

## **Modeling Health Status for Microsimulation**

Julie Topoleski  
julie.topoleski@cbo.gov  
U.S. Congressional Budget Office

Joyce Manchester  
joyce.manchester@cbo.gov  
U.S. Congressional Budget Office

The analysis and conclusions expressed in this paper are those of the authors and should not be interpreted as those of the Congressional Budget Office. This paper was prepared for the 2nd General Conference of the International Microsimulation Association, Ottawa, Canada, June 8-10, 2009. Many thanks to Patrick Bernhardt for his excellent research assistance.

## Modeling Health Status for Microsimulation

A large body of evidence indicates that an individual's educational attainment and earnings are strong predictors of health. But the relationship also goes the other way. Better health is associated with having more income, more years of education, and a higher level of well-being that translates into longer lives and fewer years with disabling conditions. When projecting the future well-being of a population at the individual level and forecasting the expected financial solvency of public pension and health programs, including health status as a factor that affects both demographic and economic outcomes is critical.

This paper describes the methods underlying health status projections in the Congressional Budget Office's (CBO's) long-term microsimulation model (CBOLT) for the United States.<sup>1</sup> CBO developed a microsimulation approach for analyzing the U.S. public pension plan known as Social Security and other long-term policy issues in order to provide the U.S. Congress with comprehensive analyses of the budgetary, distributional, and aggregate economic aspects of various policy choices. The microsimulation approach makes it possible to assess how policy affects individuals' benefits under current law and proposed alternatives.<sup>2</sup> Adding health status and additional

---

<sup>1</sup> These modules have been updated and extended from earlier work; see Amy Rehder Harris and John Sabelhaus, *How Does Differential Mortality Affect Social Security Finances and Progressivity?*, Congressional Budget Office Working Paper 2005-5 (May 2005).

<sup>2</sup> See, for example, Congressional Budget Office, *Long-Term Analysis of S.2427, The Sustainable Solvency First for Social Security Act of 2006* (April 2006); *Long-Term Analysis of the Liebman-MacGuineas-Samwick Social Security Proposal* (February 2006); and *Analysis of H.R. 3304, Growing Real Ownership for Workers Act of 2005* (September 2005).

health variables to the microsimulation model will ultimately result in a more powerful tool for policy analysis in the areas of both public pensions and health.

The methodological strategy in microsimulation is to generate realistic demographic and economic outcomes for a representative sample of the population, then apply tax and benefit rules to that sample in order to draw inferences about the effects of various policy options. A microsimulation model starts with individual data from a representative sample of the population and projects demographic and economic outcomes for that sample through time. In CBOLT, the basic demographic processes currently include fertility, educational attainment, marital transitions, marital partner assignments, and eventual death. Here we describe our attempts to add health status transitions to that list. The key economic processes are labor force participation, earnings, and retirement.<sup>3</sup>

Within CBOLT, the equations for each of the modules are designed to generate realistic patterns in the representative sample and to operate as part of an overall framework that includes macroeconomic and aggregate budgetary outcomes. The projections of health status, as discussed here, accept as given the basic demographic outcomes for the representative sample. Then, given an individual's age, sex, education, marital status, fertility history, and earnings history, we address the following question. How should health status best be projected?

---

<sup>3</sup> See Congressional Budget Office, *Projecting Labor Force Participation and Earnings in CBO's Long-Term Microsimulation Model* (October 2006); Josh O'Harra, John Sabelhaus, and Michael Simpson, *Overview of the Congressional Budget Office Long-Term (CBOLT) Policy Simulation Model*, Congressional Budget Office Working Paper 2004-1 (January 2004); Kevin Perese, *Mate Matching for Microsimulation Models*, Congressional Budget Office Technical Paper 2002-3 (November 2002); and Josh O'Harra and John Sabelhaus, *Projecting Longitudinal Marriage Patterns for Long-Run Policy Analysis*, Congressional Budget Office Technical Paper 2002-2 (October 2002).

This paper develops and compares three related approaches to forecasting health status for the representative sample. Using data from the U.S. Survey of Income and Program Participation (SIPP) matched with data from the U.S. Social Security Administration (SSA) on earnings and beneficiary type, we first observe relationships among health status and other demographic and economic variables at the individual level. We then estimate several different health transition specifications. First, we assume that health status transitions are related only to age, sex, and lagged health status. The outcomes in the microsimulation model show that assigning health status randomly, without regard for other demographic and economic characteristics, can lead to unreasonable results.

Omitting the correlation between health status and characteristics such as household earnings and education yields results that do not conform with actual outcomes. Next, we predict health status transitions as a function of age, sex, education, marital status, household earnings, and lagged health status. That approach preserves observed correlations between health status and socioeconomic variables. The results correspond to our expectations of the links between health status and social and economic well-being, but correlations between health status and mortality do not line up with actual observations in the data. Finally, we present some early results in which health status enters the mortality equation directly.

## **Previous Literature**

The links between health and demographic variables, such as educational attainment, and between health and economic variables, such as household income or earnings, are well

established at the individual or cohort level. For example, Deaton and Paxson (2001) looked at the relationship between health and economic status among American birth cohorts.<sup>4</sup> Adler and Newman (2002) cite socioeconomic disparities in health and suggest various pathways by which policy might help reduce income and educational disparities in an effort to reduce health disparities.<sup>5</sup>

Solid evidence supports the relationship between earnings or income and mortality as well (see Cutler, Deaton, and Lleras-Muney 2006), and work by Pappas et al. (1993) suggests that relationship may be growing stronger over time.<sup>6</sup> Among birth cohorts, income has a strong protective effect on mortality; the elasticity of mortality rates with respect to income is approximately -0.5. Estimates from the individual data in the National Longitudinal Mortality Study (NLMS) suggest that income is most highly protective against mortality in middle age, in the mid-40s for women and the mid-50s for men.

Deaton and Paxson also look at the respective roles of education and income in protecting health. In both cohort and individual data, they find that income and education are protective when analyzed separately. Taken together in the individual data, the effect of each is robust to allowing for the other, which is consistent with the view that both education and income promote health in different ways. Education makes it easier to use

---

<sup>4</sup> See Deaton, Angus S, and Christina Paxson, "Mortality, Education, Income and Inequality among American Cohorts," [NBER Working Paper No. 7140](#), May 1999, and in D. A. Wise, *Themes in the Economics of Aging*, Chicago: University of Chicago Press, 2001, pp. 129-65

<sup>5</sup> See Adler, Nancy E. and Katherine Neman, "Socioeconomic Disparities in Health: Pathways and Policies," *Health Affairs* 21, no. 2 (March/April 2002), pp. 60-76.

<sup>6</sup> See Cutler, David, Angus Deaton, and Adriana Lleras-Muney, "The Determinants of Mortality," NBER Working Paper 11963, 2006, and Pappas, George, Susan Queen, Wilbur Hadden, and Gail Fisher. "The Increasing Disparity in Mortality between Socioeconomic Groups in the United States, 1960 and 1986," *New England Journal of Medicine* 329, no. 2 (1993), 103-109.

and benefit from new health information and technologies and income makes life easier more generally, reducing stress and wear and tear, for example by having help to look after the children, or the money to buy first class travel.

The best-known attempt to model health status in the future is the Future Elderly Model, know as the FEM.<sup>7</sup> The FEM is a demographic-economic model framework of health care spending projections that uses microsimulations to answer questions about the effects of changes in health status on future health care costs. However, its focus is on the age 65 and over population because the purpose is to forecast Medicare spending. The model has detailed information on medical expenditures and specific diseases. In projecting future health status, demographic controls such as age, sex, and education play a role, but not economic controls such as lifetime household earnings.

More recently, RAND has developed the COMPARE microsimulation model.<sup>8</sup> The model was developed to project how households and firms would respond to changes in health care policy. The COMPARE model is not an annual microsimulation; it simulates current law and a new equilibrium under policy options. The model contains information on demographics, income, health insurance status, health status, and health expenditures for its simulated individuals. It also models the behavior of firms, and both individuals and firms make decisions following “rules”.

## **Data and Methodological Approach**

---

<sup>7</sup> See Goldman, Dana P. et al., “Health Status and Medical Treatment of the Future Elderly: Final Report,” RAND Technical Report 169-CMS, August 2004.

<sup>8</sup>See Giroi, Federico et al, “Overview of the COMPARE Microsimulation Model,” RAND Working Paper, January 11, 2009.

The merged SIPP-SSA data are available for 1984, 1990, 1991, 1993, 1994, 1996 through 1999, and 2000 through 2005. The health status measure in the data reflects self-reported health status, categorized by five choices: poor, fair, good, very good, or excellent. In the microsimulation model, we predict both the initial assignment of health status and the probability of a health status transition for each person using an ordered logit regression. We also use the SIPP data to determine marital status and educational attainment. The administrative data allow us to construct the lifetime earnings measure and to determine if an individual ever received Social Security Disability Insurance benefits.

We estimate separate equations for men and women using merged SIPP-SSA data based on an age-centering approach that allows flexibility in the relationship between health status and the underlying determinants across age groups.<sup>9</sup> The statistical approach used to estimate the health status transition equations allows maximum flexibility in the relationships between the explanatory variables and the outcome being predicted. The idea behind “age-centered regressions” is to estimate the equation – in this case an ordered logit equation for the probability of transitioning into a given health status – separately for each age and sex group. Given the limited available data, however, estimating the equations using data for just a single year of age yields imprecise results. The age-centered approach uses every observation for ages within a pre-set band around the specific age group being analyzed. The band on each side of a specific age group for the health transition equations is generally four years. For example, the sample used to estimate the health status transition equation for 25-year-olds actually includes everyone

---

<sup>9</sup> See John Sabelhaus and Lina Walker, *Econometric Flexibility in Microsimulation: An Age-Centered Regression Approach*, Congressional Budget Office Working Paper 2007-02 (January 2007).

between 21 and 29 years old. The sample for estimating the equation for 26-year-olds includes everyone between 22 and 30, and so on.

Based on those estimated relationships, each individual in the microsimulation model receives a probability of transition from the current year health status to a new one (or in the case of initial assignment, the probability of being in a given health status). The outcome of the initial assignment or the transition process depends on a combination of the deterministic probabilities and a random draw. For example, a given equation may generate a 10 percent chance that some outcome (for example, transitioning from excellent to very good health) will occur for a given individual. The model then generates a random number between 0 and 100 percent for that individual and compares it with the probability of that event to determine the actual outcome. In the example, if the random number is less than 10 percent, the individual will transition from excellent health one year to very good health in the next year. Given a large enough sample, the fraction of the population that receives a particular outcome should match the average probability of that outcome.

An additional step in the process of assigning health status ensures that in each year the “right” number of individuals fall into a given age, sex, and health status group. That additional step also eliminates some of the required calibration and micro-level variation (noise) typically associated with the approach described above. The procedure (referred to as “logit ranking”) creates a cumulative distribution using a logistic combination of an

individual's predicted probability and his or her random draw.<sup>10</sup> Given this cumulative distribution, a cut point can be assigned such that the desired number of individuals receive that outcome. Specifically, in a second loop through the sample, individuals with a logit rank value below that cut point are assigned the outcome.

Because health status, as measured in the survey data and assigned for purposes of this paper, is not a binary variable, the process must be repeated four times. Each time we exclude individuals who have already received a health status.<sup>11</sup> The process begins by assigning excellent health status and works down toward assigning the lowest health status.

Initially, we predict health status based only on age and sex using the age-centered approach described above. The equation that estimates the probability of transitioning from one health status to another includes only age, sex, and lagged health status. That initial step provides a simple starting point for the analysis. It captures the fact that the probability of transitioning to a poorer health state increases with age and depends on sex. However, it fails to capture certain strong relationships, such as those between health status and education or earnings, observed in the underlying data. Lower earners, for example, have a higher probability of remaining in or transitioning to poorer health states than higher earners at all ages. More highly educated individuals have, on average, better health status than less educated individuals.

---

<sup>10</sup> This approach is based on a suggestion from fellow microsimulation modelers at the U.K. Department for Work and Pension that was made after an exchange of ideas in 2006.

<sup>11</sup> The logit rank approach is used in the assignment of several outcomes in CBOLT. Health status is the only case where the assignment is not a binary outcome.

The next set of regressions controls for the effects of earnings and education. For individuals over age 23, we add marital status, a four category education variable, dummy variables representing the lifetime household earnings quintile for each individual, an indicator variable for whether an individual has ever received Disability Insurance benefits, and for transitions, lagged health status.

An individual's household earnings are defined as the combined earnings of a husband and wife. The measure does have limitations because to identify a spouse, we have to observe that spouse in the underlying survey data. In practice, we observe an individual for only a few years and assume that an individual has the same spouse until death. In reality, marital pairings are not static, and an individual's marital status at one point in time may not accurately reflect marital status at other points in time. The problem is particularly acute for individuals in the survey data at older ages. If a spouse is already deceased, we have no way to identify who that spouse was or to match the dead spouse's earnings records to the widow(er).

In the microsimulation, however, spouses can change over time, and household earnings reflect the lifetime household history. To reflect economies of scale, during years in which the individual is married, the sum of household earnings is divided by 2 raised to the power of 0.65.<sup>12</sup> "Lifetime" earnings are defined as earnings between the ages of 40

---

<sup>12</sup> In CBOLT, earnings are calculated based on wage and hours worked and refer to pre-tax dollars. The adjustment method described here is the one recommended by the National Academy of Science (NAS) in its 1995 Panel on Poverty and Family Assistance. The Social Security Administration microsimulation model (MINT model) also uses this method to adjust family income by family size.

and 59 for individuals over age 60. For younger individuals, lifetime earnings includes earnings for the previous 20 or fewer years beginning with earnings at age 23 and continuing through the current age minus one year.

The objective of the health modules is to predict realistic outcomes across the population through time. We evaluate that objective in this paper using tabulations that compare the projections with historical experience. Because health outcomes are related to other simulated outcomes, they should be evaluated both independently and in conjunction with other outcomes. Health status is related to mortality, disability, labor force participation, earnings, and other characteristics. If we cannot replicate the interrelationships observed in the underlying data in the simulation, we cannot expect to see reasonable projections of the important correlations in the microsimulation outcomes.

## **Results**

Not surprisingly, when health transitions depend only on age, sex, and lagged health status, no relationship exists between education and health status or between lifetime earnings and health status (see Figures 1 and 2). The random assignment of health is indeed obvious. Some relationship between health status and mortality does appear, however. In the base case, mortality rates depend on age, sex, marital status, education, and household lifetime earnings. The results show that, generally speaking, the higher the household earnings quintile, the lower the mortality rate. That result holds more strongly for men than for women (see Figure 3). A correlation also seems to exist between health

status and the probability of death. For both men and women, the average probability of dying increases as health status decreases.

The correlation between health status and probability of death is perhaps surprising given that health status is assigned based only on age, sex, and lagged health status and with no controls for health status in the mortality modules. The results presented in Figure 3 are average mortality rates from age 20 through age 99. What appears to be a correlation between average mortality rates and health status could be driven by the fact that older individuals are both more likely to die and more likely to be in poorer health. Indeed, looking at average mortality rates by ten-year age groups reveals no obvious correlation between health status and mortality (see Figures 4 and 5).

When health status depends on education and lifetime earnings, health status increases with both education and with household earnings quintile as expected. The more educated an individual is, or the higher the lifetime earnings of that individual's household, the more likely that individual is to be in better health (see Figures 6 and 7). The enhanced modeling approach also has some effect on the relationship between mortality and health. In any given age and sex group, individuals in higher-earning households have lower mortality risk than individuals in lower-earning households. That relationship, however, perhaps because it is not modeled explicitly, does not hold consistently for average mortality rates across all earnings quintiles and health status categories. Mortality rates are generally higher the lower health status is, and the negative relationship between mortality rates and lifetime earnings quintile is mostly consistent.

The poor health group for men does not display the expected relationship, however, and the expected relationship between mortality rates and lifetime earnings quintile is less clear for women across health categories (see Figure 8).<sup>13</sup>

The third approach to modeling microsimulation outcomes adds health status explicitly to the mortality regressions. Doing so has little effect on the distributions of health status by education and health status by earnings quintile, but it has a dramatic effect on mortality rates by health status (see Figure 9). Explicitly controlling for health status in the mortality regressions results in much lower average mortality rates for individuals in excellent and very good health, and it increases the average mortality rates for those in fair or poor health. For individuals in the two best health states, the relationship between mortality risk and earnings is preserved, but the reverse relationship exists for those in good, fair, and poor health. The effect of health on mortality appears to dominate the other effects, at least in the current specification.

When we compare the simulated mortality rates to the one-year mortality probabilities observed in the data by earnings quintile and health status, we find results that are similar but not precisely accurate (see Figure 10). For individuals in excellent, very good, and good health, observed and simulated mortality rates are close. Larger differences appear in the observed and simulated mortality rates for those in fair and poor health. The simulated results generally underpredict mortality in poor health states, and the difference is larger in the higher earnings quintiles.

---

<sup>13</sup> For a discussion of the relationship between lifetime earnings and mortality see Julian Cristia, *The Empirical Relationship Between Lifetime Earnings and Mortality*, Congressional Budget Office Working Paper 2007-11, August 2007.

A few explanations help to account for the differences in mortality rates between the matched survey data and the microsimulation. First, there may be underlying differences in death rates by age and sex in the two data sets. CBOLT uses mortality rates by age and sex as projected by SSA. To the extent that mortality rates in the merged SIPP-SSA data diverge from those projections, differences could evolve. The mortality equations in CBOLT determine who dies but do not determine the mortality rates. Second, CBOLT does a better job representing lifetime household earnings. Because CBOLT simulates marital transitions, household earnings are better represented than when we observe an individual's marital status at one point in time and have to assume that same pairing existed in all years.

## **Next Steps**

The work described here is a first step toward adding health status to an existing microsimulation model. Thus far, demographic and economic covariates affect health status, but health status only affects mortality directly. Additional work on refining the equations used in the microsimulation model will enable us to better replicate relationships between health status and other outcomes observed in the data.

Further, health could operate through many other pathways. Health status can affect disability status, labor force participation, earnings levels, fertility, and even marital status. An individual in poor health status, for example, is more likely to apply for Social Security Disability Insurance than an individual in excellent health. Additionally,

important correlations across many variables need to be preserved. The individual in poor health may be more likely to apply for Disability Insurance and also less likely to be in the labor force, more likely to be a low earner and also more likely to experience a higher mortality risk than an individual in better health. Capturing those correlations and integrating them into the microsimulation model is a necessary step before introducing health insurance status and health expenditures to the microsimulation.

## References

Adler, Nancy E. and Katherine Neman, "Socioeconomic Disparities in Health: Pathways and Policies," *Health Affairs* 21, no. 2 (March/April 2002), pp. 60-76.

Congressional Budget Office, *Long-Term Analysis of S.2427, The Sustainable Solvency First for Social Security Act of 2006* (April 2006).

Congressional Budget Office, *Long-Term Analysis of the Liebman-MacGuineas-Samwick Social Security Proposal* (February 2006),

Congressional Budget Office, *Analysis of H.R. 3304, Growing Real Ownership for Workers Act of 2005* (September 2005).

Congressional Budget Office, *Projecting Labor Force Participation and Earnings in CBO's Long-Term Microsimulation Model* (October 2006).

Cristia, Julian, *The Empirical Relationship Between Lifetime Earnings and Mortality*, Congressional Budget Office Working Paper 2007-11, August 2007.

Cutler, David, Angus Deaton, and Adriana Lleras-Muney, "The Determinants of Mortality," NBER Working Paper 11963, 2006.

Deaton, Angus S, and Christina. Paxson, "Mortality, Education, Income and Inequality among American Cohorts," [NBER Working Paper No. 7140](#), May 1999, and in D. A. Wise, *Themes in the Economics of Aging*, Chicago: University of Chicago Press, 2001, pp. 129-65.

Girosi, Federico et al, "Overview of the COMPARE Microsimulation Model," RAND Working Paper, January 11, 2009.

Goldman, Dana P. et al., "Health Status and Medical Treatment of the Future Elderly: Final Report," RAND Technical Report 169-CMS, August 2004.

Harris, Amy Rehder and John Sabelhaus, *How Does Differential Mortality Affect Social Security Finances and Progressivity?*, Congressional Budget Office Working Paper 2005-5 (May 2005).

O'Harra, Josh and John Sabelhaus, *Projecting Longitudinal Marriage Patterns for Long-Run Policy Analysis*, Congressional Budget Office Technical Paper 2002-2 (October 2002).

O'Harra, Josh, John Sabelhaus, and Michael Simpson, *Overview of the Congressional Budget Office Long-Term (CBOLT) Policy Simulation Model*, Congressional Budget Office Working Paper 2004-1 (January 2004).

Pappas, George, Susan Queen, Wilbur Hadden, and Gail Fisher. "The Increasing Disparity in Mortality between Socioeconomic Groups in the United States, 1960 and 1986," *New England Journal of Medicine* 329, no. 2 (1993), 103-109.

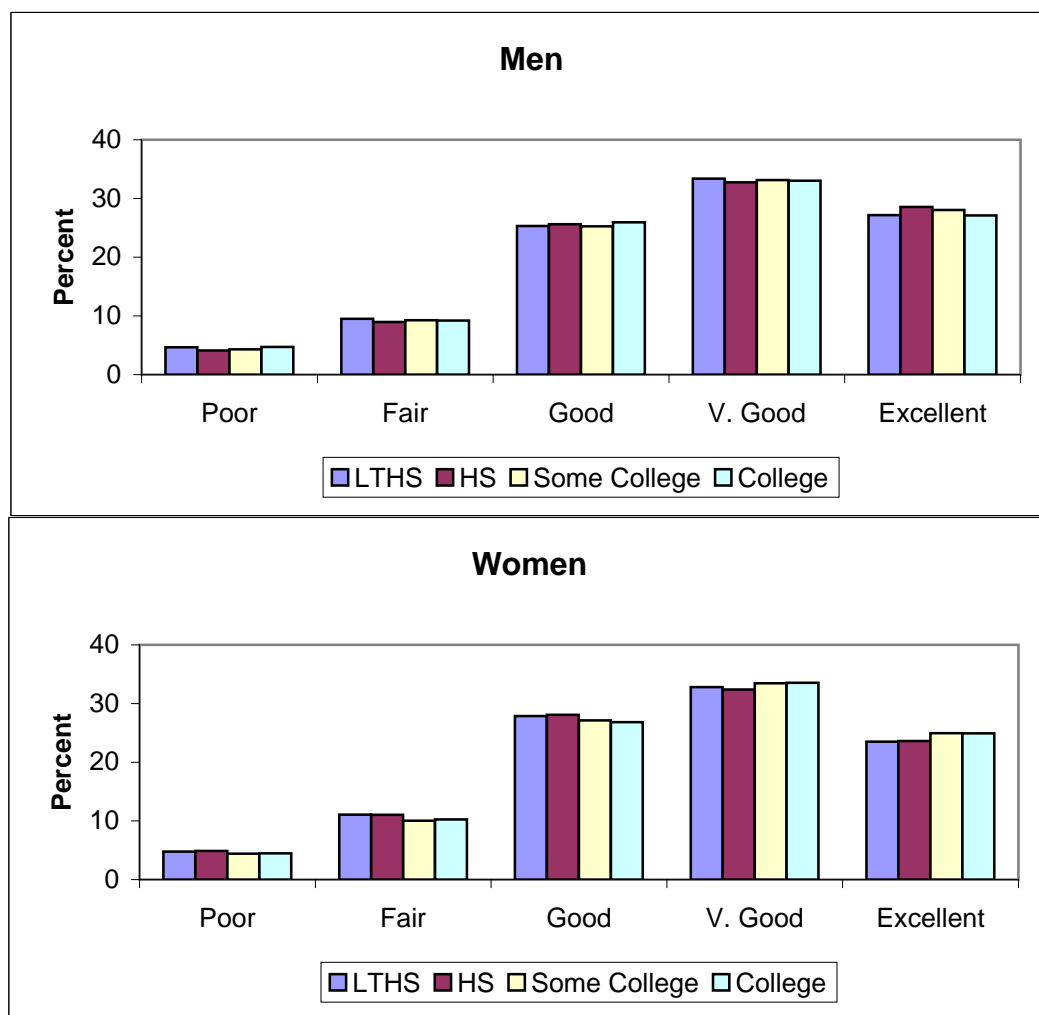
Perese, Kevin, *Mate Matching for Microsimulation Models*, Congressional Budget Office Technical Paper 2002-3 (November 2002).

Sabelhaus, John and Lina Walker, *Econometric Flexibility in Microsimulation: An Age-Centered Regression Approach*, Congressional Budget Office Working Paper 2007-02 (January 2007).

**Figure 1.**

**Health Status by Education, 2009**

(health status depends on age, sex, and lagged health only)

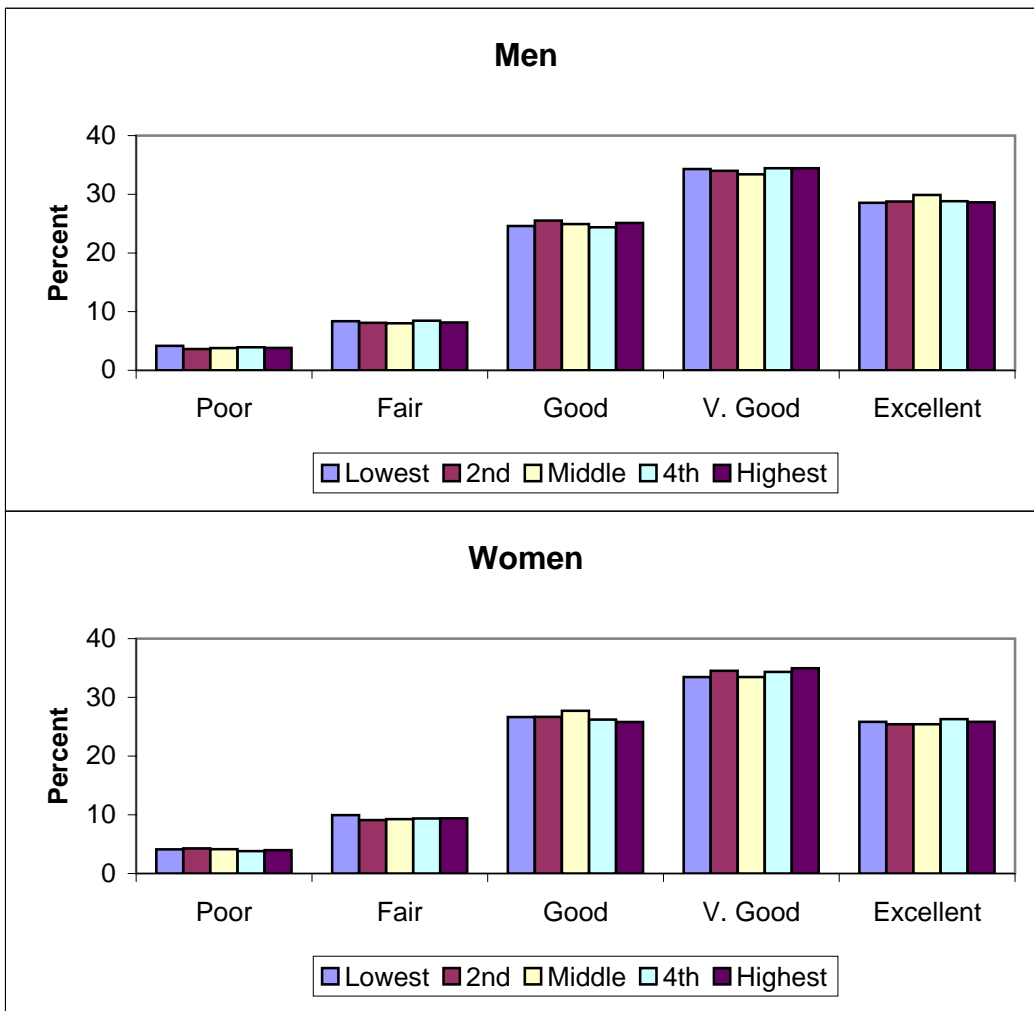


Note: Education groups sum to 100 percent.

**Figure 2.**

**Health Status by Household Earnings Quintile, 2009**

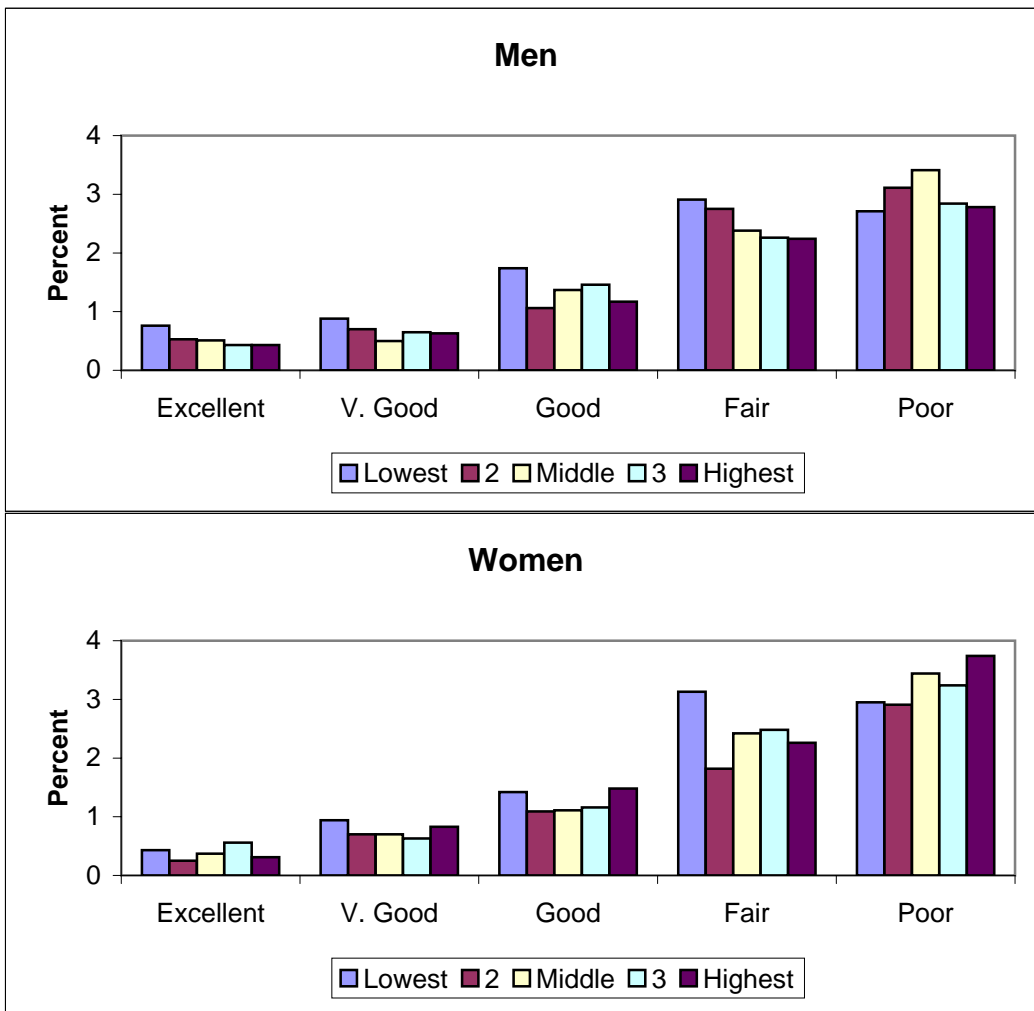
(health status depends on age, sex, and lagged health only)



Note: Earnings quintiles sum to 100 percent.

**Figure 3.**

Raw Average One-Year Mortality Rates by Household Earnings Quintile  
(health status depends on age, sex, and lagged health only)  
(Mortality does not depend directly on health status)



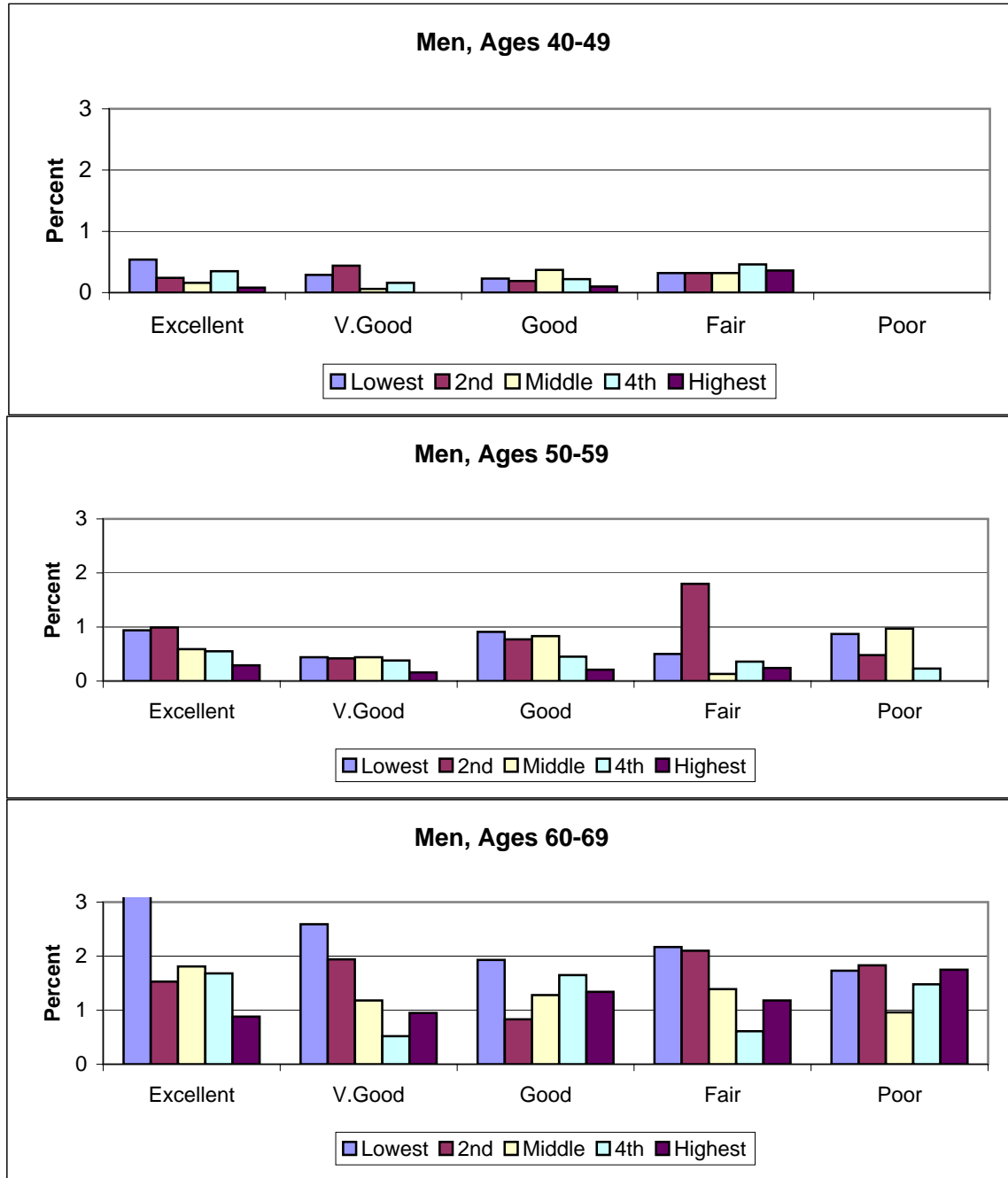
Note: One-year mortality rate defined as percentage of individuals in a given cell who die within 12 months of observation.

**Figure 4.**

**Raw Average One-Year Mortality Rates by Household Earnings Quintile, by Age Group**

(health status depends on age, sex, and lagged health only)

(Mortality does not depend directly on health status)



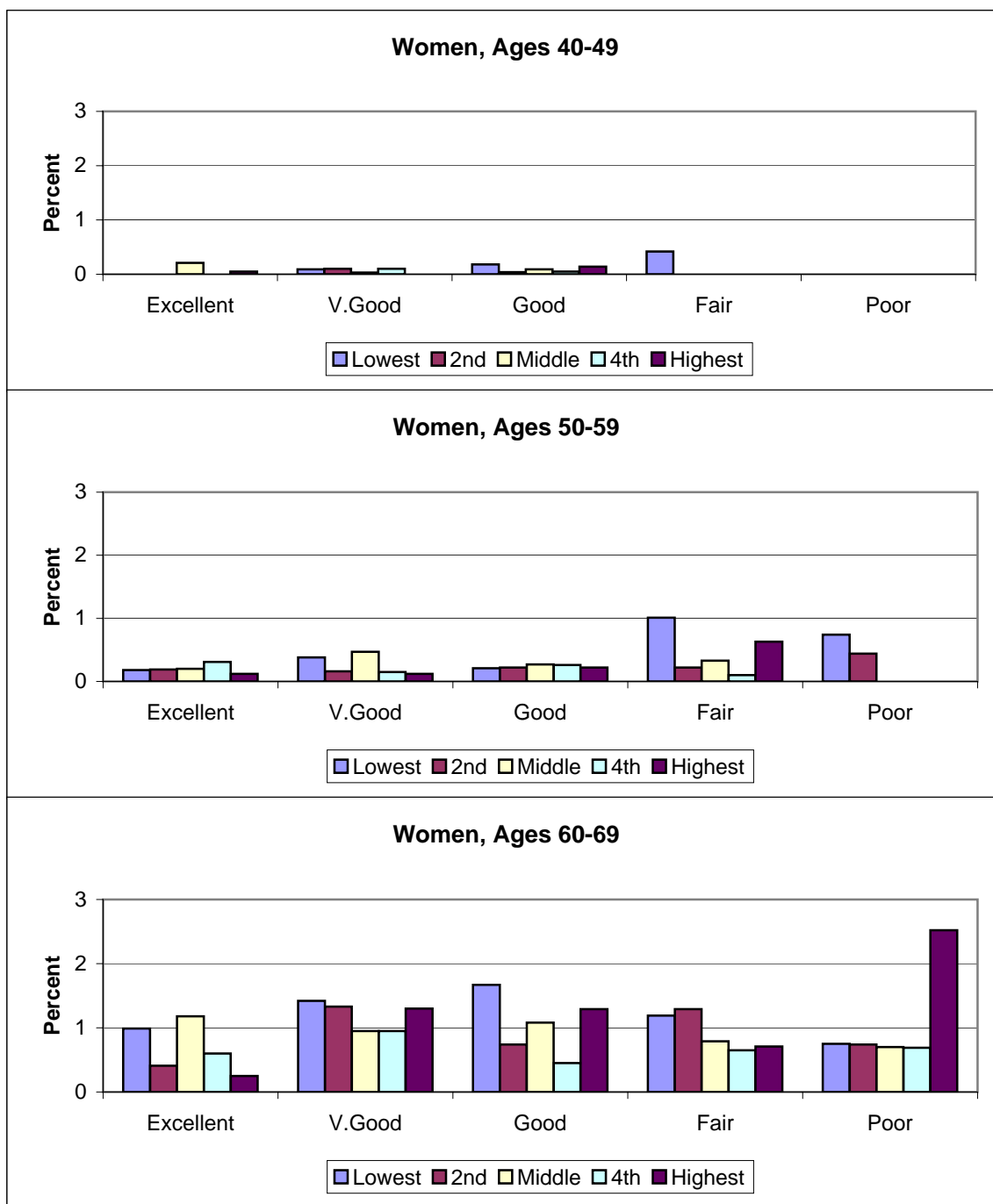
Note: One-year mortality rate defined as percentage of individuals in a given cell who die within 12 months of observation.

**Figure 5.**

**Raw Average One-Year Mortality Rates by Household Earnings Quintile, by Age Group**

(health status depends on age, sex, and lagged health only)

(Mortality does not depend directly on health status)

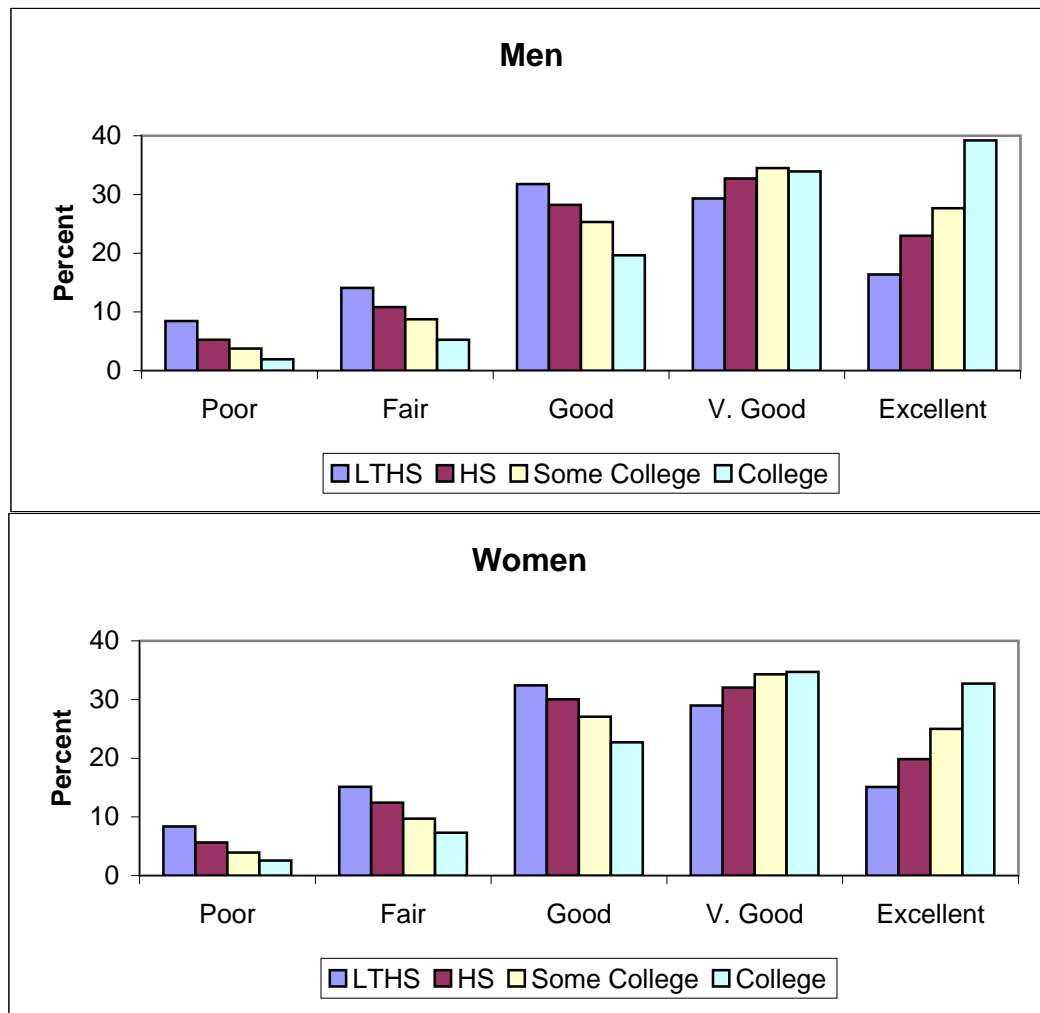


Note: One-year mortality rate defined as percentage of individuals in a given cell who die within 12 months of observation.

**Figure 6.**

**Health Status by Education, 2009**

(health status depends on age, sex, lagged health and other covariates)

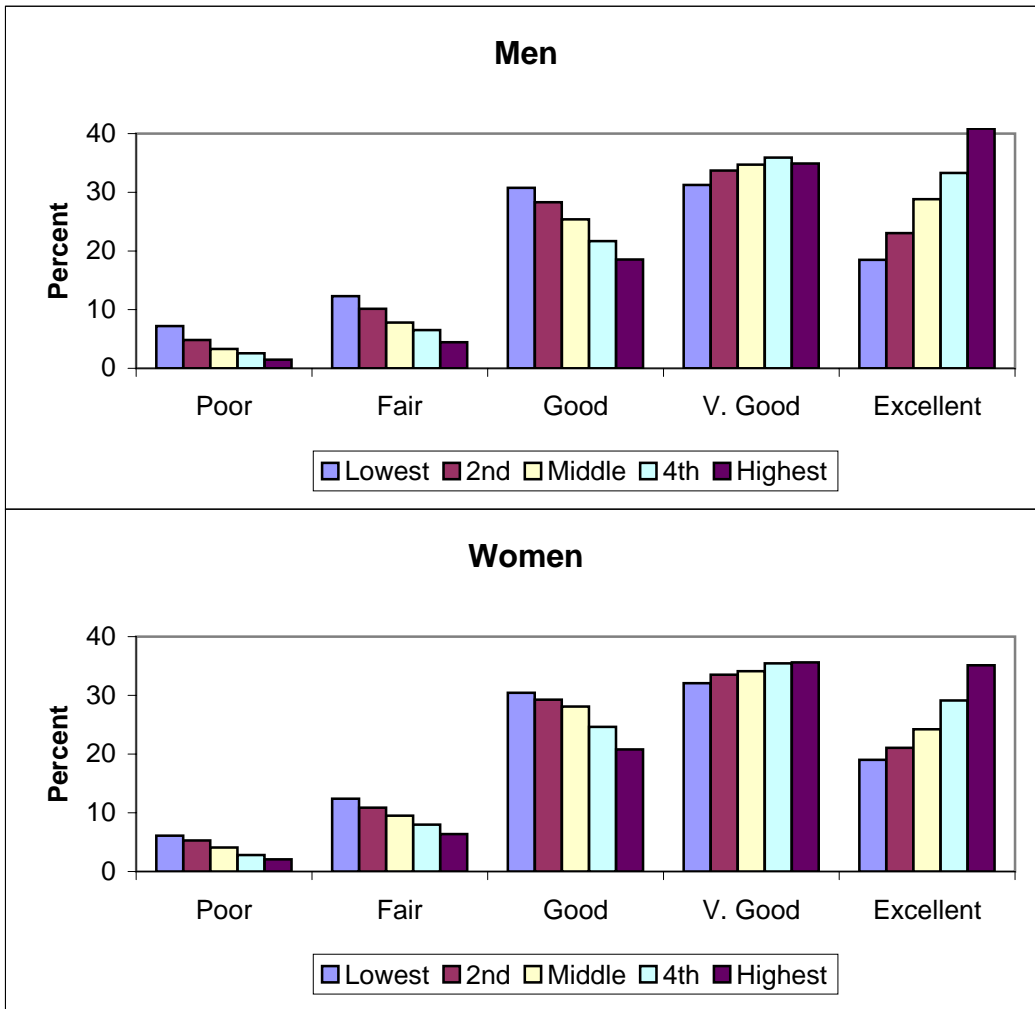


Note: Education groups sum to 100 percent.

**Figure 7.**

**Health Status by Household Earnings Quintile, 2009**

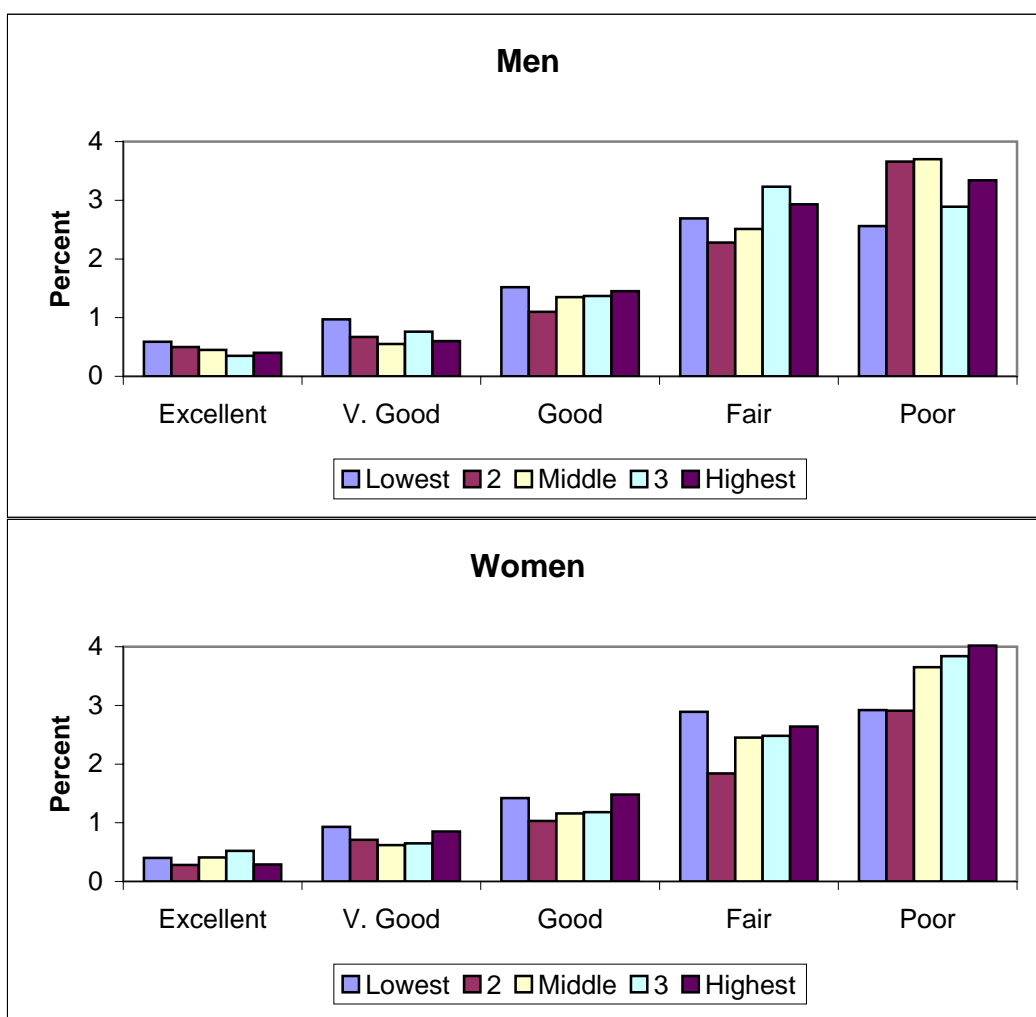
(health status depends on age, sex, lagged health and other covariates)



Note: Earnings quintiles sum to 100 percent.

**Figure 8.**

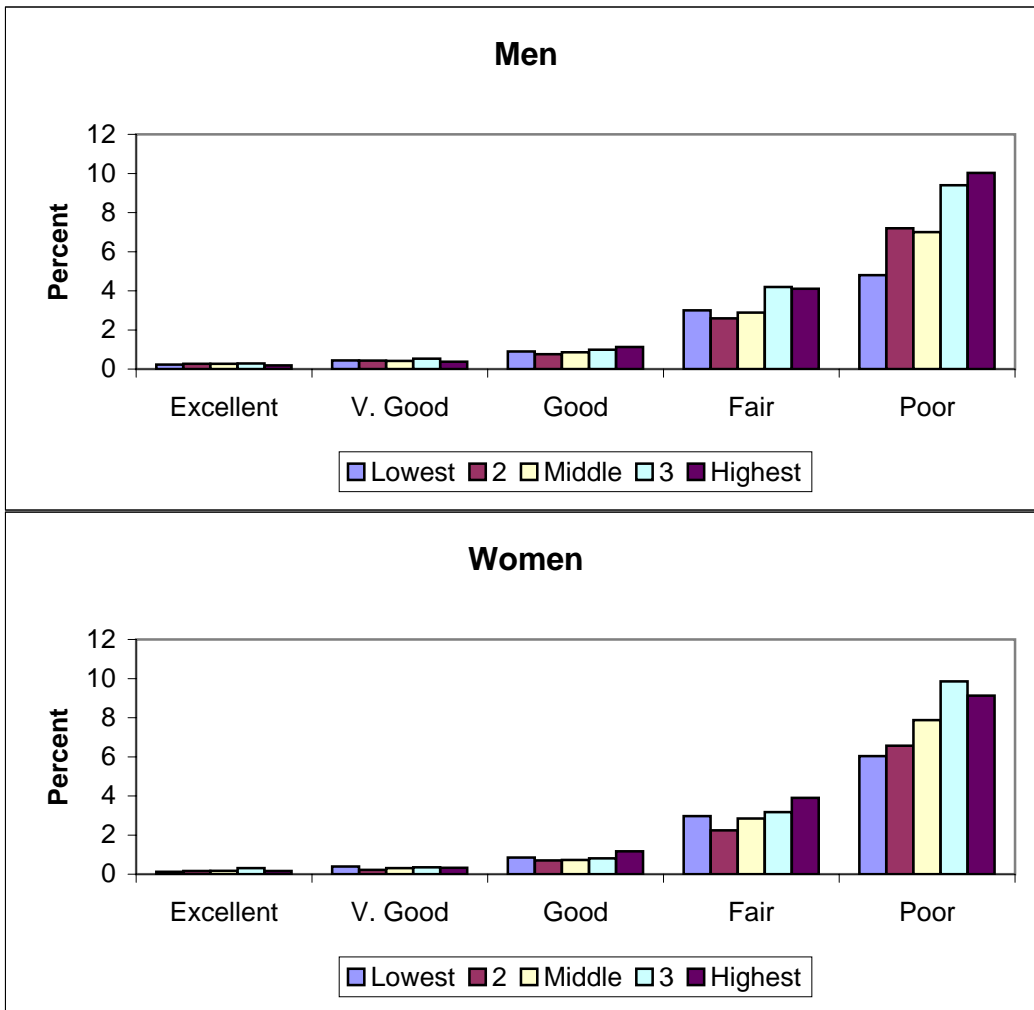
Raw Average One-Year Mortality Rates by Household Earnings Quintile  
(health status depends on age, sex, lagged health and other covariates)  
(Mortality does not depend directly on health status)



Note: One-year mortality rate defined as percentage of individuals in a given cell who die within 12 months of observation.

**Figure 9.**

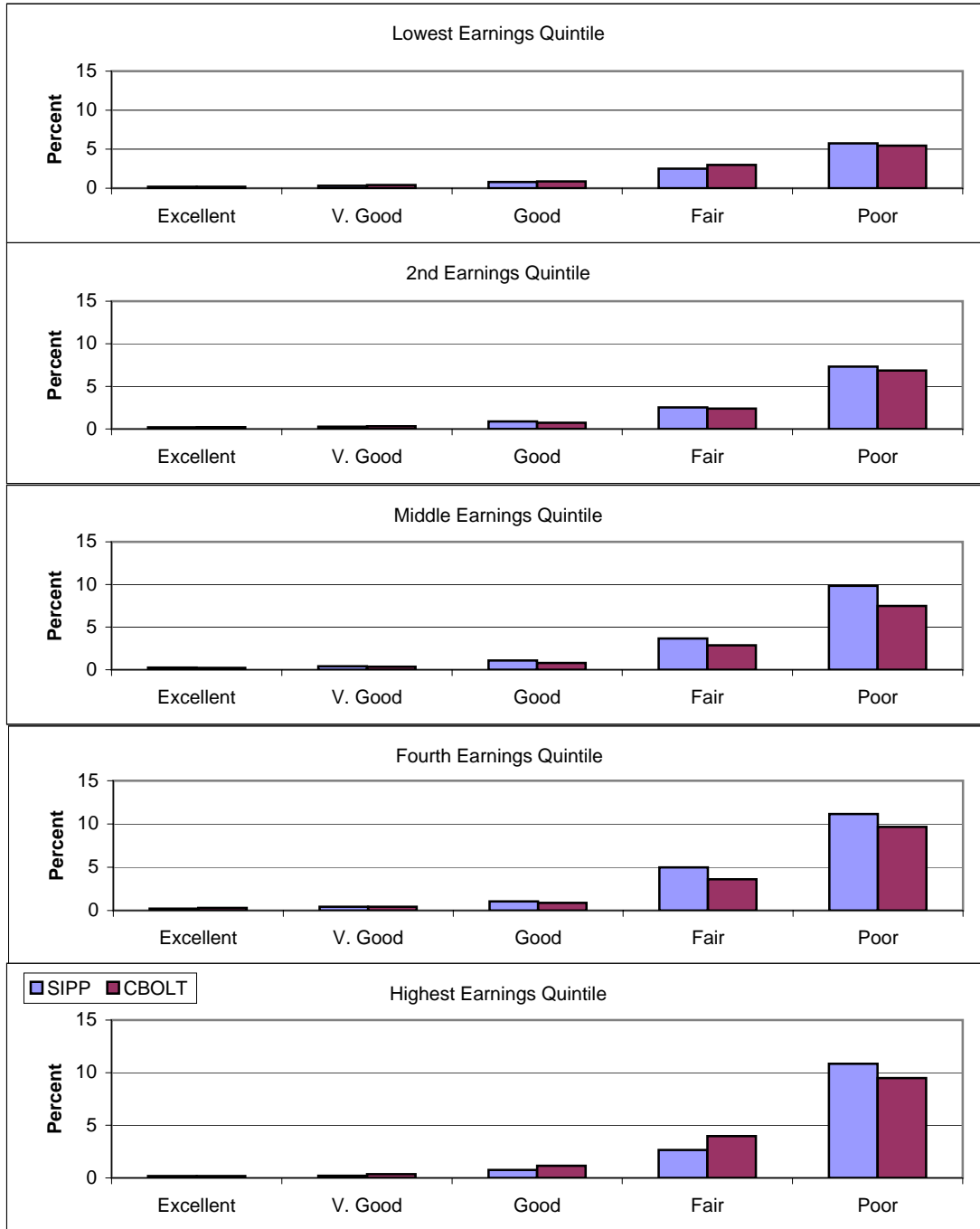
Raw Average One-Year Mortality Rates by Household Earnings Quintile  
(health status depends on age, sex, lagged health and other covariates)  
(Mortality depends directly on health status)



Note: One-year mortality rate defined as percentage of individuals in a given cell who die within 12 months of observation.

**Figure 10.**

Comparison of Average One Year Mortality Rates in the Merged-SIPP and CBOLT



Note: One-year mortality rate defined as percentage of individuals in a given cell who die within 12 months of observation.