

University of Washington

From the Selected Works of Paula Diehr

July, 2001

Transforming self-rated health and the SF-36 scales to include death and improve interpretability

Paula Diehr, *University of Washington*



Available at: https://works.bepress.com/paula_diehr/6/

Transforming Self-Rated Health and the SF-36 Scales to Include Death and Improve Interpretability

PAULA DIEHR, PHD,*† DONALD L. PATRICK, PHD, MPH,† JOHN SPERTUS, MD, MPH, FACC,§
CATARINA I. KIEFE, MD,*¶ MARY McDONELL, MS,|| AND STEPHAN D. FIHN, MD, MPH†‡||

BACKGROUND. Most measures of health-related quality of life are undefined for people who die. Longitudinal analyses are often limited to a healthier cohort (survivors) that cannot be identified prospectively, and that may have had little change in health.

OBJECTIVE. To develop and evaluate methods to transform a single self-rated health item (excellent to poor; EVGGFP) and the physical component score of the SF-36 (PCS) to new variables that include a defensible value for death.

METHODS. Using longitudinal data from two large studies of older adults, health variables were transformed to the probability of being healthy in the future, conditional on the current observed value; death then has the value of 0. For EVGGFP, the new transformations were compared with some that were published earlier, based on different data. For the PCS, how well three different transformations, based on differ-

ent definitions of being healthy, discriminated among groups of patients, and detected change in time were assessed.

RESULTS. The new transformation for EVGGFP was similar to that published previously. Coding the 5 categories as 95, 90, 80, 30, and 15, and coding dead as 0 is recommended. The three transformations of the PCS detected group differences and change at least as well as the standard PCS.

CONCLUSION. These easily interpretable transformed variables permit keeping persons who die in the analyses. Using the transformed variables for longitudinal analyses of health when deaths occur, either for secondary or primary analysis, is recommended. This approach can be applied to other measures of health.

Key words: Death; health-related quality of life; QALY; longitudinal; health status. (Med Care 2001;39:670-680)

*From the Department of Biostatistics, University of Washington, Seattle, Washington.

†From the Department of Health Services, University of Washington, Seattle, Washington.

‡From the Department of Medicine, University of Washington, Seattle, Washington.

§From the Department of Medicine, University of Missouri, Kansas City.

¶From the Center for Outcomes and Effectiveness Research and Education, University of Alabama, Birmingham.

||From the Northwest Health Services Research and Development Center of Excellence, VA Puget Sound Health Care System, Seattle, Washington.

The research reported here was supported in part by

the Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service project number SDR 96- 002, and by contracts N01-HC-85079 through N01-HC85086 from the National Heart, Lung, and Blood Institute. Dr. Diehr is the statistician for the ACQUIP project at the VA Puget Sound Health Care System.

The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs.

Address correspondence and reprint requests to: Dr. Diehr, Box 357232, University of Washington Seattle, WA 98195- 7232. Phone 206-543-8004; Fax 206-523-4680.

Received July 18, 2000; initial review September 22, 2000; accepted February 22, 2001.

Health-related interventions are often evaluated by following the health of participants over time. In studies of older or severely ill persons and in longer-term follow-up studies, data are often missing because some subjects die. Some measures of health-related quality of life (HRQOL), such as the Quality of Well-Being scale (QWB) or the Health Utilities Index, incorporate a value for death integrated with preference-based assessments of symptoms, functional status, or other domains of HRQOL.¹

Most nonpreference-weighted measures, like self-rated health or the Short Form 36-item Health Survey (SF-36), do not explicitly incorporate mortality. When deaths occur, researchers usually report only the complete cases, discarding or reporting separately the data for people who die. However, this limits the evaluation to subjects in better health (those who survived), and may miss important changes over time. The group with more deaths has an advantage under such an analysis because more of its sickest cases are removed from the analysis. Anderson et al² found that the QWB scale was more sensitive than the SF-36 to change over time in AIDS patients, primarily because the QWB gave deaths a score of 0 whereas on the SF-36 they were dropped from the analysis.

Some attempts have been made to incorporate a value for death into the SF-36. Brazier and colleagues applied preference rating to the SF-36 items³ to create a new scale. Fryback et al⁴ regressed the QWB on the eight subscales of the SF-36 to obtain an "estimated" QWB that could be used in analysis. The former method required a major research effort and the latter required a large and generalizable data set that contained both the QWB and the health measure of interest (here, the SF-36).

Another approach that can be applied to a variety of health measures is to assign a value to death in some defensible way. If death can be considered as the worst possible level in the construct being measured, then nonparametric hypothesis tests can be performed. If we can assign a numeric value to death, however, the full range of parametric analyses are also possible. One simple-seeming approach would be to assign the value 0 to death. However, this is completely arbitrary, and would invalidate the psychometric properties of the original measure. Here, we examine different approaches for transforming the

commonly used self-rated health item and the SF-36 to yield a defensible value for death.

Methods

HRQOL Variable Definitions

We used two HRQOL variables: self-rated health (Is your health Excellent, Very Good, Good, Fair, or Poor? - EVGGFP) and the SF-36. EVGGFP is a simple and well-known measure which has been studied in detail^{5,6} and is predictive of future health events.⁷ Because we are examining health status over time, we added a sixth health state, Dead. EVGGFP is probably the most commonly used single-item health status measure and is the first item on the SF-36 instrument.

The SF-36 is a widely used health status measure⁸ developed for the Medical Outcomes Study. Results are usually reported in eight subscales. Ware et al⁸ have derived a summary "physical component score" (PCS) and "mental component score" (MCS). We developed transformation equations for all subscales, but report detailed results for only the PCS.

Data

Data were taken from the Cardiovascular Health Study (CHS) and the Ambulatory Care Quality Improvement Project (ACQUIP). CHS is a population-based longitudinal study of 5,888 adults 65 years of age and older designed to identify factors related to the occurrence of coronary heart disease and stroke.⁹ CHS subjects were recruited from a random sample of the Medicare eligibility lists in four communities in the United States,¹⁰ with some health exclusions, and approximately 70% of those invited participated in the study.¹¹ Vital status was known for all subjects 9 years after baseline for the initial cohort and 5 years after baseline for a second cohort of African Americans. EVGGFP was ascertained at baseline and every year thereafter by mail, and on the half-year by telephone. Data were available 93% of the time. We imputed missing values by interpolation whenever there was a known value before and after the missing value, bringing the percent complete to 97%.

The ACQUIP study¹² was a multicenter trial to determine whether the outcomes of health care

are improved when primary care providers have access to systematic assessments of their patients' general and condition-specific health and function. Data, collected by mail, included EVGGFP and the SF-36 at baseline (T0) and 1 year later (T1). Because one study goal was to determine whether patients would provide this information on a regular basis, there were no special efforts to increase response. In this initial subset of the data, 40% of those surveyed responded at T0 ($n = 15,968$). The current research study is based on data from 9,844 veterans who responded at T0 and either had a T1 response ($n = 9495$) or died ($n = 349$). Those who responded at T0 but not at T1 and who were alive at T1 ($n = 5795$) were significantly younger than those who responded, but the T1 responders and nonresponders did not differ significantly in their physical component scores (PCS) at T0. Because the study was conducted in primary care clinics in Department of Veterans Affairs Medical Centers, only 3.5% of the subjects were women.

Respondents at T0 also reported on the presence of 23 baseline health problems: arthritis, coronary artery disease, cancer, CHF, chest pain, COPD, depression, diabetes, drug abuse, heartburn, hypertension, HIV-AIDS, kidney or liver problems, MI, osteoporosis, pneumonia, enlarged prostate, PTSD, seizures, stroke, ulcers, or thyroid problems. The number of people with a problem varied from 47 for HIV-AIDS to 9430 for "one or more health problems".

Strategies for Handling Deaths

Diehr et al¹³ reviewed several strategies for handling death. One promising strategy was to replace the observed HRQOL value, X , with Y such that Y is (approximately) the probability that the person will be "healthy" in the future, based on X . This strategy has been applied to EVGGFP, but not to the SF-36. In this paper we review the values derived for the EVGGFP categories, using the CHS data, to attempt replication of the previous transformation values, and consider varying lengths of follow-up. We also develop transformations for the SF-36, with special emphasis on the PCS.

In the earlier paper, based on volunteers older than age 64 in a trial to evaluate a health promotion program,¹³ 96% of those in excellent health at T0 were "healthy" (defined as being in excellent,

very good, or good health) 2 years later at T1. For those initially Very Good, 93% were healthy at T1; for Good, 76%, for Fair, 35%, for Poor 19%. A new variable was created, Prob(healthy), and set to the probability of being healthy at T1, multiplied by 100. Prob(healthy) is thus set to 96 if the person is in Excellent health, 93 if Very Good,, and 19 if Poor. A person who was dead in a particular time period should receive a value of zero, because he has no probability of being healthy 2 years later. Analyses can then be performed using Prob(healthy) as the dependent variable, which is known for the deaths. The interpretation of a change in Prob(healthy) is the change in the probability of being healthy in the future (multiplied by 100). This transformed variable has been used in several studies.^{14, 15, 16}

It is simple to provide a transformation value for each level of EVGGFP. For the PCS, which has a more continuous scale, we used logistic regression to estimate the probability of being "healthy" (a 0/1 variable) at T1 as a function of the PCS at T0. The logistic regression equation is: $\text{logit}(\text{healthy}) = a + bX$ (where X is the PCS value). We then used the resulting equation to transform all values of X to Y , as

$$Y = \frac{\exp(a + bX)}{1 + \exp(a + bX)} \times 100 \quad (1)$$

The dead are assigned the value of zero. There is a complete example in Appendix 1.

Definitions of "Healthy"

For EVGGFP, we defined healthy as Excellent, Very Good, or Good health (not Fair, Poor, or Dead). For the PCS, we explored three different definitions of being healthy at T1: (1) Being alive; (2) reporting EVGGFP as Excellent, Very Good, or Good; and (3) having a PCS value in the top 75% of the reference population values for men aged 65 and older (this value is 33.48 for the PCS).¹⁷ We chose the top 75% rather than the top 50% or 25% because it classifies most people in the PCS reference population¹⁷ as healthy, which is comparable to the percent in Excellent, Very Good, or Good health in CHS (also approximately 75%). We refer to these three PCS transformations as P-ALIVE, P-EVGG, and P-TOP75%.

Analysis

For EVGGFP, we used the CHS data to estimate the probability of being healthy in 2 years, 1 year, and 6 months conditional on the original value of EVGGFP. We also calculated 1-year probabilities for men only, using only data collected by mail, to compare CHS with ACQUIP. We then compared the resulting transformation values.

For the SF-36 analysis we used logistic regression to estimate the three transformation equations (for P-ALIVE, P-TOP75%, P-EVGG) for PCS, MCS, and the eight SF-36 subscales. For PCS only, we compared the distributions of P-ALIVE, P-TOP75%, and P-EVGG at T0 and T1, with the untransformed PCS values.

For each variant of PCS we noted that the mean value decreased in time and was lower for people with each health problem than for people without that problem. Although a formal evaluation of these transformed values is premature, we examined the power of each measure to discriminate between people with and without each baseline health problem, and to detect significant change in time within each subgroup of patients. One way to evaluate this performance is to estimate the effect size, which is a difference between two means divided by a SD. Because a *t*-statistic is essentially a measure of effect size multiplied by the square root of the sample size, and in addition has an associated test for statistical significance, we used *t*-statistics for this assessment.

To examine discriminatory power we compared the mean value for persons with and without each health problem using an unpaired *t*-test, and reported the percent of comparisons in which the groups differed significantly. We also ranked the four health measures by the size of the *t* statistic, for each health problem, and reported the average rank over all health problems. To determine whether all measures were equally powerful at detecting change, we calculated a paired *t*-test of the measure at T0 versus T1, within each health problem subset, and reported the number of significant differences and the mean ranks of the *t* statistics.

Findings

Transformations for EVGGFP

Table 1 shows the transformations for EVGGFP based on the different data sets. The first column

shows the original values that were proposed.¹³ For example, 96% of those in excellent health were healthy 2 years later, suggesting that people in Excellent health receive the transformed value of 96.

CHS contains data on 69,422 2-year health transitions (change in EVGGFP from year 0–2, from year 0.5–2.5, ..., and from year 7–9), which we used to estimate the probability of being healthy in 2 years based on the original health state. The estimated probabilities, shown in Column 2, are similar to those in column 1. Columns 3 and 4 contain CHS data for 1-year and 6-month transitions, respectively. The tabled values were not strongly associated with the length of follow-up.

Because CHS is population-based, and the number of available transitions very large, it seems preferable to choose values based on CHS. It is also desirable to have values that are easy to remember. A transformation of EVGGFP to values 95, 90, 80, 30, 15, and 0 for death is reasonably representative of the data in columns 2 to 4 and easier to remember than the values in Column 1. The performance of the column 1 transformation of EVGGFP has been shown to be better than some other coding approaches.^{13,16} We do not repeat such analyses here.

We also wanted to know if the ACQUIP data were generalizable. Column 5 of Table 1 shows CHS 1-year transition data for males only, using only the data collected by mail. Values from ACQUIP are in Column 6 and are fairly similar to those in Column 5, even though the CHS and ACQUIP populations were different.

Transformations for the SF-36

We used ACQUIP data to study the SF-36. The first 2 columns of Table 2 describe the untransformed PCS at T0 and T1, respectively. The mean is approximately 32, which is low (in Ware's reference population, the mean is 42, and the 25th percentile is 33.5¹⁷). The mean at T0 was slightly higher (32.6) if the 349 people who were dead at T1 were removed (data not shown).

To transform the PCS values, we estimated the probability of being healthy at T1 from the PCS score at T0, using the 3 definitions of "healthy." Figure 1 shows the percent who were healthy (in the top 75% of the reference group¹⁷) at T1, as a function of PCS at T0 (PCS0). The percent healthy a year later was close to 0 for PCS0 below 20, and

TABLE 1. Probability of Being Healthy at T1 Estimated From Health at T0

Category	1-Diehr ¹³	2-CHS	3-CHS	4-CHS	5-CHS Male Mail	6-ACQUIP Male
T1-T0	2 Years	2 Years	1 Year	6 Months	1 Year	1 Year
Excellent	96	93	95	97	96	94
Very good	93	89	92	93	93	94
Good	76	74	79	81	79	71
Fair	35	30	32	38	26	19
Poor	19	12	14	14	7	2
Dead	0	0	0	0	0	0
Mean age	73	76	76	76	77	67
No. of transitions	2,374	69,422	67,198	71,009	12,451	8,323

was close to 100% for PCS0>50. We could transform the PCS by ranges (eg, assigning persons with PCS scores of 40–49 the new value of 83). However, to take advantage of the continuous nature of the PCS, we used logistic regression to predict being healthy at T1 from the PCS at T0.

The logistic regression equations for PCS, MCS, and the 8 subscales are in Appendix 1. For example, for the PCS, the logit of P-TOP75% is estimated as $-6.57 + 0.1866 \cdot \text{PCS}_0$. Figure 1 suggests that a person with a PCS score of 45 at T0 would have a transformed score of approximately 83; the value from the regression equation is 86.2. (See Appendix 1). As always, Dead would be coded as zero because there is no possibility of becoming healthy.

We used the regression equations to create P-TOP75%, P-ALIVE, and P-EVGG, each of which represents the probability of being healthy 1 year in the future based on the PCS at T0, but using different definitions of healthy. Table 2 shows the distribution of the original and transformed variables. For example, PCS at time 1 has mean of 32.2, and takes on values from 0 to 69.7. It is approximately bell-shaped, with a mode at 20 to 29. The sample size for PCS1 is only 9,495 because the PCS has no value for death.

Table 2 next shows the distribution of P-ALIVE. Mean P-ALIVE is approximately 90, meaning that we expect approximately 90% of these people to be healthy (alive) a year later. At T0 everyone is alive, and 99.5% of the scores are above 80. At T1 the 349 people who died have the value of 0, and 96% have a value of 80 or higher. The separation between being dead and the lowest living state is large, and it is likely that P-ALIVE would perform like a binary variable denoting dead/alive. The

standard deviations at T0 and T1 are different because there were no deaths (zeroes) at T0.

Table 2 also shows the distribution of P-TOP75% at T0 and T1. The distribution is U-shaped, with modes near 0 and near 100. At T1 there are a large number of values near zero, which include the 349 deaths but also many living persons. Approximately 28% have values between 20 and 80, and 26% have values of 80 or higher. Due to its bimodal distribution, P-TOP75% might perform similarly to a binary variable that denoted healthy/not healthy.

Finally, Table 2 shows the distribution of P-EVGG, where healthy is defined as having Excellent, Very Good, or Good health. The distribution is again bimodal, but approximately 50% of the respondents had values between 20 and 80. For these reasons, P-EVGG might behave more like a continuous measure than P-ALIVE or P-TOP75%.

Relations among Variables

We tested whether the various measures could discriminate between people with and without each of the 23 health problems, at T0 and at T1. We also tested whether each measure could detect a significant change in time within each health problem sub-group. The results are summarized in Table 3.

We performed a t-test comparing people with and without each of the 23 health problems (For example, comparing the people who reported diabetes with those who did not, for all 9,844 persons). We also compared people with “any problem” with those with “no problem.” At T0, the

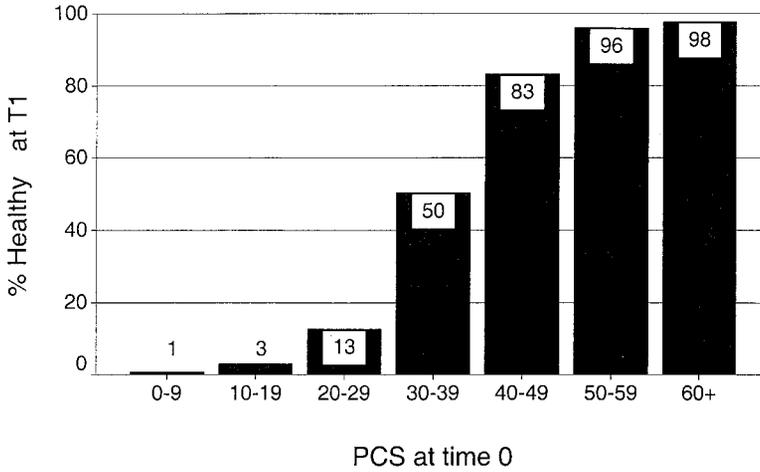


FIG. 1. Percentage healthy one year later—by PCS. Healthy: PCS >33.48

PCS, P-EVGG, and P-ALIVE detected significant difference ($P < 0.05$) for 22 of the 24 comparisons, whereas P-TOP75 detected 21. At T1, P-ALIVE detected only 18 differences, whereas the others

detected 22. The third column shows results of the 26 paired t-tests comparing T0 to T1, within each of 23 health problem groups plus “no problem”, “at least 1 problem” and “all.” P-ALIVE was always

TABLE 2. Distribution of Original and Transformed Health Measures*

	PCS	PCS	P-Alive	P-Alive	P-TOP75	P-TOP75	P-EVGG	P-EVGG
Measure								
Time	0	1	0	1	0	1	0	1
N	9844	9495*	9844	9844	9844	9844	9844	9844
Mean	32.3	32.2	92.7	89.3	41.5	40.3	44.7	43.1
SD	14.0	14.6	4.2	17.7	36.7	37.5	28.7	30.1
Min	0	0	75.7	0	0.1	0	2.3	0
Max	68.4	69.7	98.9	99.0	99.8	99.8	97.6	97.9
Distribution percentage								
0-9	2.5	3.7	—	3.5	31.9	35.4	8.3	13.7
10-19	19.7	20.2	—	—	11.8	10.7	19.3	17.4
20-29	26.8	25.9	—	—	7.1	6.8	13.8	13.0
30-39	21.2	19.5	—	—	5.5	4.6	10.0	9.4
40-49	14.7	15.0	—	—	4.9	4.0	8.7	7.1
50-59	13.3	13.9	—	—	4.4	4.1	7.4	7.0
60-69	1.8	1.9	—	—	4.1	4.2	6.5	6.1
70-79	—	—	0.5	0.9	4.6	4.1	7.2	7.3
80-89	—	—	25.8	25.5	6.2	6.5	11.0	10.7
90-100	—	—	73.8	70.1	19.4	19.5	7.9	8.3

*The PCS is undefined for deaths.

PCS is the untransformed PCS; P-Alive, P-TOP75, and P-EVGG are the estimated probabilities of being healthy, where health is defined respectively as being alive, being in the top 75% of the reference population, or being in excellent/very good/good health.

TABLE 3. Number of Significant *t* Test Results

Variable	Discrimination* at T0 (of 24)	Discrimination* at T0 (of 24)	Change [†] from T0 to T1 (of 26)
PCS	22	22	15
P-EVGG	22	22	21
P-ALIVE	22	18	26
P-TOP 75%	21	22	14

*No. of significant *t* tests (of 24 - $P < 0.05$) comparing people with each health condition to those without it.

[†]No. of significant paired *t* tests (of 26 - $P < 0.05$) comparing people at T0 and T1, in 26 health condition groups.

significant and P- EVGG was next, followed by PCS and then P-TOP75.

Most of the *t*-tests were statistically significant because the sample was so large. As a second comparison, we ranked the size of the four *t*-statistics for each health problem, and then calculated the mean rank over all the *t*-statistics. (This is similar to ranking the effect sizes, as noted above). Table 4 shows the results of these comparisons. For discrimination at T0, PCS had the lowest mean rank (corresponding to the highest *t*-values) and P-EVGG was second; for discrimination at T1 (where PCS had fewer observations) P-EVGG was first and PCS second. For *t*-tests measuring change, P-ALIVE was first, followed by P-EVGG.

Discussion

Because the new measures performed well in discriminating among groups and in detecting change within groups, the transformation method appears to be a viable approach. Some questions remain regarding its use.

Appropriateness of the Transformation Method

The method proposed here makes sense only when death is clearly "not healthy" on the scale of interest. (It is not necessary to assume that dead is the worst possible state.) This assumption is reasonable for the PCS, and for other measures of physical health. It is questionable for such measures as mental health or pain, because death could be considered peaceful and pain free. We have provided transformation equations for all SF-36 subscales, but researchers should consider

carefully the appropriateness for their specific application.

Definition of Healthy

The definition of "healthy" must be compatible with the dimension that is being measured. If a measure has a traditional threshold for "healthy," that threshold should be used. A measure such as P-Top75% would always be appropriate, because it is by definition on the same dimension. However, P-EVGG seemed to perform better than P-Top75% for PCS, in that it was more powerful (Table 3 and Table 4) and had a more continuous distribution (Table 2). Being in Excellent, Very Good, or Good health is probably a reasonable way to define "healthy" for physical and mental health, but may be less appropriate for social function or pain. Interventions that move people from "unhealthy" to "healthy" will usually show greater increase than those that move people from, say, poor to fair health. To evaluate an intervention whose goal was to move people from poor to fair health, a lower threshold for "healthy" might be preferred. The definition of "healthy" would ideally fit the goals of the intervention being evaluated. More work is needed in this area.

Other Features of Transformed Variables

Although our main goal was to provide a value for death, the transformed variables have other desirable properties as well. They have a ratio scale because they have a true zero, and a person with 25% probability being healthy in 2 years has half the probability of a person with 50% probability. The method also allows us to assign a zero for

TABLE 4. Mean Ranks of *T*-statistics

Variable	Discrimination* at T0 (of 24)	Discrimination* at T1 (of 24)	Change [†] from T0 to T1 (of 26)
PCS	1.85	2.00	3.48
P-EVGG	2.13	1.69	2.04
P-ALIVE	2.94	3.79	1.00
P-TOP 75%	3.08	2.52	3.48

*Average rank of *t* statistics comparing people with each health condition to those without it, over 24-hour conditions. (1 = most significant).

[†]Average rank of paired *t* statistics comparing people at T0 and T1, over 26 health condition groups.

some “states worse than death”¹⁸ or for some persons too sick to respond, if they are known to have zero probability of being healthy in the future. As such, this method provides an attractive adjunct to standard methods of analyzing endpoints with untransformed variables.

If the transformed variables are plotted over time, the area under the curve is (approximately) the years of healthy life (YHL), a useful summary measure for longitudinal data.¹⁴ Appendix 2 explains some details about the calculation of the area and its interpretation. We refer to this quantity as YHL rather than as quality-adjusted life years (QALY), because QALYs are usually based on preference-ratings, and maximizing QALYs can be thought of as maximizing utility-weighted population health. Maximizing YHL will maximize the amount of time the population is “healthy” but this may not agree perfectly with a preference-rated quantity.

This procedure will account for data missing because of death. It may be advisable to impute other missing data, as well, to be sure that people who die are not over-represented in the analysis.

Interpretation

The transformed variable may be interpreted strictly as the estimated probability of future health, or perhaps more generally as a measure of health status. As an example using the strict construction, we found that people with no baseline problems had an average P-EVGG of 70, whereas those with congestive heart failure (CHF) averaged 27 (data not shown). We thus expect approximately 70% of the former group to be healthy a year later, but only 27% of the latter. Having CHF is associated with a 43% point lower chance of future health. (A more detailed analysis would of course control for age and baseline

comorbidities). Changes in these measures can be thought of as the change in the probability of future health.

Using a more liberal construction, we can think of the transformed measure as a general measure of health, where 0 is death and 100 is perfect health. This operationalization conforms to the definition of health offered by Patrick et al¹⁹ that health is one’s functional status at a point in time and the possibility of health in the future. We can refer to one person or group as being twice as healthy as another person or group. Average values of most HRQOL measures do not permit such interpretations.

The transformations do not provide the best possible estimates of a person’s probability of being healthy in the future, because information about age, sex, and other variables would also help to predict future health.²⁰ Our goal here was to find a simple transformation, based only on the original score. Analyses using the transformed variables should usually include age and other covariates.

How Would the Transformed Measures be Used for Decision Making?

The transformed scales have many properties in common with preference-rated or utility scales. Decisions could be made that would increase the additional years of healthy life, estimated from the area under the curve. Such decisions might not maximize utility, because people might not value a 50% chance of future health twice as much as a 25% chance. A person with no chance of being healthy in the future contributes a value of 0 in future years of healthy life, but that state could still be preferred to death. A study whose primary goal was to maximize utility would have chosen a preference-rated HRQOL instrument and the proposed transformations would not be necessary.

Should Investigators Calculate Their Own Transformations?

Investigators could calculate transformations for their own use on their own data, if their data sets are reasonably large. To provide a general transformation that others would use, the data set would need a large range of health values, with deaths completely ascertained. Devising study-specific values means that results would not be comparable to results of others, and a common convention for coding variables is preferable. If the variable of interest is EVGGFP, and the population is older adults, we recommend using our transformation values: 95, 90, 80, 30, 15, and 0.

The SF-36 transformations, however, require replication because the VA population is sicker and older than the general population and includes few women, and the response rate was low. However, regression analysis does not require a representative sample from the population of interest; it requires only that sampled persons with a certain value of PCS be similar to other persons with the same value. This assumption may hold, based on the similarity of the ACQUIP and CHS results in Table 1; if so, the regression equations in Appendix 1 will be acceptable. The relatively poor health of persons in the ACQUIP sample is a strength, because it allowed us to estimate transform actions for low values of the PCS.

Length of Follow-up

For EVGGFP we examined 6 month, 1-year, and 2-year follow-up. It is not clear what the ideal length should be. If the follow-up period is short, then most people will be in the same health state at T0 and T1, which would result in the transformed variable being essentially a binary variable. For example, the states Excellent, Very Good, or Good would all have values near 100 because most people in them would remain healthy in the short period, and the other states would all have values near 0 because most people in them would remain unhealthy. If the follow-up period is too long, the predictive power of baseline health may attenuate, and the probability of future health would become similar for all baseline states. In the longest run, when everyone is dead, the levels of health at T0 will not discriminate at all and the transformed values for every state will be 0. Fortunately, the various probabilities for EVGGFP, shown in Table

1, were not sensitive to the length of follow-up. If one were developing new transformations for a younger population one might prefer longer follow-up times to ensure that there were enough changes. For a sick population, a shorter follow-up time might be better.

How to Use These Results

In a new study that has measured EVGGFP or the SF-36 over time, and that has ascertained deaths, the investigators would create a new variable by transforming the available data and set the new variable to zero for all persons who were dead at the time when the measure would have been made. Health of those who had no hope of future health might also be set to 0. The investigators would probably impute other missing data as well and adjust for important covariates in their analyses. Analysis of the new variable would proceed in the same manner as would have been appropriate for the untransformed PCS, except that results will be generalizable to the original population, not just to the survivors. Such measures as Years of Healthy Life can also be computed and compared among groups. If the new study was based on a measure other than the SF-36, the methods of this paper might be used to develop a transformation for that measure.

Conclusion

Studies that omit people who die can be generalized only to a healthier subset of persons (survivors), which cannot be determined prospectively, and is often not the subgroup of interest. If investigators are not using a measure that explicitly incorporates mortality, we recommend using the transformed HRQOL variables in longitudinal studies when deaths occur, to show whether the important results are the same with or without the deaths. Reporting both the original and the transformed measures would be appropriate, because experienced researchers may be able to interpret what a difference of (say) 5 PCS points means in their setting. The new measures, however, are also interpretable, and may be easier to understand for people inexperienced with the SF-36. Transformation values developed from different databases and from younger populations are of future interest.

References

1. **Patrick DL, Erickson P.** Health status and health policy: quality of life in health care evaluation and resource allocation. New York: Oxford University Press; 1993.
2. **Anderson JP, Kaplan RM, Coons SJ, et al.** Comparison of the Quality of Well-being Scale and the SF-36 results among two samples of ill adults: AIDS and other illnesses. *J Clin Epidemiol* 1998;51:755-762.
3. **Brazier J, Usherwood T, Harper R, et al.** Deriving a preference-based single index from the UK SF-36 Health Survey. *J Clin Epidemiol* 1998;51:1115-1128.
4. **Fryback DG, Lawrence WF, Maratin PA, et al.** Predicting Quality of Well-Being Scores from the SF-36: results from the Beaver Dam Health Outcomes Study. *Med Decis Making* 1997;17:1-9.
5. **Krause NM, Jay GM.** What do global self-rated health items measure? *Med Care* 1994;32:930-942.
6. **Gold M, Franks P, Erickson P.** Assessing the health of the nation: the predictive validity of a preference-based measure and self-rated health. *Med Care* 1996;34:163-177.
7. **Idler EL, Benyamini Y.** Self-rated health and mortality: a review of twenty-seven community studies. *J Health Soc Behav* 1997;38:21-37.
8. **Ware JE.** SF-36 Health Survey Manual and Interpretation Guide. Boston, MA: The Health Institute, New England Medical Center; 1993.
9. **Fried L, Nemar O, Enright P, et al.** The cardiovascular health study: design and rationale. *Ann Epidemiol* 1991;1:263-276.
10. **Ives G, Fitzpatrick A, Bild D, et al.** Surveillance and ascertainment of cardiovascular events: the cardiovascular health study. *Ann Epidemiol* 1995;5:278-285.
11. **Tell GS, Fried LP, Hermanson B, et al.** Recruitment of Adults 65 Years and Older as Participants in the Cardiovascular Health Study. *Ann Epidemiol* 1993;3:358-366.
12. **Bradley KA, McDonnell M, Bush K, et al.** The AUDIT alcohol consumption questions: reliability, validity, and responsiveness to change in older male primary care patients. *Alcohol Clin Exp Res* 1998;22:1842-1849.
13. **Diehr P, Patrick D, Hedrick S, et al.** Including deaths when measuring health status over time. *Med Care* 1995;33:AS164-AS172.
14. **Diehr P, Psaty B, Patrick D.** Effect size and power for clinical trials using years of healthy life as the primary endpoint. *Stat Med* 1997;16:1211-1224.
15. **Diehr P, Patrick DL, Bild DE, et al.** Predicting future years of healthy life for older adults. *J Clin Epidemiol* 1998;51:343-353.
16. **Diehr P, Patrick DL, Burke G, et al.** Survival versus years of healthy life: which is more powerful as a study outcome? *Control Clin Trials* 1999;20:267-279.
17. **Ware JE, Kosinski M, Keller SD.** SF-36 Physical and mental health summary scales: a users manual. Boston, MA: Health Assessment Lab, New England Medical Center; 1994.
18. **Patrick DL, Starks HE, Cain KC, et al.** Measuring preferences for health states worse than death. *Med Decis Making* 1994;14:8-18.
19. **Patrick DL, Bush JW, Chen MM.** Toward an operational definition of health. *J Health Soc Behav* 1973 Mar;14:6-23.
20. **Diehr P, Patrick D.** Probabilities of transition among health states for older adults. Technical Report NO. 168. Seattle, WA: University of Washington; 2000.

Appendixes

Appendix 1

Logistic Regression Results for SF-36 Subscales

The appendix table shows the logistic regression coefficients for the three definitions of healthy for PCS and MCS and for the 8 usual subscales. Two definitions (being Alive and in E/VG/G health) are the same for all subscales. When healthy is defined as being in the top 75% of the reference group, the reference group was men 65 and older for the PCS and the MCS,¹⁷ and for the other subscales was men and women aged 65 to 75.⁸

The regression coefficients were taken from a logistic regression of "healthy" (defined in the

column heading) on the T0 value of the particular subscale (defined in the row name). For example, the logit of being in the top 75% of PCS at T1 as a function of PCS at T0 is estimated as $-6.57 + 0.1866 \cdot \text{PCS}_0$. As expected, all slopes are positive, indicating that people with better health at T0 are more likely to be healthy at T1.

These equations can be used to transform the original variable to a new variable that has a valid value for death. Consider a hypothetical person who had a PCS score of 45 at T0. Figure 1 suggests that his transformed score would be approximately 83. This person's P-TOP75% score at T0 is calculated using the following equation.

APPENDIX A. TABLE 1. Logistic Regression Coefficients

"Healthy"	P-Alive		P-E/VG/G		P-Top 75%	
	a	b	a	b	a	b
PCS	1.36	0.0496	-3.77	0.1089	-6.57	0.1866
MCS	1.97	0.0178	-3.44	0.0674	-6.66	0.1480
Role-emotional	2.38	0.0078	-1.28	0.0174	-2.34	0.0192
General health	1.80	0.0240	-4.14	0.0820	-3.59	0.0455
Mental health	2.22	0.0086	-3.03	0.0402	-4.06	0.0463
Bodily pain	2.16	0.0125	-2.17	0.0354	-3.08	0.0327
Physical function	1.81	0.0238	-2.28	0.0404	-3.42	0.0379
Role-physical	2.41	0.0150	-1.11	0.0265	-1.88	0.0176
Social function	1.94	0.0150	-2.62	0.0366	-2.75	0.0284
Vitality	2.06	0.0191	-2.71	0.0564	-2.94	0.0401

a = intercept; b = slope.

$$\begin{aligned}
 Y &= \frac{\exp(a + bX)}{1 + \exp(a + bX)} \times 100 \\
 &= \frac{\exp(-6.57 + 0.1866(45))}{1 + \exp(-6.57 + 0.1866(45))} \times 100 \\
 &= 86.14
 \end{aligned}
 \tag{2}$$

Appendix 2

Years of Healthy Life and Area Under the Curve. In a typical study, one measures health status at baseline and then "K" times thereafter.¹⁴ In our notation, at times t_0, t_1, \dots, t_K , during the study one measures the average health score, h_0, h_1, \dots, h_K . For simplicity, we let $t_0 = 0$ and assume that the measurements occur at regular intervals. The total length of the study is t_K . The simplest estimate of the area under the curve (AUC) comes from a trapezoidal or "connect the dots" strategy, with

$$YHL = AUC = \frac{\left(h_0 + h_K + 2 \sum_{j=1}^{K-1} h_j \right) t_K}{2 \bar{K}}
 \tag{3}$$

For example, in a pretest/posttest design that lasts 2 years ($K = 1$ and $t_K = 2$), AUC reduces to $h_0 + h_K$. To understand the interpretation of the AUC, first suppose that we have measured EVGGFP at each time and have recoded it in a simple way as "healthy yes/no," with Excellent/Very Good/Good set to 1, and Fair/Poor/Dead set to 0. The height of the curve at each time will be the proportion who were "healthy" at that time and the AUC will be the portion of the "K" years in which

people were healthy on average, or the years of healthy life. The curve will usually slope down to the right, and is similar to a survival curve. It is different, however, in that people who are not healthy at one time may be healthy later.

This system is straightforward but has the unfortunate characteristic that dead is treated as equivalent to fair and poor (all are coded as 0). If we use the transformed variable P-EVGG, as defined in the body of the paper, then each state will have a unique value. At each time, the mean health value is the mean probability of being healthy, which is also the estimated proportion who will be healthy in the future. The mean of the transformed variable thus has the same interpretation as the mean of "healthy yes/no" in the previous paragraph, except that it applies to the future rather than to the current time. That is, if the transformation was created using a follow-up period of 2 years, then the AUC would be the expected number of YHL the group experienced beginning 2 years after baseline (rather than at baseline).

The two estimates of YHL (based on healthy yes/no and on P-EVGG) would be approximately the same if the follow-up time used in creating the transformation was small. We noted above that the coefficients did not depend strongly on the length of follow-up, and so this may be a reasonable assumption. We have calculated YHL both ways in a variety of data sets and found them to be comparable, with P-EVGG perhaps a little more sensitive than healthy yes/no.¹⁶ It is probably reasonable to think of the AUC simply as a measure of YHL without being too concerned about the follow-up time used to compute the transformation values. When in doubt, YHL can be calculated both ways and compared.