Methods for incorporating death into health-related variables in longitudinal studies

Paula Diehr, University of Washington
Methods for incorporating death into health-related variables in longitudinal studies

Paula Diehra,b,*, Laura Lee Johnstone, Donald L. Patrickb,c, Bruce Psatyb,c,d

aDepartment of Biostatistics, Box 357232, University of Washington, Seattle, WA 98195, USA
bDepartment of Health Services, University of Washington, Seattle, WA 98195, USA
cDepartment of Epidemiology, University of Washington, Seattle, WA 98195, USA
dDepartment of Medicine, University of Washington, Seattle, WA 98195, USA
eOffice of Clinical and Regulatory Affairs, The National Center for Complementary and Alternative Medicine (NCCAM), National Institutes of Health, 6707 Democracy Blvd, Suite 401, Bethesda, MD 20892, USA

Accepted 5 April 2005

Abstract

Background and Objectives: Longitudinal studies of health over time may be misleading if some people die. Self-rated health (excellent to poor) and the SF-36 profile scores have been transformed to incorporate death. We applied the same approaches to incorporate death into activities of daily living difficulties (ADLs), IADLs, mini-mental state examination, depressive symptoms, blocks walked per week, bed days, the timed walk, body mass index and blood pressure.

Study Design and Setting: The Cardiovascular Health Study of 5,888 older adults, was followed up to 9 years. Mean age was 73 at baseline, and 658 had an incident stroke during follow-up.

Methods: We recoded each variable as the probability of being healthy 1 year in the future (PHF), conditional on the current value of the variable. This was done for 11 health variables, using three definitions of healthy, and two estimation models. Deaths were set to zero, and mean PHF was plotted in the 3 years before and after an incident stroke.

Results: Analyses without the deaths were too optimistic. The effect of stroke was greatest on hospitalization, self-rated health, and IADLs. Alternative transformation approaches had slightly different results.

Conclusion: These methods provide an additional approach for handling death in longitudinal studies. © 2005 Elsevier Inc. All rights reserved.

Keywords: Death; Longitudinal; Years of healthy life; ADL; Quality of life; Stroke

1. Introduction

Research on the health of older adults is increasingly based on outcomes other than survival. Such health-related outcome variables include morbidity, health status, and health-related quality of life. Study designs using these measures typically obtain a value for each subject at baseline, and then multiple times thereafter. Typical analyses of such outcome variables may be based on the comparison of the baseline measure to the final measure, the slope of the change over time, or the area under the curve of the measure plotted over time. Such summary measures may be used descriptively, to compare one treatment group to another, or to measure the change in the health-related variable due to some health event such as a stroke. Unfortunately, if people die during the study, their health-related variable is “missing” after death.

The usual analytic approach is to perform either a complete case analysis (dropping the people who died) or an available data analysis (using fewer measures for persons who died). Both of these approaches effectively focus the analysis on the healthiest subgroup of the population, the survivors, and may give an incorrect picture of the changes in health over time for the original cohort. Analyses that omit the deaths have been used to “prove” that younger persons age faster than older persons [1], and that the dying process has little effect on physical health [2]. A companion survival analysis is often performed, in which the effect of the “treatment” on survival is estimated for all persons, and the effect of the treatment on the health-related variable such as difficulties with activities of daily living (ADL) is estimated only for those who survived, even though most of the ADL difficulties occurred in those who eventually died. The ADL results would be strictly
applicable only to those who were alive at the end of the study, which could not be known in advance. Such results cannot be used to predict future health trajectories for new subjects, whose future survival is not known.

Revicki et al. [3] have suggested using missing value techniques to impute the value that would have been attained had the person not died. This approach may be appropriate for some studies of older or healthier populations, where the death may have occurred independent of the person’s health, but it does not seem appropriate for studies of older adults. Pauer et al. [4] have pointed out that the use of pattern-mixture models to account for data missing due to death also has the unfortunate effect of implicitly imputing the values of (for instance) ADL after death as though the person had not died. Other techniques, such as random effects regression or hierarchical modeling also implicitly impute values for the dead persons as though they had not died, and so do not account properly for deaths.

A final approach is to transform the variable of interest into a new variable that has a clear meaning and is also defined for death. Several approaches have been presented for the ubiquitous measure “how is your health? excellent, very good, good, fair, or poor” (EVGGFP) [5,6]. When this measure is used longitudinally, one can add a category for the dead. The proportion of people in each health state over time can then be examined, using stacked bar graphs [1,7]. EVGGFP can also be dichotomized into healthy/not healthy, with healthy defined as excellent, very good, or good (E/VG/G) and not healthy defined as fair, poor, or dead. This transformation permitted plotting the trajectories of health over time before and after sentinel events such as stroke, myocardial infarction, and death [1,7]. Finally, the variable can be transformed to the probability of being healthy 1 year in the future, conditional on the current value, with deaths logically set to zero because there is no possibility of being healthy in the future [5]. All of these measures are interpretable and understandable.

Similar transformations have been studied for the subscales and summary scores of the SF-36 [2,5], but not for other commonly used health-related variables. Here, we use these approaches to transform the 11 health-related variables shown in Table 1 into new variables that are defined for death. We address whether it makes sense to include death as part of these variables, and examine properties of the transformed variables by comparing the effect of stroke on each variable. We also consider new definitions of “healthy,” and predicting future health from age and sex as well as from current health. Due to space constraints, this article provides detailed results only for ADL, although the transformations are presented for all 11 variables. Some additional results are provided in an online technical report [8]. The resulting transformation equations may be useful to investigators in longitudinal studies using these variables.

2. Methods

2.1. Data and context: the Cardiovascular Health Study

The Cardiovascular Health Study (CHS) is a population-based longitudinal study of 5,888 adults aged 65 and older.
at baseline [9]. Subjects were recruited from a random sample of the Medicare eligibility lists in four U.S. counties, and extensive baseline data were collected for all subjects. After baseline, subjects had an annual clinic visit and provided additional information by mail and telephone. Two cohorts were followed, one with 9 years of follow-up \((n = 5,201)\) and the second (all African American, \(n = 687)\) with 6 years of follow-up. Data collection began in about 1990, and follow-up for longitudinal variables was virtually complete for all surviving subjects in 1999 [10]. At baseline the mean age was 73 (range 65–100); 58% were women, and 84% were White. Morbidity and mortality outcomes were identified through patient or family member self-report, review of hospital and physician records and death certificates, and were adjudicated by a physician review panel [10]. By the year 2000, 658 CHS participants had suffered an incident stroke. We used their information in the 3 years before and after the stroke (about six measures per person) to illustrate the effect of stroke on each variable, while accounting for death.

About 30% of participants died during the first 9 years of follow-up. We examine 11 longitudinal health variables that were used elsewhere [11]. As noted in Table 1, these include measures of function, behaviors, clinical variables or risk factors, some of which seemed like reasonable candidates for transformation and some that did not. The variables also represent a mix of categoric and continuous distributions. The CHS study had little loss to follow-up. We imputed values for the missing data from the known longitudinal data, either by interpolation from the closest known values before and after the missing data or from a person-level regression of the variable on time, different for each person. These methods have been shown to be appropriate for the CHS data [12], and are explained in more detail elsewhere [7,11].

2.2. Analysis

First, we categorized each continuous variable (\(Y\)) into five categories, added a category for death, and created stacked bar graphs over time. (The choice of five categories was arbitrary.) Next, we chose a definition of “healthy” for each variable (see Table 1), and plotted the percentage who were healthy over time (with dead considered as not healthy). In some cases the definition of “healthy” was obvious or traditional (e.g., no days spent in bed); otherwise we chose a threshold that allowed a reasonable number of persons to be classified as “healthy” (e.g., 85% of the subjects could walk 15 feet in less than 10 sec). Finally, we estimated the probability of being healthy 1 year later as a function of \(Y\), using logistic regression. We used three definitions of “healthy”: (1) being healthy on the variable of interest 1 year later (with “no ADLs 1 year later” predicted from ADL now); (2) being in excellent, very good, or good health (E/VG/G) 1 year later (with E/VG/G 1 year later predicted from ADL); or (3) in E/VG/G health measured at the same time (with E/VG/G at one time predicted from ADL at the same time). A transformation including age and sex was also developed, because those variables would usually be available (predict “healthy” 1 year later from ADL, age, ln(age) and sex).

We used the transformed variables to illustrate change before and after stroke in the 11 variables of interest. A comparison group for the stroke patients was constructed by assigning each CHS participant a random date, and treating that as the date of a “comparison event” if he/she was still alive on that date \((N = 4,511)\). Mean ages for the stroke and comparison groups were 75 and 77, respectively, and the percentages male were 42 and 40%. Analysis is primarily descriptive, based on graphs and tables. A sample regression analysis is also presented, which predict future health from baseline characteristics, using the different transformations to account for the deaths. This is explained more fully further down.

3. Findings

The three transformation approaches (add a category, dichotomize, probability of being healthy) have been described elsewhere for EVGGFP and the SF-36 [5]. Here, we illustrate them in detail for ADL, which is the number of difficulties the person reported in response to the following question: “Do you have any difficulty walking around your home? Getting out of bed or a chair? Because of health or physical problems do you have any difficulty or are you unable to … eat, including feeding yourself? Dress yourself? Bathe or shower? Use the toilet, including getting to the toilet?” The variable is coded as the number of difficulties (0–6). In the following we abbreviate “ADL difficulties” as “ADLs.” Findings for the other variables are described briefly here and in more detail in the technical report [8].

3.1. Approach 1. Add a category for death to the longitudinal variable

3.1.1. ADL

Figure 1 is a stacked bar graph of the number of ADLs in the CHS population before and after the stroke. About 75–80% had no ADLs before the stroke, and only a few had four to six ADLs. The percent who were dead after the stroke naturally increased over time. The percent healthy (having no ADLs) was fairly flat in the 3 years before the stroke, dropped by 25–30 percentage points at the time of the stroke, and declined further after the stroke. This drop was caused in part by the deaths and in part by an increase in ADLs. A qualitatively different picture emerges in Fig. 2, which does not account for the deaths. This figure suggests a drop of only 10–20 points at the time of the stroke, followed by a small decline, which understates the devastating cumulative effect of stroke on the cohort.
Figure 2 also suggests that a large number of people had ADLs, even though Fig. 1 shows that relatively few ever had ADLs. Figure 2 refers only to the persons alive at each time, and so does not represent the trajectory of the entire cohort who experienced a stroke.

3.1.2. Other variables

This approach is not limited to categoric variables, because any continuous variable can be categorized. For example, the CESD depressive symptom scale can be divided into five categories, plus death. The stacked bar graphs for the other variables were similar to Fig. 1. More information on variables other than ADL is available in the on-line technical report [8].

3.2. Approach 2. Dichotomize Y as Healthy yes/no, and define death as “not healthy”

3.2.1. ADL

The stacked bar graphs show the distribution of ADL health states over time, but are inconvenient if the goal is to compare the trajectories of different groups. We next transformed ADL, with seven categories plus death, into a binary variable that takes on the value of 100 if the person is healthy (no ADLs) and 0 if not healthy (1 or more ADLs). Death can be considered an extreme form of having ADLs, and so can also be coded as zero. The mean of this new variable is interpreted as the percent of the cohort who were healthy (alive and no ADLs). The lower line in Fig. 3 shows the mean of the new dichotomous variable, for the stroke cohort, and is interpreted as the percent who were healthy (no ADLs) before and after the stroke. Trends are similar to those in Fig. 1.

The decline in health was probably not due entirely to the stroke, because there is also a general increase in disability over time [11]. The percentage healthy for the random-date comparison group is also shown. The comparison group was a little healthier than the stroke group before the event, and declined only slightly over time. The area between the two curves is 0.26 years of healthy life before the stroke and 1.13 years after the stroke. Adjusting crudely for prior differences in the percent healthy, stroke accounted for a loss of about 1.13 – 0.26 = 0.86 years of healthy life in the 3 following years.

3.2.2. Other variables

The definitions of “healthy” for the other variables are given in Table 1. For example, for the CESD, persons with a score of 10 or lower would be healthy (100) and those with a score above 10 or dead would be not healthy (0). All of the variables but systolic blood pressure and number of blocks walked showed an abrupt drop in the percent healthy after the stroke, with the biggest drops for EVGGFP and for being hospital-free (not shown). Most variables declined further after the stroke, with the exception of bed days and hospitalizations, which after an initial decline showed improvement approximately 1 year after stroke. These results are detailed further in the on-line technical report [8].

3.3. Approach 3. Probability of being healthy in the future

The mean of the dichotomized variable is the percent of the people who are “healthy” (percent healthy, or PCTH). PCTH is completely interpretable, but we discarded some...
information in going from the seven original ADL categories (+ death) to two. Further, some interventions such as one that kept people from dying but left them with two ADLs would not show any benefit using PCTH. Finally, some find it unsatisfying to group persons with a few ADLs together with the dead. All of these objections can be satisfied by recoding the variable as the percent probability of being healthy 1 year in the future, conditional on the current health status. The probability that a dead person will be healthy 1 year in the future is zero. This transformation allows the new variable to have a different value for each unique value of the original variable, and also has a unique (and obvious) value for death [5]. The mean of this transformed variable is the percent expected to be healthy 1 year later, or EPCTH. For categoric variables, the probability can be estimated by calculating the percent in each category who were “healthy” 1 year later, with dead set to 0.

3.3.1. Predict probability of no ADLs 1 year in the future from the number of ADLs now

In the complete CHS dataset, 88% of those with no ADLs were disability-free 1 year in the future, and so “no difficulties” is coded as 88; one difficulty is coded as 42; two as 21; three as 14; and four to six difficulties is coded as 7 because only 7% of those with four to six ADLs had no ADLs 1 year later. Dead persons have no chance of being ADL-difficulty-free 1 year in the future, and so are coded as 0. The area under the PCTH curve over time is the years without ADL disability, analogous to disability-adjusted life-years. Similarly, the area under the EPCTH curve is the expected years without ADL disability starting 1 year in the future. The transformed variable can be graphed or used as the dependent variable in regression analyses.

The probability of being healthy may also be estimated using logistic regression, which is illustrated here because of its applicability to the other, continuous variables. For example, for ADLs, we estimated:

\[
\text{logit} \left( \text{No ADLs 1 year later} \right) = 1.97 + 0.55 \times \text{ADLs now} - 4.064 \times \ln(\text{ADLs now} + 1),
\]

or more succinctly = \( a + b \times \text{ADL} + c \times \ln(\text{ADL} + 1) \).

(The fit was not as good without the logarithm term.) The estimated probability of being healthy (no ADLs) 1 year in the future for a given ADL value is then:

\[
\text{Prob}(\text{No ADLs}_{1yr}) = \frac{\exp[a + b \times \text{ADL} + c \times \ln(\text{ADL} + 1)]}{1 + \exp[a + b \times \text{ADL} + c \times \ln(\text{ADL} + 1)]} \times 100.
\]

The resulting probability estimates for 0–6 ADLs, plus death, are: 88, 43, 20, 12, 9, 7, 7, and 0, which are quite close to the values obtained directly that were shown in the previous paragraph. Other worked-out examples of the logistic regression transformation are available [5]. Regression coefficients for the other variables are in Appendix Table A1.

3.3.2. Predict probability of no ADLs 1 year in the future from number of ADLs, age, and sex

It may be possible to improve the estimate of the probability of being healthy by adding other commonly available variables, such as age and sex, to the logistic regression equations. For example, we estimated:

\[
\begin{align*}
\text{logit} \left( \text{no ADLs 1 year in the future} \right) &= -3.81 + 0.562 \times \text{ADLs now} \\
&\quad - 3.94 \times \ln(\text{ADLs now} + 1) + 0.126 \times \text{male} \\
&\quad - 0.117 \times \text{age now} + 3.39 \times \ln(\text{age}).
\end{align*}
\]

Appendix Table A2 provides the necessary regression coefficients for the other variables. Preliminary work showed that this transformation did improve longitudinal graphs when a long time horizon was considered, because it allowed for aging (not shown).

3.3.3. Predict future health using a different definition of healthy

We have defined “healthy” as a threshold on the scale of interest (e.g., no ADLs or no bed days). For the physical component score (PCS) of the SF-36, we found that defining “healthy” as being in excellent, very good, or good health, rather than as a threshold of the PCS itself, gave transformed variables with a less bimodal distribution and better ability to discriminate between groups with different health problems, suggesting that external definitions of being healthy should be considered [5]. Here, we defined healthy as E/VG/G for all variables, and estimated:

\[
\begin{align*}
\text{logit} \left( \text{E/VG/G health 1 year later} \right) &= 1.19 + 0.22 \times \text{ADLs now} \\
&\quad - 1.96 \times \ln(\text{ADLs now} + 1).
\end{align*}
\]

The estimated probabilities of being healthy (E/VG/G) 1 year in the future for the ADL categories are then: 77, 51, 37, 30, 25, 23, 21, 0, which decrease less abruptly than in the previous transformations. In particular, the distance between “0” and “1” ADLs is only 77 – 51 = 26 percentage points using this definition of healthy, as opposed to 45 points under the previous transformation. The mean of the new transformed variable would be interpreted as the expected percent healthy (defined here as E/VG/G) 1 year later as a function of ADL today. Regression coefficients for all variables are in columns 4–6 of Appendix Table 1, labeled as transformation #2.

Similarly, we can estimate probability of being healthy (E/VG/G) at the same time as the original variable was measured, which avoids the time shift in the other probability of being healthy transformations. The equation is:

\[
\begin{align*}
\text{logit} \left( \text{E/VG/G now} \right) &= 1.39 + 0.239 \times \text{ADLs now} \\
&\quad - 2.00 \times \ln(\text{ADLs now} + 1).
\end{align*}
\]
In the technical report, the mean values of all the transformed variables are shown to be about 75 3 years before the stroke, meaning that persons were in a sense equally healthy on all 11 variables then [8]. This transformation, in a sense, standardizes across different variables, which is desirable if the effect of stroke on a variety of variables is to be compared. Regression coefficients for this transformation are in the last three columns of Appendix Table A1, labeled as transformation #3. (The probability of being E/VG/G now given EVGGFP now is equivalent to the dichotomous transformation, but here is calculated using logistic regression.)

Figure 4 shows the mean trajectory before and after the stroke using the five different transformations of ADL. The curves differ a little, particularly before the stroke. Before the stroke, the percent healthy (with no ADLs now, line 1) is higher than the others because it refers to health now rather than a year later. The logistic regression transformation accounting for age and sex (line 3) gives lower values than the transformation that does not (line 2), but the differences are small. The curves for the probability of being E/VG/G are lower than the others because in CHS the percentage E/VG/G was lower than the percentage with no ADLs (lines 4 and 5). After the stroke, the lines are closer together, with Prob(E/VG/G now) the highest and % with no ADLs the lowest. Interestingly, the greatest change associated with the stroke was shown by the simple dichotomous variable, percent with no ADLs. The other transformations, which used all of the ADL information rather than simply dichotomizing, did not show as much change. More work is needed to compare the properties of these transformations.

Figure 5 shows the mean value over time for six of the variables, each transformed to the probability of being in excellent, very good, or good health now. The height of the lines after the stroke is in the same order as the legend, with CESD and mini-mental state examination highest and very similar, with ADL, timed walk, and IADL intermediate and quite similar, and with EVGGFP the lowest. The EVGGFP value plotted is equivalent to the dichotomized value. EVGGFP shows the largest change at the time of the stroke. Because all of the variables have similar transformed values 3 years before the stroke, results from this transformation are in a sense standardized across the measures. The effect on hospitalization was the largest of all (not shown here) [8]. Additional research is needed to help users choose among the definitions of healthy.

3.4. Example: effect of stroke on ADLs 1 year after the event

Recent studies have used these methods to examine years of healthy life as a function of smoking [13], of body mass index [14], of drinking [15], and of using Adult Day Health Care [16]. Here we present a brief regression example to illustrate the use and interpretation of the transformed variables. The goal is to predict health (defined by ADLs) 12–23 months following stroke. We transformed the ADL value 12–23 months after stroke using the various methods, and regressed each transformed variable on group membership (stroke/comparison), then on group and ADL before the event, and finally on group, prior ADL, age, and sex.

The regression results are shown in Table 2. The first line shows the results of predicting whether the person would have no ADLs 12–23 months after the event, and this line includes only the persons who were alive 12–23 months after the event as in Fig. 2. The unadjusted difference between controls and strokes is 26 percentage points, which drops to 20.1 points after adjustment for prior ADL, age, and sex. The second line is for the same dependent variable, but includes the dead as “not healthy,” and the difference in the % healthy in the two groups is much larger (41 percentage points unadjusted, dropping to 32 percentage points with full adjustment). Lines 3 and 4 show results predicting the probability of no ADLs 1 year later (in this case, 24–35 months after the event), where line 4 included age and sex as part of the transformation. The effect on line 4 is slightly larger than that on line 3 until age
and sex are controlled in the regression, after which the direction changes. The effect of including age and sex in the transformation is unclear. Line 5 predicts the probability of being E/VG/G one year later (24–35 months after the event), and line 6 the probability of being E/VG/G “now” (12–23 months after the event). These two lines are very similar. Table 2 thus shows that including deaths had a large effect on the outcome (line 1 vs 2), and that results from different transformations would have different interpretations, since they refer to different dependent variables. In every case, the effect of stroke was highly significant ($P < .0005$). Lines 2–6 all represent prospective analyses, while line 1 applies only to persons known to be alive 12–23 months after the event, and so cannot be used prospectively.

### 4. Discussion

We proposed and operationalized several approaches for incorporating death into longitudinal health variables. We then used the transformed variables to compare the effect of a stroke on 11 different health-related variables, and also demonstrated their use in a regression analysis. We next discuss which variables appear to “make sense” when they are transformed, which transformations to choose, other approaches for incorporating death, and limitations of this research.

#### 4.1. Can all variables can be transformed to include death?

It is possible to add a category for death to any variable (e.g., number of books read lately), and also in many cases to know what value to assign the dead on this variable (no books). Here, however, we wish to consider health-related variables whose underlying or latent construct may sensibly include death. Death is probably a part of the construct for measures of function and health status, and perhaps for health-related mobility, but probably not for risk factors such as weight, blood pressure, or the number of blocks walked. Table 1 notes the category in which we consider each variable.

With respect to dichotomization, it is always possible to separate people into two groups, with and without a particular characteristic. The question here is whether we feel comfortable classifying dead as “not healthy” on the variable of interest. Again, this seems reasonable for measures of health status and function, but perhaps less so for mobility and risk factors. We are particularly uncomfortable with considering the dead to have “unhealthy weight” or “high blood pressure” or “insufficient walking.” All of the variables except systolic blood pressure have been shown to become worse close to death, independent of age, which suggests that death may be part of their construct [11]. Investigators must decide whether the construct underlying their variable of interest includes death.

#### 4.2. Which transformation should we choose?

Adding a category for death can yield useful graphs and some ordinal analyses. Incorporating a value for death (PCTH or EPCTH) allows for more flexible analyses. If the goal is to compare groups, or to conduct regression analyses with the health measure as the dependent variable, a definition of “healthy” must be chosen. PCTH can then

---

Table 2

<table>
<thead>
<tr>
<th>Line</th>
<th>Dependent variable (transformation No.)</th>
<th>Interpretation of regression coefficient is the difference in:</th>
<th>Nothing</th>
<th>ADL before event</th>
<th>ADL before event, age, and sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ADL no/yes (dichotomous), dead excluded</td>
<td>% with no ADLs 12–23 months after event given that they are alive 12–23 months after event.</td>
<td>$-26.0$</td>
<td>$-21.3$</td>
<td>$-20.1$</td>
</tr>
<tr>
<td>2</td>
<td>ADL no/yes (dichotomous)</td>
<td>% with no ADLs 12–23 months after event</td>
<td>$-41.0$</td>
<td>$-36.2$</td>
<td>$-32.0$</td>
</tr>
<tr>
<td>3</td>
<td>P(no ADLs 1 year later</td>
<td>ADL) (#1)</td>
<td>% predicted to have no ADLs 24–35 months after event</td>
<td>$-17.9$</td>
<td>$-14.7$</td>
</tr>
<tr>
<td>4</td>
<td>P(no ADLs 1 year later</td>
<td>ADL, age, sex) (#4)</td>
<td>% predicted to have no ADLs 24–35 months after event</td>
<td>$-18.7$</td>
<td>$-15.5$</td>
</tr>
<tr>
<td>5</td>
<td>P(E/VG/G 1 yr later</td>
<td>ADL) (#2)</td>
<td>% predicted to be in excellent, very good, or good health 24–35 months after event</td>
<td>$-11.2$</td>
<td>$-9.3$</td>
</tr>
<tr>
<td>6</td>
<td>P(E/VG/G now</td>
<td>ADL) (#3)</td>
<td>% predicted to be in excellent, very good, or good health 12–23 months after event</td>
<td>$-11.0$</td>
<td>$-9.1$</td>
</tr>
</tbody>
</table>

* Coefficient of “stroke” in regression of dependent variable on age, sex, ADL before the stroke, and an indicator variable for Stroke vs. Comparison groups.
be calculated directly, or EPCTH can be calculated either from the regression equations in the Appendix tables or from equations developed elsewhere. PCTH has the advantage of simplicity, but EPCTH has the theoretical advantages of not discarding information and of allowing a distinct value for the dead, and so may permit more powerful analyses than PCTH [17]. In the stroke example, however, the biggest change was seen for PCTH, suggesting that using a more detailed variable will not necessarily improve power. Using an external definition of healthy (e.g., E/VG/G for all variables) eliminates the large difference between the recoded values just above and just below the “healthy” thresholds that were found here for ADLs and have also been found for the EVGGFP and the SF-36 scales [5]. Using “E/VG/G now” also removes the awkwardness of the shift in time to “1 year later.” This may make the external definition worth consideration, although its interpretation is a little less straightforward.

In practice, we usually present stacked bar graphs or graphs of mean PCTH, after checking that the results do not change when some form of EPCTH is used. It is usually advisable to impute values for data missing for other reasons than death, because values for the dead will never be missing, and the dead might go from having no influence at all to having too much influence [2].

4.3. Other approaches for incorporating death

The alternative approaches mentioned in the Introduction basically ignored deaths. One way to examine the effect of deaths on a particular analysis is to assign death some extreme value on the variable of interest, such as the worst observed value or the worst possible value, and repeat the analysis. Stability of the results indicates that the particular finding is not sensitive to what is done with the deaths. The new variable cannot be interpreted in the usual way, however, because the value assigned to death is arbitrary, and because the published values, the psychometric characteristics, and the meaning of, say, a five-point change, were all based only on living persons. Our transformation approach also yields a new variable, but the beauty is that it can always be interpreted as the % healthy (or the expected % healthy 1 year later). The definition of being healthy can vary according to the variable of interest. A related analytic approach by Bozette et al. [18], that considers many definitions of being healthy simultaneously, is not addressed specifically here.

4.4. Limitations

The transformation coefficients presented here were calculated from a single dataset of older adults. Other data for older adults has provided similar transformations for EVGGFP, and transformations of the SF-36 were shown to hold true in a middle-aged population even though they were developed in a population of aging (mostly male) veterans [2,5]. We believe that the transformations developed here can reasonably be used in other middle-aged and older populations, but that their applicability to younger persons has yet to be verified. Imputation of the data missing just after the stroke may have dampened slightly the estimated drop at the time of the stroke. In data sets with fewer deaths and more loss to follow-up, death may not be the most important threat to validity, and methods presented here may be less salient. This methodology may be more useful for health policy research and a public health perspective than for individual patient counseling and education.

4.5. Conclusion

Analyses and graphs of available longitudinal data may give misleading pictures of the trend over time for a cohort if some people die during the study. Incorporating death into the variable of interest can provide a better picture of the cohort’s trajectory over time, or before and after an event such as a stroke. Transformation of health-related variables to interpretable measures that naturally include a value for death is reasonable, and should be considered at least for supporting or sensitivity analyses in longitudinal studies where deaths occur.

Acknowledgments

The research reported in this article was supported by contracts N01-HC-85079 through N01-HC-85086, N01-HC-35129, and N01 HC-15103 from the National Heart, Lung, and Blood Institute. Wake Forest University School of Medicine: Gregory L. Burke, MD. Wake Forest University—ECG Reading Center: Pentti M. Rautaharju, MD, PhD. University of California, Davis: John Robbins, MD, MHS. The Johns Hopkins University: Linda P. Fried, MD, MPCTH. The Johns Hopkins University—Magnetic Resonance Imaging Reading Center: Nick Bryan, MD, PhD, Norman J. Beauchamp, MD, University of Pittsburgh: Lewis H. Kuller, MD, DrPCTH. University of California, Irvine—Echocardiography Reading Center (baseline): Julius M. Gardin, MD. Georgetown Medical Center—Echocardiography Reading Center (follow-up): John S. Gottdiener, MD. New England Medical Center, Boston—Ultrasound Reading Center: Daniel H. O’Leary, MD. University of Vermont—Central Blood Analysis Laboratory: Russell P. Tracy, PhD. University of Arizona, Tucson—Pulmonary Reading Center: Paul Enright, MD. Retinal Reading Center—University of Wisconsin: Ronald Klein, MD. University of Washington—Coordinating Center: Richard A. Kronmal, PhD. NHLBI Project Office: Jean Olson, MD, MPH.
### Appendix Table A1

Logistic regression coefficients for estimating the probability of being healthy 1 year in the future, for 3 definitions of “healthy” for 11 variables

<table>
<thead>
<tr>
<th>Definition of “healthy”:</th>
<th>(1) Healthy 1 year later (see section 3.3.1)</th>
<th>(2) E/VG/G 1 year later (see section 3.3.3)</th>
<th>(3) E/VG/G NOW (see section 3.3.3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>a</td>
<td>b</td>
<td>c</td>
</tr>
<tr>
<td>Activities of daily living</td>
<td>1.97</td>
<td>0.55</td>
<td>−4.06</td>
</tr>
<tr>
<td>Bed days</td>
<td>2.55</td>
<td>0.113</td>
<td>−1.73</td>
</tr>
<tr>
<td>Blocks(^6)</td>
<td>−2.47</td>
<td>3.29</td>
<td>−1.48</td>
</tr>
<tr>
<td>Body mass index</td>
<td>−352.7</td>
<td>−6.06</td>
<td>155.</td>
</tr>
<tr>
<td>Depression</td>
<td>3.25</td>
<td>−0.254</td>
<td>−0.024</td>
</tr>
<tr>
<td>EVGGFP</td>
<td>−7.448</td>
<td>0.139</td>
<td>5.96</td>
</tr>
<tr>
<td>(5 = excellent, 4 is very good, ..., 1 = poor)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital</td>
<td>2.009</td>
<td>−0.011</td>
<td>n/a</td>
</tr>
<tr>
<td>Instrumental activities of daily living</td>
<td>1.51</td>
<td>0.56</td>
<td>−3.90</td>
</tr>
<tr>
<td>Modified mini-mental state exam(^5)</td>
<td>−19.709</td>
<td>236</td>
<td>372</td>
</tr>
<tr>
<td>Systolic blood pressure</td>
<td>31.198</td>
<td>−0.042</td>
<td>−5.22</td>
</tr>
<tr>
<td>Timed walk</td>
<td>13.36</td>
<td>0.126</td>
<td>−5.90</td>
</tr>
</tbody>
</table>

\(^a\) Logit (healthy\(_1\)) = a + b(variable\(_0\)) + c ln(variable\(_0\) + 1).

\(^b\) Uses Y = lg10(bk) instead of bk and ln(lg10(bk+1)) instead of ln(bk+1).

\(^c\) Uses Y = ln (101 − MM SCORE) instead of ln(mmscore+1).

### Appendix Table A2

Logistic regression coefficients for estimating the probability of being healthy in 1 year in the future conditional on current health, age, and sex for 11 variables [Regression coefficients\(^a\) for transformation (4) (see section 3.3.2)]

<table>
<thead>
<tr>
<th>Variable</th>
<th>constant (a)</th>
<th>Y (b)</th>
<th>ln(Y) (c)</th>
<th>Male (d)</th>
<th>Age (e)</th>
<th>ln(age) (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activities of daily living</td>
<td>−3.81</td>
<td>0.562</td>
<td>−3.94</td>
<td>0.126</td>
<td>−0.117</td>
<td>3.391</td>
</tr>
<tr>
<td>Bed days</td>
<td>−38.60</td>
<td>0.136</td>
<td>−1.79</td>
<td>−0.013</td>
<td>−0.242</td>
<td>13.780</td>
</tr>
<tr>
<td>Blocks(^6)</td>
<td>0.379</td>
<td>3.20</td>
<td>−1.56</td>
<td>0.321</td>
<td>−0.064</td>
<td>0.469</td>
</tr>
<tr>
<td>Body mass index</td>
<td>−492.2</td>
<td>−6.17</td>
<td>158.6</td>
<td>−0.179</td>
<td>−0.574</td>
<td>40.975</td>
</tr>
<tr>
<td>Depression</td>
<td>−51.68</td>
<td>−0.257</td>
<td>0.014</td>
<td>0.024</td>
<td>−0.270</td>
<td>17.423</td>
</tr>
<tr>
<td>EVGGFP</td>
<td>−2.13</td>
<td>0.154</td>
<td>5.84</td>
<td>−0.066</td>
<td>−0.047</td>
<td>−0.359</td>
</tr>
<tr>
<td>(5 = excellent, 4 is very good, ..., 1 = poor)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital</td>
<td>−18.31</td>
<td>−0.010</td>
<td>n/a</td>
<td>−0.256</td>
<td>−0.141</td>
<td>7.204</td>
</tr>
<tr>
<td>Instrumental activities of daily living</td>
<td>−10.41</td>
<td>0.63</td>
<td>−3.91</td>
<td>0.223</td>
<td>−0.125</td>
<td>4.929</td>
</tr>
<tr>
<td>Modified mini-mental state exam(^5)</td>
<td>−20.95</td>
<td>0.229</td>
<td>0.411</td>
<td>−0.178</td>
<td>−0.090</td>
<td>2.015</td>
</tr>
<tr>
<td>Systolic blood pressure</td>
<td>25.95</td>
<td>−0.038</td>
<td>−5.56</td>
<td>0.060</td>
<td>−0.063</td>
<td>2.601</td>
</tr>
<tr>
<td>Timed walk</td>
<td>45.21</td>
<td>0.096</td>
<td>−5.29</td>
<td>−0.333</td>
<td>0.029</td>
<td>−8.043</td>
</tr>
</tbody>
</table>

\(^a\) Logit (healthy\(_1\)) = a + b(variable\(_0\)) + c ln(variable\(_0\) + 1) + d * male + e * age + f * ln(age).

\(^b\) Uses Y = lg10(bk) instead of bk and ln(lg10(bk+1)) instead of ln(bk+1).

\(^c\) Uses Y = ln (101 − MM SCORE) instead of ln(mmscore+1).

### References


