Mapping and Visualizing Demographic Information in Structured and Unstructured Clinical Data

Clair A. Kronk, BSc[1], Danny T. Y. Wu, PhD, MSI[1]
[1]Department of Biomedical Informatics, University of Cincinnati, Cincinnati, Ohio

Abstract

Demographic information—race, age, gender, etc.—is becoming increasingly vital in medical care in the context of social determinants of health (SDOHs). However, not many electronic health record (EHR) systems include systematic methods of collecting, managing, and utilizing such information. Here, we present a way to glean SDOH data from clinical notes and present it in an aesthetically pleasing manner. In our application, Intersect, we allow identification and visualization of values such as gender identity, sex assigned at birth, sexual orientation, family status, and ethnicity. This allows researchers to pan through clinical notes to create viable test populations and identify potential risk factors.

Introduction

Galea et al. determined that, in 2000, approximately 874,000 deaths were attributable to social factors such as racial segregation and income inequality (1). Unfortunately, many electronic health record (EHR) systems do not have the infrastructure necessary to collect and manage such data systematically (2,3). Despite this, many social factors may be recorded in clinical notes, either purposefully, inadvertently, or accidentally, and can be retrieved using text mining techniques (4).

For example, an EHR may note that a patient is a transgender woman, telling us that the patient was assigned male at birth (AMAB) and self-identifies as female. Another EHR may note "transgenderism" and "hysterectomy." Given that hysterectomy is removal of the uterus, this patient is more than likely a transgender man, assigned female at birth (AFAB). A prostate examination would indicate to us that the patient was AMAB, without any other mention of the patient's sex. These "hidden" features or rules can be utilized to help clarify vulnerable or otherwise notable populations for different disease or disorder groups. Here we propose a system to extract SDOH data from clinical notes and present the information using an interactive visual dashboard.

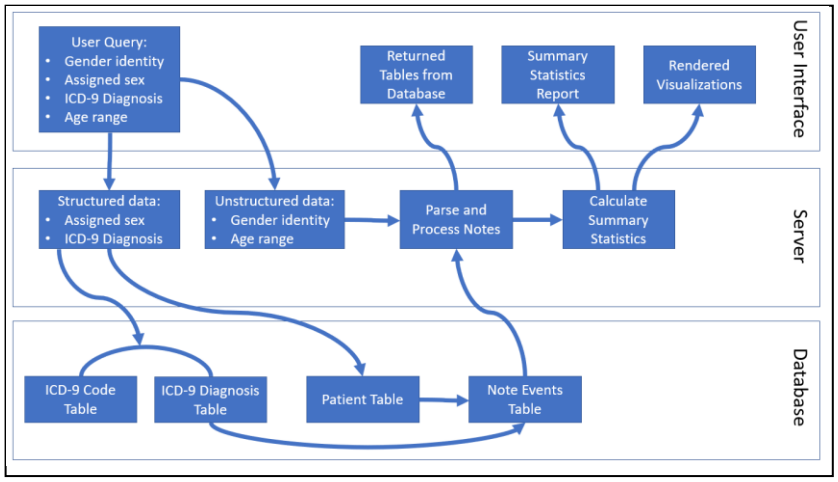

Figure 1. Outline of application structure dissecting user query response. Both UI and server built using R Shiny. Database system was built using PostgreSQL.

Method

The system combines a PostgreSQL backend version of the MIMIC-III database (5) and an R Shiny frontend which can be run in browser (Fig. 1). The frontend makes calls to the backend to request both structured and unstructured data before combining them into both visual and tabular results. This allows not only for test population mapping but also for diagnosis incidence comparison—i.e. does *x* diagnosis affect men and women equally in this given database. This is especially important when considering mental health disorders as significant differences in prevalence rates abound in depression, panic disorder, social phobia, antisocial personality disorder, and alcohol and drug dependence, to name a few (6). Incongruence of gender identity may contribute to mental health as well, although this may be due to external social factors (7).

One issue is that gender identity may be recorded in unstructured notes and cannot be identified without techniques such as regular expression or name entity extraction. Since there is a lack of standardized terminology to capture gender identity, it is difficult to develop a set of rules to extract that information. In our pilot exploration of clinical notes, at least 30 mappings, determined via manual exploration of clinical notes, for "transgender woman" and "transgender man" were deemed required to map over 90 percent of such persons in MIMIC-III successfully. However, additional rules involving mentions of

procedures (hysterectomy, penectomy, etc.) and medications (spironolactone, estradiol, etc.) help catch the last 10 percent or so. There is a significant room for improvement to identify gender identity and other SDOH information in clinical notes.

Proof-of-Concept: Mapping a Test Population

Our system works by allowing one to select multiple criteria such as gender identity, age range, and/or ICD-9-CM diagnosis (see Fig. 2). One can query for 30-40 year old male patients with breast cancer, a disease in which differentiation between designated male at birth and cisgender male identity are important—transgender women using hormone therapy, but marked as assigned male could potential confound results as the risk factors are still considered undefined (8). Several lessons were learned through this proof-of-concept. Over the course of multiple runs, various false positives were found and corrected with additional rule structures. For example, phrases such as "x-year-old" could not be taken at face value as some notes included "x-year-old son" or "x-year-old daughter." Sexual orientation was also non-specific, in that, if a patient was determined to be AFAB and to have a female gender identity, the word "wife" would seem to indicate that the patient was a lesbian. However, bisexuality could not necessarily be entirely eliminated in this case, leading us to return "lesbian or bisexual" instead. Finally, certain structural issues such as physician discrimination or "copy-paste" notes, could not effectively be accounted for in the MIMIC-III dataset at this point in time.

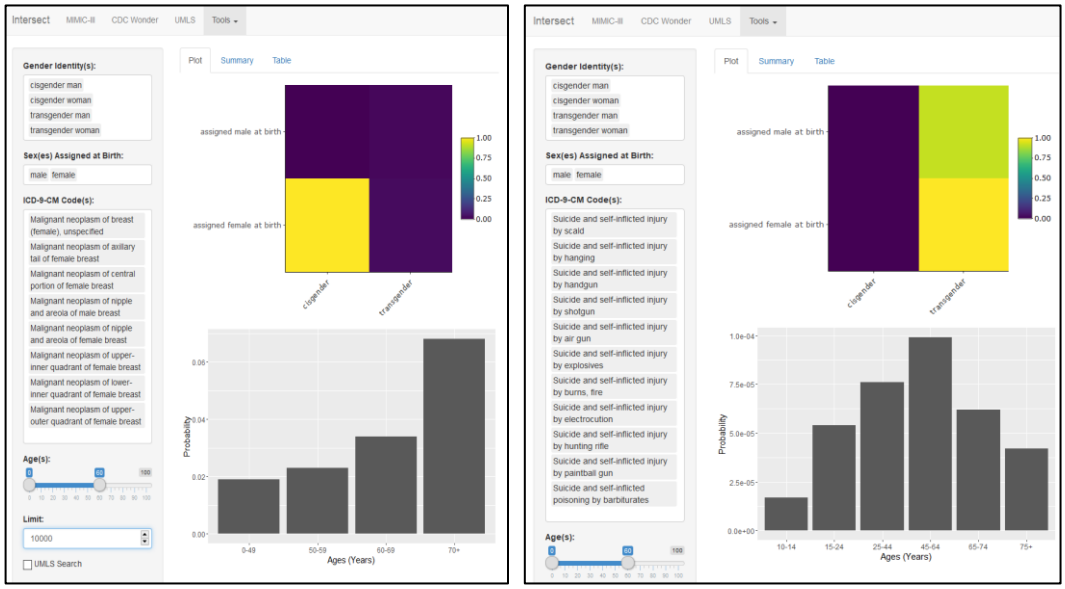

Figure 2. Example mockup of basic selection criteria for population identification and summary of results, showing age, gender identity, and assigned sex at birth distributions. Left: breast cancer incidence. Right: suicide rates.

Conclusion

Visualizing patient populations and elucidating patient cohorts based on demographics and SDOHs such as gender identity are essential to understanding how patients are affected by diseases and disorders in general. We designed and developed an interactive visual dashboard to address this issue. It is our hope to refine this system and promote its use by having researchers identify specific patient groups without violating patient privacy or displaying unnecessary or irrelevant records.

References

1. Galea S, Tracy M, Hoggatt KJ, Dimaggio C, Karpati A. Estimated deaths attributable to social factors in the United States. Am J Public Health. 2011 Aug;101(8):1456–65.
2. Cantor MN, Thorpe L. Integrating Data On Social Determinants Of Health Into Electronic Health Records. Health Aff (Millwood). 2018 Apr;37(4):585–90.
3. Palacio A, Suarez M, Tamariz L, Seo D. A Road Map to Integrate Social Determinants of Health into Electronic Health Records. Popul Health Manag. 2017 Dec;20(6):424–6.
4. Bejan CA, Angiolillo J, Conway D, Nash R, Shirey-Rice JK, Lipworth L, et al. Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. J Am Med Inform Assoc. 2018 Jan 1;25(1):61–71.
5. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016 May 24;3:160035.
6. Eaton NR, Keyes KM, Krueger RF, Balsis S, Skodol AE, Markon KE, et al. An invariant dimensional liability model of gender differences in mental disorder prevalence: Evidence from a national sample. J Abnorm Psychol. 2012;121(1):282–8.
7. Dhejne C, Van Vlerken R, Heylens G, Arcelus J. Mental health and gender dysphoria: A review of the literature. Int Rev Psychiatry. 2016 Jan 2;28(1):44–57.
8. Sonnenblick EB, Shah AD, Goldstein Z, Reisman T. Breast Imaging of Transgender Individuals: A Review. Curr Radiol Rep. 2018;6(1):1.