4-29-2016

Methods for Dealing with Death and Missing Data, and for Standardizing Different Health Variables in Longitudinal Datasets: The Cardiovascular Health Study

Paula Diehr
University of Washington, pdiehr@u.washington.edu

[Tech report Missing_death_standardization_postadjustment_longitudinal_13.doc]

Methods for dealing with death and missing data, and for standardizing different health variables in longitudinal datasets: the Cardiovascular Health Study

Paula Diehr

Abstract

Longitudinal studies of older adults usually need to account for deaths and missing data. The study databases often include multiple health-related variables, whose trends over time are hard to compare because they were measured on different scales. Here we present a unified approach to these three problems that was developed and used in the Cardiovascular Health Study. Data were first transformed to a new scale that had integer/ratio properties, and on which "dead" logically takes the value zero. Missing data were then imputed on this new scale, using each person's own data over time. Imputation could thus be informed by impending death. The new transformed and imputed variable has a value for every person at every potential time, accounts for death, and can also be considered as a measure of "standardized health" that permits comparison of variables that were originally measured on different scales. The imputed variable can also be transformed back to the original scale, which differs from the original data in that missing values have been imputed. Imputed values near death required an addition "post-adjustment". One approach is shown in sections 5 and 6. In the resulting tidy dataset, every observation is labeled as to whether it was observed, imputed (and how), or the person was dead at the time. The resulting "tidy" dataset can be considered complete, but is flexible enough to permit analysts to handle missing data and deaths in other ways. This approach may be useful for other longitudinal studies as well as for the Cardiovascular Health Study.

Methods for dealing with death and missing data, and for standardizing different health variables in longitudinal datasets: the Cardiovascular Health Study

Table of Contents

Methods for dealing with death and missing data, and for standardizing different health variables in longitudinal datasets: the Cardiovascular Health Study

Introduction: Death, missingness, and multiple measures

They say that nothing is certain but death and taxes. Longitudinal datasets don't pay taxes, but they are often challenged by deaths and by other missing data. Another issue arises when there are several different longitudinal variables, all measured on different scales, where the interest is in comparing time trends for the different variables. Our goal is to present a unified approach to these 3 problems: death, missingness, and comparison of multiple variables. As part of the first two goals we created a rectangular dataset, with K records per person, where K is the maximum number of periods in which the person could potentially have contributed data (although she may in fact have died or gone missing). Even at the times after the person has died, there is still a record for that person, with an indication that the person is dead. For each observation of each variable, we include an auxiliary variable called "status" which indicates whether this value was observed, the person was dead, or how it was imputed. The analyst may choose to use all the imputed values, or may easily eliminate or re-impute some of the values, since they are clearly labeled.

We illustrate this approach using data from the first 10 years (1990-1999) of the Cardiovascular Health Study (CHS),[1] based on a sample of Medicare enrollees, who were followed annually from 1990 to 1999. (Baseline data were actually collected in 1989-1990, but we refer to baseline year as 1990 for simplicity. We occasionally refer to "CHS year" in which baseline for cohort 1 is year 2 and baseline for cohort 2 is year 5). A second cohort, all African American, was followed from 1993 to 1999. The "tidy" dataset (a more pronounceable version of tdie) thus has 10 records for each person in cohort 1, and 7 for those in cohort 2. Fortunately, deaths were completely ascertained in CHS, and the amount of missing data and persons lost to follow-up was small. (Sections 5 and 6 deal with 3 variables that had 20-year follow-up, with peculiar missing patterns that required post-adjustment).

The primary goal of this report is to provide documentation for users of the longitudinal CHS data. Although it is unusual to have such a long series of measures, we hope that some of these methods will also be useful to researchers using other datasets. Sections 1-3 deal with deaths and missing data. Section 4 deals with standardization of the various health measures so they may be compared. Section 5 refers to problems for the 20-year datasets where no data were collected for 4 (or 7) interim years, and post-adjustment was required, primarily for persons who died in this interim period and whose imputed values were noted to be too pessimistsic. Section 6 uses the approach in section 5 to post-adjust all of the data missing in the period just before death.

Unified Approach

The unified approach to the 3 problems is described here. For each variable "X" that was potentially measured 10 times (cohort 1) or 7 times (cohort 2), we created a series of auxiliary variables, as explained in Table 1. We refer to the resulting dataset as a "tidy" dataset, which refers to the subscripts "tdie".

	Table 1
	Definitions of Auxiliary Variables in Unified Approach
Variable	Definition
X	Longitudinal Health Variable, such as instrumental activities of daily living (IADL), which has values from 0 to 6 difficulties.
X_t	A transformation of X that is on an integer/ratio scale, and for which the value for death is logically 0.
X_td	X_t with the deaths set to zero.
X_tdi	X_td with missing data imputed using linear interpolation over time. (X_tdi is thus complete for persons who died)
X_tdie	X_tdi with any terminal missing data imputed from the average of last available observation of X and from self-rated health at that time (because SRH was measured for >10 yrs)
(X_tdi_pa)	(X_tdi post-adjusted in 20-year followup. See section 5)
(X_tdie_pa2)	(X_tdiepa with the remaining 20-year data missing just before death post-adjusted. See section 6).
X_back	X_tdie transformed back to the original scale
Status_X	A marker for whether X_tdie is observed, missing because dead, imputed using interpolation, or extrapolation, or a missing baseline measure imputed as the next observation

	carried back (NOCB)
Standardized X	X_tdie, relabeled as “standardized X”

0 Reference dataset

We will refer to a “reference dataset”, which is used to assist with the problems of death, missingness, and different scales. It could in theory be any large datset, not necessarily the same as the dataset that will be analyzed. Here, the reference dataset is all of the available longitudinal data collected in CHS, from 1990 to 1999, with no distinction as to the person’s age, sex, or which year it was collected. (Although CHS enrolled both men and women, we often use the female gender to replace unpleasant grammatical constructions -- she is used instead of “he/she” and her instead of “his/her” or “their” when referring to an individual. All examples include men and women.) Everyone in cohort 1 could have contributed 10 records, and everyone in cohort 2 could have contributed 7 records. Much of this discussion will deal with self-rated health – is your health excellent, very good, good, fair, or poor? – usually abbreviated as EVGGFP. Unlike the other variables, EVGGFP was collected every semester (6 months) and is still being collected at this date, along with mortality and a few other variables.

1 Transform X to a new, integer/ratio scale (X_t)

The first step is to transform the variable "X" from its original scale (which is often ordinal) to a different scale (X_t) that is interpretable, has an integer/ratio property, and where death would have a natural value. One approach we considered was to replace each observed value with the probability that a person with this value would be "healthy" in the following year. This probability was estimated from the reference dataset. That is, for t from 1 to 9 years, we dichotomized the data in year t to "healthy/sick", and then used logistic regression to predict the probability that the person would be "healthy" in each year (t+1) based on their observed value in year t. Specifically, we dichotomized self-rated health in year t+1 as "excellent/very good/good" = 1, and "fair/poor" = 0. (Later we refer to these to these combined categories as EVGG and FP, or as healthy and sick). Earlier research using several different reference datasets [2] found that persons whose self-rated health was excellent in year t had about a 95% chance of being healthy in year t+1, persons whose self-rated health was very good in year t had about a 90% chance of being healthy in year t+1…., and that persons whose self-rated health was poor had about a 15% chance of being healthy. For that reason, we recommended recoding excellent/very good/ good/ fair/ poor to 95/90/80/30/15.[2]

The large difference between the values for good (80) and fair (30) is due in large part to the fact that "healthy in year 2" was dichotomized between good and fair. If we had dichotomized at some other point, say between very good and good, a different large gap would have occurred (in this case between very good and good). Some empirical work suggested that, where possible, it was better to define "healthy" using some other variable than the one being transformed. [3] In a different dataset, we transformed the SF-36 scales according to the probability of being in excellent/very good/ good health (EVGG), and also to the probability of having a "healthy" SF-36 score in the following period. [4] The former method had fewer large gaps between the values, because it was not based on dichotomizing the SF-36 score itself. In addition, we did not need to estimate the probability of being healthy in the following year, but could estimate the probability of being healthy in the same year, which simplified the interpretation. This variant was used for the CHS variables, with the exception of self-rated health (EVGGFP).

Specifically, we used logistic regression to predict EVGG from the logarithm of "X+ 1" (e.g.,ln(1+ IADL))in the same year. The logarithms were used to minimize the effect of outliers. We added 1 before calculating the logarithm because 0 was a valid value for many of the variables. One variable, 3MSE, was negatively skewed and we instead used log(101-3MSE) in the standardization regression.

To illustrate these calculations we will use IADL, which refers to the number of instrumental activities of daily living (heavy or light housework, shopping, meal preparation, money management, or telephoning) which the person has some difficulty in performing. IADL takes on values from 0 (no difficulties) to 6. If we transform IADL as to the probability of a "healthy" IADL (having no IADL difficulties) in the following year, the codes are as follows: 0/1/2/3/4/5/6 / dead ➔ 82/35/15/9/7/8/6/0. Notice the big gap between 82 and 35, due to our definition of "healthy IADL" as having 0 IADL difficulties.

Here, however, are the probabilities of being EVGG in the same year for different IADL values: 0/1/2/3/4/5/6 / dead ➔ 84/64/47/39/35/31/27/0. Thus, transformation using the probability of being EVGG gives a more uniform set of values with no large gaps. (When we repeated the transformation regression including the imputed IADL data the means were 85/61/41/34/31/26/20/0, which is the version used in some places in this documentation for convenience.) We will use this kind of transformation for every variable but EVGGFP itself, which was transformed to 95/90/80/30/15/0, as noted before. (We expect transformed EVGGFP (probability of being healthy next year) to be a little lower on average than the other variables (probability of being healthy this year) because a person has some probability of being dead in the following year.)

We shall refer to a variable "X" that is recoded in this way as X_t, where the "t" stands for being transformed to the "probability of being healthy" scale. (We often multiply the X_t by 100, to facilitate discussing it in terms of percentages points). X_t has an interpretable value (the (%) probability of being EVGG conditional on X). In addition, a (say) 5-point change in the scale has the same meaning (5 percentage points change in the probability of being healthy) everywhere in the scale, and there is a true 0 (0 probability of being healthy). This means that the new variable is on an integer/ratio

scale. This property means that it is "proper" to calculate means and other summary statistics that might have been questionable on the original (often ordinal) scale.

Table 2 shows information for person A. The first column shows that he had 0 IADL difficulties in 1990, 1991, 1994, and 1995. He had one difficulty in 1992. Data were missing in 1993 and 1996. He was dead in 1997. The observed values are shown in column 1. IADL_t, the probability of being EVGG for this number of IADL difficulties, is either 84 (no IADL difficulties) or 64 (1 IADL difficulty).

Table 2
Person A missing and death by time

		IADL (observed)	IADL_T	IADL_TD	IADL_TDI	IADL_TDIE	IADL-back	E/VG/G_tdie
Calendar Year	1990	0	84	84	84	84	0	80
	1991	0	84	84	84	84	0	80
	1992	1	64	64	64	64	1	85
	1993	.	.	.	74	74	0	48
	1994	0	84	84	84	84	0	80
	1995	0	84	84	84	84	0	80
	1996	.	.	.	42	42	2	80
	1997	.	.	0	0	0	-12	0
	1998	.	.	0	0	0	-12	0
	1999	.	.	0	0	0	-12	0

prob(e/vg/g this year) -12 means dead

2 Add a value for Death (X_td)

When deaths occur, the analyst must think carefully about how to address them in the analysis, as different approaches can yield profoundly different results. [5] For example, the subset of persons that had the most deaths could seem to have better outcomes than other subsets, because its sickest cases were removed by death. While there are many approaches for handling deaths at the time of analysis, our goal here is to provide a dataset with a reasonable value for death. Since the deaths are clearly identified in Status_X, they may be handled in different ways at the time of analysis, if desired. (The labels X_Status and Status_X are sometimes used interchangeably).

Our goal was to create a new variable that had a reasonable value for dead. This approach is referred to elsewhere as the "joint model". [5] Since a person who is dead is not healthy now, and has no probability of being healthy next year, the natural value to assign to X_t for dead is 0, which is what was done. The new variable is referred to as X_td, which stands for X, transformed and with deaths set to zero. Table 2 shows the transformations for person A. IADL_td is set to 0 for the three years when he was dead.

The assignment of 0 as the value for death will always be at least speciously accurate, since a dead person has no probability of doing or being anything. More seriously, the approach has the effect of conceptualizing the underlying construct as having dead as the worst possible value of the construct. We feel that this is reasonable for most measures of health, quality of life, and function; for example, the worst level of function is being dead. Some have felt that this might not be appropriate for measures of mental health (e.g., should we think of death as extreme depression, or alternatively does death cure depression?). The reasonableness of the approach to death is probably context-specific. For example, suppose we rated piano playing. If we conceptualize piano playing as a measure of physical dexterity, it may be reasonable to consider dead as an extremely low ability to play the piano. Alternatively, if piano playing is conceptualized as a measure of musicality, it is probably not appropriate to think of death as being extremely unmusical, and deaths will need to be handled in some other way. Because X_t usually declines near to death, it may not matter whether it is very low after death, but this should be considered carefully for each analysis.

3 Missingness (X_tdi, X_tdie)

There are many approaches to handling missing data at the time of analysis [5 Kurland] [6], which are not reviewed here. Our goal is to create a "complete" dataset that does something reasonable about missing data. One study of four CHS variables found that estimating a person's missing data from her own available longitudinal data had the best performance of the methods considered.[7] We suggest that, rather than imputing missing data from available X data on the original scale, that X_td should be the basis for imputation, because X_td is on an integer/ratio scale, making it "more appropriate" to calculate means and conduct regression analyses. X_td also has a value for dead, meaning that data missing just before death will be imputed using the information about impending death.

A recent approach for imputation is "multiple-imputation", in which the data are imputed multiple times and the comparison of parameter estimates from the different imputations provide information about the variability of the estimates. We have not found such an approach convenient for longitudinal data[16], although an in-depth study may find otherwise. We noted that the best predictors for a person's X at time t may be that same person's values of X at different times.[7] These predictors tend to be highly correlated with one another, which kept the MI software we were using from converging. For this reason, we used a different and more straightforward approach.

Our goal is to impute a person's missing data from her available data, while using the time of death to inform this imputation. One possible approach is to regress the person's available X_td data on time, and to use the regression equation to predict values for times when the person's data were missing.[8] (For dead persons, we found it best to use only one or two of the 0's after death in the regression calculation). We did not use the regression imputation approach for CHS, but rather used linear interpolation and extrapolation of the person's own observed data.

3.1 Interpolation (X_tdi)

In the CHS data we imputed the missing data using linear interpolation of the person's own X_td over time. This simple approach is (probably) locally optimal under

some assumptions. We refer to variables that are transformed, have death set to zero, and having missing data imputed within the range of the available data (that is, by interpolation) as X_tdi. Because dead has a value, missing data for every person who died can be completely filled in by interpolation. (Any terminal missingness for persons who were still alive at the end of the study, however, still needs to be imputed, as is explained in section 3.2.)

Table 2, in the column for IADL_tdi, shows that the two missing values for person A were imputed (IADL_tdi) as 74 (1993) and 42 (1996). The missing value in 1993 was imputed as the mean of the value in 1992 (64) and 1994 (84). For 1996, the imputed value was 42, the mean of 84 and 0. All missing data were thus imputed by interpolation because person A died during followup. (Sections 5 and 6 describe an additional step (post-adjustment) used for the 3 variables with 20-year follow-up).

3.2 Extrapolation (X_tdie)

Often there is monotone missingness, in which all of the values after a certain time are missing but the person is known to be alive. Last-observation-carried-forward (LOCF) is often used but may be risky, especially for older adults, where missingness is likely to be associated with worse health (i.e., informative). In one sensitivity analysis that considered different approaches for the terminal missingness, we found that 3 of the 4 approaches yielded the same analytic results, but that use of LOCF changed the study findings slightly. [8]

For the CHS longitudinal data, we used a variant of LOCF when the person did not die. CHS was fortunate to have one variable (self-rated-health, EVGGFP) that was measured for a much longer time than the others (1990 to present), and measured more frequently than the others (every semester). For one study, we calculated X_td for all the variables to be analyzed, according to the probability of being in excellent/very good/good health. [9] We then used the mean of the LOCF estimate of X_td (call it X_locf) and the value of self-rated health (EVGGFP_tdie) at the same time, as the estimate for X_td, for values missing at the end of the sequence for persons who were still alive. This was appropriate since both EVGGFP_tdie and X_td were on the same scale (probability of being EVGG). It also incorporated information about health and

death that occurred after 1999, because EVGGFP was measured for a longer time. We chose to average in EVGGFP_tdie only if EVGGFP_tdie was lower than the X_locf, because our main concern was that using LOCF alone could cause the person to appear too healthy. In addition, the Engels study suggested that most imputed data were optimistic, on average.[7] Different choices may be made at the time of analysis. We refer to the version of X in which X is transformed, death is added, data missing between two available time points were interpolated, and monotone missing data for survivors were extrapolated as X_tdie. An example is shown in Table 3 for person B, who lived throughout the study but had one missing observation at the end, in 1999.

Table 3
Person B missing and death by time

		IADL (observed)	IADL_T	IADL_TD	IADL_TDI	IADL_TDIE	IADL-back	E/VG/G_tdie
Calendar Year	1990	0	84	84	84	84	0	95
	1991	0	84	84	84	84	0	90
	1992	0	84	84	84	84	0	88
	1993	0	84	84	84	84	0	90
	1994	.	.	.	84	84	0	85
	1995	0	84	84	84	84	0	95
	1996	.	.	.	84	84	0	95
	1997	0	84	84	84	84	0	90
	1998	0	84	84	84	84	0	30
	1999	.	.	.	.	54	2	25

prob(e/vg/g this year) -12 means dead

Table 3 is similar to Table 2, but is for a person whose final observation (1999) was missing, meaning that interpolation could not be used to impute that point. The last observed value was 84 in 1998, and the EVGG_tdie value in 1999 was 25. Because 25<84, we averaged the two. The mean of 84 and 25 is 54.5 (rounded to 54 in Table 3). Thus, IADL_tdie = (84+25)/2. (Why did EVGGFP have a 1999 value while IADL did not? In fact, EVGGFP was also missing in 1999, but it could be imputed by interpolation from known values in 1998.5 (80) and 1999.5 (30) because of the longer time series and more frequent measurements made for EVGGFP). Person B died in year 2002.5 (not shown).

3.3 X_back

Note that if desired, X_tdie may be transformed back to the original scale, to take advantage of the imputed missing data. (Dead would have to be treated as a separate

category, indicated by a value of -12 here). We refer to this variable as X_back. Table 2 and Table 3 have a column for IADL_back. A logistic equation was used to transform X to X_t. If we instead solve that equation for X, given X_t, the result is a value on the original scale. For values which were originally observed, the back-transformed value was the observed value. Where the person was dead, the back-transformed values was set to -12. In other cases, the back-transformed value often did not result in a valid value on the original scale (it was often not an integer). When this occurred, we added a small amount of random noise and then rounded the result to the nearest valid living value. No missing value was ever back-transformed to dead; such low values of X_tdie were instead set to the worst possible living value.

Note also that for person B (Table 3), IADL_back = 2 in 1999, even though he had never reported any IADL difficulties prior to that year. This occurred because the very low value for EVGGFP_tdie was used (along with the last available value of IADL) to impute that value. The low value of EVGGFP is consistent with a person having IADL difficulties, but it is possible to have low EVGGFP without any IADL difficulties. Therefore, such an imputed value may not be useful for every analysis. For some analyses, such as those predicting the date of a person's first IADL difficulty, the analyst may prefer not to use the person's extrapolated data. This type of analytic option is always possible in a tidy dataset.

We also created a separate variable, Status_X, with a value for each person/year, describing whether the value was observed, the person was dead, the data were interpolated, or were extrapolated (in our example, Status_IADL). The analyst may choose, at times, to use only the observed data, to use interpolated but not extrapolated data, to impute the missing data in some other way, etc. Such options might also be considered to conduct sensitivity analyses. The distribution of Status_IADL is in Table 4. It shows that 79.1% of the potential observations were in fact observed, 12.3% were missing because of death, only 5.4% were interpolated, and only 3.1% extrapolated. In addition, 10 persons had missing baseline IADL data, which was imputed by substituting the first available IADL value (referred to as NOCB, for next observation carried back).

Table 4. Status of IADL Values in Reference Dataset

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 VALID	44928	79.1	79.1	79.1
	2 DEAD	7001	12.3	12.3	91.4
	3 INTERPOLATED	3095	5.4	5.4	96.8
	4 EXTRAPOLATED USING LASTREAL AND VG DATA	1785	3.1	3.1	100.0
	5 BASELINE MISSING, NOCB	10	.0	.0	100.0
	Total	56819	100.0	100.0	

The following sections describe two special cases, 4 -- analysis comparing across different variables ; 5 and 6 -- post-adjusting data missing for everyone over long gaps; and 6 pos-adjusting all of the data missing just before death. These sections may be skipped for a person interested only in how missing data were imputed before year 2000, Such readers may proceed to section 7 for a recap of the tidy dataset method.

4 Standardized health variables

Longitudinal datasets of the health of older adults commonly include many different measures of health, quality of life, and function. In one analysis, our goal was to compare the 5-year decline across 13 different measures of health, to see which declined the fastest.[9] (A more detailed version of the paper, with detailed examples of standardization, is available as a technical report. [10]) This was a challenging comparison because all the variables were measured on different scales, and change over time would also be on different scales. Instead of trying to compare changes over time for the various X's, we looked at changes in the various X_tdie's, estimated as shown in sections 1 - 3. All variables were then on the same scale, because they were estimates of the probability of being in excellent, very good, or good health, conditional on X. In that study, we referred to X_tdie as "standardized X", or X standardized by self-rated health, in the sense that all of the X_tdie's were all estimates of the same thing, the probability of being in EVGG health.

The strict interpretation is that X_tdie is the probability that a person in the reference dataset is EVGG, estimated only from X. Less strictly, X_tdie can be thought of as "standardized X". The mean of X_tdie will be about the same for all X's. But for an individual, or over time, the X's may be quite different. For example, for a person with "good" ADL (no ADL difficulties) but "bad" (slow) gait speed, standardized ADL would be better than standardized gait. If a person's ADL changed very little, but her gait speed slowed over time, then the change in standardized ADL would be smaller than the change in standardized gait.

We will illustrate the use of standardized health from that study, that compared 5-year change in 13 different health-related variables: hospitalization, bed days, cognition, extremity strength, feeling that life was worthwhile, grip strength, satisfaction with the purpose of life, depression, digit symbol substitution test, ADL, IADL, and gait speed.[10] We found that the standardized (X_tdie) versions of HOSP, BED, and COG declined the least, while standardized ADL, IADL, and GAIT declined the most. Table TR2, from the technical report, is reproduced here. Each row represents a different variable, and the columns present the mean standardized value in each year. The 5-year decline is labeled

"slope". The top 3 variables (that declined the least) were HOSP, BED, and COG which declined 12 or 13 percentage points. The bottom 3 (the most decline) were ADL, IADL, and GAIT, which declined 16 or 17 points.

Technical Report Table TR2[10]

Table 2 Mean Standardized Health by Year (N=5688)

Mean Standardized Health over Time for 13 Measures of Health

	YEAR						Slope	
	1	2	3	4	5	6	(y 6 - y1)	(s
HOSP	77.4	75.8	73.3	71.0	68.0	65.1	-12.3	
BED	77.4	75.3	73.3	70.6	67.7	64.7	-12.7	
COG	77.4	75.6	72.8	70.5	66.6	63.5	-13.9	
XSTR	77.4	75.7	72.4	69.8	66.4	63.1	-14.3	
FLW	77.4	75.2	72.3	69.4	66.1	63.0	-14.4	
GRIP	77.4	75.9	72.9	70.0	66.8	62.8	-14.6	
SPL	77.4	75.1	72.1	69.4	66.0	62.6	-14.8	
DEP	77.4	75.6	71.9	69.6	65.6	62.1	-15.3	
EVG	77.4	76.2	73.0	69.7	65.7	62.1	-15.3	
DSST	77.4	75.1	72.1	68.8	65.5	62.0	-15.4	
ADL	77.4	74.8	72.3	68.6	64.9	61.2	-16.2	
IADL	77.4	74.2	71.8	69.1	65.1	61.0	-16.4	
GAIT	77.4	74.2	70.4	67.6	64.0	60.2	-17.2	
MEAN	77.4	75.3	72.4	69.6	66.0	62.6	-14.8	
HOSP MINUS GAIT							5.0	

4.1 Consistency of ranks across different subsets of the data

Tables TR3 and TR4 in the technical report[10] (not shown here) reported that the top 3 (least decline) and bottom 3 (most decline) variables were substantially the same whether the comparison was based on the entire sample or on subsets defined by age and sex. The same variables were in the top 3 and the same in the bottom 3 if we looked at the 8-year change available in cohort 1 only (95% white), and if we looked only at the subset that survived throughout the study period. More detail about these assertions now follows.

Table 5 summarizes the results from tables TR2-TR4 in the technical report.[10] It shows the ranks of the slopes for HOSP, BED, COG, ADL, IADL, and GAIT. For example, the ranks from Technical Report Table TR2 for those 5 variables are 1, 2, 3 and 11, 12, 13, as already noted. Table TR2 results are shown in the first line of ranks.

Technical Report Table TR3 showed the same information as Table TR2, but broken down by age and sex. For example column 1 of Table TR3 (which is not reproduced here) which contained the youngest females (F-young). Row 3 of Table 5 shows that for the youngest women, the ranks of those 6 variables were 1,3,2,11,12,13 (the ranks of BED and COG were interchanged). Table TR4 (not shown), which was restricted to persons who were alive at the end of the analysis, also showed similar but not identical results, as shown in Table 5. (For young females, the ranks of BED and COG were interchanged). In all of the rows of Table 5, HOSP, BED, and COG have low ranks, while ADL, IADL, and GAIT have high ranks. This indicates that results from the standardized data were relatively independent of age, sex, and mortality. The rank of COG was less consistent than the others, especially for the oldest persons.

Table 5
Ranks of slopes for 6 variables in different tables of tech report[10]

	HOSP	BED	COG	ADL	IADL	GAIT
Table TR2	1	2	3	11	12	13
Table TR3						
F-young	1	3	2	11	12	13
F-middle	2	1	5	11	12	13
F-old	1	2	4	12	11	13
M-young	2	1	3	10	11	13
M-middle	2	1	3	10	11	13
M-old	1	2	3	11	12	13
Table TR4						
F-young	1	3	2	11	12	13
F-middle	2	1	5.5	11	12	13
F-old	1	2	8	11.5	11.5	13
M-young	2	1	3	10	11	13
M-middle	2	1	3.5	10	11	13
M-old	1	2	8	10	11	13

4.2 Comparison with Different Standardization Methods

There was some concern that the results might be specific to the variable used for standardization (EVGGFP). We chose EVGGFP, because it was the strongest longitudinal variable in CHS. It was measured more often, and for a longer time period. (That was why we used EVGGFP to create X_tdie, as explained in sections 1 - 3). We could, however, have standardized by some other variable (a different Y) as long as it was monotonically related to all 13 of the study health variables. If the standard were completely uncorrelated with X, then the standardized X would take the identical value for every level of X, and be essentially a measure of mortality rather than of X. Therefore, it is desirable to have a significant correlation between Y and each of the X's.

We thought it likely that the same general study results would obtain under a different standardization. To see whether this was correct, we standardized the X's in two additional ways, and performed the analysis of Technical Report Table TR2 [10] using the differently standardized variables. First, we standardized according to the probability of having no ADL difficulties (instead of by the probability of being EVGG). Second, we standard by age, substituting for each value of X the mean age of persons in the reference dataset who had that particular value of X.

Table 6 shows the transformed (standardized) value for the variable "Number of IADL difficulties", under each of the standardization approaches. For example, column 2 shows the estimated probability of being in EVGG health. In the reference dataset, a person with 0 difficulities had an 84% chance of being EVGG, while only 27% of those with 6 difficulties were EVGG. (Dead was always coded as zero.) These are the values that were used in our study of decline [9].

In column 3, Table 6 shows the values assigned to IADL if No ADL was used as the standard. These values indicate that, in the reference dataset, the probability of No ADL for a person with no IADL difficulties was 94% but for a person with 6 IADL difficulties the probability of no ADL was only 4.7%. The range of the scale (5 to 94) is larger than the range when EVGG was the standard (27 to 84).

Column 4 of Table 6 shows the results of age standardization, where each value of IADL was replaced by the mean age of persons in the reference dataset who had that

IADL value. The mean age of persons with no IADL difficulties was 75.4, while the mean age of persons with 6 difficulties was 81.6. To further differentiate this standardization approach from the others, these means were estimated by linear (not logistic) regression of age on untransformed X (not log X), while the other two transformations/standardizations were based on logistic regression using the log of X. The value for "dead", 82.6, was the mean age of all the records in the reference dataset where the person was set to "dead". (Persons who died contributed a value to this mean for every year in the dataset when they were dead).

Table 6
Standardized values of IADL under three different standardizations

1	**2**	**3**	**4**
Standard:	**EVGG**	**No ADL Difficulties**	**Age**
# of IADL diffs			
0	84	94.4	75.4
1	64	68.1	77.6
2	47	38.7	78.9
3	39	21.0	79.8
4	35	12.0	80.6
5	31	7.3	81.1
6	27	4.7	81.6
dead	0	0	82.6

Figure 1 is a graph of the 3 sets of coefficients inTable 6. The X axis is the value of IADL (0-6 or dead). The lines for EVG and NoADL are fairly similar, especially for 0-2 IADL difficulties. In contrast, there is relatively little variation under the AGE standardization, primarily because age did not vary much with IADL. (Correlation of IADL with EVG was -.354, with NOADL was -.588, and with age was +.278).

Figure 1.

Graphs of Standardization Coefficients of IADLunder 3 different standardizations

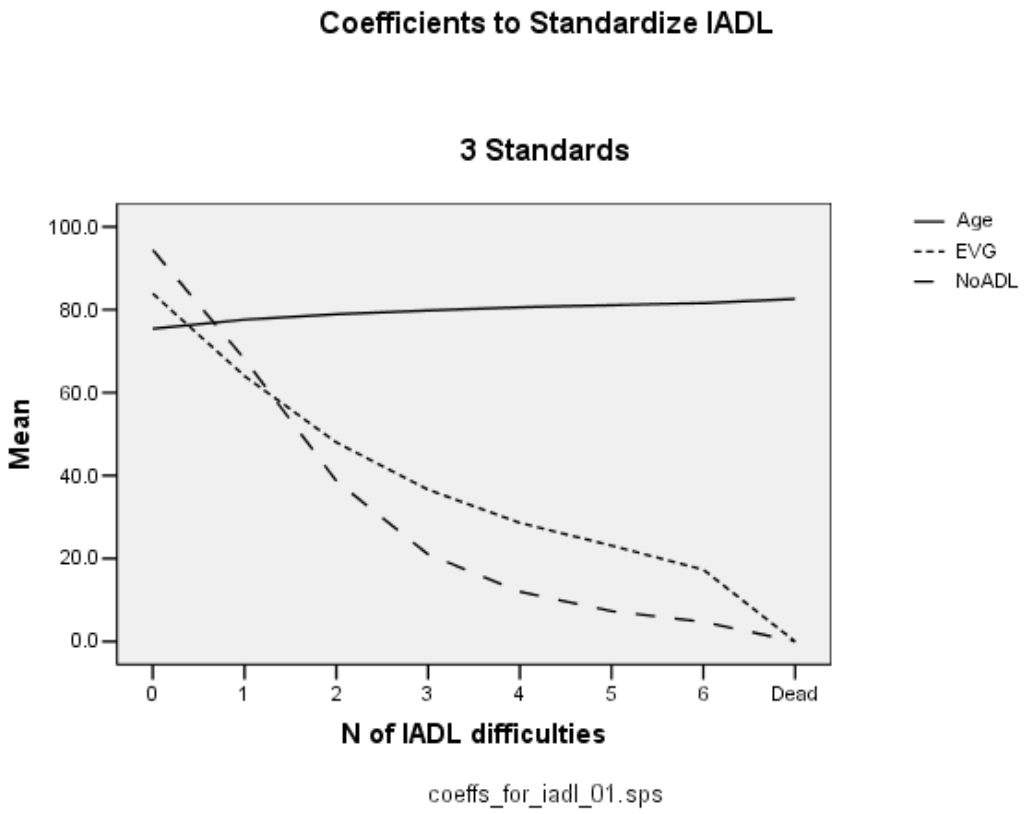


Table 6 and Figure 1 show that the standardized values of IADL are different if a different standard is used. This is not surprising, and is not a cause for concern. Our question here was whether the different standardization methods affected the results of

the study.[10] That is, were the relative rankings of the slopes over time for the different variables the same if a different standard was used? Table 7 shows the ranks of the top 3 and bottom 3 slopes from Technical Report Table TR2.[10]

The rankings using EVGGFP as a standard (column 2) were identical to the rankings using ADL as a standard (column 3). (There were some differences in ordering among the middle ranks, not shown here).

When Age was used as the standard, the same 3 variables were in the top 3, but their ordering was changed slightly. For age, the bottom 3 had high ranks, but the very highest rank was for grip strength, which was not the highest under the other two standardizations. Therefore there was good consistency for the extreme rankings, but the rankings were not identical when different variables used for standardization. We show only the extreme rankings because the technical report [10] indicated that the less extreme variables were not significantly different from one another, and variation in their ranks would not be important here.

Table 7

Ranks of slopes for top 3 and bottom 3 variables, all cases, yr 1-6 (all persons, years 1-5)

1	2	3	4
Standard:	**EVGG**	**No ADL Diffs**	**Age**
HOSP	1	1	2
BED	2	2	3
COG	3	3	1
ADL	11	n/a (the standard)	12
IADL	12	12	10
GAIT	13	13	11
			13 was GRIP

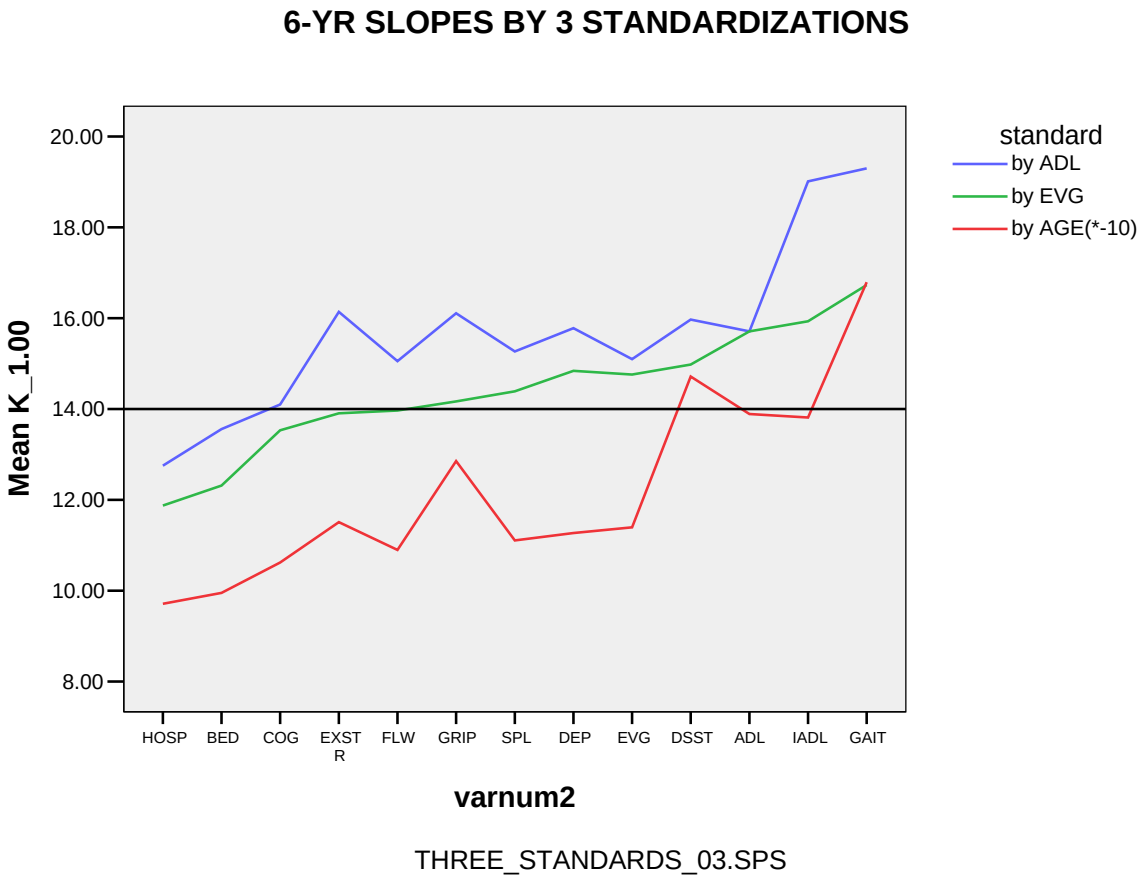
Table 8 shows the same information as Table 7, but for cohort 1 only, for which there were 9 measurement times (year 1 plus 8 years of follow-up).[10] The rankings are consistent with the main analysis for standardization by EVGGFP and ADL, but when age was the standard, DSST had rank 12.

Table 8			
Ranks of slopes for top 3 and bottom 3 variables, Cohort 1 only, years 1-9			
1	2	3	4
Standard:	**EVGG**	**No ADL Diffs**	**Age**
HOSP	1	1	2
BED	2	2	3
COG	3	3	1
ADL	11	n/a (the standard)	10
IADL	12	13	11
GAIT	13	12	13
			12 = DSST

Figure 2 is a plot of the slopes under the 3 standardizations. The X axis indicates which variable the slopes are for (from HOSP at the left to GAIT at the right). (The slopes for X standardized by Age were multiplied by -10 to put them on approximately the same scale as the other two standard variables). The middle line is slope of each of the 13 variables when standardized by EVGG. The upper line is standardized by ADL, and the lowest standardized by AGE. There is strong agreement as to the variables with the highest and lowest slopes, but there were some discrepancies in the middle.

Therefore, we found that standardizing by self-rated health, ADL, or age had similar effects in identifying the variables with extreme slopes. Most of the differences among methods occurred for the intermediate slopes, which may not have been significantly different from one another. Standardization therefore allows conversion of disparately-scaled variables to the same scale, with generally consistent results regardless of the standard. In other words, the specific standard used may not be of great consequence. However, this needs to be verified in any analysis.

Figure 2
Estimated Slopes using 3 Standardizations



4.3 Features of standardized health

Standardized X is an estimate of the probability that a person in the reference dataset with a particular value of X is in EVGG health. It should be clear that standardized health based on (say) IADL is a value entirely based on a person's IADL, and not on her EVGGFP at that time. In fact, standardized IADL is a poor estimate of an individual's

EVGGFP, which could be estimated much more accurately from her earlier or later values of EVGGFP. Standardized health refers to means in the reference dataset, not to an individual. Similarly, a person's IADL standardized by age is not directly related to her actual age, but only to her IADL value.

Standardized variables tend to be more highly intercorrelated than the original variables, because of the inclusion of 0 for death which is the same for every variable. The deaths can be removed from this calculation by use of the status variable if this is a question of interest.

Standardization also creates a complicated type of dependence among the standardized variables. Let Y be the standard, and X_i be the various variables to be standardized. (Y is healthy yes/no and X in IADL is most of our examples). Consider a much simpler setting, in which the standardized values were calculated from a simple linear regression of Y on X (not logistic regression of Y on log X). A feature of linear regression is that the regression line passes through ($\overline{X}$, $\overline{Y}$). Therefore, the estimated value of Y at $\overline{X}$ is $\overline{Y}$. In standardization, we regress Y on all the X's that are to be standardized. For each X_i, the mean standardized value at $\overline{X}_i$ must be $\overline{Y}$. Therefore $\overline{Y}$ = $\overline{X}_{1\_tdie}$ = $\overline{X}_{2\_tdie}$ = $\overline{X}_{3\_tdie}$, and so on. Standardization thus equates the mean of each variable on the original scale, equating the mean of each transformed variable to $\overline{Y}$.

Suppose we further assume that Y declines linearly over time (t). Then the value at $\overline{t}$ will be $\overline{Y}$. In our reference dataset, the average year is 1995.6. Therefore in 1995.6, y= $\overline{Y}$, and (under monotonicity assumptions for all variables) the standardized values of all the variables will also be $\overline{Y}$. That is, in a graph of mean standardized X over time (with deaths removed), that includes a separate straight line for each X, the lines should all intersect somewhere between 1995 and 1996.

This indicates that there is a constraint on the standardized X's in that they all converge at $\overline{t}$. We should thus not be surprised to find that the means of the standardized variables are all approximately the same, any more than we would be surprised to find that z-scores of all the variables all had mean zero. Since 2 points determine a line, and the plots of all standardized variables over time must intersect at $\overline{t}$, the various lines can

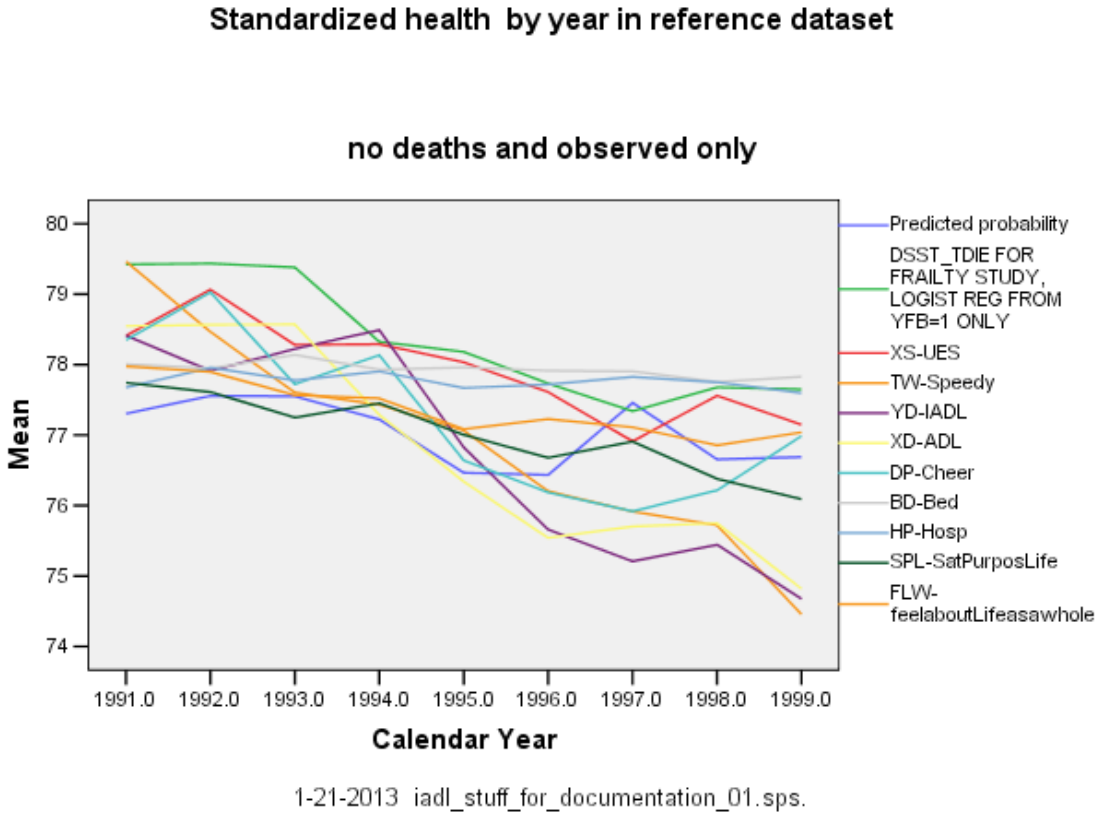
differ only in their slopes or, alternatively, in their intercepts. That is, there is only 1 degree of freedom in a comparison of slopes of different standardized variables.

This was a much simplified example. The actual graph of standardized X over time, shown in Figure 3, does not show a distinct intersection in 1995-1996. This is in part because we did not create the standardized variables using linear regression - we used logistic regression - and because Y and the X's did not decline linearly with time. This simplified example was used only to motivate the existence of a constraint, and should not be taken literally.

Figure 3

Mean standardized health over time in reference dataset.

(deaths removed)



4.4 Discussion of Standardization

We proposed a method to standardize diverse health measures, recoding each value as the probability that a person with that value would be in EVGG health in the reference dataset. In the analytic example, we identified the three (standardized) variables that declined the most, and the 3 that declined the least. Subsetting the analytic sample by age and sex made little difference in the order of the rankings. Using the longer time range available for cohort 1 also made little difference in the findings. Standardizing by ADL or AGE instead of by EVGG gave strongly consistent results, but there were a few differences in the orderings of the slopes.

The standard should ideally have a significant monotonic relationship to all of the variables that need to be standardized. Age was less correlated with the health variables than were EVGG and ADL, which may make it less desirable as a standard. In CHS, the standard itself (EVGGFP) was not standardized in this way, which suggests using a variable that is not crucial to the study goals as the standard. Given the importance of ADL in the aging literature, our choice of EVGGFP as the standard for the paper in the technical report[10] seems reasonable.

Standardizing the health variables in this way worked well, in that the study findings were surprisingly robust. It may be wise, however, to consider more than one way of standardizing the variables as a type of sensitivity analysis, depending on the purposes of the study.

Standardization has some similarities to item response theory, which equates individual items based on the expected response of a person with a given underlying "latent health" status.[11] We instead effectively equated values of variables according to expected self-rated health. For example, from Table TR1 in the technical report, having 2 bed days, having a 3MSE score of 60, feeling unhappy about life as a whole, being extremely unsatisfied with the purpose of life, or having a CESD score of 15 can be "equated" because they all correspond to a standardized score of about 50 (only about half the persons with those values were expected to be in excellent, very good, or good health). An item response analysis would not have accounted for death, and was not necessary for our purposes.

Variables are often standardized by transforming them into z-scores. We did not do that here because the z-scores are still all estimates of different quantities, even though they all have mean 0 and standard deviation 1. In addition, the spaces between the different levels are the same on the original and transformed scales, while the method used here re-scales some of those differences to agree with differences in EVGGFP. Similarly, dichotomizing each variable to "high/low" would still result in variables being estimates of different quantities.

4.5 Standardization of variables in a different dataset

Since the publication of the original paper [9] we attempted to use standardization on a different dataset, with less satisfying results.[12] That dataset, for persons with clinical depression, consisted of 6 highly correlated depression scales, 3 general health variables, and 5 scales measuring quality of life. There was more variation in the results under different choices of the standard than in the CHS study described here. We may continue to examine this dataset, to provide additional suggestions for use of the standardization method. Until then, sensitivity analysis is strongly recommended.

5. Post-adjustment of 20-year data with 5 or 7 missing yrs

The examples above referred to data collected between 1990 and 1999. For a few variables, there is currently 20-year follow-up. (Actually there is 23 years of follow-up for cohort 1, but the last three years are sometimes omitted here so that both cohorts have 20 years of follow-up). The 3 variables for which a tidy dataset is available are self-rated health, number of ADL difficulties, and cognition (measured early-on by 3MSE and later by TICS).

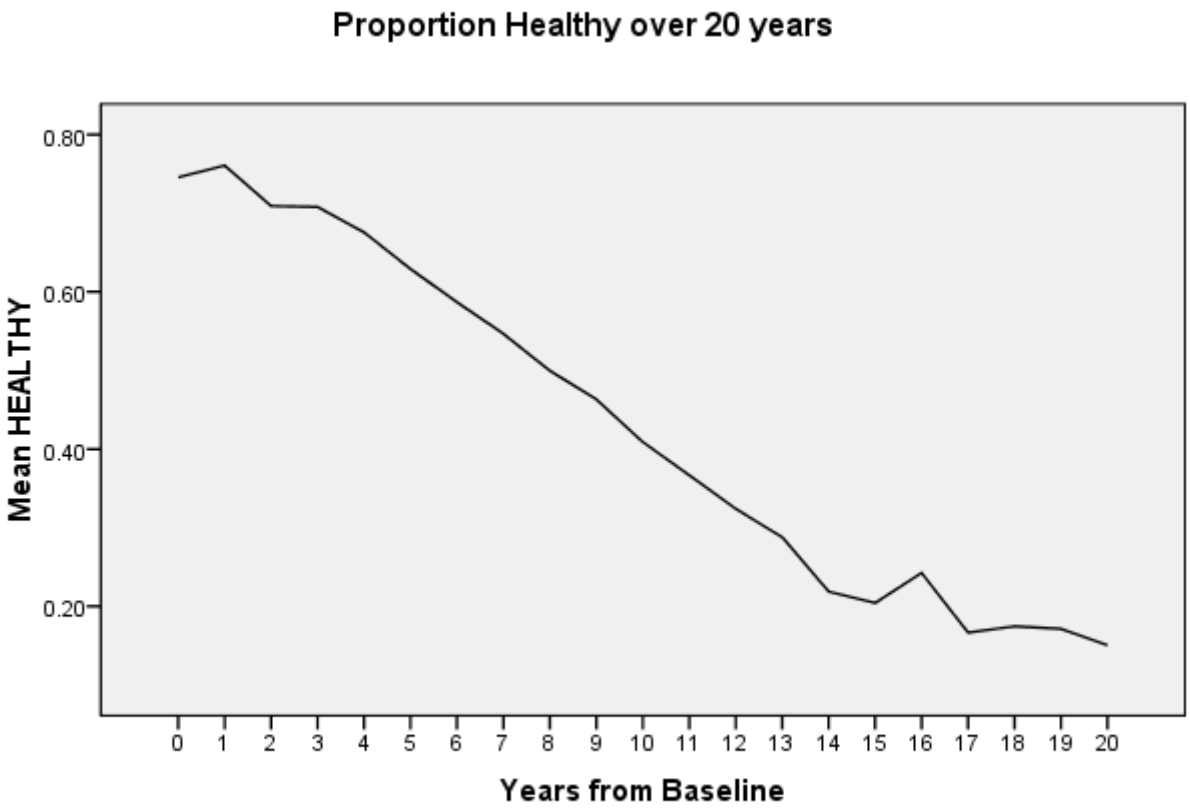
5.1 **Self-rated health** is dealt with exactly as described above. Year 2.0 was not measured for anyone in cohort 1, and was imputed as above. The following table shows the Status of observations in the 23-year EVGGFPD dataset (21*5888+3*5201 = 139251 potential observations). About 50% of potential observations were observed, and another 41% were unobserved because the person was dead, 8.1% were interpolated and -0.7% were extrapolated. Clearly, EVGGFP is an excellent variable, with very little missing data even though year 2 was not measured.

Table 9 Missingness of EVGGFPD, status over 20 years

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00 Observed	69585	50.0	50.0	50.0
	2.00 Dead	57331	41.2	41.2	91.1
	3.00 B/L missing, nocb	13	.0	.0	91.2
	4.00 Interpolated	11284	8.1	8.1	99.3
	5.00 Extrapolated	1038	.7	.7	100.0
	Total	139251	100.0	100.0	

In the following graph, the proportion healthy (excellent, very good, good) as opposed to not healthy (not fair, poor, or dead) is also shown. This proportion declines smoothly across the 20 years, with an unexplained high point 16 years after baseline. Note that extrapolated data were set equal to LOCF, without being averaged with the "extra years: of EVGGFP, since there are no extra years.

Figure 4. Proportion healthy over 20 years



tables_for_imputation_tech_report_01.sps

5.2 Activities of Daily Living (ADL), which had values from 0 to 6 activities with difficulties, was followed for 20 years except for years 2000-2004 (referred to as the gap, or the "bad years"), when measurements were made in an incompatible way. We deleted the incompatible data for those years, and imputed the five years from the remaining data as described above. We discovered that the imputed number of people with no ADL difficulties in year 2004 was much lower than the observed number in 2005, which seemed inconsistent. Further investigation showed that the discrepancy was caused by the persons who were missing and then died during the gap, for whom interpolation between the last previous known value and death provided values that were too pessimistic. Decline near death is a well-known phenomenon called "terminal drop", in which a person declines at one rate but then, closer to death, declines at a faster rate. [13] But the bad years dropped further than seemed sensible.

To solve this problem, we applied "post-adjustment" to the imputed values in the gap, on the ADL_tdie scale. (The tdie scale was used because it is an integer/ratio scale, and so linear imputation is more reasonable.) Post-adjustment involved determining the proportion with no ADL difficulties in 1999 and 2005 (when data were collected using compatible instruments). The "expected" value in 2000-2004 was determined by assuming a linear decline in the mean from 1999 to 2005. ADL_tdie from 2000-2004 was then multiplied by a factor which brought the average ADL_tdie up to the expected value, and the new variable was called ADL_tdiePA (for post-adjustment). The adjustment was done for all values in 2000-2004, in subsets determined by sex, age category, race, and whether or not the person died from 2000 to 2005. The adjustment factors were very close to 1 for persons who were still alive in 2005, but considerably larger for those who died from 2000 to 2005. ADL_tdie_pa was that back-transformed to the original scale.

[More specifically, post-adjustment started with a file I can't open right now due to the new version of spss: 'C:\aging\YAL_data\for Paula\ADLSUM_04.sav'. In 03_adlsum_postadjustment_12.spss, chs years 12-16 were bad years and chsyears 11 and 17 were good. A person who was alive at year 11 and dead at year 17 was labelled as "died in between". Agecattemp was agecat5, with 85+ coded as 85. I calculated

adlsumtdie11_mean through adlsumtdie17_mean. 11 and 17 should have real values, the other years only 0 for dead and missing for everyone else. This was aggregated separately by age, sex, perstat, and diedinbetween.

For each subgroup, the expected value was calculated as the straight line between the mean for year 11 and the mean for year 17. "Factor" is set to 1 for each bad year. If person is alive, factor is mean/expected. That is, the imputed value divided by the expected value. Factor can go as low as .09, meaning we'd want to divide the imputed value by .09. Extreme values are possible. I set any post-adjusted values above 1 to 1. Only 4203 cases had any adjustment at all. Others had factor = 1.]

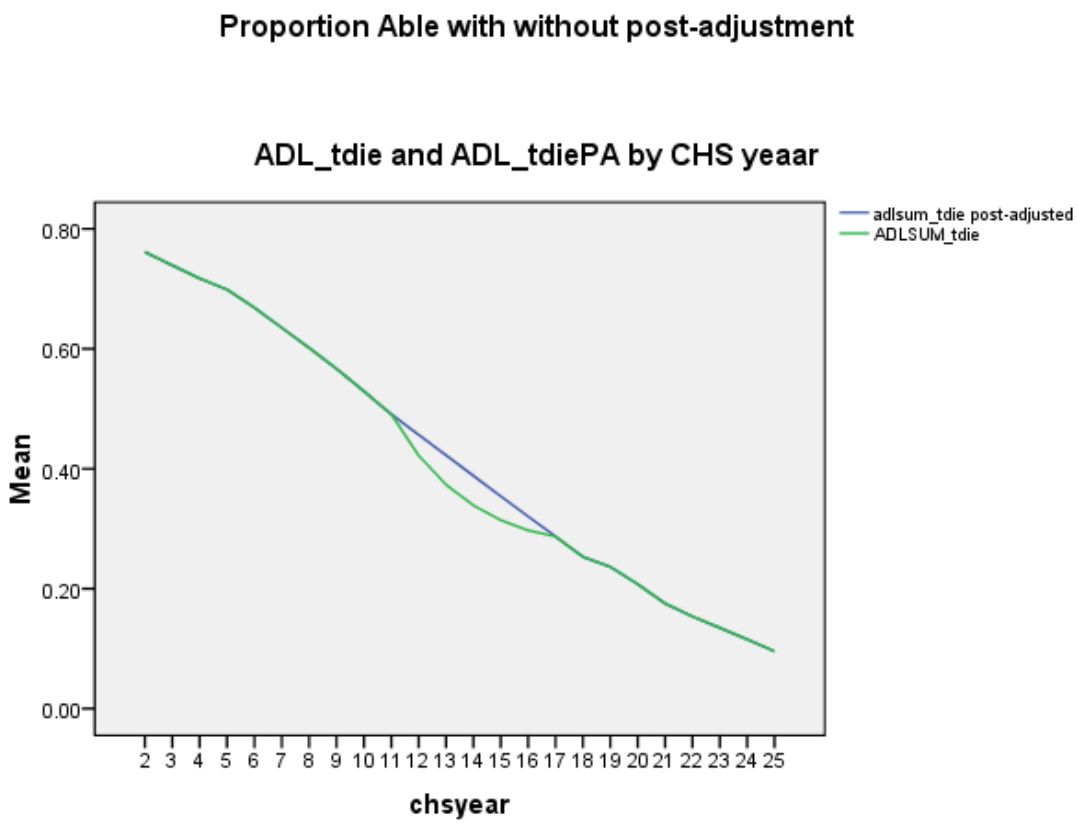
The Status table below shows that 41% of potential ADL observations were observed, and 32% + 9.0% = 41% were unobserved because of death, (the same deaths as for EVGGFPD). The "bad years" refer to the gap, and death, interpolated, and extrapolated cases are categorized by whether they were in the bad year or not. About 12.1% were imputed and then post-adjusted.

Table 10 ADL6_status ADL6, status

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00 Valid	57305	41.2	41.2	41.2
	2.00 Dead	44793	32.2	32.2	73.3
	3.00 Interpolated	6664	4.8	4.8	78.1
	4.00 Extrapolated	1040	.7	.7	78.9
	6.00 missing baseline	9	.0	.0	78.9
	12.00 dead in badyear	12527	9.0	9.0	87.9
	13.00 interpolated and post-adjusted	16613	11.9	11.9	99.8
	14.00 extrapolated and pa	300	.2	.2	100.0
	Total	139251	100.0	100.0	

The following figure illustrates the effect of post-adjustment on the ADLsum_tdie scale (ADLsum refers to the 6-point ADL scale which is the sum of 6 potential difficulties), for Cohort 1 only. Note how ADLsum_tdie has a dip during the bad years, while ADLSUM_tdiePA is linear over time during the bad years. At least on average, the post-adjusted value has a more reasonable trend over time.

Figure 5



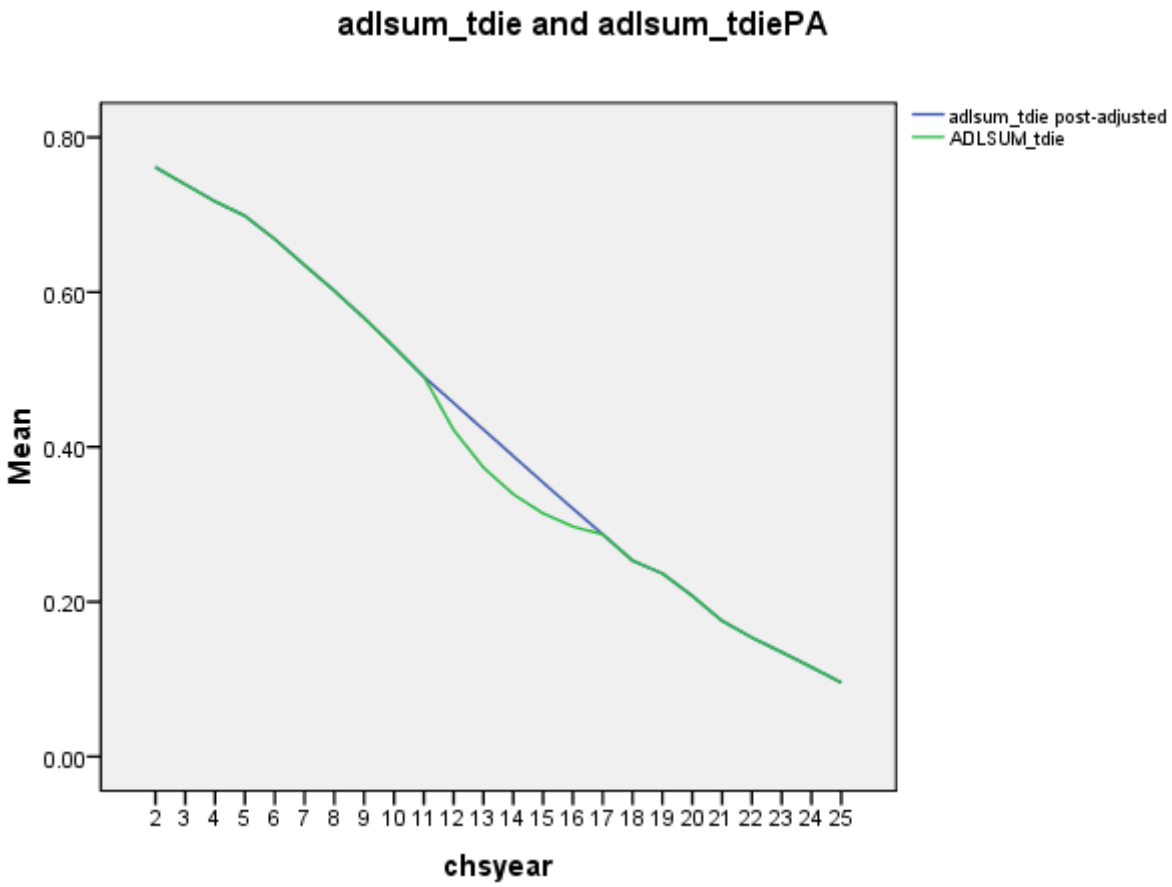
5.3 Cognition was measured by one instrument early on (3MSE scale) and another instrument later (Telephone Interview Cognition Scale, or TICS). We used a published algorithm for transforming the TICS values to the 3MSE [14] scale, and then created Cognit_tdie as above. There was a 6-year gap (bad years) in which no cognition data were collected at all, and which required post-adjustment. The status table for cognition is fairly complicated because it divides the observations into those observed (and whether from the 3MSE or TICS instrument), and the dead or imputed as to whether they were in a good year or a bad year (bad years were post-adjusted if the person died during a bad year) and which instrument was missing.

Table 11 Cognit_status

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2.00 dead	41456	29.8	29.8	29.8
	10.00 valid 3mse	28380	20.4	20.4	50.2
	20.00 valid tics or 3mse	16894	12.1	12.1	62.3
	30.00 valid tics	4699	3.4	3.4	65.7
	41.00 interp 3mse	3016	2.2	2.2	67.8
	42.00 interp tics or 3mse	3552	2.6	2.6	70.4
	43.00 interp tics	1819	1.3	1.3	71.7
	44.00 interp none planned	1948	1.4	1.4	73.1
	51.00 extrap 3mse	50	.0	.0	73.1
	52.00 extrap tics or 3mse	228	.2	.2	73.3
	53.00 extrapolated tics	1732	1.2	1.2	74.5
	54.00 extrap none planned	149	.1	.1	74.6
	102.00 dead in bad year	15873	11.4	11.4	86.0
	144.00 interp and post-adjusted	18825	13.5	13.5	99.5
	154.00 extrapolated and post-adjusted	630	.5	.5	100.0
	Total	139251	100.0	100.0	

The following graph shows adl_tdie (green line) as well as the post-adjusted version (blue line) for Cohort 1 only. (All on the transformed, "probability of being healthy" scale). As in Figure 5, there was a dip during the bad years, which was corrected (or at least made more reasonable) by post-adjustment.

Figure 6



6. Post-adjustment for bias in imputed values near death.

6.1 Section 5 found that the X_tdie method gave values that were "too low" for data missing just before death, if the missing data were in CHS years 12-16 (ADL) or 12-17 (COG). This adjustment was necessary only when the missing value occurred between the last observed value and death, and was done separately by sex and age groups. We reasoned that this type of adjustment could be used for values missing just before death in other ("good") years. In section 5, we post-adjusted X_tdie in the bad years by dividing it by a factor that was created to bring the grand mean up to the "expected" value (linear between CHS year 12 and CHS year 17 or 18). The mean factors are shown for ADL and COG in the following table, by years before death (YBD), limited to values in the correct time period that were partially imputed from the "0" for dead. (EVG is explained later).

Table 12

Report

YBD YRS BEFORE DEATH (YOD-YFB)		evg_factor	ADL_FACTOR	COG_FACTOR
1.00	Mean	.4551	.4492	.4812
	N	1403	1403	1403
	Std. Deviation	.20508	.20549	.19507
2.00	Mean	.6082	.6029	.4926
	N	1354	1354	1354
	Std. Deviation	.24077	.24316	.18985
3.00	Mean	.7370	.7335	.6293
	N	1291	1291	1291
	Std. Deviation	.22257	.22462	.21940
4.00	Mean	.8533	.8513	.7578
	N	1231	1231	1231
	Std. Deviation	.17399	.17582	.20577
5.00	Mean	.9474	.9464	.8704
	N	1173	1173	1173
	Std. Deviation	.09985	.10132	.16153
6.00	Mean	1.0000	1.0000	.9529
	N	1090	1090	1090
	Std. Deviation	.00000	.00000	.09101
7.00	Mean	1.0000	1.0000	1.0000
	N	1034	1034	1034
	Std. Deviation	.00000	.00000	.00001
Total	Mean	.7811	.7784	.7207
	N	8576	8576	8576
	Std. Deviation	.26074	.26333	.26119

For both ADL and COG, the factor for 1 year before death was about .5, and the factors increased with YBD, eventually reaching a value of 1 (no post-adjustment needed). The factors for ADL and COG were fairly similar. (Note, it is not clear that averaging the factors makes more sense than averaging the inverse of the factors – that is worth exploring).

There is also a column for EVG, even though EVG had no bad years (except for CHS year 4, for some reason) and so did not need the post-adjustment of the previous section. To derive a factor for EVG, we temporarily set all of the data from CHS years 12-16 to missing (except for the deaths, which were coded as zero) and then used the methods above to impute those "missing" values. We then calculated the necessary post-adjustment factor for EVG, shown in Table 12. Note that it, too, is quite similar to the factors for COG and ADL. Presumably these similarities are because the standardized variables are all estimates of the same thing (probability of being healthy) and because time is years from death in all cases. It would be nice to derive this algebraically. A special case is in section 6.2 (We of course used <u>all</u> of the observed EVG data for further analyses. We also set EVG_tdiepa to EVG_tdie, for consistency in notation).

We next identified all of the values of X_tdie that were imputed from the "0" for dead. (That is, missing between death and the last observed value, and not in the time periods that had already been post-adjusted). For each of those cases, we calculated X_tdiepa2 = X_tdiepa/factor, where the factor was taken from table 12. (Otherwise, X_tdiepa2 was just set to X_tdiepa).

Table 13 is an example of the calculations. It shows ADL information for person C, who was chosen to illustrate all of the steps in imputation. Person C had no ADL difficulties from CHS year 2-9, was missing ADL in years 10-14, and died in year 15. The "tdie" column is 0.77 (the probability that a person with no ADL limitations is "healthy") for the observed values. This column was set to zero for the dead years, and is linearly interpolated from year 10 to 14 (the shaded years, with values from .64 to .13).

Table 13

ADL values for Person C

		Transformed Scale			Original Scale		
CHS Year	ADL observed	tdie	PA	PA2	BACK	PA	PA2
2	0	0.77	0.77	0.77	0	0	0
3	0	0.77	0.77	0.77	0	0	0
4	0	0.77	0.77	0.77	0	0	0
5	0	0.77	0.77	0.77	0	0	0
6	0	0.77	0.77	0.77	0	0	0
7	0	0.77	0.77	0.77	0	0	0
8	0	0.77	0.77	0.77	0	0	0
9	0	0.77	0.77	0.77	0	0	0
10	.	0.64	0.64	0.68	1	1	0
11	.	0.52	0.52	0.61	3	3	2
12	.	0.39	0.52	0.52	4	3	3
13	.	0.26	0.47	0.47	5	3	3
14	.	0.13	0.31	0.31	6	5	5
15	dead	0	0	0	-12	-12	-12
16	dead	0	0	0	-12	-12	-12

ADL_tdiepa (PA column) shows the post-adjustment of the values near death during the "bad" years. Years 12, 13, and 14 were already increased in value by the first post-adjustment. ADL_tdiepa2 shows the adjustment for years 10 and 11, which were not during the "bad" years, but which were missing within 5 years of the date of death and were estimated in part from the 0 for dead.

The last 3 columns show the transformed values back-transformed to the original scale. (The deaths were arbitrarily set to -12 so it would be clear that this are not valid imputed values). Years 10-14 were all originally missing. Although person C never had an ADL difficulty in years 2-9, she was imputed to have ADL difficulties during the missing years. The imputed number of ADL difficulties was somewhat lower after post-adjustment.

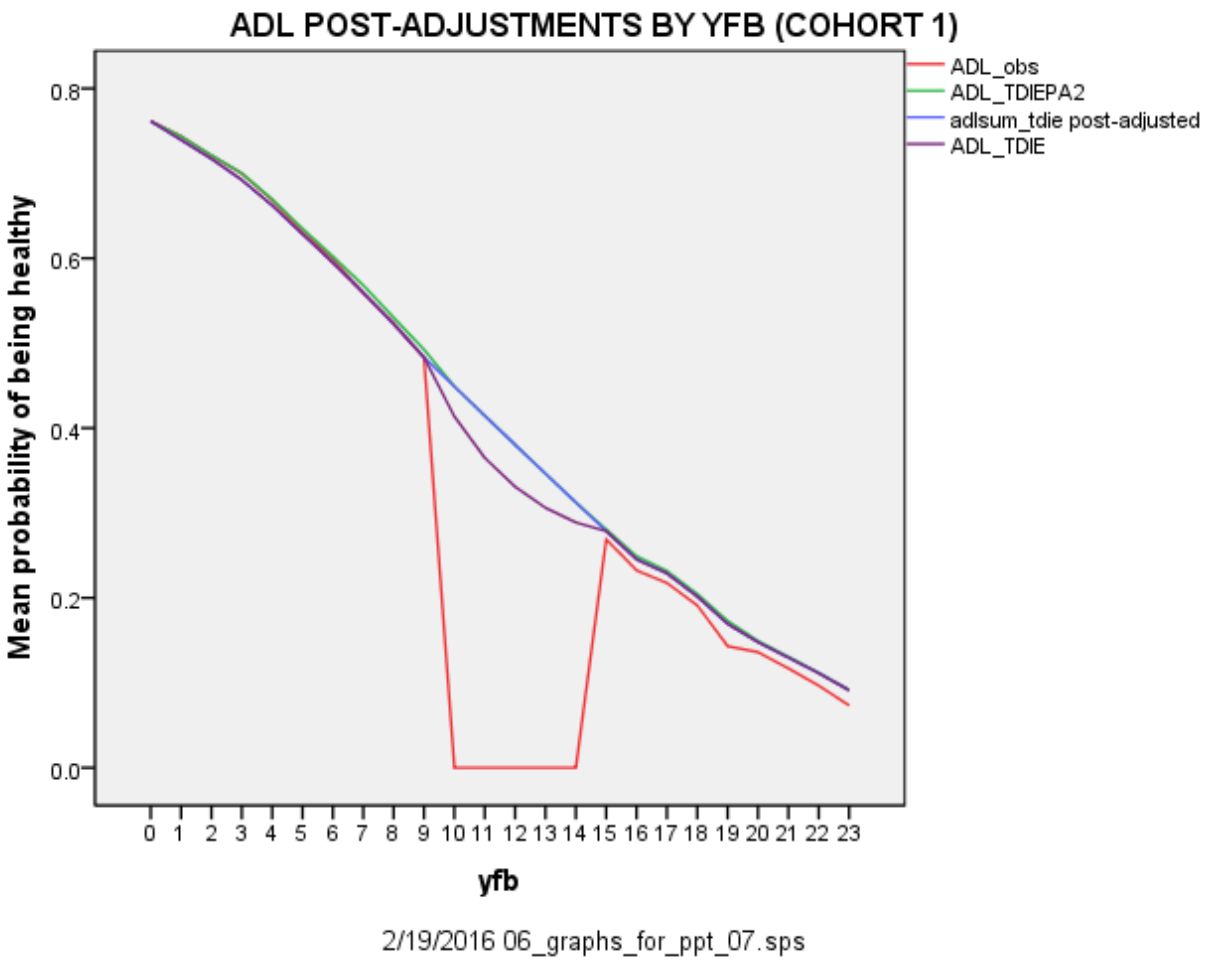
Table 13 illustrates the methods, and also illustrates why analyses involving person-level data might be risky. Person C was never observed to have an ADL

difficulty, which occurred only in the imputed data. It seems reasonable that a person near death, who was too impaired to provide data even over the telephone, might have had some ADL difficulties, but this assumption, which may make sense on average, may not be correct for certain individuals. Sensitivity analysis is of course recommended in all analyses involving missing data.

The following graphs show different versions of ADL (on the tdie scale) over time. The lines have different colors, and may best be examined on-line. In general, the "best" line represents the observed values, the worst represents ADL_tdie, next worst ADL_tdiepa and the closest to the observed is ADL_tdiepa2.

In figure 7, the red (lowest) line shows the mean of the observed values of ADL_td. The sample sizes are probably different at the various times because no imputation was done. Note that in the "bad" years the mean drops to 0 because only the dead had observed data. The Y axis represents the mean "probability of being healthy|ADL" both observed and under different types of imputation: tdie, tdiePA1 and tdiePA2. The purple (curvy) line is for ADL_tdie which is inconsistently low during the bad years. The blue (3rd lowest) line is for ADL_tdiepa, which post-adjusted the bad years. And the green line is for ADL_tdiepa2 (second from top), which post-adjusted the missing values just before death, outside of the bad years.

Figure 7



It is difficult to see the effect of the second post-adjustment in Figure 7. Figure 8 shows the values in the 7 years before death (only for those who died). It is quite clear that using only the observed data (red line) gives the most optimistic view of ADL just before death, presumably because persons too sick or unable to respond did not have a value and so are omitted. The bottom-most (purple) line shows ADL_tdie, and it is clear that imputation made the trend much more pessimistic. Post-adjustment of the bad years (blue line) gave more optimistic estimates, and the second post-adjustment of the good years (green) gave the most optimistic results of all (though still lower than the observed-only means, and preferable because all values are imputed).

Figure 8

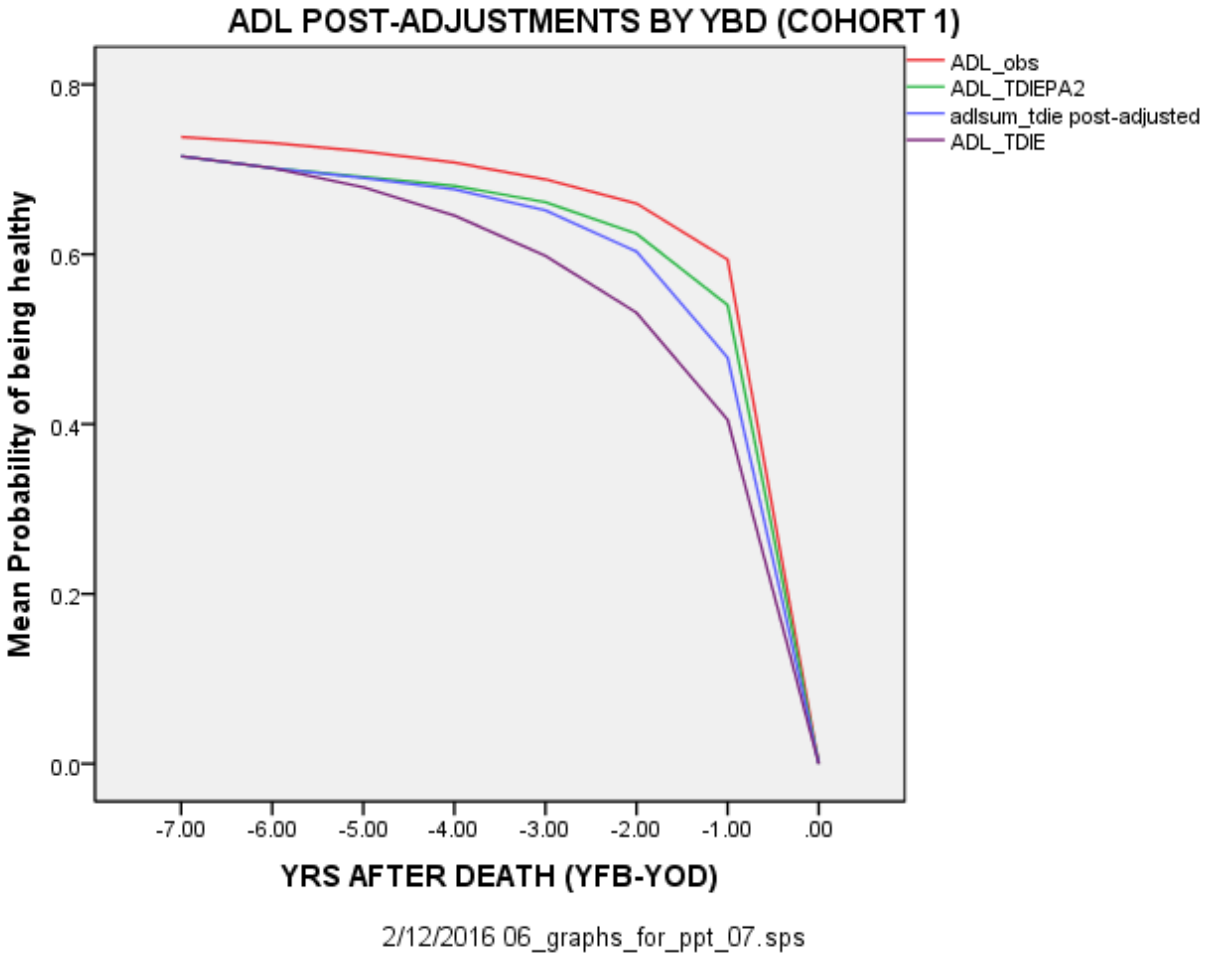
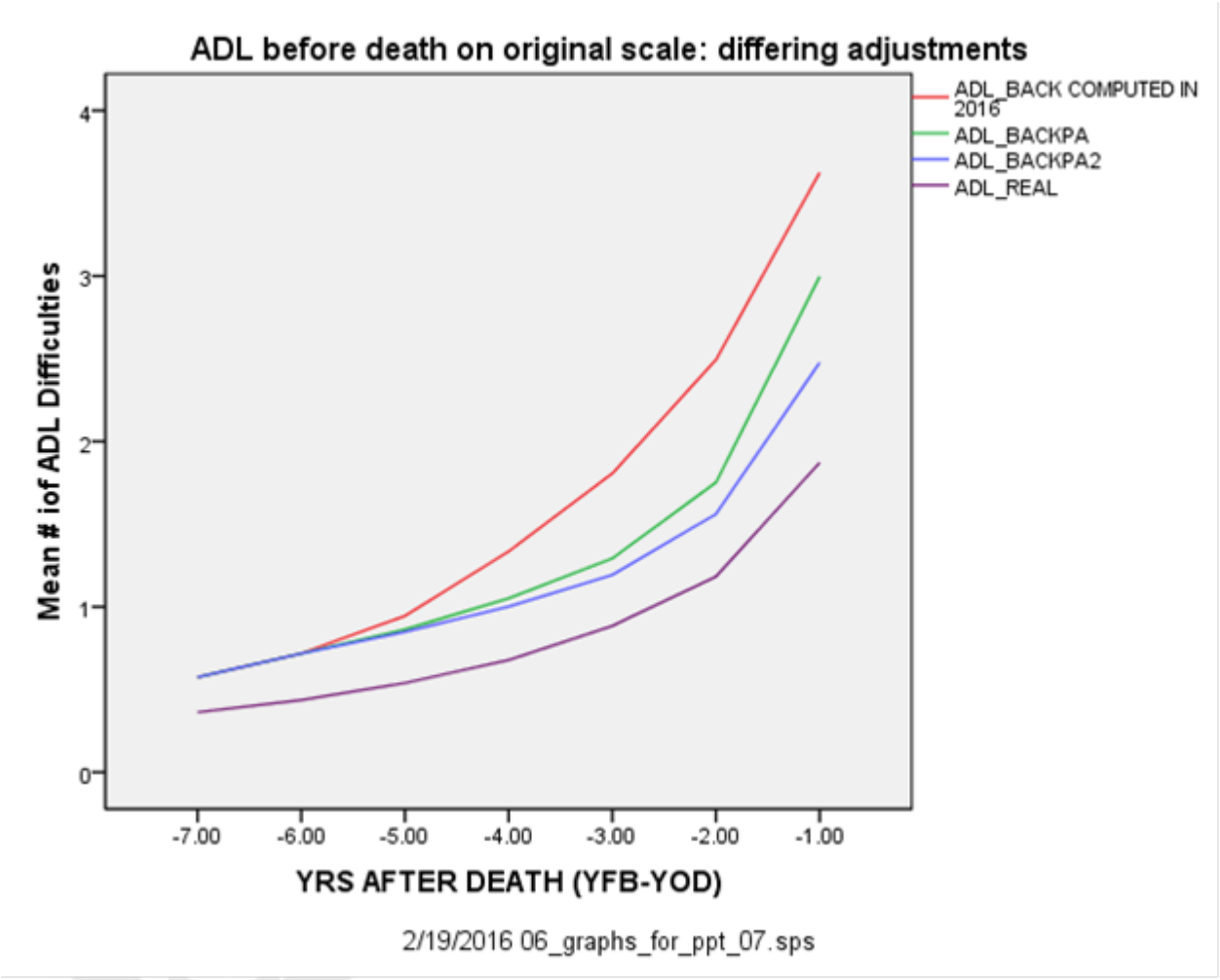


Figure 9 shows a similar figure to Figure 8 , except that it is transformed back to the original scale. On the original scale, death has no value, so time goes from -7 to -1 years. For ADL, low values are the best, and so the ordering of the lines is reversed. But, as expected, the observed values are the most optimistic, the unadjusted the least optimistic, and the second post-adjustment looks the best.

Figure 9



6.2 Example of Calculations: a special case

The following discusses the spreadsheet table, Figure 10. The non-shaded parts are intermediate calculations, which can be ignored. We show how the post-adjustment factors were calculated, and then how the average factors were calculated to allow post-adjustment in the "good years". (Ignore the non-shaded areas).

In Figure 10, we look at hypothetical values of the interpolated results and show how they result in the adjustment factor. Column 11 refers to the observed status in year 11 (always 100 for convenience) and column 17 refers to the observed status in year 17 (always 0, because everyone in this example was alive in year 11 but dead in year 17, and their imputed values thus require post-adjustment).

The first row of Block A shows that 200 persons (arbitrary number), with initial value of 100, had already died by year 12, and so had a zero in years 12-15 (no missing data). The second row shows that 200 persons died in year 13, and had an imputed value of 50 at year 12, and zeroes thereafter. The last row refers to the persons who alive in years 12-16, but dead in year 17. Their imputed values (83.33, etc.) were obtained by linear interpolation between the value for year 11 and year 17.

The second block shows "mean imputed", which is the average value in Block A, over the deaths and the interpolated values. For example, the mean imputed value in year 12 is 59.17, and the mean in year 16 is 2.78.

The following block shows "mean expected". Here, we take advantage of the fact that the missing data are uninformative. We should therefore expect a smooth decline in the average line from years 11 to 17. For example, we should expect the values from year 11 to 17 to have approximately linear decline, as was shown in the real data in Figure 6. The row labelled "mean expected" shows the "expected" value, if there was linear decline. Note that the expected values decline more slowly than the mean imputed values in the previous line. The following shaded line shows that "factor" was calculation as the ratio of the imputed to the expected. For the post-adjustment of the bad years, we divided each imputed value by the factor, which increased them so that the average trend over time was the smooth line. The Year 16 would be increased a good deal (1/.41 = .24, so nearly 4 times as high), but year 12 not so much.

In the second post-adjustment, of the values missing just before death but not in a bad year, we used the average factor depending on how far before death the missing value was. What factor should be used 5 years before death? Examine Block A. The only factor available for 5 years from death was for year 12 the last line of in Block B, and the factor used for year 12 was 0.71. Therefore, the factor for 5 years before death is the factor for year 12, 0.71. There are two factors calculated 4 years before death; the last line of Block A (year 13) and the next-to-last line of Block A (year 12). We used the average factors for those two years, (.71 + .53)/2 = .62. And so on. Every row in Block A but the first needed an imputed value in the year before death, and so the average factor for such years was the average of all the factors in the year before death, .41.

These average factors by years before death (for values imputed between the last observed value and the 0 for death and not in a bad year) are shown in Block B.

The average factors actually used for ADL are shown in the rightmost column of Block B. The "general trend" of the factors by years before death is similar for the special case and the ADL data. Differences were caused by violations of the assumptions that all initial values were the same, every line in Block A had the same number of cases, and so on. The actual factors used were calculated within in age and sex groups, as well. We used no actual ADL parameters in this special case, which suggests that the factors may well be similar for the various variables that are post-adjusted. We hope that this example will give the users at least a feeling for the type of calculations done in the second post-adjustment.

Figure 10, special case of calculating the factor

	Block A	means						
Y	n	**11**	**12**	**13**	**14**	**15**	**16**	**17**
100	200	100.00	0.00	0.00	0.00	0.00	0.00	0.00
100	200	100.00	50.00	0.00	0.00	0.00	0.00	0.00
100	200	100.00	66.67	33.33	0.00	0.00	0.00	0.00
100	200	100.00	75.00	50.00	25.00	0.00	0.00	0.00
100	200	100.00	80.00	60.00	40.00	20.00	0.00	0.00
100	200	100.00	83.33	66.67	50.00	33.33	16.67	0.00
	1200							
n by mean	products	20000.00	0.00	0.00	0.00	0.00	0.00	0.00
		20000.00	10000.00	0.00	0.00	0.00	0.00	0.00
		20000.00	13333.33	6666.67	0.00	0.00	0.00	0.00
		20000.00	15000.00	10000.00	5000.00	0.00	0.00	0.00
		20000.00	16000.00	12000.00	8000.00	4000.00	0.00	0.00
		20000.00	16666.67	13333.33	10000.00	6666.67	3333.33	0.00
	SUM	120000.00	71000.00	42000.00	23000.00	10666.67	3333.33	
mean imputed		100.00	59.17	35.00	19.17	8.89	2.78	0.00
mean expected		100.00	83.33	66.67	50.00	33.33	16.67	0.00
	factor		0.71	0.53	0.38	0.27	0.17	

Block B

factor by ybd

ybd	**Special**	**ADL**
1	0.41	*0.45*
2	0.47	*0.60*
3	0.54	*0.73*
4	0.62	*0.85*
5	0.71	*0.95*

6.3 The second post-adjustment was, let us admit, a rather ad hoc approach. While we can see clearly that the first post-adjustment improved the trend in the bad years, there is no firm evidence that the second post-adjustment improved data missing before death in the good years. As always, the user might test key findings with different levels of imputation and adjustment to make sure that a key finding was not just the result of the imputation or post-adjustment method.

Longitudinal data provide some challenges, but also some opportunities. The approach used here, of deleting a range of values (or missing values or death) and imputing them all by the method under consideration allows a way of estimating bias that is not available to shorter time series. If imputation is done by transforming to a new scale with a value for death, and then interpolation, what we have done here will provide a post-adjustment. If, after transformation, the values are fit by regression of the person's own values, some other post-adjustment factor might be needed.

7. Summary and Conclusion: the Tidy Dataset

There are relatively few papers in the literature about imputation of data missing near to death. Several papers have described (multiple) imputation as part of analysis [15] [16] [17] [18] including several that used multiple imputation. Some discuss treatment of death, but conclude only that death "cannot safely be assumed to represent noninformative censoring."[19] None of these gives advice on a plausible way to impute data missing before death.

Here we presented the approach used in the Cardiovascular Health Study to create a "tidy" longitudinal dataset, which has a separate record for every potential measurement. Each record contains either the person's observed value, a notation that she was dead at that time, or an imputed value if the data were missing at that point. Variables were first transformed to a new scale with integer/ratio properties, and on which "dead" takes the value zero. Missing data were then imputed on this new scale, using each person's own data over time. Imputation was thus informed by impending death. (Those who didn't die had any missing data at the end imputed by extrapolation). Data missing between the last observed value and death were post-adjusted to be more consistent on average (20-year data only).

The new transformed and imputed variable has a value for every person at every potential time, and accounts for death. It can also be considered as "standardized health," permitting comparison of time trends for variables that were originally measured on disparate scales. The standardized variable can also be transformed back to the original scale, which has the missing values imputed from the person's own longitudinal data and informed by impending death. Each observation is labeled as to whether it was observed, imputed (and how), or the person was dead at the time.

The resulting "tidy" dataset may be considered complete, but is flexible enough to permit analysts to handle missing data and deaths however they want. This approach may be useful for other longitudinal studies as well as for the Cardiovascular Health Study. Ideally, those datasets would have 3 or more measures per person, a low rate of missing data, and complete ascertainment of deaths.

Sections 4 and 5 dealt with (4) a special comparison of trends among different variables and (5) post-adjustment of imputed values over a long unmeasured internal gap. Section (6) used this method to post-adjust the remaining values calculated near death. These methods may be of interest to some other readers.

References

[1] Fried LP, Borhani NO, Enright PL, et al. The Cardiovascular Health Study: design and rationale. Annals of Epidemiology 1991. 1:263-276.

[2] Diehr P, Patrick DL, Spertus J, Kiefe CI, McDonell M, Fihn SD. Transforming self-rated health and the SF-36 Scales to include death and improve interpretability. *Medical care* 39:670-680, 2001. (PMID: 11458132)

[3] Diehr P, Johnson LL, Patrick DL, Psaty B. Methods for incorporating dath into health-related variables in longitudinal studies. J Clilnical Epidemiology 2005; 58:1115-1124.

[4] Diehr P, Patrick DL, McDonell MB, Fihn SD. Accounting for deaths in longitudinal studies using the SF-36: the performance of the Physical Component Scale of the Short Form 36-Item Health Survey and the PCTD. Medical Care 2003; 41:1065-1073. (PMID: 12972846)

[5] Kurland BF, Johnson LL, Diehr PH. Longitudinal data with follow-up truncated by death: match the analysis method to research aims. Statistical Science 2009; 24:211-222. (PMID: 20119502)

[6] Kurland BF, Johnson LL, **Diehr PH**. Accommodation of missing data in supportive and palliative care trials. Curr Opin Support Palliat Care 2012;6.

[7] Engels JM (Jean Mundahl), Diehr P. Imputation of Missing Longitudinal Data: a comparison of methods. Journal of Clinical Epidemiology 2003; 56:968-976. (PMID: 14568628)

[8] Diehr PH, Lafferty W, Patrick DL, Downey L, Devlin S, Standish LJ. Quality of life at the end of life. Health and Quality of Life Outcomes. 2007; 5:51. (PMID: 17683554)

[9] Diehr, **PH**, Thielke SM, Newman AB, Hirsch CH, Tracy R. Decline in health for older adults: 5-year change in 13 key measures of standardized health. J Gerontol A Biol Sci Med Sci. 2013 May 10. [epub ahead of print] doi:10.1093/gerona/glt038 5

[10] Diehr, Paula H.; Thielke, Stephen M.; Newman, Anne B.; Hirsch, Calvin H.; and Tracy, Russell, "Decline in Health for Older Adults: 5-Year Change in 13 Key Measures of Standardized Health" (October 2012). *UW Biostatistics Working Paper Series.* Working Paper 385. http://biostats.bepress.com/uwbiostat/paper385.

[11] McHorney CA. Equating health status measures with item response theory: illustrations with functional status items. Med Care. 2000;38(9 Suppl):II43-59.

[12] Diehr PH, Derleth AM, McKenna SP, Martin ML, Bushnell DM, Simon G, Patrick DL, the LIDO group. Synchrony of change in depressive symptoms, health status, and quality of life in persons with clinical depression. Health Qual Life Outcomes 2006; 4:27.

[13] Diehr P, Williamson J, Burke G, Psaty B. The aging and dying process and the health of older adults. *Journal of Clinical Epidemiology* 55:269-278, 2002. (PMID: 11864798)

[14] Arnold AM, Newman AB, Dermond N, Haan M, Fitzpatrick A. Using telephone and informant assessments to estimate missing Modified Mini-Mental State Exam scores and rates of cognitive decline. Neuroepidemiology 2009;33:55-65.

[15] Kurland BF, Johnson LL, Egleston BL, Diehr PH. Longitudinal data with follow-up truncated by death: match the analysis method to research aims. Stat. Sci. 2009; 24(2):211.

[16] Ning Y, McAvay G, Chaudhry SI, Arnold AM, Allore HG. Results differ by applying distinctive multiple imputation approaches on the longitudinal cardiovascular health study data. Exp. Aging Res. 2013;39(1);27-43.

[17] Biering K, Hjollune NH, Frydenberg M. Using multiple imputation to deal with missing data and attrition in longitudinal studies with repeated measures of patient-reported outcomes. Clin Epidemiol. 2015 Jan 16;7:91-106. doi: 10.2147/CLEP.S72247. eCollection 2015.PMID: 25653557

[18] Barzi F, Woodward M. Missing data on the Center for Epidemiologic Studies Depression Scale: a comparison of 4 imputation techniques. Am J Epidemiol 2004;160:34-35.

[19] Murphy TE, Lan L, Allore HG, Peduzzi PN, Gill TM, Lin H. Treatment of death in the analysis of longitudinal studies of gerontological outcomes. J Gerontol A Biol Sci Med Sci,2001. January; 66A(1);109-114. Doi 10.1093/Gerona/glq188 .