LA--9462-MS

DE84 011016

UC-41
Issued: March 1984

Quality Control Activities in Support of the Plutonium Workers Study

Assessment of Coding Consistency for Data Collected at Rocky Flats

Michele Reyes
Gregg S. Wilkinson
John F. Acquavella

Los Alamos National Laboratory
Los Alamos, New Mexico 87545

QUALITY CONTROL ACTIVITIES IN SUPPORT OF THE PLUTONIUM WORKERS STUDY

Assessment of Coding Consistency for Data Collected at Rocky Flats

by

Michele Reyes, Gregg S. Wilkinson, and John F. Acquavella

ABSTRACT

The Plutonium Workers Study is a multifaceted epidemiologic investigation of workers at six Department of Energy (DOE) facilities: Los Alamos, Rocky Flats, Mound, Savannah River, Oak Ridge, and Hanford. Information from a variety of record sources has been collected and abstracted for these studies. This report considers the accuracy of the demographic, occupational, and radiation exposure data collected for studies at Rocky Flats. The majority of the information was accurately abstracted, and analyses based on these data may be conducted.

I. INTRODUCTION

The health study of workers in the nuclear industry is an epidemiologic investigation of health effects among workers potentially exposed to plutonium and other radionuclides at six Department of Energy facilities: Los Alamos, Rocky Flats, Mound, Savannah River, Oak Ridge, and Hanford. In order to conduct these studies, a data base was constructed for each study site which included certain demographic, occupational, and exposure data. Information was often abstracted from records and transformed into machine-readable form. This report documents the data collection procedures and the quality control measures conducted to validate the accuracy of the demographic, occupational, International Classification of Disease (ICD) Code, and radiation exposure data for studies of workers at Rocky Flats (RF).

II. METHODS

We identified and reviewed record sources at RF, which contained information required for the study. Employee work history records contained occupational and demographic information (name; race; sex; job title; education; Social Security number; an RF identification number; the dates of birth, hire, and termination; and if known, the date of death). Radiation exposure histories were available from health physics records. We microfilmed all records available from these sources and transferred the film to Los Alamos for data abstraction. In addition, partial personnel and health physics records were retrieved in computerized form.

A. Employee Personnel Records

We merged the RF personnel files into a single file which included name; Social Security number; sex; an RF identification number; and the dates of birth, hire, and termination. Employee work histories were then referenced to validate the computer data and to supply missing information.

Detailed coding protocols were developed for abstracting information from the employee work history records onto paper copies of the computer-generated information for each individual employee. Each record was then independently edited following the same instructions used for the original coding. After completion of the original coding effort, approximately 10% of the employee records were systematically sampled (every 9th employee was selected from an alphabetical listing of all 9539 employees) and recoded/re-edited using the original instructions.

The original edited information was compared with the re-edited information to assess the percentage of error for each of the variables coded. Ninety-five percent confidence intervals (CI) were calculated to assess the boundaries surrounding these error measurements.[1] We accepted variables when the upper 95% CI was less than 2%.[2] These data were then double keypunched, verified, and stored electronically into a study data base.

Detailed work history information was manually abstracted from the microfiche for a 25% random sample of the RF employees (N = 918). Job titles were condensed into a three-digit code according to a strict protocol, developed by an industrial hygienist. The first digit denoted broad job categories: machinists, craftsmen and maintenance workers,

service workers, office workers, laborers, professionals, technicians, and manufacturing/production workers. Within each of these eight categories, the second digit further defined a job (i.e., the type of professional--chemist, physicist, etc.). A third column was used to record potential for nonradiation exposure (ever versus never). In addition, the date of service was recorded for each job title abstracted. A 5% systematic sample of these data (N = 190) was then edited using the same protocol. Ninety-five percent CIs were calculated for each error, and the 2% upper CI set the acceptable error limit.

In the course of editing job titles for this 25% sample, we were able to compare the original microfilmed information with the data in the newly created study data base. Inaccuracies in the computerized information work history record were corrected and tallied.

B. Health Physics Records

Cumulative external radiation exposures (mrem) were available in computerized form from RF. These data were supplemented with yearly exposure information which required transforming the microfilmed record information into machine-readable form.

A software program was developed for direct computer entry of these data. A detailed protocol provided instructions for carrying out this procedure in a standardized fashion. A computer file identifying the cohort members and the employment period for each worker was created from the study data base of all personnel. External exposure values (mrem) were then entered for each year an individual was employed at RF. Cumulative lifetime totals were then calculated and incorporated into the file.

Blind double entry of each record was conducted. Data were stored when the first and second entries agreed. When discrepancies occurred, the original record was compared with the computer data, and the correct information was entered. After completion of the data entry, we manually compared a random sample of 250 hard-copy data base records with the microfiche source to assess the quality of the data entered. The statistical comparison procedures used on the personnel data were repeated for the health physics data maintaining the 2% upper CI restriction for error. In addition, the cumulative totals generated in our coding were compared with the RF computerized totals.

C. Vital Status Information

The follow-up of these 9539 employees required much effort. Vital status was initially determined by a Social Security Administration (SSA) record search in 1978. A second search, two years later, was made because our first query was incomplete. Of the 6777 white males, 419 were found to have died during the study time period 1951-1977. As of December 31, 1977, 5731 males were alive and 568 were determined lost to follow-up, that is, ascertained by the SSA as unknown, impossible to locate, or mismatched. All of the mismatches and a 10% sample of the remaining lost to follow-up were actively traced for vital status determination. Nine additional deaths were thus identified. Additional deaths were also identified by the RF medical benefits department. Death certificates for 425 of the 428 deaths were obtained. The underlying cause of death was then independently coded by two nosologists to 8th ICD code. The cause-of-death code and death dates were then entered into the study data base and 100% edited.

III. RESULTS

Table I presents the coding errors in the RF personnel data which were identified in the comparison of the edited/re-edited sample. The variables name; Social Security number; sex; RF ID number; and dates of hire, termination, and birth demonstrated differences less than 2% (95% CI = 0.0-1.1). The field for education exhibited an unacceptable error rate of 4% (95% CI = 2.6-5.0).

We were able to assess the usefulness of "double coding" by comparing the percentage of error demonstrated in the original coding/editing of the personnel record variable with those found in the 10% re-edited sample (Table II). In the first edit, the percentage of error for name and Social Security number, RF identification number, and sex was found acceptable. The percentage of error for education and all date fields demonstrated upper 95% CI greater than 2%. In the 10% re-edited sample, errors were reduced to acceptable levels. An additional coding/editing of the personnel variables was necessary to create an accurate data file for our analytic purposes.

TABLE I

ERRORS IDENTIFIED IN THE CODING OF PERSONNEL DATA RECORD INFORMATION FOR ROCKY FLATS

Variable	Number of Errors	% Error	95% CI
Last name	2	0.2	0.1-0.8
Title	0	0.0	0.0-0.4
First name	1	0.1	0.0-0.6
Middle initial	1	0.1	0.0-0.6
Social Security number	0	0.0	0.0-0.4
Sex	1	0.1	0.0-0.6
RF ID number	0	0.0	0.0-0.4
Education	34	4.0	2.6-5.0
Hire date	4	0.4	0.2-1.1
Term date	2	0.2	0.1-0.8
Birth date	1	0.1	0.0-0.6

TABLE II

COMPARISON OF EDITING VS RE-EDITING OF PERSONNEL DATA INFORMATION

	Editing			Re-editing		
Variable	Number of Errors	% Error	95% CI	Number of Errors	% Error	95% CI
Last name	4	0.4	0.2-1.1	2	0.2	0.1-0.8
Title	0	0.0	0.0-0.4	0	0.0	0.0-0.4
First name	3	0.3	0.0-0.9	1	0.1	0.0-0.6
Middle Initial	4	0.4	0.2-1.1	1	0.1	0.0-0.6
Social Security number	3	0.3	0.0-0.9	0	0.0	0.0-0.4
Sex	4	0.4	0.2-1.1	1	0.1	0.0-0.6
RF ID number	0	0.0	0.0-0.4	0	0.0	0.0-0.4
Education	56	6.0	5.9-7.7	34	4.0	2.6-5.0
Hire date	17	1.8	1.0-2.9	4	0.4	0.2-1.1
Term date	13	1.4	0.7-2.4	2	0.2	0.1-0.8
Birth date	9	1.0	0.4-1.8	1	0.1	0.0-0.6

N = 938.

Each personnel record contained multiple job title entries. Of the total 769 job codes edited, an error rate of 0.88% (95% CI = 0.3-1.7) was

demonstrated. No errors for dates of service were observed. In comparing the 190 original records with the study data base information, no errors were found in the fields for name; RF identification number; or in the dates of birth, hire, or termination. Two discrepancies were found in the coding of Social Security number which resulted from differences in the data base due to the later entry of complete Social Security numbers. All 9539 records were recoded for education only, using the same instructions, and a 10% sample was edited (N = 918). An improved but unacceptable error rate was demonstrated (95% CI = 0.9-2.5).

Computerized race information was received later from RF but could not be validated from any other source. Race was merged directly into the study data base, classifying individuals as white, black, Indian, oriental, and unknown. All persons for whom race was "unknown" were assumed to be white.

Table III presents the errors identified in the coding of the health physics data. The differences in the cumulative exposure values among the coded and edited sample of 250 records were 0.4% (95% CI = 0.0-2.2). The re-edited cumulative values were also compared with the totals generated in the RF computerized data. A disagreement of 7.2% (95% CI = 4.6-11.1) was demonstrated. For each of the 250 health physics records coded and edited, multiple entries were made, one for each year of an individual's employment. These yearly external exposure values were compared for 4122 entries demonstrating an error rate of 0.2% (95% CI = 0.0-1.0).

TABLE III

ERRORS IDENTIFIED IN THE EDITING OF HEALTH PHYSICS DATA

Type of Comparison	Number of Errors	% Error	95% CI
Yearly Values			
Coded/Edited (N = 4122)	7	0.2	0.0-1.0
Cumulative Values			
Coded/Edited (N = 250)	1	0.4	0.0-2.2
RF Computerized/Edited (N = 250)	18	7.2	4.3-11.1

There were no errors made in the double entry of ICD code or date of death. Errors in the coded cause of death could not be evaluated. However, because all death certificates were coded twice, the effect of this type of error was assumed to be negligible.

Since the ascertainment of vital status through the SSA was less than 90% complete, a third submission is currently in progress. The impact of this lost-to-follow-up mortality on the analysis has been modeled and described elsewhere.[3]

IV. DISCUSSION

With the exception of education, all variables necessary to conduct studies of the RF cohort were accurately coded, as demonstrated by error rates less than 2% (95% CI range = 0.0-1.1). An additional recoding effort will be necessary to correct the deficiencies in the coding of education. The detailed yearly external exposure information was extremely accurate (0.2% error, 95% CI = 0.0-1.0) and suitable for analysis. However, the error in the cumulative totals may range from 0.0% to 2.2%. Since we have thoroughly reviewed the measurements which contributed to our summary totals, we have judged them more suitable for analysis than the RF computerized cumulative data.

All data sets are imperfect no matter what the source or how collected. The sources available for our study were far from ideal. Records originally maintained for other purposes were often the only sources of information available, and these data were often incomplete and difficult to read. In addition, error was potentially introduced by the activities we employed in abstracting, transforming, and reducing the data into a machine-readable form. Therefore, it was important that we assessed the accuracy of these data transformation activities and estimated the degree of error in the data used for analyses.

We have demonstrated the results of quality control exercises designed to measure the accuracy of information coded from employee, health physics, and death certificate records into a form allowing storage in a computerized data base. This involved both the laborious manual coding of information from hard-copy records onto paper code forms, which were then keypunched, verified, and finally computer stored, and the

more efficient direct entry of data. Once all data were computerized, a study data base was created. In addition, range checks were run for all fields and illogical entries were corrected. Therefore, the created data base contained less error than we have reported here, thus contributing to the integrity of the data before they were analyzed and subsequently to the validity of our study results.

REFERENCES

1. K. J. Rothman and J. D. Boice, Jr., Epidemiologic Analysis with a Programmable Calculator (US Government Printing Office, Washington, DC, 1979).

2. C. P. Chamblee, "Use of Statistical Sampling in Validating Health Effects Data," Data Validation Conference Proceedings, EPA-600/9-79-042 (September 1979), pp. 31-38.

3. J. F. Acquavella, G. L. Tietjen, and G. S. Wilkinson, "Lost to Follow-Up Bias in an Occupational Mortality Study: A Quantitative Consideration," Los Alamos National Laboratory report LA-9530 (December 1982).