Methods for Ensuring High Quality of Coding of Cause of Death

The Mortality Register to Follow Southern Urals Populations Exposed to Radiation

N. Startsev; P. Dimov; B. Grosche; F. Tretyakov; J. Schüz; A. Akleyev

1Urals Research Center for Radiation Medicine, Chelyabinsk, Russia;
2Chelyabinsk Oblast Clinical Hospital №3, Chelyabinsk, Russia;
3Federal Office for Radiation Protection (BfS), Neuherberg, Germany;
4Southern Urals Biophysics Institute (SUBI), Ozorsk, Russia;
5International Agency for Research on Cancer (IARC), Lyon, France

Keywords
Mortality register, cause of death, International Classification of Diseases (ICD), radiation, Russia

Summary

Background: To follow up populations exposed to several radiation accidents in the Southern Urals, a cause-of-death registry was established at the Urals Center capturing deaths in the Chelyabinsk, Kurgan and Sverdlovsk region since 1950.

Objectives: When registering deaths over such a long time period, quality measures need to be in place to maintain quality and reduce the impact of individual coders as well as quality changes in death certificates.

Methods: To ensure the uniformity of coding, a method for semi-automatic coding was developed, which is described here. Briefly, the method is based on a dynamic thesaurus, database-supported coding and parallel coding by two different individuals.

Results: A comparison of the proposed method for organizing the coding process with the common procedure of coding showed good agreement, with, at the end of the coding process, 70–90% agreement for the three-digit ICD-9 rubrics.

Conclusions: The semi-automatic method ensures a sufficiently high quality of coding by at the same time providing an opportunity to reduce the labor intensity inherent in the creation of large-volume cause-of-death registries.

1. Introduction

In the 1950–1960s, several radiation accidents associated with the Mayak Production Association (Mayak PA) activities took place in the Southern Urals. The most well-known of them were contamination of the Techa River due to scheduled discharges of liquid nuclear waste, and the 1957 nuclear accident in a Mayak PA waste storage facility which resulted in the formation of the East-Urals Radioactive Trace (EURT). As a consequence of these accidents, large territories in three administrative regions (“oblasts”) of the Urals were contaminated with radionuclides: the Chelyabinsk, Kurgan and Sverdlovsk oblasts. The contamination of the environment caused a significant overexposure of the population resident in these areas [1, 2].

As is generally known, studies of the mortality patterns and rates provide insight into the risks in populations related to environmental factors. About 95% of the total population exposed to radiation due to the radiation accidents described above lived in the five administrative territories (“raions”) in the Chelyabinsk oblast (Kaslinsky, Krasnoarmeysky and Kunashaksky raions) and in the Kurgan oblast (Dalmatovsky and Kataisky raions) (Figure 1). To facilitate the follow-up of chronic radiation exposure effects in the Techa River cohort (TRC), both for members and for their offspring, paper copy information has been collected on all deaths registered in the above mentioned five raions for the period 1950 through 1996 [3], and recently up to 2009. Collection of information on causes of death in other administrative territories of the Chelyabinsk and Kurgan oblasts was not systematically done and covered only members of the TRC and their offspring [4].

When starting the cohort studies, cohort members had to be identified in the hard copies of death certificates, both being laborious and leading to some loss of information. This initiated the setting up of an electronic cause-of-death registry. The inclusion of unexposed populations in the cause-of-death registry allowed the calculation of baseline mortality rates typical for the regions, of people living under similar conditions as the cohort members apart from the chronic radiation exposure.

The cause-of-death registry is based on the International Classification of Diseases (ICD; Russian 9th edition) [5]. The quality of coding is influenced by several factors,
in particular exhaustive medical terminology and, as coding depends heavily on experience, following closely instructions and knowledge on causes of death. Lack of knowledge or non-compliance with coding rules could lead to inaccuracies in the records. In order to reduce the rate of such errors and to increase the quality of coding it is often recommended to apply parallel coding by different coders [6]. However, for large cause-of-death registries of 50,000 or more cases, the above-indicated method would require too many specialist resources. The task of ensuring a similar quality of coding over long time periods is also challenging, due to changes in staff or classification systems.

2. Objectives

Here, we describe the procedures implemented in the Urals Research Center for Radiation Medicine of the Federal Medical Biological Agency of Russia (URCRM of FMBA, RF) cause-of-death registry to ensure high quality of coding of causes of death according the Russian edition of ICD-9. In addition, we present results of a validation study of coding done by three independent centers.

3. Methods

3.1 Data Entry and Preparation for Coding

The process of creating a cause-of-death registry was divided into two steps: data entry and cause-of-death coding. At the data entry stage, we tried to ensure correct entry of personal identifiers and textual statements on the causes of death cross-linked to a dynamic reference handbook (see below). To achieve this goal, a special screen editor was developed. Identifiers were entered into the fixed input fields. The fixed input fields have the same meaning as the identifiers already contained in the registry for exposed residents and their offspring. This ensures completeness of the search for causes of death for deceased members of the cohorts. For the cause of death, the task was solved through coding of only unique textual entries of causes of death. If a good quality of coding unique textual entries is ensured, and all the recurrent entries are automatically assigned to the already acquired code, the registry will harbor a unified systemic error. In order to implement this, it was necessary to separate the process of data entry from the process of data coding. At the data entry stage it was suitable to engage staff without special training in the coding of causes of death. At the coding stage, the trained coder only considered unique textual entries on causes of death which allows speeding up the coding process, but not entries of individual persons. To improve the quality of coding, conditions were created to apply the method for cause-of-death
coding performed in parallel by two or three coders. As a result, a handbook was created due to which all the recurrent entries were automatically assigned to the codes available in the manual. To reduce the number of spelling mistakes in text of the diagnosis at the entry stage, a reference book of terms widely applicable in medicine and consisting of 8 or more characters has been created. The operator could choose a term in the reference book, e.g., “atherosclerotic”, and it was automatically inserted in the alphabetic string. In addition, a reference of authorized abbreviations which included 27 most frequently used abbreviations has been created for operators. The reference book was augmented in the process of the work involving creation of a cause-of-death registry, and currently contains over 1100 medical terms. Further, the synonyms of the terms designated to the same disease were identified and replaced by one term. The next step involved taking away text that was not used for coding; for example, information on disease stage or extent were deleted, but only for diseases for which this information did not affect the coding. In case of diseases which required taking into account such parameters they were retained. All these processes were carried out automatically using special reference books which were constantly updated.

### 3.2 Coding

Causes of death were split into two subgroups depending on the age at death: up to 28 days and older. This subdivision resulted from the specific features of coding causes of death for newborns.

The following procedures were carried out for each subgroup: 1) Textual entries on causes of death are alphabetically ordered. 2) Only unique entries were selected. 3) Coding was done by two independent coders.

The results of the work done by the two specialists engaged in coding of unique causes of death were compared. In case of a mismatch of the cause-of-death codes, the issue is discussed with the participation of a third specialist in coding, and a concerted decision is prepared. If one of the coders insists on his/her coding, a protocol is prepared including the textual formulation of the cause of death, the decision made, and a reference to the decision rationale. Also, a record is made regarding the existence of a different opinion about the code and a reference to the ICD-9 rules.

The cause-of-death registry was programmed in C++ and the format of the database is RDB (“relational database”).

### 3.3 Methods of the Inter-institutional Comparison

In the course of the project implementation, to test the method of automatic coding, an inter-institutional comparison of coding was conducted. To achieve this aim, a sample of 500 deceased persons (250 from the URCRM cause-of-death registry and 250 from the cause-of-death registry from the city of Ozyorsk [7]) was randomly selected and maintained at two centers, URCRM and the Southern Urals Biophysics Institute (SUBI) in Ozyorsk. The distributions of causes of deaths in the sample corresponded to the overall distributions in the two registries. The coding of the random sample was conducted in three centers: in addition to URCRM and SUBI, a coding specialist from an Institute in Germany was invited to take part in the coding (German Federal Office of Radiation Protection (BfS)) (with the text translated from Russian into German). To do the coding, the URCRM researchers applied the method of automatic coding described above. The coding of causes-of-death was done manually at SUBI, given the smaller number of death certificates occurring in a city of 70,000 inhabitants. The BfS coded the sample manually. For the inter-institutional comparison all listed causes of death on the death certificate were provided (immediate cause-of-death, underlying cause-of-death, major disease that lead to death, reasons if external cause of death, accompanying diseases (e.g., diabetes, hypertension)), for the coder to determine the main cause of death.

To conduct the comparison, the codes on the death certificates were grouped in several ways. First, codes were grouped according to the ICD-9 main groups I—XVII. Second, comparison was made for those certificates which were coded as malignancy (ICD 9 140 – 208) by any of the three coders. Among those, the following analyses were conducted: i) is the death certificate coded as solid cancer (ICD9 140–199), or as leukemia or lymphoma (200–208); ii) within the solid cancers, codes were grouped according to ICD-9 subgroups (140–149, 150–159, 160–165, 170–175, 179–189, 190–199), with Kaposi's sarcoma (176) not taken into account, and agreement compared; iii) leukemias and lymphomas were grouped according to whether they are radiation-related or not radiation-related (chronic lymphocytic leukemia and Hodgkin’s disease), and agreement was compared; iv) leukemias and lymphomas were compared by type (each individual code within the subgroup 200–208). Third, comparison was made for those certificates which were coded as cardio-vascular disease (ICD 9 390 – 459) by any of the three coders. Among those, the following analyses were conducted: i) is the death certificate coded as cardio-vascular disease or not; ii) amongst those death certificates which were coded as cardio-vascular diseases the comparisons were made for heart diseases (390 – 429), cerebrovascular diseases (430 – 438) and others (440 – 459), respectively.

For all the comparisons, the percentage of agreement was calculated as well as the Cohen’s unweighted kappa coefficient as indicator for inter-rater reliability were calculated, with the comparisons being made between SUBI and URCRM, SUBI and BfS, and URCRM and BfS. When calculating the kappa values, only those certificates were included in the calculation that were coded as any of the diseases of interest by at least one of the two respective coders, but irrespective of the coding of the third coder. Cohen's kappa values above 0.6 are considered to reflect good agreement and those exceeding 0.8 to reflect very good agreement, respectively. Lower values indicate weak (≤ 0.20), light (≤ 0.40), or moderate (≤ 0.60) agreement [8]. SAS Software was used for the calculation of the kappa values.
4. Results

Overall, 500 death certificates were coded and used for the inter-institutional comparison. A good to very good overall agreement was achieved in coding by main ICD-9 classes of diseases: 84.2% for URCRM-SUBI (kappa = 0.80), 85.6% for URCRM-BfS (kappa = 0.81), and 80.4% for SUBI-BfS (kappa = 0.75). Since in the framework of the epidemiological projects the death rates from neoplasms (ICD-9 class II) and cardio-vascular diseases (ICD-9 class VII) are the main focus, the agreement of the coding of these selected classes of diseases was assessed separately (▶Table 1). The percentage of agreement of coding was high. For 111 death certificates coded as malignancies by any of the three coders, the kappa values were 0.46, 0.47 and 0.67, respectively. The agreement was better for those 99 death certificates which were coded as solid cancer when taking ICD-9 subgroups into account; the respective kappa values were 0.71, 0.73 and 0.77. Based on only 14 cases of leukemia or lymphoma and distinguishing between radiation-related and not radiation-related diagnoses, the agreement was light to moderate with kappa values of 0.48, 0.31 and 0.51, respectively. For 239 death certificates coded as cardio-vascular diseases by any of the three coders, the agreement on coding of these diseases was almost very good with 89.4% between URCRM and SUBI (kappa = 0.78), 90.0% between SUBI and BfS (kappa = 0.87) and 93.4% between URCRM and BfS (kappa = 0.79). The inter-rater reliability with respect to whether this cardio-vascular disease was a heart disease, a cerebrovascular disease or any other cardio-vascular disease was good, with kappa values of 0.62, 0.61 and 0.66, respectively.

At the same time, a full agreement in coding diseases assigned to the three-digit rubrics of ICD-9 was only observed for 176 (35.2%) out of the 500 death certificates (data not shown).

In order to test the proposed coding method, each of the participating institute was offered to conduct revisions of the initial codes. In 91 cases (18.2%) all three institutes had assigned different codes, and in 232 cases (46.4%) two coders were in agreement but disagreeing with the third one. For clarifying the discrepancies, coders were offered the following: If a coder insisted on having the coding his/her way, he/she was asked to give reasons for having assigned that specific code, otherwise coders were allowed to change their mind. After this revision process of the coding, all three institutes assigned the same code in 350 cases (70.0%), in 8 cases (1.6%) all three centers indicated different codes, and in 142 cases (28.4%) one of the coders stayed in the minority (▶Table 2).

Thus, after the revision of the coding results, full concurrence in coding cases using three-digit rubrics increased from 176 (35.2%) to 350 (70.0%). ▶Table 3 presents the percentage of agreement achieved by the institutes after the revision of the coding.

The cause of death registry contains information on more than 240,000 deaths registered in the above mentioned five regions over the period from 1950 through 2009. The applied method of automatic coding provided high quality of coding of causes of death assuring the data can be used studying the late effects of radiation exposure in the cohort studies.

5. Discussion

The extension of the cause-of-death registry both in terms of catchment area and time period has offered an opportunity to establish a system for conducting automatic coding of the causes of death entered into the registry, to deal with the huge

<table>
<thead>
<tr>
<th>Institutes participating in the coding comparison</th>
<th>N</th>
<th>URCRM – SUBI</th>
<th>SUBI – BfS</th>
<th>URCRM – BfS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD-9 main groups (17 groups)</td>
<td>500</td>
<td>84.2 0.797</td>
<td>80.4 0.746</td>
<td>85.6 0.809</td>
</tr>
<tr>
<td>Cancer</td>
<td>111</td>
<td>78.9 0.457</td>
<td>57.1 0.468</td>
<td>87.3 0.672</td>
</tr>
<tr>
<td>1) solid cancer vs. leukemia or lymphoma (140–199 vs. 200–208)</td>
<td>99</td>
<td>75.8 0.710</td>
<td>78.4 0.731</td>
<td>91.4 0.770</td>
</tr>
<tr>
<td>2) solid cancer by organ (140–199), subgroups</td>
<td>14</td>
<td>71.4 0.482</td>
<td>72.7 0.312</td>
<td>63.6 0.509</td>
</tr>
<tr>
<td>3) leukemia (excluding chronic lymphocytic leukemia) and Non-Hodgkin’s lymphoma vs. chronic lymphocytic leukemia and Hodgkin’s disease (200, 202–208 ex. 204.1 vs. 201 and 204.1)</td>
<td>239</td>
<td>89.4 0.780</td>
<td>90.0 0.866</td>
<td>93.4 0.794</td>
</tr>
<tr>
<td>Cardio-vascular diseases</td>
<td>239</td>
<td>74.1 0.780</td>
<td>73.3 0.866</td>
<td>78.6 0.794</td>
</tr>
<tr>
<td>coded as heart diseases (390–429) (n = 148)</td>
<td>91.4 0.780</td>
<td>90.0 0.866</td>
<td>93.4 0.794</td>
<td></td>
</tr>
<tr>
<td>cerebrovascular diseases (430–438) (n = 83)</td>
<td>89.4 0.780</td>
<td>90.0 0.866</td>
<td>93.4 0.794</td>
<td></td>
</tr>
<tr>
<td>other cardiovascular diseases (440–459) (n = 26)</td>
<td>89.4 0.780</td>
<td>90.0 0.866</td>
<td>93.4 0.794</td>
<td></td>
</tr>
</tbody>
</table>

Table 1  Agreement (%) in coding based on groups selected from ICD-9 disease classes and un-weighted kappa coefficients of agreement

Table 2  Results of coding revision

<table>
<thead>
<tr>
<th>The original code was changed</th>
<th>SUBI</th>
<th>URCRM</th>
<th>BfS</th>
</tr>
</thead>
<tbody>
<tr>
<td>The percent of changes made</td>
<td>159</td>
<td>51</td>
<td>92</td>
</tr>
<tr>
<td>The number of changes which resulted in assigning the cause of death to another ICD-9 class</td>
<td>31.8</td>
<td>10.2</td>
<td>18.4</td>
</tr>
<tr>
<td>The number of cases when the coder was in the minority</td>
<td>14</td>
<td>27</td>
<td>101</td>
</tr>
</tbody>
</table>
amount of death certificates in an efficient way when little resources are available. In case there is a full concurrence between the case and the reference book text, a code from the reference book is assigned automatically. This procedure allowed conducting automatic coding for 75–85% of the total newly-entered causes of death. The causes of death that remained un-coded were examined by two independent coders. New synonyms of the terms and the textual formulations which could not be associated with coding were identified. The updates of the reference books used for preparing the textual formulations for coding are ongoing, thus continuously increasing the percentage of automatic coding.

From the inter-institutional comparison it can be inferred, that the method proposed for automatizing the process of cause-of-death coding ensures a good quality of coding and offers the following advantages: i) subjective errors resulting from different levels of competence of the coders are minimized; ii) a unified systematic error determined by the error contained in the reference book is ensured, if errors are unavoidable; iii) the method allows revisions of questionable cause-of-death codes included in the protocol on discrepancies by engaging the participation of new specialists in doing the coding and this approach enables reduction of the systematic error; iv) provides an opportunity to pass, in case of necessity, from coding based on ICD-9 to ICD-10. Agreement was usually good to very good, with the exception of hematological malignancies, a phenomenon documented previously [7].

6. Conclusions
Currently, the cause-of-death registry maintained at the URCRM contains over 240,000 deaths. The total volume of the two unique reference books used in cause-of-death coding makes up more than 19,000 entries. The reference books have been reviewed by five independent specialists in coding causes of death. The use of the automatic method for coding causes of death ensures a common systematic error that may result from expansion of the cause of death registry both in terms of coverage and the time period, but inter-institutional comparisons suggest the quality of this efficient way of coding results in the same high quality as laborious manual coding.

Acknowledgments
The development of the method for ensuring the quality of coding was started within project N° FP6–516478 (SOUL) and completed within project N° 249675 (SOLO). Both projects received financial support from the European Commission. Special thanks go to Nora Fenske, Federal Office for Radiation Protection (BfS), for her support in the statistical analysis.

References

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Agreement in coding, based on ICD-9 classes and on three-digit rubrics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>URCRM – SUBI</td>
</tr>
<tr>
<td></td>
<td>After coding</td>
</tr>
<tr>
<td>Agreement on main classes</td>
<td>84.2%</td>
</tr>
<tr>
<td>Agreement on three-digit rubrics</td>
<td>54.2%</td>
</tr>
</tbody>
</table>