SOME RESULTS FROM THE 1979-83 OCCUPATIONAL MORTALITY STUDY

Peter Sailer, Barry Windheim, and Mario Fernandez
Internal Revenue Service

For over 10 years, the Statistics of Income Division of the Internal Revenue Service (IRS) has been involved in putting together a file of taxpayers, coded by age, sex, and occupation, and showing when they died, and of what cause. This study was sponsored in part by the National Cancer Institute (NCI) and the Social Security Administration (SSA), both of which were interested in the use of occupation-coded administrative records for the purpose of studying possible links between occupation and high mortality rates--in the case of NCI, particularly mortality rates from cancer.

Our plans for this project were first outlined in a paper given at the 1980 Meetings of the American Statistical Association. In retrospect, the title may have been somewhat unfortunate: "Coming Soon, Taxpayer Data by Occupation." [1]

Subsequent papers at the 1983 [2] and 1984 [3] annual meetings described our progress--as well as our trials and tribulations--and in 1989 we presented some preliminary results. [4]

With this paper, we are announcing that "soon" is "now." The file we worked on for so long is ready to be used by the sponsors for research on occupational mortality issues. It should prove useful, as well, to many other researchers interested in relationships among such variables as occupation, industry, income, gender, taxation, etc.

Organizationally, the paper is divided into four main sections: first, we will provide some methodological background; next, will be a comparison between occupation codes from this study and those derived from death certificates; then, we will give a brief description of the population covered by the study, as it is only a portion of the whole U.S. population; finally, we will present our first analysis of occupational mortality statistics based on individual income tax returns.

PUTTING TOGETHER THE OCCUPATIONAL MORTALITY FILE

The basis of our study was a sample of individual income tax returns which was pulled and transcribed as part of the program to create the 1979 version of the annual report *Statistics of Income--Individual Income Tax Returns.* [5] In addition to the usual transcribing of income and tax items -- including, of course, each spouse's social security number (SSN) -- we also picked up two items not collected routinely for the Statistics of Income sample: sex and occupation. We asked our editors to enter a sex code based on the taxpayers' names. Joint returns were coded according to whether the male or female taxpayer was listed as the primary taxpayer. The editors were further asked to transcribe the entry in the occupation box (in both boxes, where applicable). We left space on our edit sheet for up to 20 alpha characters for each taxpayer's occupation entry.

As expected, many taxpayers gave us occupation titles (for example, "repairman") which are meaningful only when paired with an industry code (for example, "automobile dealership"). So, to make the connection to the appropriate industry, two steps were taken:

- If the taxpayer was a wage earner, we used his or her SSN to obtain the corresponding Wage and Tax Statement or Form W-2. It, in turn, gave us the employer's Employer Identification Number or EIN, which could be matched to the Social Security Administration's employer file, to obtain industry information.

- For sole proprietors, many of whom did not have EINs, we could use the SSN to go to our own Schedule C file to obtain industry information.

Thus, with the help of SSNs and EINs, it was possible to effect a number of matches to files both at IRS and within other Government agencies. These matches, it must be stressed, were done with strictest regard for maintaining the taxpayers' right to privacy. The Statistics of Income files can be used for statistical purposes only, never to identify individual taxpayers. Furthermore, IRS guaranteed that, for all matches, no links to other IRS files would be permitted and that, upon completion of the occupational mortality file, all identifying information would be eliminated, so that future linkage attempts would not be possible.

At this point, we were in a position to construct Standard Occupational Classification (SOC) codes for the file. Every time we made a decision on how to code a combination of a taxpayer's entry in the occupation box and industry code, obtained with the help of a Form W-2 or Schedule C, we entered the result in a computerized occupation coding dictionary. [2] This way, all decisions on occupation coding could be made once only and applied automatically to subsequent returns with the same entries.

Having thus taken care of occupation, we had to get mortality information. For that we needed to match in death certificates. Death certificates are maintained separately by each State, and we did not want to ask each State to search its files for all of the taxpayers in our sample. Luckily, the National Center for Health Statistics maintains a computerized file known as the National Death Index (NDI), with entries (including death certificate number) for each decedent in the United States. The NDI is maintained for precisely the kind of medical and health research we were doing.

To improve the likelihood of making a correct match to the NDI, we needed as much identifying information as possible. We started by going to our own Master File to retrieve the last known name and State of residence used by the taxpayers in our sample. We next went to Social Security's Year-of-Birth File, to determine the date of birth of each taxpayer.

We then transmitted the taxpayer's SSN, first and last name, date of birth, sex, and State of residence to NCHS, to be matched against the National Death Index (NDI). The NDI, in turn, gave us date of death and the death certificate number for decedents in our study. As was detailed in an earlier paper, [4] not all of these death certificates were "true" matches, and we had to develop an extensive selection process to identify which taxpayers really were deceased.

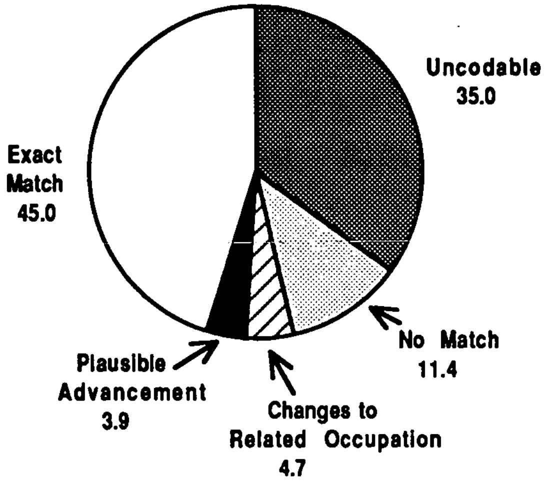
Finally, the death certificate -- which we obtained after extensive negotiations with all 50 States and the District of Columbia -- gave us the cause of death, as well as the "usual" occupation and industry of the taxpayer. After masking all identifying information, we sent the cause of death to be coded by the National Center for Health Statistics. This, theoretically, gave us all we needed. However, our sponsor, the National Cancer Institute, wanted an independent verification of the comparability of occupation codes obtained from the taxpayer and decedent sources, so the last step was to "mask" the death certificates, leaving only the occupation and industry entries showing, and send them to the Census Bureau for coding. At long last we had the file we had set out to build 8 to 10 years before. It contained two independently-coded industry-associated occupation codes, mortality information, and a cause of death code. The remainder of this paper will describe some of our findings about the resulting sample.

COMPARISON OF OCCUPATION CODES FROM TWO SOURCES

To begin with, we were interested in the comparability of occupation codes from both sources. Discrepancies were to be expected. For example, it is important to point out that the occupation as coded from the 1979 tax return is not necessarily wrong if it disagrees with the one coded from the subsequent death certificate. The taxpayer may, in fact, have changed occupations between 1979 and 1983 (the last year of deaths we added to our file). And since the information on the death certificate is generally provided by someone other than the decedent, it may not be accurate either.

Despite these limitations, we were surprised by our results. As is shown in Figure A, when we compared 2-digit occupation codes derived from

Figure A.--Occupation on the 1979 Individual Tax Return vs. Occupation on the Death Certificate



the tax returns to those derived from death certificates, 45 percent of the codes were the same. A surprisingly high 35 percent of the death certificates came back from Census labelled "uncodable." This meant that the entry had been left blank, was illegible, or contained non-helpful information, e.g., an entry of "retired." This high non-codable rate is particularly surprising because the comparison was limited to taxpayers with codable labor force occupations for 1979. Taxpayers who had written "retired" on their 1979 tax returns had already been removed from the comparison.

Of the taxpayers included in the comparison, 3.9 percent seemed to have advanced to a more exalted profession, but we felt the advancement was plausible. More athletes were teachers in their final year than were athletes, but this is not unexpected. We considered any advancement from do-er to teacher, manager, or supervisor to be plausible.

A slightly larger portion of the sample -- 4.7 percent -- showed different occupations on their tax returns and their death certificates, but the shifts at least did not appear unreasonable. A logger might well be qualified to operate other heavy equipment --say, in road construction. And a teacher might easily retire to the more contemplative life of writing.

The 11.4 percent that seemed implausible to us contained such changes of occupations as:

-- nurse to garbage collector;
-- administrator to truck driver;
-- pharmacist to kitchen worker; and
-- policeman to barber.

This is not to say that these shifts could not have taken place, but the proportion seems a bit high. It is our plan to review both our selection criteria for the matches and our occupation coding scheme to make sure that we are not making any mistakes. As of this writing, it does not appear that we made any systematic errors in either area.

POPULATION COVERED BY THE OCCUPATIONAL MORTALITY STUDY

Later in this paper, we will compare mortality rates for certain occupational groups to those of all taxpayers in this study. Therefore, it is important to describe the parameters of the study. In order to be selected for the 1979 Statistics of Income sample, an individual obviously had to file a tax return. This fact alone eliminated 39 percent of the population from our study. Furthermore, in order to become part of the Occupational Mortality Study, the tax filer had to give us a codable entry in the occupation box which indicated a labor force job (in a few cases, we were able to impute an occupation code from other information provided on the tax return, especially Schedules C and F).

In Figure B, our population (i.e., taxpayers with codable working force occupations) is compared to the total U.S. population and to the U.S. labor force. Our study is weighted up to 91.0 million taxpayers, of whom 1.7 million died over the 5-year period.

This means that the Occupational Mortality File represents 41 percent of the population and 87 percent of the work force. More importantly, in the 25 to 65 years age group, we have 65 percent of the population and 90 percent of the work force (Figure C).

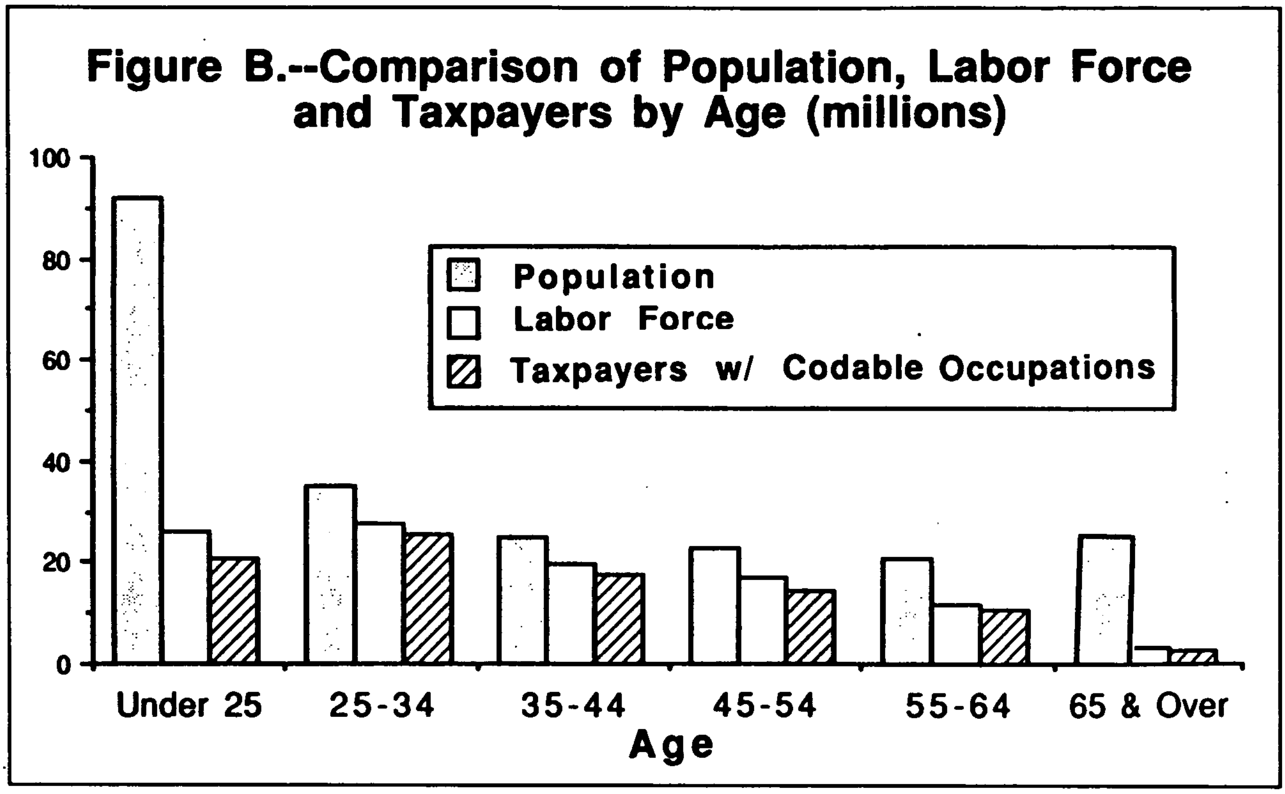


Figure B.--Comparison of Population, Labor Force and Taxpayers by Age (millions)
100
80
60
40
20
0
Population
Labor Force
Taxpayers w/ Codable Occupations
Under 25
25-34
35-44
45-54
55-64
65 & Over
Age

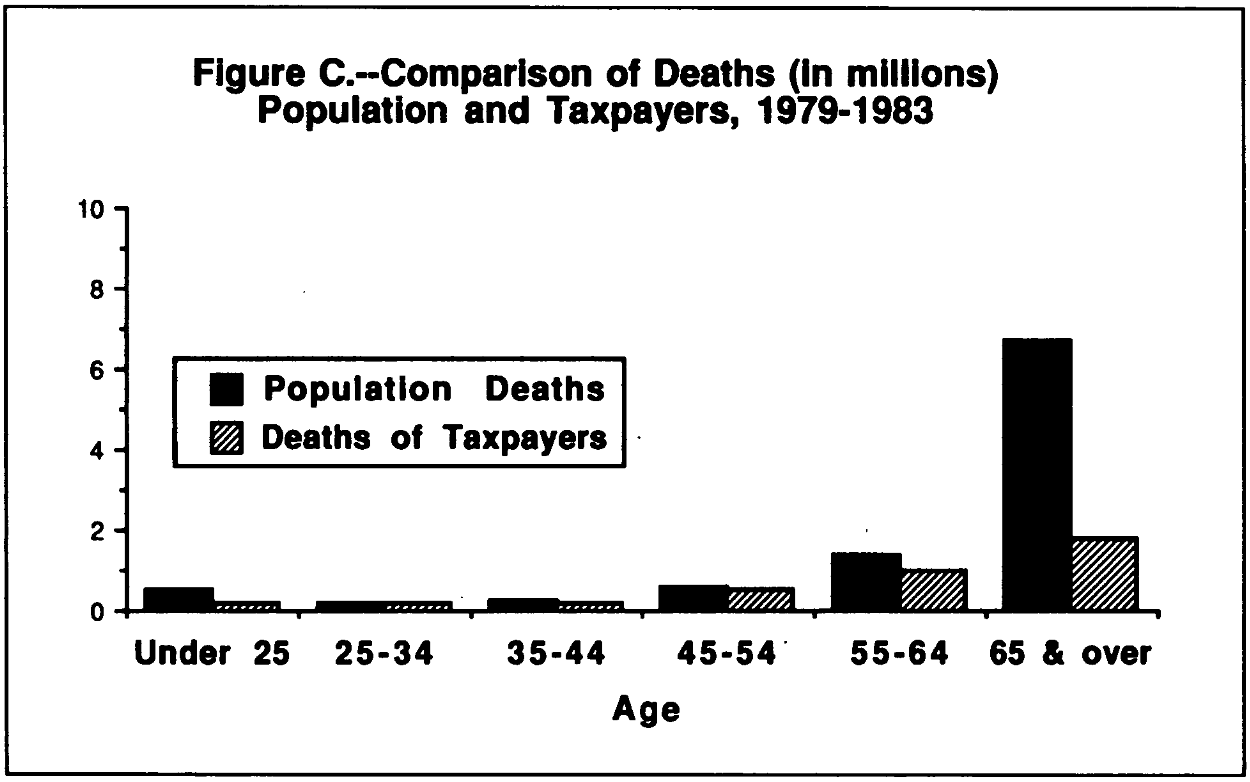


Figure C.--Comparison of Deaths (in millions) Population and Taxpayers, 1979-1983
10
8
6
4
2
0
Population Deaths
Deaths of Taxpayers
Under 25
25-34
35-44
45-54
55-64
65 & over
Age

Overall, the deaths in our study represent only about 37 percent of all deaths. However, in the 25 to 65 years age group, we do much better--getting about 61 percent of all deaths. In summary, the population we are studying is more middle-aged and noticeably healthier than the population as a whole. However, for the purpose of studying occupation-related mortality, this may, in fact, be a good population to study.

SOME OCCUPATIONAL MORTALITY STATISTICS

In order to check whether the occupational distribution of these deaths was reasonable, we divided up our file into 19 occupational groups-- some of which might be expected to have high mortality rates, some low. Then we computed standardized mortality ratios (SMRs) for each group. These are derived by dividing the observed number of deaths by the number you would expect to see, given the age/sex distribution of the group. Expectations were based on the observed death rates for all working taxpayers in each age/sex group. An SMR greater than 100 means this occupational group is dying off faster than the rest of the working/ taxpaying population; an SMR of less than 100 means it has a lower mortality rate.

Table 1 shows SMRs for all 19 groups, some further divided by industry. What follows are a few random observations about these statistics:

- Teachers (in spite of any protestations that their students are driving them to an early grave) had the lowest SMRs of these broad occupational groups. Social scientists and librarians did quite well, too. Engineers and technologists did well as a group, but those in the petro-chemical and plastics industries did not.

- Machine operators and tenders (basically skilled workers) had SMRs just slightly above average. However, those in the metal-working and textile industries had much higher SMRs.

- The group with the highest SMRs included "agricultural workers" and "helpers and laborers." Laborers in certain industrial groups stand out as having particularly high mortality rates. Those in the wood and paper industry, for example, show almost three times the expected value (although it should be noted that the sampling variability on this item is rather high). As was truc of engineers and scientists in the petro-chemical industry, laborers in this industry are dying off faster than are their peers in other industries. The same pattern holds for the semi-skilled (hand-working) occupations in this industry.

The purpose of presenting these statistics is not to put forward any new theories on links between occupation and mortality. This work will be performed in the future by experts in the field; it is our intention to create a disclosure-proofed public-use file from this project for use by the National Cancer Institute and others. At least, we can say that the results gained from this series of matches of files from IRS, SSA, and NCHS -- an undertaking never before attempted -- appear reasonable enough to warrant using the file for serious research. If examination of this relatively small sample yields valuable results, we may have paved the way for important studies on a larger scale in the future.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Gilbert Beebe and Dr. Sheila Zahm of the National Cancer Institute for their help and guidance during the lifetime of this project. The following members of the Statistics of Income Division also deserve acknowledgment: Jeri Mulrow, for help with the charts; Bobby Clark and Dodie Riley, for their technical assistance; and Wendy Alvey and Beth Kilss, for editorial review.

NOTES AND REFERENCES

[1] Sailer, Peter; Orcutt, Harriet; and Clark, Phil (1980), "Coming Soon: Taxpayer Data Classified by Occupation," *1980 American Statis-*

tical Association Proceedings, Section on Survey Research Methods, pp. 467-471.

[2] Crabbe, Patricia; Sailer, Peter; and Kilss, Beth (1983), "Occupation Data From Tax Returns: A Progress Report," *Statistics of Income and Related Administrative Record Research: 1983*, Internal Revenue Service, pp. 59-64.

[3] Crabbe, Patricia; Sailer, Peter; and Kilss, Beth (1984), "Taxpayer Data Used to Study Wage Patterns by Sex and Occupation, 1969, 1974, and 1979," *Statistics of Income and Related Administrative Record Research: 1984*, Internal Revenue Service, pp. 43-48.

[4] Clark, Bobby; Riley, Dodie; and Sailer, Peter (1989), "1979 Occupation Study/1979-1983 Mortality Study," *Statistics of Income and Related Administrative Record Research: 1988-1989*, Internal Revenue Service, pp. 181-187.

[5] Internal Revenue Service, *Statistics of Income -- Individual Income Tax Returns*, U.S. Government Printing Office, Publication 79 through 1982, Publication 1304 from 1983-present.

Table 1.-- Number of Taxpayers (Tax Year 1979), Number of Deaths (1979-1983), and Standardized Mortality Ratios,* by Occupation

Occupation (SOC Codes)	Number of taxpayers	Expected number of deaths	Observed number of deaths	Standardized mortality ratio
Administrators (11, 12, 13)	7,514,949	183,980	161,200	88
Management support (14, 45-47)	17,901,125	250,038	260,625	104
Engineers and technologists (16-18, 37-39)	5,037,887	101,381	80,806	80
In metal working industry	1,084,680	21,188	17,722	84
In petroleum, chemicals plastics, & rubber industry	206,845	3,811	4,380	115
Social scientists, librarians (19-21, 25, 32-34))	2,960,874	60,792	46,859	77
Teachers & counselors (22-24)	4,261,829	62,505	44,695	72
Health practitioners (26-30, 36)	3,616,706	57,177	66,891	117
Sales occupations (42-44)	5,481,826	99,959	99,144	99
Service occupations (50-52)	9,345,741	151,865	188,133	124
Farm operators and managers (55)	1,249,381	52,241	50,502	97
Agricultural & related workers (56-58)	1,644,134	42,366	54,969	130
Mechanics and repairers (61)	4,589,715	99,883	100,417	101
Construction trades (64)	4,027,345	86,789	97,623	112
Extractive occupations (65)	286,350	5,241	5,900	113
Precision production (68-69)	2,907,528	62,633	69,856	112
In metal working industry	1,250,580	25,139	24,166	96
Machine operators/tenders (75-76)	5,697,059	99,432	107,477	108
In metal working industry	1,450,854	26,924	37,679	140
In textile industry	1,483,747	22,298	31,329	141
Hand working occupations (77-78)	2,168,459	36,222	56,544	156
In petroleum, chemicals, plastics, & rubber industry	69,476	1,649	2,952	179
Transportation & material moving (82-83)	4,046,821	91,321	115,779	127
Helpers and laborers (86-87)	5,815,291	95,009	125,326	132
In construction industry	665,037	10,448	17,306	166
In wood & paper industry	94,424	1,567	5,131	327
In food & tobacco industry	403,287	6,204	11,412	184
In petroleum, chemicals, plastics, & rubber industry	169,274	2,543	5,657	222
Military and Government (91, 97)	3,574,992	56,230	48,982	87

* Ratio of observed number of deaths to expected number of deaths, times 100. Expected number of deaths based on distribution by age and sex of taxpayers in the occupational groups.