# Digitizing Directories: Lessons Learned from Digitizing Historical Directories of Physicians

Sean Morey Smith, Rice University

## Abstract

Digitizing historical directories of physicians into a database allows deeper analysis than previous manual methods but presents novel data integrity issues due to processing errors in comparison to previous sampling methods..

## Project Overview

- Prototype process for digitizing historical medical directories
  - Used 1918 *American Medical Directory*
- Optical Character Recognition (OCR) and parsing introduce errors that lead to incomplete and incorrect data
- Researchers need to be mindful of sources of error when interpreting data

## Background

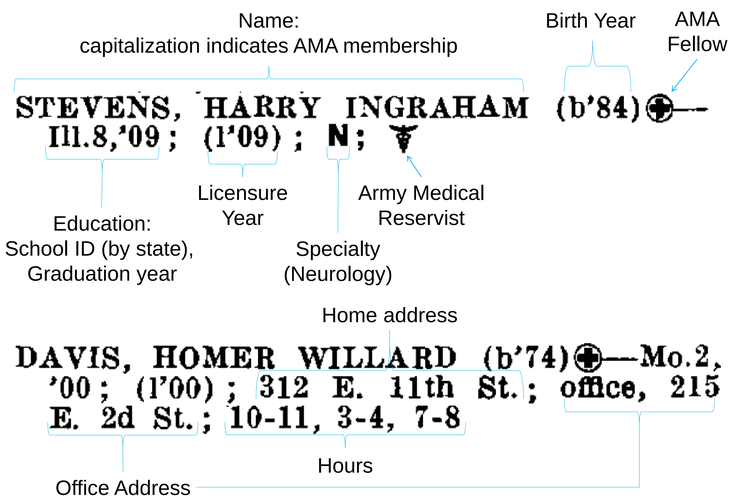
Medical directories are rich sources of information about the historical state of the medical profession. However, their availability as printed text has limited their usefulness to historians of medicine who could more readily delve their contents in a digital format. Consisting of a list of physicians, usually along with their addresses and their professional and specialty affiliations, these directories have been used by historians to explore the consolidation of the medical profession and the emergence of specializations. However, because researchers have been limited by the print form of the directories, their work has been based on relatively small sample sets of entries from them, focused only on select major urban areas. Scholars who have used *American Medical Directories* previously have only examined specific urban centers, with non-random samples of between 20% and 50% of chosen cities' physicians.

Digitizing these directories into a database allows for a broader-based historical analysis of physicians, that can easily cover whole states or the entirety of the United States. However, this prototype effort to digitize one such directory demonstrates that scholars must consider the potential for errors in the digitization process when interpreting its output.

## Acknowledgements


- Thanks to Monica Rivero, Jean Aroom, and Lisa Spiro of Rice University's Fondren Library for technical and logistical support
- Work undertaken while a Graduate Student Fellow of the John E. Sawyer Seminar on the Comparative Study of Cultures funded by the Andrew W. Mellon Foundation at Rice University


## The Source Data



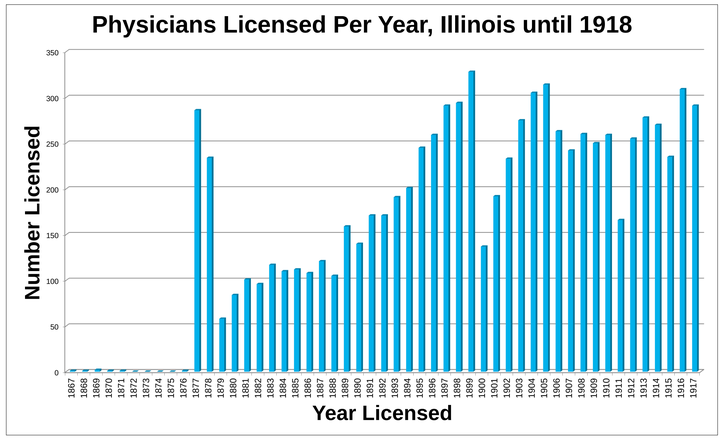

## Procedure

- **Step 1** • Text scanned, then batch cropped and straightened
- **Step 2** • Images consolidated into PDF and OCRed with heavy user oversight
- **Step 3** • Lines of text heuristically grouped into entries
- **Step 4** • Entries parsed into data fields based on separators
- **Step 5** • Parsed data output as spreadsheet or database

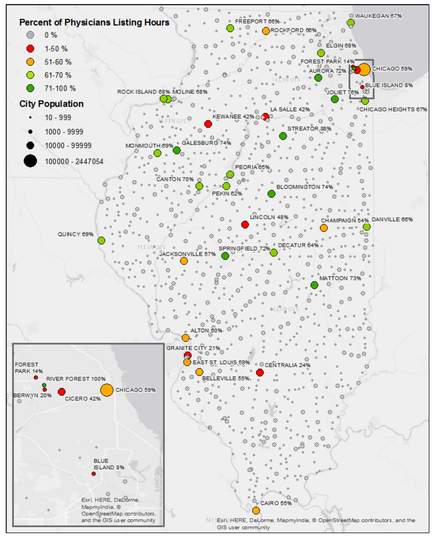
## Types, Sources, and Rates of Error

- Recognized errors result in missing entries (<30% of input published entries)
  - OCR errors due to characters being misrecognized (often incorrect punctuation separators)
  - Occasional oddities in the print format
  - Programmatic errors in line grouping program
- Unrecognized errors result in incorrect data (<8% of output database entries)
  - OCR errors due to characters being misrecognized (often numerals or letters substituted for another numeral or letter)
  - Programmatic errors in line parsing program, especially free text being confused with an address

## Interpreting the Data




The database includes a few clearly erroneous license years from before Illinois started licensing in 1877. However, the overall patterns reflect the states' tightening of licensure by requiring a diploma and exam after 1899 and by requiring more education from 1911.




Even without perfect and complete data, clear patterns are visible:

- The AMD does not list hours for physicians in cities with a population less than 10,000, except for Chicago suburbs
- For smaller cities, especially those in Illinois's more rural south, the AMD lists hours at a lower rate

## Conclusions

- This prototype method allows for the acquisition of much larger datasets than previous methods, allowing analyses of entire states
- OCR quality biggest single determiner of output data quality
- Human oversight and intervention between steps reduces error propagation greatly
- Even with digitization errors, volume of data outweighs problems