FURTHER RESULTS FROM THE 1979-1983 OCCUPATIONAL MORTALITY STUDY

Peter Sailer and Dodie Riley, Internal Revenue Service

For about 11 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, occupation, date of death and cause of death. This study was sponsored, in part, by the National Cancer Institute (NCI) and, early in its life, by 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 Annual Meeting 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 [4] and 1990 [5] we finally presented some results.

Background

This year's paper expands on last year's findings and attempts to answer questions raised by the results we found. Organizationally, it is divided into five parts: First of all, we will provide a brief background on the Study, after which we will discuss some problems with our matching procedures, and present some ideas for improvements; then, we will present revised mortality and occupation data, based on the corrections we made; next, we will analyze some differences in the reporting of occupations between death certificates and tax returns; and we will conclude with some of the lessons learned in this project.

It is not possible to reiterate here all the file matching that was necessary in order to create the Occupation Mortality File -- interested parties should consult the papers cited above. The following is a brief summary of the process: We started by abstracting the occupation entries for a sample of 203,000 individual tax returns for Tax Year 1979; obtained earnings data for each taxpayer by matching to Forms W-2; secured industry data by matching the W-2's to employer records; developed a computerized occupation-coding dictionary and used it to code the file; acquired date of birth information from the Social Security Administration; matched the file against the National Death Index; developed a computerized program to sort out the good matches from the bad; ordered the death certificates we considered good matches and abstracted cause of death, as well as occupation and industry; performed various validation studies to check the quality of our work; reorganized our data on a taxpayer (rather than a tax return) basis; and weighted the file to represent the universe of taxpayers.

Last year (1990) we presented what we expected to be the final report on this project. However, because we were not completely satisfied with the results we had obtained, we decided to continue our research on the quality of our efforts to create this merged file. First, we were concerned that a large number of death certificates had proven uncodable as to occupation (about 35 percent); second, we were even more concerned that, of the codable death certificates, less than half received codes that matched those on the tax returns. The next section describes our most recent research into these problems.

Matching Death Certificates to Tax Returns

The 1979 Occupational Mortality File was built on a sample of 203,000 tax returns for Tax Year 1979, filed by 354,000 taxpayers (joint returns are always filed by two taxpayers). Either from the tax returns, themselves, or from one of the matches

described above, we had first name, last name, sex, birthdate, marital status, state of residence, and social security number for each of these taxpayers. Our match to the National Center for Health Statistics' National Death Index informed us that 118,000 of these taxpayers were possible decedents during the period 1979-1983, although only 4,056 matched on all seven items. In other words, somewhere between 1 and 34 percent of our taxpayers had died. Our job was to narrow this percentage down a bit.

With the help of our seven matching criteria, we developed a computer program which assigned a score (from 2 to 100) to each match. Table 1 summarizes the program.

Using the matching scores, we established the following rules:

Maximum score (perfect match)	100
Minimum score for automatic selection	89
Minimum score for ordering and manual review	50
Minimum score for ordering and manual review if SSN missing	28.

This formula cut down our number of death certificates to be matched from 118,000 to just over 21,000. Had all been valid matches, this would have meant a death rate of 6.0 percent. National mortality figures indicated a death rate of 4.3 percent for the same 5-year period, so we were reasonably confident that we were getting enough death certificates. There were only three problems, one of which we had anticipated, two we had not. We knew we would have to check over some of the matches with low scores very carefully, to be sure we really had the correct taxpayer. We did not realize that many of our high-scored matches were also suspect, and we certainly did not think we might have a lot of death certificates matched to the wrong spouse of a married couple.

Let us start with the last problem. We assigned a very high score to any match on SSN -- 54 out of the 89 points needed to be accepted without any manual review. What we did not realize was that the "SSN" question is not answered all that well on the death certificate. Whether out of misunderstanding or desperation to give some answer when

Table 1.--Score Assigned for Match, by Matching Criteria

Matching Criterion	Score	
	Complete Agreement	Partial Agreement
Social security number	54	6*
Last name match	8	4**
First name match	8	4**
Birthdate (month) match	6	
Birthdate (day) match	5	
Birthdate (year) match	6	4***
Sex match	6	
Marital status match	4	
State of residence match	3	

*If 4 through 8 digits matched, 6 points were assigned for each matching digit.
**Four points were assigned for match on NYSIIS [6] phonetic code.
***Four points assigned for 1-year discrepancy, three for 2-year discrepancy, two for 3-year discrepancy.

they don't know the right one, a lot of informants give their own SSNs, when asked for the decedent's SSN. Since the informant is often the surviving spouse, we did, indeed, have a death certificate for one of the taxpayers in our sample -- not the one we matched to, however. The fact that the two spouses shared the same last name, state of residence, and often the same (or a close) birth year, all helped raise the score of these matches.

Of course, the issue of mixing up spouses could have been avoided if we had given more weight in our matching algorithm to agreement on sex code. The reason we did not had to do with the perceived lack of quality in the tax return sex code, which was, after all, only an editor's guess based on the first name. Indeed, several matches were correctly achieved in spite of a discrepancy in sex code. (We subsequently changed the sex code assigned to the tax return.) In retrospect, we probably should have given more weight to the sex code on joint returns. First of all, the editors had two names to check on those returns, and, secondly, even if they had been unable to figure out either one with certainty, the assumption that the primary taxpayer was male and the secondary female would have been correct about 98 percent of the time.

Another variable which, in retrospect, could have been given a greater weight was date of birth. Here again, we were under the impression (from the Social Security Administration) that there could be a fair amount of inaccuracy in our data, especially in the case of older taxpayers born before the social security system was in effect. Again, the solution, obvious in retrospect, was to give this variable a higher weight for some taxpayers than for others -- in this case, a higher weight to younger taxpayers.

Interestingly, it turns out that month and day of birth may be more accurate than year of birth, especially in the case of women. There appears to be a tendency to report a woman's age as being a few years younger than she was according to social security records. We leave it to our readers to decide whether this means that men (who tend to be the informants) are likely to delude themselves that they were married to younger women or that women sometimes do not give their husbands their correct age.

The problems enumerated so far just deal with matches where the death certificate had a social security number. Our problems were much greater, of course, when the National Death Index record lacked a social security number. Absence of data on the NDI generally meant that the data were not present on the death certificate. Missing data appear to be a regional phenomenon. The southeastern states appear to be much more lax in demanding that undertakers supply complete information than are other parts of the country. We found "n/a" (not available) as the answer to all sorts of questions on death certificates from the South, including not only SSN, but also cause of death, parents' names, and occupation. In northern states, especially the Midwest, death certificates with missing or vague information were generally stamped "pending," indicating that more information was being sought. Frequently, by the time we received the copies, the missing information had already been received and was attached.

With the SSN missing in large numbers of cases in certain states, our remaining six matching criteria -- first name, last name, sex, marital status, state, and birth date -- became even more important. The limitations of some of these items have already been discussed earlier in this paper. But the item that assumed prominence -- name -- needs some more discussion. Our algorithm gave a great deal of weight to an exact match on spelling; it even gave some weight to a match on the NYSIIS [6] phonetic code -- i.e., a variant in spelling that would yield a similar pronunciation. This system undoubtedly did a wonderful job of identifying the Bedrich Ziblonskis and Tuschnelda Unteroberdorfers of this country, but it was far too eager to declare a match when it saw a John Smith or a Mary Doe. Again, the solution is obvious in retrospect -- increase the weight given to a certain name based on how seldom it appears in the files.

Revised Mortality and Occupation Data

Having found what we believe to be all our mistakes the hard way -- through manual examination of every match, as well as a large proportion of the non-matches -- we corrected our file and re-ran the results presented in previous papers. Tables 2 and 3, below, show death rates for male and female taxpayers before and after our file corrections. For comparison purposes, data for the nation as a whole are shown, as well.

The data show that the death rates did go down slightly as a result of correcting our matching errors.

Table 2.--Death Rates for Males, 1979-1983

Age	U.S. Population	IRS Mortality Study	
		Original	Revised
Total	5.0	3.9	3.6
Under 25	0.7	0.9	0.7
25 - 34	1.0	1.1	0.8
35 - 44	1.5	1.4	1.2
45 - 54	3.6	3.7	3.3
55 - 64	9.2	8.0	7.4
65 and over	32.8	19.9	19.1
Not specified	--	1.4	1.6

Table 3.--Death Rates for Females, 1979-1983

Age	U.S. Population	IRS Mortality Study	
		Original	Revised
Total	4.0	2.1	1.9
Under 25	0.4	0.4	0.3
25 - 34	0.4	0.4	0.3
35 - 44	0.8	0.7	0.6
45 - 54	2.0	1.9	1.6
55 - 64	4.9	4.3	3.8
65 and over	23.4	11.3	10.6
Not specified	--	1.7	1.3

In other words, we discovered more false positive matches than we did false non-matches. Most notably, death rates for young male taxpayers, which were originally somewhat higher than the comparative data for the nation as a whole, are now below the national figures in the revised data. As a result, the general thesis presented in our earlier paper [5] still holds -- is, in fact, fortified: taxpayers have slightly lower death rates than does the population as a whole, during the early working years (ages 25 to 45) and much lower death rates during retirement years. This is because of the "healthy worker effect" [6], which would suggest that disabled and retired people are less likely to meet the filing requirements than are able-bodied workers -- not only because they have less income, but also because they have less stringent filing requirements. And, of course, the disabled and retired are much more likely to die than are healthy workers.

As was mentioned above, the impetus for continuing our research was the large number of uncodable death certificates in our file and the relatively low number of exact matches between the two files. So, what effect did these changes have on the correspondence rates beween the two files?

The first thing that is obvious from Table 4 is that the percentage of uncodable death certificates went down from 35 percent (which we thought at the time was unbelievably high) to 10 percent (which is in line with what we experienced when coding tax returns). In part, this reduction is a result of getting the right spouse matched up with the right tax return. If, under our previous matching scheme, the death certificates of a husband with earnings and a wife without earnings had been mixed up, the husband's death certificate (which was codable) would have been lost to this comparison (since the wife's W-2 showed no earnings for 1979); on the other hand, the wife's death certificate, which had the entry "housewife" and was, therefore, considered uncodable by Census, was in the table due to the presence of earnings on the husband's 1979 W-2.

A second factor in the reduction of "uncodables" is a processing error we discovered during our review. A portion of the death certificates from one state arrived late, after coding and processing of occupation and industry codes had been long completed for the others. The new batch of records was coded and keyed by a different person, who reversed the order in which the two codes were entered. Flipping them back into the proper order did wonders for reducing the number of invalid occupation codes on our file.

Table 4.--Correspondence Between Occupation Codes Derived from Tax Returns and Derived from Death Certificates

Match Status	Percent in Original Study	Percent in Revised Study
Total	100	100
Death certificate uncoded or uncodable	35	10
Exact match (2-digit level)	45	45
Plausible changes	9	12
Non-match	11	33

Note: Taxpayers with labor force occupations in 1979 only.

Finally, we found a number of occupation entries, previously uncoded by the coders at the Census Bureau, which we felt could, in fact, be coded -- we probably had a greater spirit of adventure than did the editors at Census.

Differences in Reporting Occupations: Death Certificates Versus Tax Returns

The second line of Table 4 represents our greatest disappointment. In spite of our best efforts to code more returns, and in spite of the fact that we are quite sure that we have many more correct matches than we did before and have eliminated switches between occupation and industry codes, the percentage of exact matches of occupation codes from tax return to death certificate stubbonly stayed at 45 percent. Most of the changes from uncodable went either to plausible changes in occupation or to nonmatch on occupation code. Let me explain briefly what we mean by "plausible."

Remember that, when we compare the tax return occupation to that on the death certificate, we are talking about entries made up to five years a part. The tax return entry is always for Tax Year 1979; the death certificate could be for any year between 1979 and 1983. In general, we considered a plausible change to be one from do-er to supervisor or teacher in the same general field of endeavor. Changes we considered implausible (and, therefore, nonmatches) were changes to completely unrelated fields, such as lawyer to laborer or physician to truck driver.

Actually, further research has turned up evidence that we may have to broaden our concept of plausibility. A number of these apparent descents from exalted to humbler professions were experienced by individuals for whom the underlying cause of death was alcoholism or drug abuse. Tragically, the immediate cause of death for some of these individuals was suicide.

However, the single most important source of nonmatching occupation codes was a difference in perceptions between the decedent and the survivors on what that person actually did during his or her career. As it turns out, men frequently see their professions as more exalted than do their widows, who are the informants on their death certificates. For example, one gentleman reported on his tax return that he was the president of his company in the services industry; his widow said he cleaned floors. Another taxpayer told IRS he was a rancher. His widow told the undertaker that his occupation had been "feeding the cows." We had many examples of men who considered themselves presidents or executives, whose widows felt they were window washers, cleaning men, or junk dealers. Actually, both descriptions may be accurate, to a point -- unfortunately, they yield different occupation codes.

On the other hand, men get their revenge on their spouses, as well -- if they manage to outlive them. We found some evidence that no matter how exalted the position the wife reported on her tax return, no matter how much income her W-2 or Schedule C showed she earned, all her husband remembered when asked by the undertaker was that the little lady was a housewife. It is probably good for the accuracy of occupation statistics from death certificates that women tend to live longer than men.

Conclusion: Lessons Learned

Peters and Waterman [7] have pointed out that it is important for employers to create a work atmosphere where people can make mistakes and thereby learn and grow. The Statistics of Income Division has fostered such an atmosphere over most of the 11 years we worked on this project, and we have, indeed, made mistakes -- and learned a lot.

The first of these lessons is summarized by Scheuren and Winkler in a recent paper, as follows: "Linkers should resist the temptation to design and develop their own software." [8] What we developed actually worked reasonably well in narrowing down the 118,000 possible decedents identified by the NDI to 21,000. Furthermore, we would have

done no significant damage to the data by accepting without further review, the 4,000 death certificates that matched to tax returns on all variables. Our system of narrowing down the remaining 17,000 death certificates (of which we finally accepted about 6,000) was, however, a mistake. Perhaps one of the many software packages already developed [9] would have given us the correct decisions -- this would certainly be a fruitful area for further research. In the meantime, we should have done what we ended up doing anyhow -- made the final selection manually.

We learned a lot more along the way -- for example, how to get permission to match administrative files. Each agency contributing data to the match will demand many safeguards to protect the confidentiality of their files, but these demands can be accommodated -- at least when the match is undertaken for a benign purpose, such as research on the prevention of cancer. We also learned a lot about the nature of self-reported and survivor-reported occupation titles. It is our hunch that what taxpayers reported to IRS is more accurate than what informants report on death certificates. The taxpayer is, after all, the person actually working the job. However, in both cases, more detailed questions or instructions would probably improve the quality of the reporting.

The one thing we most feared we would find during this, our latest review of the occupational mortality study, did not end up happening: we did not find any flaws in our computerized system of occupation-coding tax returns. Where there were differences between the death certificate and tax return occupation codes, they appear to have resulted from reporting differences. Therefore, the legacy of this project is not only a file (soon to be made available to the public) useful for studying occupation-related mortality issues, but also a computerized occupation-coding dictionary, which can be used for many future occupation projects.

Acknowledgments

The Occupational Mortality Project took place over an 11-year period, and it would be impossible to acknowledge everyone who contributed to the development of the ideas and findings put forth in this paper. Special thanks must go to the following IRS employees: Mario Fernandez, for programming support; Barry Windheim, for research; and Beth Kilss and Wendy Alvey, for editorial comments.

References

[1] Sailer, Peter; Orcutt, Harriet; and Clark, Phil (1980), "Coming Soon: Taxpayer Data Classified by Occupation," *1980 American Statistical 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] Sailer, Peter; Windheim, Barry; and Fernandez, Mario (1990), "Some Results from the 1979-1983 Occupational Mortality Study," *1990 American Statistical Association Proceedings, Section on Survey Research Methods.*

[6] For a further discussion of the NYSIIS (New York State Intelligence and Identification System), see Lynch, B.T. and Arends, W. L., "Selecting of a Surname Coding Procedure for the SRS Record Linkage System, U.S. Department of Agriculture, *Statistical Reporting Service*, 1977.

[7] Peters, Thomas J., and Waterman, Jr., Robert H., *In Search of Excellence*; New York: Warner Books, 1984, p. 287.

[8] Scheuren, Fritz, and Winkler, Williams E., (1991), "Regression Analysis of Data Files that are Computer Matched," Proceedings for the 1991 Census Annual Research Conference, Bureau of the Census.

[9] For an excellent summary, see Scheuren, Fritz, "Methodological Issues in Linkage of Multiple Data Bases," *Record Linkage Techniques -- 1985*, Internal Revenue Service.