Perfecting Data

HISTORICAL METHODS, Winter 2004, Volume 37, Number 1

Integration of Specificity Variation in Cause-of-Death Analysis

JANET PADIAC

*Department of Anthropology
University of Toronto*

Abstract. Integration of mortality data by cause of death is typically problematic for researchers because of the inadequacies of historical records. Two aspects of cause-of-death (or disease) data processing are discussed here: vocabulary and specificity. By developing a system of processing in which causes of death are in nested tables that are linked with a relational database program, the researcher can integrate highly specific and sensitive disease data with data that are less specific. Vocabulary variations can also be maintained by the system. The result of integrating data of differing specificity levels and vocabularies allows one to use mixed data for longitudinal analyses. The author developed the present system to allow the use of nineteenth-century British army statistical data, and an example of the application is presented.

Keywords: cause of death, classification, relational database, specificity, vocabulary

Much space has been devoted in these pages to the classification and aggregation of causes of death, and many methods of processing and extracting the meaning of historical mortality data have been developed (Alter and Carmichael 1996; Kintner 1986; Williams 1996). In this article, I discuss the problems associated with using health records from nineteenth-century British military archives, in particular those associated with vocabulary and with different levels of disease specificity. Because the military data included both causes of morbidity and causes of mortality, the list of potential diseases and conditions contains many nonfatal conditions. The extent of the causes of disease forced me to confront the problems of data processing and classification, and in this article I outline the system I developed to retain maximum integrity in the original material. In particular, the problem of variation in the levels of disease specificity and its resolution are highlighted. The solution required the integration of a relational database program with a classification system from the nineteenth century, which together make it possible not only to accommodate the heterogeneity of the original material but also to permit robust longitudinal analysis over changing medical theory climates.

Vocabulary

After missing data, unfamiliar, misapplied, or imprecise vocabulary are probably the most frequently encountered problems for those investigating historical morbidity and mortality. In some cases, the disease name has become outmoded and been replaced. The replacement of *jaundice* for the earlier term *icterus* is an example. If the name of the disease changes several times, then the chain of terminology can become obscure and the meaning of the term lost. Another common type of problem is the allocation of the same term that today is a symptom to what in the past was considered a disease. *Diarrhea* and *constipation* serve as examples; they were major illnesses in the past and occasionally were causes of death. Today, they are seen as symptoms of a number of conditions and are, therefore, not present in modern classifications. A third type of problem arises from the cultural determination of a disease. A cause of illness or death in the past may not be considered a bona fide illness under today's biomedical system in which an underlying physiological cause (preferably bacterial or viral) for each affliction is determined. *Hysteria* and *stuttering* are examples from Victorian society, and both were encountered in the military data. Yet another problem is language; many diseases, and especially injuries, were recorded in Latin.

In whatever way the changing disease vocabulary came about, the means of processing nineteenth-century military data required understanding of what the regimental medical officer believed to be the ailment when he recorded data about the individual. Many conditions, particularly those that are nonfatal, are unfamiliar to the modern medical lexicon. Fortunately, a number of historical reference materials are available (e.g., the *Cambridge World History of Disease* [Kiple 1993]), which assist in giving some identity to unfamiliar terms. However, scholars have tended to concentrate on killer diseases, such as tuberculosis or fevers, or on those most unlikely to be misdiagnosed, such as smallpox, so

chronic, nonfatal conditions have often been overlooked. Skin and eye ailments, for example, are rarely causes of mortality, but they constitute a large portion of the chronic problems from which the troops suffered. Because vocabulary changes in those sorts of conditions have attracted little attention over the years, one requires a different method to uncover the type of condition to which they referred.

Any of a number of medical source books from the last decade of the nineteenth and the early part of the twentieth centuries is an excellent tool for reverse-tracking vocabulary changes (e.g., William Osler's *The Principles and Practice of Medicine* [1892]). Such texts were penned for practitioners who may still have used older terminology, but they were written after the germ theory revolution. Authors such as Osler were writing for an audience that was adjusting to the new etiology, and the multiple editions of his reference works spanned decades and provided a meeting place for old and new terms. Another excellent tool is a copy of an early classification system, such as the one published in 1885 by the Royal College of Physicians. In bringing together a number of disease terms for like conditions, that classification gives the researcher some sense of the placement of the condition in the medical past. The disease nomenclature published by the Royal College of Physicians is especially helpful because it presents the system in four languages.¹ It is important to repeat the consensus that at no time should the goal be retrodiagnosis, that is, to place today's diagnosis on yesterday's symptoms or disease. An attempt to do so is fraught with methodological and cultural pitfalls (Alter and Carmichael 1996; Arrizabalaga 1999). Instead, one's goal should be to gather sufficient information about the condition to be able to integrate yesterday's diagnosis into a system that allows comparison with coeval material accumulated by both past and current researchers.

Specificity

In addition to the vocabulary problems, the raw military data presented the problem of varying specificity. Although smallpox and measles are both of similar levels of specificity, smallpox, eruptive fevers, and fevers are not; the latter two categories contain many more fevers than just those

resulting from smallpox. The data were determined to be composed primarily of three levels of specificity. Level-1 specificity contains terminology that delineates a disease or condition, as commonly understood, such as hepatitis, jaundice, gastritis, and dyspepsia (see figure 1). But in the military data reports, hepatitis and jaundice may have been grouped together and reported as diseases of the liver, whereas gastritis, dyspepsia, and other intestinal ailments may have been grouped together and reported under diseases of the stomach. The latter group represents level-2 specificity. In level 3, the level of least specificity, hepatitis, jaundice, gastritis, and dyspepsia were occasionally grouped with many other conditions and reported under diseases of the digestive system. Variants of specific level-1 diseases such as acute and chronic dysentery were also found, but those were seen as subcategories of level-1 specificity rather than as specificity levels of their own.

Had the military data always been at the same level of specificity, development of an innovative way to integrate the material would not have been necessary. But there was considerable variation in the specificity level of the raw material. One solution would have been to aggregate the data upward, gathering it all into level-3 categories, but in so doing, much of the sensitivity available in cause of illness and mortality analyses would have been lost. My goal in the research design was to maintain the flexibility in the data by incorporating all levels of specificity and most variation in the vocabulary. Integration of the data and the differing levels of specificity requires a system of ordering diseases that is computer generated, yet compatible with the nineteenth-century British data. Ultimately, the solution lay in integrating a classification system over 100 years old with a twenty-first-century relational database software program.

Classification

A system of disease classification had existed in the military since 1805 with the publication of Robert Jackson's *A System of Arrangement and Discipline for the Medical Department of Armies* (1805), in which he suggested the organization of disease causes for the tabulation of all hospital admissions and deaths. Jackson included sample forms

for data recording in his publication. Thirteen years later, when data collection in the military began on a large scale, forms almost identical to those Jackson had published were used (Padiak 2004). When disease classification for the entire British population was overhauled in 1837 by William Farr, the structure of his classification was similar to that found in Jackson's (1805) work. In 1851, Farr took his system to the First International Sanitary Congress, and it was eventually adopted by that body in 1864 (Alter and Carmichael 1999). Farr's system was further developed by a committee of the Royal College of Physicians in 1869 partly because of the changing needs of the Army Medical Department (AMD). The committee worked closely with the AMD, and because the army reported morbidity as well as mortality, the committee included diseases, such as eye conditions, that were unlikely to be returned as causes of mortality. The result was a numerical system in which the afflictions were ordered by type of affliction (constitutional or local), body area (generative organs or digestive system), and type of cause (poisons or injuries). In 1885, that system was revised, as the board adjusted to improved terminology and new developments in the germ theory. In 1895, the first International List of Causes of Death (ICD) was printed. Today, the ICD is used internationally by the nations belonging to the World Health Organization for the collection and compilation of morbidity and mortality statistics. Other systems are in use, for example, by those who are interested in a system that is based on etiology or pathology rather than on a compromise between physiological, anatomical, and etiological considerations, as is the ICD (for a discussion, see Meslé 1999).

The current ICD is considered unsuitable for use by historical demographers for several reasons. For example, the system is too detailed for unspecific data, such as that of specificity levels 2 or 3 (or even level 1). Some criticize the ICD's foundation on physiological and etiological considerations as unsuitable for diagnoses based on symptomatic terminology (Kintner 1986). Furthermore, constant revisions make it difficult to compare data collected over different decades and under different revisions of the ICD. Historical demographers are generally required to devise their own systems (e.g., Gage 1993; Kintner 1986; McKeown 1977; Williams 1996) or to report an excerpted list of causes of death (Landry and Lessard 1996; Rogers 1999).

For the present data, I chose to go to neither of those extremes. Instead, I appropriated a nineteenth-century precursor to the ICD, the system presented in the 1885 edition of *The Nomenclature of Diseases*, published by the committee of the Royal College of Physicians. My reasons for doing so are as follows:

- It is symptomatically and historically based rather than theoretically or etiologically based.
- It was developed in conjunction with AMD directors, so it is very much aligned with extant nineteenth-century data.

- It foreshadows today's ICD in its hierarchical structure.
- It uses nineteenth-century terminology but is sufficiently modern to be mostly comprehensible to research today.
- It is numerical.

Given those reasons, the revision of 1885 was used as the basic classification system, and its original numerical system, disease causes, and original hierarchy were retained. The use of a nineteenth-century classification system solved many of the vocabulary problems encountered in the data. With a combination of early-nineteenth-century terms and multiple languages integrated under a system responding to the newly understood germ theory, this classification permitted the ready identification of almost all conditions. It retained, as much as possible, the ideology with which the practitioner of the time made his original diagnosis. Use of the 1885 classification meant that there was little temptation for retrodiagnosis and that the late-nineteenth-century medical paradigm remained somewhat pristine.

Relational Database Method

Once a nineteenth-century classification scheme had been adopted and the effect of anachronisms in vocabulary minimized, the problem of the variation in specificity remained. In this section, I discuss its solution: the use of a relational database program. Although standard spreadsheet programs work well for within-level data processing, they cannot cope well when the aggregation changes or the terminology shifts. With a relational database (Microsoft Access was used for this study), one can link multiple tables that allow causes of disease or death to be nested by levels of specificity.² The nested cause-of-disease tables are of decreasing specificity, and each succeeding table has fewer listings because the diseases are aggregated according to classifications. Before developing the linkage system, the researcher prepares the tables, and they can be edited and expanded as need arises once the system is operational.

The database for the military is made up of seven tables (figure 2 shows a screen shot of the relationships diagram of the database). The tables follow:

- Three nested cause-of-disease tables: level-1 specificity, level-2 specificity, and level-3 specificity.
- Place: The data are assigned a variable from the location of the garrison, such as Gibraltar or Malta.
- Corps: The variable is assigned a category of troops, officers, civilians, and so forth.
- Corps 2: This table allows the data to be assigned to a specific regiment. The default is *all*.
- Source of data table: Each row of data is linked to the original source.

In addition to the previously noted tables, each row is assigned a date of entry and a unique ID number. A notes column was also incorporated when required. The arrangement

Record ID: 16 Place: Malta Year: 1874 Corps: 1 corps3: 0 Source: annual report 1874 Date: 6-Jul-01 Notes:

1DCat: 473 1Disease: diarrhea 2DCat: 456 2Disease: intestines 3DCat: 373 3Disease: digestive

Admitted: 14 rate: 4.0 Deaths: 0 rate: 0.0

Record: 14 of 123

FIGURE 3. Data-entry form showing three levels of specificity.

or number of tables is not fixed and can be customized for specific projects.

Practically speaking, the program allows one to use a customized form to enter the data, which is particularly useful if a number of individuals are entering the data, because the forms can be managed so that missing or mistyped entries are not permitted. The entry form for data (pictured in figure 3) shows the form used for the military data. If the disease was of level-1 specificity, as in figure 3, all three levels were entered on the single entry of the cause of disease, because of the table linkage. If the entry was for diarrhea, for example, whose numeric code is 473, then the remaining two entries are automatic. The entry may be made numerically or by using a drop-down table. If the datum was less specific, the first level is left at default (for *no detail at this level*) and the second box is entered, in the case of diarrhea, disease of the intestines, whereupon the third level would be automatic. If, however, the datum was general, informative only for a disease of the digestive system, then only the third level of specificity would be entered (with levels 1 and 2 at default). A datum row, then, was made up of the disease, the number of admissions and deaths, and the other descriptors such as place and corps.

Relational database programs have many features, but two were very useful in this project. The first is the ability to expand the tables. For example, if one wished to add data for, say, India, then *India* simply had to be added to the place table, enabling one to continue using the form. This feature was valuable because it identified specific diseases that were not on the original list. For example, dysentery (no. 22) was listed in the original disease classification from 1885; but the army reported acute and chronic dysentery separately, and it was desirable to retain that distinction. To accommodate that requirement, subsections under dysentery were added (22.1 acute dysentery, 22.2 chronic dysentery). Sometimes a new section was appended at the end of all three lists: for example, more information was necessary for the entire category of self-inflicted causes of death (to which were added subcategories 1213.1 cut throat, 1213.2 hanging, 1213.3 gunshot, etc.). The flexibility was especially invaluable when a cause of disease arose that was unfamiliar (e.g., phlegmon) and could not be readily translated into a cause already on the list. Through the course of data collection, 138 specific illnesses were added to the starter list to make a total of 1,360, although only 439 of the total diseases were required for the members of the military. The other important feature of a relational database is its querying capabilities, which

allow one the freedom to select and aggregate at will from the data. For example, when I included influenza with respiratory diseases for my analysis, following Timothy Gage (1993) and Thomas McKeown and R. G. Brown (1955), rather than under the category of fevers, following the 1885 nomenclature of diseases, it was then possible to adjust the aggregated categories with little fuss.

Level-1 aggregation of data allows for the investigation of specific diseases, such as alcoholism or smallpox, which are easily diagnosed and may be of interest to current researchers. But many analyses are more suited for aggregation of data at levels 2 or 3, which is particularly true of diagnoses from the past, when observation, touch, and the appearance of expelled bodily substances were often the only tools for differentiating many conditions. Grouping causes of death into large heterogeneous categories solves the problems of ill-defined causes of death common in historical data (Gage 1993) but, of course, also eliminates some of the sensitivity of analysis. The approach developed for use of the present data allowed the inclusion of both aggregated and specific-level cause-of-death tracking. An example of the type of analysis that resulted follows.

Diseases of the Digestive System, 1859 to 1872

I used the method to analyze changes in the causes of morbidity of the digestive system among the troops of Gibraltar from 1859 to 1872. The data consisted of hospital admissions from the *Army Medical and Statistical Reports*, published annually by the British government.³ The data were inconsistent in quality, and the years 1864, 1865,

1869, and 1870, were lacking in level-1 and level-2 detail. To reduce year-to-year variation and to approximate the data of the missing years, I used five-year averaging for all morbid conditions (see resulting graph in figure 4). Four specific categories (level 1) of diseases and one level-2 category are included. The remaining level-3 category contains the residual digestive diseases.

Because all three levels of specificity are retained in the analysis, the resulting morbidity pattern communicates the changes in several gastrointestinal conditions—constipation, diarrhea, colic, and dyspepsia—common during the late nineteenth century, while minimizing the possible compounding effect of faulty categorization from liver diseases that might occur in an analysis that used level-3 categories. The drop in diarrhea shows the effects of public health initiatives occurring in the garrison during that decade; attention was paid to the barracks' sanitary arrangement for the men. Colic, which had much higher rates of morbidity early in the century, had dropped to a minor level after the Crimean War. Dyspepsia morbidity rates rose throughout the decade, supporting Denis Gibbs's (1997) suggestion that its growth was rooted in the Victorians' fascination with their abdominal symptoms. Diseases of the liver remained constant, perhaps reflecting the high levels of alcohol consumption common among troops of that era or, alternately, results of hepatitis exposure.

Conclusion

Clearly, an analysis technique that combines fine-grained data with those that are less specific has potential for many applications. When used for the study of morbidity and

mortality, it is especially attractive for longitudinal analysis; in such analyses, changing terminology and inconsistency in data are the norm, not the exception. The example presented for the digestive morbidity for the troops shows that the improvement in the detail offered can contribute to the understanding of health trends in the nineteenth century. Such a technique can also facilitate the study of disease terminology itself, because the changes in vocabulary are readily captured and tracked. Whatever the specific application, such an approach allows flexibility until the moment of analysis, thus readily accommodating a shift in the researcher's design and mode of querying.

NOTES

1. The nomenclature appears in English, German, French, and Latin.
2. Access[®] is a trademark of the Microsoft Corporation.
3. The reports are part of the British Parliamentary Papers, published annually.

REFERENCES

- Alter, G., and A. Carmichael. 1996. Studying causes of death in the past. *Historical Methods* 29:44–48.
- . 1999. Classifying the dead: Towards a history of the registration of causes of death. *Journal of History of Medicine and Allied Sciences* 54:114–32.
- Arrizabalaga, J. 1999. Medical causes of death in preindustrial Europe: Some historiographical considerations. *Journal of the History of Medicine* 54:241–60.
- Gage, T. B. 1993. The decline of mortality in England and Wales, 1861 to 1964: Decomposition by cause of death and component of mortality. *Population Studies* 47:47–66.
- Gibbs, D. 1997. The demon of dyspepsia: Some nineteenth-century perceptions of disordered digestion. In *Gastroenterology in Britain*, edited by W. F. Bynum, 29–42. London: Wellcome Institute for the History of Medicine.
- Jackson, R. 1805. *A system of arrangement and discipline for the medical department of armies*. London: John Murray.
- Kintner, H. J. 1986. Classifying causes of death during the late nineteenth and early twentieth centuries: The case of German infant mortality. *Historical Methods* 19:45–54.
- Kiple, K. F. 1993. *The Cambridge world history of human disease*. Cambridge: Cambridge University Press.
- Landry, Y., and R. Lessard. 1996. Causes of death in seventeenth- and eighteenth-century Québec as recorded in parish registers. *Historical Methods* 29:49–57.
- McKeown, T. 1977. *The modern rise of population*. New York: Blackwell.
- McKeown, T., and R. G. Brown. 1955. Medical evidence related to English population changes in the eighteenth century. *Population Studies* 9: 119–41.
- Meslé, F. 1999. Classifying causes of death according to an aetiological axis. *Population Studies* 53:97–105.
- Osler, W. 1892. *The principles and practice of medicine*. Edinburgh: Young J. Pentland.
- Padiak, J. 2004. Modeling military morbidity and mortality: An analysis of the military population of 19th-century Gibraltar. Ph.D. thesis, University of Toronto.
- Rogers, J. 1999. Reporting causes of death in Sweden, 1750–1950. *Journal of the History of Medicine* 54:190–209.
- Royal College of Physicians. 1885. *The nomenclature of diseases*, 2nd ed. London: Harrison.
- Williams, N. 1996. The reporting and classification of causes of death in mid-nineteenth-century England: The example of Sheffield. *Historical Methods* 29:58–71.