Policy, Division of Health Services Research, Santa Monica, California (EBK); and Division of General Internal Medicine and Health Services Research, UCLA Department of Medicine, Los Angeles, California (CMM). Revision accepted for publication 12 January 2005.

Address correspondence to Arash Naeim, 10945 Le Conte Avenue, Suite 2333, Box 951687, Los Angeles, CA 90095-1687; e-mail: anaeim@mednet.ucla.edu.

DOI: 10.1177/0272989X05275153

among older patients with mild cognitive deficiencies or low educational levels. The HUI3 instrument is composed of 40 questions measuring 8 health attributes (vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain) and 1 question that reflects overall health status. The attributes were selected to be structurally independent (each independently affects overall health), and the system as a whole defines 972,000 different health states.[9,10] The HUI3 includes categorical information in the form of attribute levels (for vision and ambulation: best score for level = 1, totally disabled = 6). These attribute scores can then be converted to single-attribute utility scores, ranging from 0 to 1. For example, the single-attribute vision utility ranges from 0.00, representing blindness, to 1.00, representing a perfect vision state. Furthermore, a weighted-scoring algorithm is applied to combine the scores for each attribute to derive a multiattribute utility score that represents the overall health state utility (death = 0.00, perfect health = 1.00).

In the HUI, most questions are of the "yes/no" categorical variety and all questions allow the respondent to legitimately answer "Don't Know" (DK) or to refuse to answer. Because skip patterns, that is, jumps between questions based on responses, are used in determining the attribute level of many HUI3 domains (Table 1), DK responses cause a disruption in the scoring algorithm. When a response is DK, the interviewer is required to move on to the next question in the sequence rather than perform a skip. The failure to follow the appropriate skip pattern eliminates the possibility of accurately scoring the attribute (Table 1).

DK responses to well-defined questions asking about capabilities are very different from typical nonresponse missing data, for which no response is given at all. Little research has been conducted on the best way to handle DK responses. One article reviewed DK responses as they applied to a survey of Slovenians regarding Slovenian independence.[11] In election polls, DK means "undecided" and is valuable information to political parties. In the HUI3, DK is a valid response and may have similar utility in estimating the proportion of a survey population that cannot clearly answer a question one way or another.

In addition to DK responses in the HUI3, traditional nonresponse or data input errors cause disturbances in the skip patterns and complicate scoring. Because the HUI3 is a widely used preference measure and use of the DK option is unlikely to be random, developing strategies to handle DK responses and impute nonresponses so that attribute levels can be assigned to virtually all participants in a clinical trial is very important for subsequent cost-effectiveness analysis.

HANDLING MISSING DATA RESULTING FROM ITEM NONRESPONSE

In the HUI3 interview, item nonresponse may be attributed to 3 primary factors: 1) refusal or inability to answer a question (e.g., use of DK), 2) failure to ask the question or to record the answer (generally rare), and 3) recording logically inconsistent responses. Item nonresponse can be dealt with in several ways. Many software packages delete the records with missing items, which has potentially significant consequences for introducing bias in the analysis.[12] Alternatively, records with missing items can be analyzed separately from records with complete data. Finally, by using information that is known about a participant, one can impute, or fill in, plausible values for missing items.

REVIEW OF IMPUTATION TECHNIQUES

Imputations are drawn from a predictive distribution of missing values, which can be created by a variety of methods using observed data. There are 2 groups of methods used to generate such predictive distributions: 1) explicit modeling and 2) implicit modeling.[13] Explicit modeling is based on a formal statistical model with explicit assumptions. Examples of explicit modeling include 1) mean imputation and 2) regression imputation.[13]

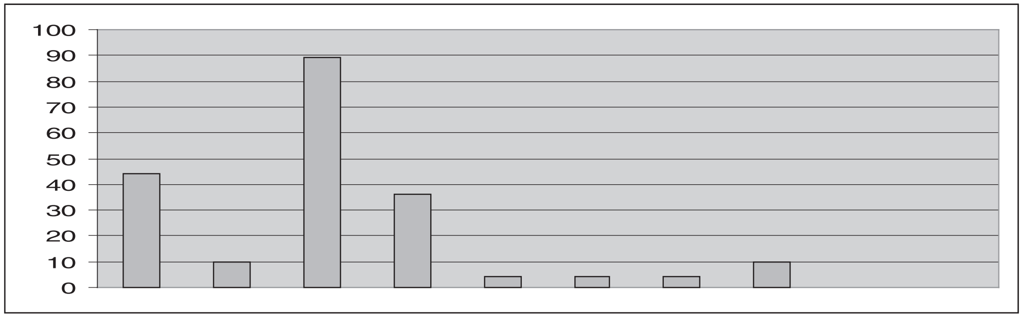
Mean imputation, which substitutes the sample mean for each of the missing values, is an example of an explicit model, in which the assumption is that nonrespondents are similar to respondents on the item in question, but just failed to respond.[13] In addition to this assumption being suspect, this method underestimates the variance by imputing missing values at the center of the distribution.[14] Regression imputation replaces missing values using predicted values from a regression. The regression of the missing items on observed items uses data from both missing and observed variables.[15] For example, complete data from the Activities of Daily Vision Scale (ADVS) can be used in conjunction with existing vision levels from the HUI3 in a regression to predict missing HUI3 vision attribute levels.

Implicit modeling relies on an underlying model based on implicit assumptions. Examples of implicit modeling include 1) hot deck imputation and 2) multiple imputation.[12,16] Hot deck imputation involves substituting values drawn from similar individuals. The schema used to determine "similarity" can range from very simple (all respondents with complete data) to very complex (based on responses to specific items).[13] The disadvantage of a single hot deck imputation is

Table 1 HUI3 Vision Attribute: Possible Scoreable Patterns If the Respondent Does Not Use the "Don't Know" Choice (*N* = 201)

	Response Patterns[a]									
1. During the past 4 weeks, have you been able to see well enough to read ordinary newsprint without glasses or contact lenses? ("Yes" → go to question 4; "No"; "Don't Know"; "Refused")	Y	Y	N	N	Y	N	N	N	N	N
2. Have you been able to see well enough to read ordinary newsprint with glasses or contact lenses? ("Yes" → go to question 4; "No"; "Don't Know"; "Refused")	—	—	Y	Y	—	Y	N	N	N	N
3. During the past 4 weeks, have you been able to see at all? ("Yes"; "No" → finished; "Don't Know"; "Refused")	—	—	—	—	—	—	Y	Y	Y	N
4. During the past 4 weeks, have you been able to see well enough to recognize a friend on the other side of the street without glasses or contact lenses? ("Yes" → finished; "No"; "Don't Know"; "Refused")	Y	N	Y	N	N	N	N	Y	N	—
5. Have you been able to see well enough to recognize a friend on the other side of the street with glasses or contact lenses? ("Yes", "No"; "Don't Know"; "Refused")	—	Y	—	Y	N	N	Y	—	N	—
Attribute level	1	2	2	2	3	3	4	4	5	6
Distribution of scoreable responses	44	10	89	36	4	4	4	10	0	0
Total = 201										

— = Appropriate skip
Y = Yes
N = No



Note: The questions and coding algorithms of the HUI3 are copyright of HUInc and should not be used or reproduced without written permission of HUInc.[8,21]

a. Response patterns should be read vertically.

that it does not reflect sampling variability. Multiple imputation, combining the results of multiple rounds of imputation, helps to better reflect sample variance.[17]

Even more sophisticated imputation methods than those briefly discussed exist. However, imputation methods for handling missing data have been described as "both seductive and dangerous" by some of the foremost experts in the field of imputation.[18] Oftentimes, using imputation techniques lulls the user into believing that the data generated are both complete and legitimate, even when the imputed data may have substantial biases.[13,19] A good imputation procedure meets the objectives of 1) imputing values that are consistent and 2) reducing the nonresponse bias while preserving the variance and the relationships among items as much as possible. In addition, imputation procedures can be set up prior to data collection and evaluated in terms of their impacts on the bias, precision of estimates, and conclusions drawn from the study.[19,20] The HUI3 is a great example because a variety of imputation techniques can be used for missing data and are recommended in the scoring manual.[21]

ITEM NONRESPONSE IN THE HUI3

Nonresponses to the questions in the HUI3 need not result in the loss of attribute-level scores. In many cases, the question lacking a response does not impact the standard scoring. For example, case A in Table 2 demonstrates that whether the answer to question 4 is assumed to be yes or no, the resulting attribute level is 2. In many cases, analysis of all of the questions involved in the attribute score will allow assignment of an appropriate score or dramatically narrow the range of possible attribute levels. Such missing data may be termed "missing without consequence" (MWC) because its absence may have no practical consequences.

This report presents and evaluates several approaches for estimating HUI3 attribute-level scores when the within-attribute item level has some nonresponse. Our hypothesis is that a substantial portion of HUI missing data are actually MWC. Our analysis demonstrates the degree to which the range of attribute levels can be narrowed. Although it may seem obvious that the 1st step in any analysis should be to determine which values are MWC, many users of the HUI3 overlook the importance of this step, skipping to more elaborate imputation techniques. We compare the use of formal techniques of imputation, such as mean substitution, hot deck, multiple imputation, and logistic regression, to a simpler method of inspection and deduction.

METHODS

Data Collection

The data for these analyses are derived from the Cataract Management Trial (CMT). The CMT is a randomized trial that compares the benefits of immediate cataract surgery with those of watchful waiting, using inclusion criteria defined to select older patients who have a low predicted probability of improvement in vision-specific functioning from surgery. Eligible participants had to be older than 64 years, have a diagnosis of bilateral age-related cataracts with no previous history of cataract surgery, be considered candidates for eye surgery by their ophthalmologists, and have a less than 30% predicted probability for improvement in visual functioning after surgery.[22] Additional eligibility criteria included ability to speak English, intact hearing, and sufficient cognitive function to provide informed consent and fully participate in the prerandomization and follow-up interviews. A complete description of inclusion and exclusion criteria is provided with the main trial results.[23] All relevant institutional review boards approved the study protocol.

Prerandomization assessments were conducted on all 248 participants. Measurement instruments included 1) the ADVS, a vision-targeted measure of functional status[24]; 2) the SF-12, a short generic measure of HRQOL[25]; and 3) the HUI3 (HUI23S1.40Q).[9,21] Two HUI3 attributes were chosen for nonresponse analysis: baseline or prerandomization HUI3 Vision and HUI3 Ambulation. We chose those 2 attributes because they accounted for a significant proportion of the missing data and were key outcome measures for the intervention (cataract surgery) in the randomized trial.

Analysis

The DK response to a single question within a specific attribute results in a missing total attribute level and a missing total weighted HUI3 score. Thus, attribute level was considered to be the most important component to impute, because it is used to derive both single-attribute and multiattribute utility scores. We calculated the number of missing items per attribute and the proportion of DK responses. Our analysis took a 2-step approach. As a 1st step, we inspected patterns to fill in attribute values that are either independent of the value of the missing question or can be assigned by logical deduction. In the 2nd step, we compared alternative imputation schemes, including mean substitu-

Table 2 HUI3 Vision Attribute Inspection and Logical Deduction

	Raw Responses						Possible Attribute Patterns and Levels					
Inspection	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**		**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Score**
"Don't Know" (19 cases)						If Q4=Y →	N	Y	—	Y	—	2
A (18 cases)	N	Y	—	??	Y	If Q4=N →	N	Y	—	N	Y	2
						If Q1=N →	N	Y	—	N	N	3
B (1 case)	??	Y	—	N	N	If Q1=Y →	Y	—	—	N	N	3
Logical Deduction	**Raw Responses**						**Possible Attribute Patterns and Levels**					
"Don't Know" (2 cases)												
C (2 cases)	N	Y	—	??[a]	N	If Q4=N →	N	Y	—	N	N	3
Result of a skip error (2 cases)												
D (1 case)	N	N	○	N	N	If Q3=Y →	N	N	Y	N	N	5
E (1 case)	N	Y	Y	○	—	If Q4=Y →	N	Y	—	Y	—	2

?? = Don't know or refused.
— = Appropriate skip.
○ = Missing due to administrative error.
Y = Yes.
N = No.

Note: The questions and coding algorithms of the HUI3 are copyright of HUInc and should not be used or reproduced without written permission of HUInc.[8,21]
a. In case C, answering yes to question 4 would be inconsistent with an answer of no to question 5.

tion, model scoring, hot deck, multiple imputation, and regression imputation.

Inspection, Deduction, and Model Formation

We inspected the pattern of responses to questions for those individuals who were missing vision and ambulation scores. Those patterns were compared with the scoring scheme for the particular attribute. The patterns were identified as MWC nonresponses if the missing response did not alter the attribute-level score, regardless of the answer. For those patterns with relevant nonresponses, an attempt was made to use the internal logic in the sequence of questions within each attribute to score the question correctly and then to provide an attribute-level score. For those patterns for which inspection and deduction were useful, 2 independent judges assessed interrater consistency.

In circumstances where neither inspection nor deduction allowed for definitive scoring, a scoring model was created for use in the 2nd (imputation) part of the analysis. For example, if someone answered DK for question 4 in the vision attribute, 2 cases could be created with a yes and a no, respectively (Table 3, Model 1). If more items in a single attribute were answered DK, there would be more cases reflecting different combinations of possible answers. For example, if both questions 4 and 5 are answered DK, 3 possible combi-

Table 3 HUI3 Vision Attribute Imputation Patterns and Models

Model						All Possible Combinations[a]						# of Participants with Complete Data with pattern
Q1	**Q2**	**Q3**	**Q4**	**Q5**		**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Score**	
Model 1 (2 cases)												
						Y	—	—	Y	—	1	44
Y	—	—	??	Y	→	Y	—	—	N	Y	2	10
(the random draw or regression performed on those individuals with scores who said "yes" to Q1 and did *not* say "no" to Q 5 [not "no" means an answer of yes or the question was skipped])												
Model 2 (9 cases)												
						N	Y	—	Y	—	2	89
					→	N	Y	—	N	Y	2	36
N	Y	—	??	??		N	Y	—	N	N	3	4
(the random draw or regression performed on those individuals who said "no" to Q1 and "yes" on Q2)												
Model 3 (3 cases)												
					→	Y	—	—	Y	—	1	44
??	Y	—	Y	—		N	Y	—	Y	—	2	89
(the random draw or regression performed on those individuals who did *not* say "no" to Q2 and said "yes" to Q4)												
Model 4 (3 cases)												
						Y	—	—	Y	—	1	44
						Y	—	—	N	Y	2	10
??	Y	—	??	??	→	N	Y	—	Y	—	2	89
						N	Y	—	N	Y	2	36
						Y	—	—	N	N	3	4
						N	Y	—	N	N	3	4
(the random draw or regression performed on those individuals who did *not* say "no" to Q2 and did *not* say "no" to Q3)												
Model 5 (2 cases)												
						Y	—	—	Y	—	1	44
					→	Y	—	—	N	Y	2	10
Y	—	—	??	??		Y	—	—	N	N	3	4
(the random draw or regression performed on those individuals who said "yes" to Q1)												
Model 6 (1 case)												
						N	Y	—	Y	—	2	89
N	??	Y	Y	—	→	N	N	Y	Y	—	4	10
(the random draw or regression performed on those individuals who said "no" to Q1, did *not* say "no" to Q3, and said "yes" to Q4)												
Model 7 (2 cases)												
					→	Y	—	—	Y	—	1	44
						N	Y	—	Y	—	2	89
??	??	Y	Y	—		N	N	Y	Y	—	4	10
(the random draw or regression performed on those individuals who did *not* say "no" to Q3 and said "yes" to Q4)												

?? = Don't know or refused.	— = Appropriate skip.
Y = Yes.	N = No.

Note: The questions and coding algorithms of the HUI3 are copyright of HUInc and should not be used or reproduced without written permission of HUInc.[8,21]

a. All possible combinations display the pattern of results if we assume that "Don't Know" is either yes (Y) or no (N).

nations exist (Table 3, Model 2), whereas if questions 1, 4, and 5 are answered DK, 6 possible combinations exist (Table 3, Model 4).

Imputation Technique

The 2nd part of the analysis employed a variety of previously published imputation techniques. We first assumed that a DK response was equivalent to missing. Two sets of imputations were performed, prior to and after using inspection and deduction. We compared 5 imputation techniques on the post-inspection-and-deduction data: mean substitution, model scoring, hot deck, multiple imputation, and regression imputation. This analysis was performed on both single-attribute utility scores and attribute levels. Single-attribute-level scores were calculated based on the scoring algorithm provided with the HUI3.[21]

Mean Substitution[1,13]: This technique uses the single mean of the observed data from individuals with complete data. The mean single-attribute visual utility score and attribute level were used for those with missing scores.

Scoring Models[2]: HUI3 attribute levels range from 1 to 5 or 1 to 6. The use of inspection and deduction reduced the possible attribute levels to either 2 or 3 choices from the 5 to 6 possibilities (Table 3). A total of 7 models were needed to account for the pattern of responses among those individuals who still had missing HUI3 vision-level attribute scores.

Weighted-Mean Imputation: Using these scoring permutations, we then imputed a weighted mean using the proportion of individuals who had complete data with attribute levels allowable in each model.

Hot Deck[3,18,19]: Two sets of analysis were performed. The 1st set, prior to inspection, was a random draw using scores from any of the complete cases. In the 2nd set, postinspection, we used the HUI3 vision-attribute models to identify individuals with complete survey responses whose response pattern was similar to individuals with missing data. Pattern similarity, based on the models in Table 3, was used to define the pool of possible replacement scores. Attribute levels were then selected randomly with replacement from this pool and used to impute levels for those with missing levels. Other covariates, such as age, gender, visual acuity, and SF-12 scores, were considered to help define the pool of similar individuals but were not helpful once the patterns defined from the response models were used.

Multiple Imputation[12,16,17,26]: Because single imputation, using the hot deck procedure above, has potential problems (confidence intervals that are too narrow, high type I error rates),[12] we also performed multiple imputation using the hot deck method above, both prior to and after inspection and deduction. Ten imputed datasets were developed. Previously published techniques for combining the datasets and combining within and between variance were used.[16,17]

Logistic Regression[4,21]: We performed 3 sets of logistic imputation analysis using ADVS scores. Initial analysis was performed using an ordered-logit model with ADVS as the only independent variable and HUI3 Vision attribute levels as the dependent variable, providing probabilities for each level.[1–6] We used a uniform distribution to generate random deviates that were added to the regression probabilities to create additional "noise" in the imputation process. This imputation analysis was performed prior to and after inspection and deduction. In addition, a 3rd analysis used ordered logit regression with ADVS as the independent variable on only "similar" individuals, defined using the previously built response models. These models limited the analysis to either 2 or 3 response categories. If the model had only 2 categories, a simple logistic regression model was used, and for those with 3 categories, a cumulative logit model was used. Other covariates, such as age, gender, visual acuity, and SF-12 scores used to further specify the regression model, did not alter our results (data not shown).

Unweighted Mean Imputation-Using Models: Finally, we assumed that a DK response indicated that respondents were caught in the middle of a dichotomous answer, between yes and no. In that case, it seems reasonable to impute an unweighted mean of the scores associated with yes and with no on that choice and the rest of their responses. For example, if a yes response on an item would lead to an attribute level of 2 based on the other responses and a no response would lead to an attribute level of 4, DK would be assigned a level of (2 + 4)/2 = 3, independently of the pattern of responses and attribute levels among people with complete data.

RESULTS: INSPECTION AND DETECTION

HUI Vision

The vision-attribute portion of the HUI is composed of 5 questions (Table 1). Complete responses and scores were available for only 201 of the 248 patients enrolled. A review of baseline HUI vision scores revealed that 39 of the 47 missing attribute levels were a result of a DK response. With inspection, 23 of these 47 cases could be assigned attribute levels with a high degree of certainty. In 19 of the 23 cases, the missing data were not needed to assign appropriate attribute levels (Table 2, A and B). In 2 cases, internal logical deduction was needed to assign a score (Table 2, C). Here, the main question was how to interpret a DK response to question 4 when the person answered "no" to question 5. It is reasonable to assume, as verified by our 2 independent judges, that those who answered no to question 5 should answer no to question 4 as well.

Two cases of missing data resulted from a skip or input error (Table 2, D and E). In those cases, logic could also be used to assign an attribute level. For case D, the skip pattern dictated that a "no" to questions 1–3 should result in skipping questions 4 and 5. Because questions 4 and 5 were answered, it is logical to assume that the answer to question 3 was "yes," which would have forced the interviewer to ask questions 4 and 5. In case E, a "yes" to question 2 required a skip to question 4. In this case, no data were entered for question 4. Thus, the "yes" response to question 3 most likely was entered in the data file incorrectly and represents the answer to question 4.

HUI Ambulation

The ambulation portion of the HUI comprises 7 related questions (Table 4). The initial assessment of baseline HUI ambulation-level scores revealed 24 missing values, 21 of which were due to DK and 3 to input errors. Inspection led to the immediate assignment of appropriate-level scores in 16 cases, where the missing data were not relevant (Table 5, A, B, and F). In 5 cases, logical deduction was required (Table 5, C, D, and E). If an individual answered "yes" to question 20, then a "no" would be required to questions 16, 17, and 18. Both of our independent judges assigned the level of 3 to all 5 questions. Three cases remained for which neither inspection nor internal deduction was sufficient to assign a level score.

SUMMARY OF INSPECTION AND DEDUCTION

The use of inspection and deduction dramatically reduced the missing attribute-level problem for both the HUI vision and ambulation domains. There was 100% agreement between the 2 independent judges in score assignment for missing scores resulting from DK responses. The only area of disagreement between the judges occurred for the 2 cases resulting from skip errors or transcription error by the interviewer (Table 2, D and E). One judge felt comfortable assigning specific scores, whereas the other judge felt that trying to assign the etiology of the error was outside the scope of logical deduction. Nevertheless, of the 44 cases scored by inspection and deduction, there was agreement on 42, representing an overall 95% interrater agreement rate. The number of missing items was reduced by 49% in the HUI vision attribute and 88% in the HUI ambulation attribute (Table 6). Because only 3 HUI ambulation attribute scores remained missing, the rest of the analysis focused on the remaining 24 missing scores in the HUI vision attribute.

IMPUTATION WITHOUT PRIOR INSPECTION

Without inspection, there were only 201 complete HUI3 vision attribute levels with 47 missing values. Substituting the mean single-attribute utility score, 0.927, for the 47 missing cases provided a total sample mean (standard deviation) of 0.927 (0.095) (Table 7, middle), a reduction in standard deviation of 0.01 (0.105–0.095). Using a random draw of an attribute level from completed cases to fill in the missing cases, a simple hot deck approach yielded a mean single-attribute utility score of 0.919 (0.116). Multiple imputations using hot decking resulted in a mean of 0.922 (0.105). Using ADVS scores via logistic regression to impute missing scores yielded a mean single-attribute utility score of 0.930 (0.102) (Table 7).

There were 23 cases that were "imputed" by inspection and deduction perfectly because these cases had data MWC. The mean single-attribute utility score for these cases (n = 23) was 0.906 (0.131; Table 7, top). Without the inspection and deduction step, these cases were imputed using one of the imputation methods described above. The differences between the imputed values from inspection and deduction and the values imputed using another method represent errors that would then be carried into subsequent analyses if one did not follow the 2-stage approach this article outlines.

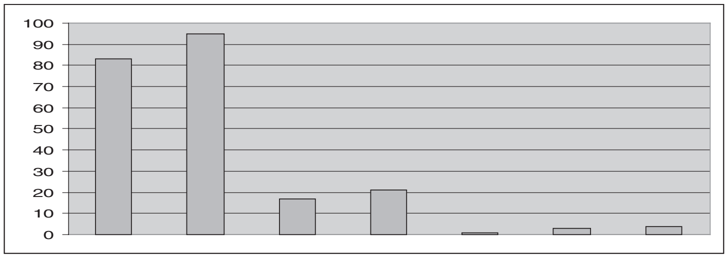
IMPUTATION AFTER INSPECTION

After the initial inspection and logical deduction, 24 cases of missing data remained to be analyzed in the HUI vision attribute. For the 224 complete cases (deleting cases with missing information), the mean single-attribute utility score was 0.925 (0.101). Substituting this mean in the 24 cases yielded a total mean for the 248-case sample of 0.925 (0.102) (Table 7), resulting in more robust standard deviation and variance compared with mean imputation with no inspection. Logistic regression, using ADVS scores as the only independent variable, yielded a complete sample mean of 0.926 (0.108).

Table 4 HUI3 Ambulation Attribute: Possible Scoreable Patterns If the Respondent Does Not Use the “Don’t Know” Choice (N = 224)

	Response Patterns[a]						
16. During the past 4 weeks, have you been able to bend, lift, jump, and run without difficulty and without help or equipment of any kind? (“Yes” → finished; “No”; “Don’t Know”; “Refused”)	Y	N	N	N	N	N	N
17. Have you been able to walk around the neighborhood without difficulty and without help or equipment of any kind? (“Yes” → finished; “No”; “Don’t Know”; “Refused”)	—	Y	N	N	N	N	N
18. Have you been able to walk around the neighborhood with difficulty but without help or equipment of any kind? (“Yes” → finished; “No”; “Don’t Know”; “Refused”)	—	—	Y	N	N	N	N
19. During the past 4 weeks, have you been able to walk at all? (“Yes”; “No” → go to question 22; “Don’t Know”; “Refused”)	—	—	—	Y	Y	Y	N
20. Have you needed mechanical support, such as braces or a cane or crutches, to be able to walk around the neighborhood? (“Yes”; “No”; “Don’t Know”; “Refused”)	—	—	—	Y or N	Y or N	Y or N	—
21. Have you needed the help of another person to walk? (“Yes”; “No”; “Don’t Know”; “Refused”)	—	—	—	N	N	Y	—
22. Have you needed a wheelchair to get around the neighborhood? (“Yes”; “No”; “Don’t Know”; “Refused”)	—	—	—	N	Y	Y or N	Y or N
Attribute level	1	1	2	3	4	5	6
Distribution of scoreable responses	83	95	17	21	1	3	4
Total = 224	→				Poorer function		

— = Appropriate skip
Y = Yes
N = No



Note: The questions and coding algorithms of the HUI3 are copyright of HUInc and should not be used or reproduced without written permission of HUInc.[8,21]

a. Response patterns should be read vertically.

Table 5 HUI3 Ambulation Attribute Inspection and Logical Deduction

	Result of "Don't Know" (20 cases)								Feasible Answers							
	Q16	Q17	Q18	Q19	Q20	Q21	Q22		Q16	Q17	Q18	Q19	Q20	Q21	Q22	Score
Inspection (15 cases)																
								↗	Y	—	—	—	—	—	—	1
A (14 cases)	??	Y	—	—	—	—	—	→	N	Y	—	—	—	—	—	1
B (1 case)	N	N	N	Y	??	N	N	→	N	N	N	Y	Y or N	N	N	3
Logical deduction (5 cases)																
C (2 cases)	N	N	??	Y	Y	N	N	↘								
D (2 cases)	N	??	??	Y	Y	N	N	→	N	N	N	Y	Y	N	N	3
E (1 case)	??	N	N	Y	Y	N	N	↗								
	Result of a Skip Error (1 case)								Feasible Answers							
Inspection (1 case)							○	→								
F (1 case)	N	N	N	N	—	—			N	N	N	N			Y or N	6

?? = Don't know or refused.

— = Appropriate skip.

○ = Missing due to administrative error.

Y = Yes.

N = No.

Note: The questions and coding algorithms of the HUI3 are copyright of HUInc and should not be used or reproduced without written permission of HUInc.[8,21]

Table 6 Impact of Inspection and Deduction on Reducing Missing Data

	Before Inspection/ Deduction		After Inspection/ Deduction	
	Vision	Ambulation	Vision	Ambulation
Complete responses	201	224	224	245
Missing responses	47	24	24	3
Don't know	39	21	22	3
Refuse/ Skip error	8	3	2	0

From the 24 missing HUI vision scores, 7 models were needed to represent the pattern of missing information from which the sets of "similar" individuals could be defined (Table 3). These models were used for the subsequent imputation techniques. Weighting the feasible pattern scores in each model by the distribution of the scores among those with complete answers and then substituting the weighted mean for missing values yielded a mean single-attribute utility score of 0.927 (0.103).

Using the more complex imputation hot deck approach provided a complete sample mean single-attribute utility score of 0.930 (0.103) (Table 7). Multiple imputations using the hot deck procedure yielded a mean of 0.929 (0.103). The ADVS scores and the models with internal information on responses were used to perform an ordered-logit regression across a similar population. The complete sample mean was 0.929 (0.104) (Table 7).

If a DK response reflects an individual being caught in the middle between a "yes" and "no," we can substitute the unweighted mean of the possible scores for each missing pattern for the missing attribute level

Table 7 HUI3 Vision Single-Attribute Utility and Level Imputation: A Comparison of Imputation Methods with and without Prior Inspection and Deduction

	Single-Attribute Utility Score		Mean Attribute Level	
Complete Data	**Mean**	**(*s*)**	**Mean**	**(*s*)**
n = 201/248 (prior to inspection)	0.927	(0.105)	1.960	(0.730)
n = 224/248 (postinspection)	0.925	(0.101)	1.986	(0.734)
Comparison of imputation strategies for those assigned by inspection and deduction (*n* = 23)				
Inspection and deduction	0.906	(0.131)	2.217	(0.671)
Mean substitution	0.927	(0.000)	1.960	(0.000)
Simple hot deck	0.875	(0.171)	2.304	(0.703)
Multiple imputation (10 draws)	0.919	(0.123)	1.992	(0.779)
Logistic regression using ADVS	0.932	(0.110)	1.957	(1.022)
Imputation strategy *without* step 1 inspection (*n* = 248)				
Mean substitution	0.927	(0.095)	1.960	(0.660)
Simple hot deck	0.919	(0.116)	1.971	(0.781)
Multiple imputation (10 draws)	0.919	(0.123)	1.997	(0.788)
Logistic regression using ADVS	0.930	(0.102)	1.964	(0.743)
Imputation strategy *after* step 1 inspection (*n* = 248)				
Mean substitution	0.925	(0.102)	1.986	(0.694)
Logistic regression using ADVS	0.926	(0.108)	1.972	(0.716)
Simple model scoring scheme A				
Mean (weighted) score	0.927	(0.103)	1.963	(0.686)
Hot deck using models	0.930	(0.103)	1.931	(0.702)
Multiple imputation (10 draws)	0.929	(0.103)	1.937	(0.708)
Logistic regression using ADVS and models	0.929	(0.104)	1.935	(0.700)
Simple model scoring scheme B				
Mean unweighted score	0.922	(0.105)	2.000	(0.709)

Note: ADVS = Activities of Daily Vision Scale; *s* = standard deviation.

value. Using this procedure, the mean of the total sample was 0.922 (0.105) (Table 7, bottom). The mean is greater for those with complete data, as expected if a DK response has the assumed meaning.

DISCUSSION

Missing information is a potential pitfall of the HUI3 because survey subjects have the option to answer questions with a DK response. It is not clear if the high rate of DK responses in this trial was due to the older age of the participants. However, without a method to handle valid DK responses, there would have been a loss of 20% of the data endpoints in the cost-effectiveness analysis (data not shown). This amount of missing data may be unacceptably high. Differences in utility of just 0.02 over a 1-year time horizon could change the incremental cost-effectiveness analysis for cataract surgery from $50,000/QALY to $100,000/QALY (unpublished data).

In this article, we compared a 1-step (direct imputation) and 2-step approach for handling the missing items. In the latter approach, the 1st step inspects patterns and uses logical deduction to fill in attribute levels that are independent of the missing data. It may seem self-evident that there should always be an inspection and deduction step when using the HUI3, but our experience is that many users of this instrument do not perform this step because it is not a formal part of the scoring algorithm[21] and it is potentially time consuming. However, if the logical inference could be coded and applied to the data automatically, it could ultimately save a substantial amount of time. Inspection and deduction successfully imputed 49% to 87% of the missing attribute levels. The 2nd step imputed the remaining 13% to 51% of missing values, employing a variety of published methods.

Pattern inspection and internal logical deduction for each attribute with missing data proved to be valuable. In many cases, the nonmissing items within the attribute determined unique attribute-levels so that the data were missing without consequence. Even in circumstances where missing data did affect the attribute score, pattern inspection was useful in limiting choices to only 2 or 3 levels consistent with the observed responses.

We found that when the amount of missing data is small, results may be insensitive to imputation methods, and simple methods, such as mean substitution or model scoring, will suffice for the 2nd imputation step. Furthermore, disregarding the internal response patterns from which we perform our inspection and deduction, and pursuing straight imputation from the start may be unreliable and provide incorrect attribute levels, erroneous group means, and unwanted variance reduction, which in turn can lead to inappropriately narrow confidence intervals and type I errors.

More sophisticated imputation methods may preserve variance estimates in the sample. The use of within-attribute responses in the imputation of missing attribute levels reduces the range of feasible scores. As a result, sophisticated imputation methods do not perform better from the standpoint of preserving variance estimates than simpler methods. Using a weighted mean value from the patterns of missing data yielded results very similar to those of mean substitution, hot deck, multiple imputation, and logistic regression methods. Logistic regression imputation yielded standard deviations that corresponded better with the overall sample's, but the method ignores appropriate answers to other questions within the attribute exploited by other methods. Additionally, in many clinical trials that use the HUI3, data will not have been collected with other condition-specific measures such as the ADVS.

After taking the pattern of responses into account, regression using ADVS scores did not contribute significantly, in terms of mean and variance estimates, in predicting vision-attribute scores. Other scales administered to the same subjects at the same time for other attributes could also be considered for use in modeling. For example, the SF-12's physical functioning items might be used in a regression imputation approach for ambulation, but such additional scales would offer little improvement to the imputation if their ability to discriminate between HUI3 states were limited. The SF-12 asks individuals if they can walk but is not able to discriminate whether they need help or use equipment, essential for imputing the ambulation-attribute score. Additionally, respondent burden and interview length limits this approach.

In summary, pattern inspection and logical deduction can greatly mitigate problems with DK responses and missing values in the HUI3. The initial task of developing algorithms for scoring missing data by inspection and deduction can be time consuming but can be coded and implemented automatically to reduce time in the long run. Those algorithms can be verified and standardized easily using a panel of independent judges. In the long run, such algorithms will save significant amounts of time and provide for more accurate estimates of the incremental cost-effectiveness of various treatments by ensuring that virtually all participants randomized will have an estimate of their QALYs both before and after an intervention. After an initial stage of pattern inspection and deduction, the missing data problem may become so small that simple imputation methods may suffice for the remaining missing data. This 2-step strategy alleviated the HUI3 missing data for vision and ambulation attributes cases resulting from DK responses and could help with other HUI attributes as well.

ACKNOWLEDGMENTS

The authors especially acknowledge Honghu Lu, PhD, who reviewed the statistical methodology and results of this work prior to its submission. The authors also thank the following ophthalmologists who donated time and effort to enroll the participants into the cataract management trial: Drs David Aizuss, Adrienne Alston, Arnett Carraby, Richard Casey, Donald Dickerson, Brad Elkins, Jay Gross, John Hofbauer, Cristina Kenney, Ezra Maguen, Samuel Masket, Kevin Miller, Anthony Nesburn, Michael Olson, James Peace, George Rajacich, James Salz, Lawrence Schwartz, Nancy Wang, Jeffrey Winston, and Peter Zeegan. The authors also acknowledge M. Roy Wilson, MD, the lead ophthalmologist during the field period, and Steve Whipple, PhD, the project director during the field period, Ron Hays, PhD, and David Reuben, MD, for reading the article and providing editorial guidance, and Neraj Sood, MS, MPhil, for serving as one of the independent scorers. This research was supported by National Institute of Aging (NIA) grant #AG10415-12, NIA Claude D. Pepper Older American Independence Center #AG10415-01, and through a Hartford Center of Excellence Fellowship Award.

REFERENCES

1. Paltiel AD, Fuhlbrigge AL, Kitch BT, et al. Cost-effectiveness of inhaled corticosteroids in adults with mild-to-moderate asthma: results from the asthma policy model. J Allergy Clin Immunol. 2001;108(1):39–49.
2. Goldie SJ, Kuhn L, Denny L, Pollack A, Wright TC. Policy analysis of cervical cancer screening strategies in low-resource settings: clinical benefits and cost-effectiveness. JAMA. 2001;285(24):3107–15.
3. Zimmerman RK, Jackson RE. Vaccine policy decisions: tension between science, cost-effectiveness and consensus? Am Fam Phys. 2001;63(10):1919, 1923.
4. Tsevat J, Dawson NV, Wu AW, et al. Health values of hospitalized patients 80 years or older. HELP Investigators. Hospitalized Elderly Longitudinal Project. JAMA. 1998;279(5):371–5.
5. Lenert L, Kaplan RM. Validity and interpretation of preference-based measures of health-related quality of life. Med Care. 2000;38(9 Suppl II):138–50.
6. Brazier JE, Walters SJ, Nicholl JP, Kohler B. Using the SF-36 and Euroqol on an elderly population. Qual Life Res. 1996;5(2):195–204.
7. Coast J, Peters TJ, Richards SH, Gunnell DJ. Use of the EuroQoL among elderly acute care patients. Qual Life Res. 1998;7(1):1–10.
8. HUInc. Health Utilities Index: an overview. Available from: http://www.fhs.mcmaster.ca/hug/
9. Feeny DH, Torrance GW, Furlong WJ. Health Utilities Index. In: Spilker B, ed. Quality of Life and Pharmacoeconomics in Clinical Trials. 2nd ed. Philadelphia: Lippincott-Raven; 1996. p 239–52.
10. Grootendorst P, Feeny D, Furlong W. Health Utilities Index Mark 3: evidence of construct validity for stroke and arthritis in a population health survey. Med Care. 2000;38(3):290–9.
11. Rubin DB, Stern HS, Vehovar V. Handling “don’t know” survey responses: the case of the Slovenian plebiscite. J Am Stat Assoc. 1995;90(431):822–8.
12. Rubin DB. Multiple Imputation for Non-response in Surveys. New York: John Wiley; 1987.
13. Little R, Rubin DB. Statistical Analysis with Missing Data. Hoboken (NJ): John Wiley; 2002.
14. Figueredo AJ, McKnight PE, McKnight KM, Sidani S. Multivariate modeling of missing data within and across assessment waves. Addiction. 2000;95(Suppl 3):S361–80.
15. Belin TR, Diffendal GJ, Mack S, Rubin DB, Schafer JL, Zaslavsky AM. Hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation. J Am Stat Assoc. 1993;88(423):1149–66.
16. Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. Stat Med. 1991;10(4):585–98.
17. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res. 1999;8(1):3–15.
18. Dempster AP, Rubin DB. Introduction. In: Madow G, Olkin I, & Rubin D, eds. Incomplete Data in Sample Surveys: Volume 2. New York: Academic Press; 1983. p 3–10.
19. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods. 2002;7(2):147–77.
20. Oshungade IO. Some methods of handling item non-response in categorical data. The Statistician. 1989;38(4):281–96.
21. Furlong W, Feeny D, Torrance GW. Algorithm for Determining HUI Mark 2 (HUI2)/Mark 3 (HUI3) Health Status Classification Levels, Health States, Health-Related Quality of Life Utility Scores and Single Attribute Level Utility Scores from HUI23S1.40Q Health Status Questionnaire. Ontario: Health Utilities; 1999.
22. Mangione CM, Orav EJ, Lawrence MG, Phillips RS, Seddon JM, Goldman L. Prediction of visual function after cataract surgery. A prospectively validated model. Arch Ophthalmol. 1995;113(10):1305–11.
23. Naeim A, Keeler E, Mangione C. Cost effectiveness of cataract surgery in marginal patients (abstract). J Am Geriatr Soc. 2004;52:S9.
24. Mangione CM, Phillips RS, Seddon JM, et al. Development of the ‘Activities of Daily Vision Scale’. A measure of visual functional status. Med Care. 1992;30(12):1111–26.
25. Ware J Jr, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care. 1996;34(3):220–33.
26. Belin TR, Hu MY, Young AS, Grusky O. Performance of a general location model with an ignorable missing-data assumption in a multivariate mental health services study. Stat Med. 1999;18(22): 3123–35.