The Intramural Research Program of the US National Library of Medicine, National Institutes of Health, supported the research and writing of these reflections and the editing of the articles which have appeared in the series 'Teaching and Researching the History of Medicine in the Era of (Big) Data'.

Frederick W. Gibbs[1] and **Jeffrey S. Reznick**[2]
[1]University of New Mexico, USA
[2]History of Medicine Division, US National Library of Medicine, National Institutes of Health, USA

doi:10.1017/mdh.2017.71

Digitizing Doctors: Methodologies for Creating a Database from Historical Directories of Physicians

Medical directories are rich sources of historical information about the medical profession. However, the labour required to extract information from them as printed text has limited their usefulness to historians of medicine who could more readily delve into their hundreds of thousands of entries in a digital database format. These directories consist of a list of physicians, usually including information such as their professional and speciality affiliations, and historians have used them to explore the structure of the medical profession and the emergence of specialisations. Despite the directories' widespread availability, the difficulty and inefficiency of using these sources has limited both the number of historians using them and the number of physicians included in studies. Here I describe a process for creating a database of physicians from a historical directory to provide historians with easier access to these data.

For historians of physician professionalisation, education and specialisation in the United States and Canada, the *American Medical Directory* (*AMD*) is a particularly promising resource because of its completeness and the regularity of its editions. Purporting to list all the registered physicians in those countries and their territories, the American Medical Association (AMA) published the first *AMD* in 1906, with new, updated editions following every two to three years. Although historians often spend time discovering and compiling data, the *AMD* represents a unique opportunity whereby nearly complete data have already been compiled and are ready for analysis. The information included about each doctor changed in nearly every edition, but physicians' addresses (both home and office), their medical school and year of graduation, their year of licensure, their speciality affiliations, their military affiliation (if any) and their birth year were reliably listed. Other details of their practice, such as their hours and hospital or medical school appointments, also regularly appeared. Currently, scholars have to comb through tens of thousands of records in print *AMD*s, counting and collating figures, to reach statistical conclusions about historical physicians and their practices. However, if these directories were made available in a database format, information from historical *AMD*s could be queried almost instantaneously, leaving time for more in-depth and complementary analyses involving a wider range of physicians. Additionally, the greater completeness of the data would allow for the study of subsets of the medical profession, such as rural, female or immigrant physicians.

The method of manually counting or entering data by hand from medical directories has limited historians' research to samples of physicians in hand-picked cities. Using the *AMD*

and other directories, George Weisz has studied categories of specialisation, their changes over time and their variations in different countries.[1] Additionally, James A. Schafer, Jr., used physicians' addresses from the *AMD* to map their uneven geographical distribution in Philadelphia in the first half of the twentieth century in order to explain increasingly uneven access to medical care as suburbs developed.[2]

Although Weisz's and Schafer's works successfully gathered information from medical directories, the printed form of the directories limited the statistical sample set of physician data that these studies could practically use. Weisz, for instance, recorded only the categories of specialisation and derived the rates of individual specialisations for specific cities. Although his sampling of French directories was generally more exhaustive and he supported his data with contemporary estimates of American specialists, Weisz examined relatively few records of individual doctors in the *AMD*. He counted only the specialisations listed for the first 2000 doctors in New York City and the first 1250 in Boston in the 1934 *AMD*, covering approximately 20 per cent and 45 per cent of all the physicians listed in those respective cities. Schafer considered more information about each doctor such as the location of their offices as well as their listed specialities, but he included only one in every five physicians in a single city, Philadelphia, in his research. With those limited data sets, these scholars were able to make interesting and convincing conclusions about the physicians they examined. However, if they could have collected that information in minutes or hours from a database rather than manually collecting and collating that data over months or years, their studies could have included many more doctors, in a broader range of locales and over more years. More inclusive findings would have been possible if the effort-cost of data entry had been reduced.

Making *AMD*s and other directories publicly available in a database format would enable historians of medicine to explore processes of professionalisation and specialisation in greater detail and in novel ways. The Trans-Atlantic Slave Trade Database is an interesting example of how historians have made innovative use of data originally compiled for other reasons. Emerging from the demographic history moment of the 1980s, the Slave Trade Database was envisioned as a way to more specifically map the African origins of slaves in the Americas.[3] However, since its creation and publication, historians have used the database to explore a variety of other related topics, including slave names and environmental and cultural limitations on the trade. In short, the availability of easily accessible data has encouraged scholars to pose and answer new questions. Undoubtedly, making a database of historical physician data would have a similar impact on the history of medicine.

In an effort to make historical *AMD*s more accessible for analysis, I have developed a process to extract the print data and enter it into a spreadsheet or relational database. Using the 1918 edition as a test case, I have prototyped a process that involves first scanning the printed *AMD*, text recognising the scans and then running the now digital text through custom programs in order to consolidate printed lines into single-line entries for each

[1] Although Weisz has used a variety of medical directories from multiple countries in a handful of published works, I will limit my analysis of his use of directories to the *AMD* in one article and a subsequent monograph: George Weisz, 'Medical Directories and Medical Specialization in France, Britain, and the United States', *Bulletin of the History of Medicine*, 71, 1 (1997), 23–68; George Weisz, *Divide and Conquer: A Comparative History of Medical Specialization* (New York: Oxford University Press, 2006).

[2] James A. Schafer Jr., *Business of Private Medical Practice: Doctors, Specialization, and Urban Change in Philadelphia, 1900–40* (New Brunswick, NJ: Rutgers University Press, 2014).

[3] See 'Voyages: The Trans-Atlantic Slave Trade Database,' 1 August 2017, http://www.slavevoyages.org/.

https://doi.org/10.1017/mdh.2017.71 Published online by Cambridge University Press

physician and to parse those entries to populate database fields, which are akin to cells in a spreadsheet. In more detail, these steps involve the following.

1. The 1918 *AMD*'s pages were scanned at 300 pixels per inch and then batch cropped and straightened using an Adobe PhotoShop script. Straightening the images helps text recognition programs to better identify lines of text, creating a more accurate output. Finally, these images were consolidated into a common pdf file for easier management.

2. The pdf of images was then text recognised using Nuance OmniPage. Because the *AMD* text contains numerous abbreviations, shorthand notation and unique symbols, the text recognition process required heavy user oversight to pick between possible interpretations of the more complex lines of text. Especially because the syntax mixes numerals and letters, the program has difficulty knowing how to interpret characters. OmniPage also had to be taught to interpret unusual characters used by the *AMD* as more easily handled letters. For instance, '(l†)' which represents a physician licensed in an unknown year is recognised as '(lt)'. ASCII characters (essentially Latin letters without accents, Arabic numerals and basic punctuation) were chosen to replace non-alphanumeric symbols because not every symbol has an equivalent character, because OmniPage often mistakenly recognises symbols as alphanumeric characters and because it simplified the development of the programs that convert the textual data into a database. Despite the difficulties posed by the unusual symbols and abbreviations, users can successfully text recognise roughly seven pages, containing approximately 900 entries, per hour.

3. The text generated by the recognition process is then run through two custom programs, which will eventually be made publicly available. The first program consolidates city and physician entries that can span multiple lines into single-line entries. Based on formatting and context, the program determines whether a new line of text should be appended to an entry started on an earlier line or whether it is the start of a new entry. The second program then splits the single-line entries into database fields by looking for specific characters (usually punctuation) that separate the different elements in the printed entry, and it determines what kind of information is contained based on the syntax of the element. The output of this process is a spreadsheet or relational database containing between 70 and 90 per cent of the single-line entries created by the first program, varying by state. These success rates are further supported by comparison with the 1918 *AMD*'s printed counts of physicians in the largest cities. Manual validation of the generated database indicates over 92 per cent of its physicians' entries are perfect replications of the print entries.

While earlier statistical uses of medical directory data introduced errors through the possibility of erroneous data entry and limited sample sizes, this programmatic method can incorporate incorrect data because of text recognition errors or irregular formatting in the source directory. Currently, the program that parses the recognised text into database fields ignores all data from malformed or incomplete entries in order to reduce the possibility of including incorrect data. Despite these precautions, other errors creep in as well. Common text recognition mistakes that a computer cannot readily identify include misidentifying characters, especially numerals and punctuation whose meaning is otherwise ambiguous. For instance, the year a physician was licensed could easily be misread, with 1898 being misinterpreted as 1896. However, other data can help identify and correct these situations.

https://doi.org/10.1017/mdh.2017.71 Published online by Cambridge University Press

In the licensure year example, the doctor's medical school graduation year can provide an early boundary for licensure.

Despite needing to recognise the potential problems with a program-created database, the completeness of its data enables researchers to recognise connections that would otherwise not be apparent. For instance, an examination of which doctors list hours in the 1918 *AMD* quickly shows that doctors' listings include hours only in cities with a population over 10 000 or in suburbs of major cities. Additionally, doctors listing specialities were overrepresented amongst those listing hours, relative to the profession as a whole. These business practices could not have been recognised with previous methodologies for analysing medical directories, and they suggest new research questions about physicians' business practices. For instance, the lack of listed hours for general practitioners implies that they were effectively always on call or summonable whereas specialists were not. This suggests that better work-life balance, to borrow a modern phrase, may have incentivised specialisation. Initial analysis of hours listed further supports this suggestion because specialists generally listed a more limited range of hours, usually between 9am and 5pm, when compared with non-specialists. The data contained in *AMD*s can help pose and answer new questions like these, but any conclusions need to be further supported with qualitative analysis. Conversely, a more complete dataset allows researchers to use more sophisticated statistical techniques that rely on larger samples, such as regression analysis, to explore this information in greater depth. Furthermore, as more *AMD* editions are processed, individual physicians could be tracked longitudinally, following their career trajectories and adding elements of temporal change to these analyses.

By prototyping a method to digitise one *AMD* into a database format, this paper demonstrates the potential for creating databases from *AMD*s and other directories. Though this process takes more effort than scanning and automatically text recognising printed works like some digitisation projects, the potential benefit to researchers is similarly greater. Not only would a publicly available historical database of physicians allow scholars easier access to these records, but they could also perform more complex analyses. These new methods could then help to pose and answer new questions about the history of medicine and about the medical profession in particular.

Sean Morey Smith
Rice University, Houston, TX, USA

https://doi.org/10.1017/mdh.2017.71 Published online by Cambridge University Press