Visual Analytics to Combat Selection Bias in Retrospective EHR Data Analyses

David Gotz, PhD[1], Jonathan Zhang, BS[1], Smiti Kaul, BS[1], Georgiy Bobashev, PhD[2], David Borland, PhD[1,3]
[1]University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [2]RTI International, Research Triangle Park, NC, USA; [3]RENCI, Chapel Hill, NC, USA

Introduction

Retrospective analyses of electronic health records (EHRs) and other health data sources are increasingly common as investigators seek to employ data collected during routine health care delivery to learn about health practices and outcomes. The data analysis process typically includes cohort selection, in which a group of patients (and their corresponding data) are identified from within a health organization's overall population of patients. Such cohorts are typically defined using sets of inclusion and/or exclusion criteria that are applied as filters within a data query; several interactive tools (such as i2b2 [1] and our own Cadence cohort selection system [2]) have been developed to facilitate this cohort definition process.

However, with such systems, the lack of randomization combined with the high level of expected interdependence between variables can produce cohorts that are highly skewed in ways that are unexpected to the analyst. For example, a cohort focused on a high-priced medication may inadvertently bias a cohort toward the privately insured. This in turn may skew the prevalence of various interventions within the selected cohort such that it is no longer representative of the population which the analyst intends to study. Moreover, these bias effects can be invisible to analysts who do not have access to all of the data within a health system required to detect these shifts.

Methods

To make these selection bias effects more transparent to analysts, we are developing a selection bias report feature as an extension to the Cadence cohort selection system. The existing Cadence system enables users to select and explore cohorts using a combination of advanced visual analytics techniques that depend upon prevalence rates and correlations between different types of medical events (e.g., ICD or CPT codes), providing for various data-driven analysis capabilities [2, 3].

The selection bias report capability calculates prevalence rate statistics for all known variables in the current cohort. In addition, baseline prevalence rates are computed from a representative baseline population. Our prototype includes four broad condition-specific populations (Diabetes, Cancer, Cardiology, Obstetrics) computed from the full UNC Health population over a two-year span. For each baseline population, we compute the prevalence rate of every unique ICD10 code as well as parent codes (to enable aggregation). At runtime, users can request a selection bias report for any cohort (see Fig. 1). We are developing ways to interactively and visually compare the focused cohort in Cadence with a selected baseline population to estimate differences in representation. These differences are prioritized to focus attention on the largest bias effects and communicated via a dynamically generated web page. While we do not

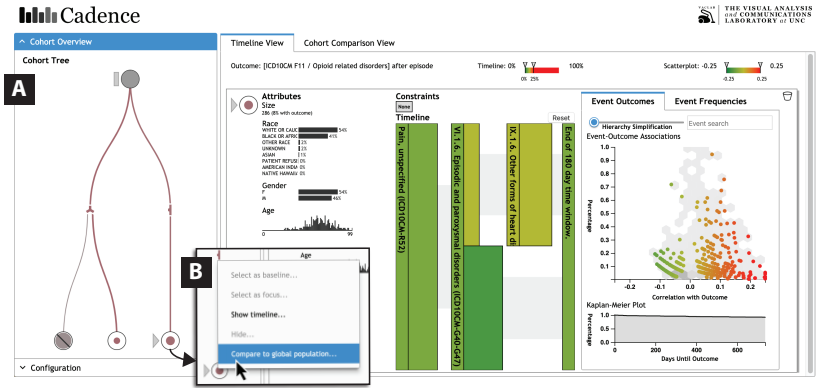

Figure 1: Cadence being used to analyze a cohort of heart failure patients with a pain diagnosis. (A) From the cohort icons in the left sidebar, (B) a context menu enables users to request a baseline population comparison (see Fig. 2).

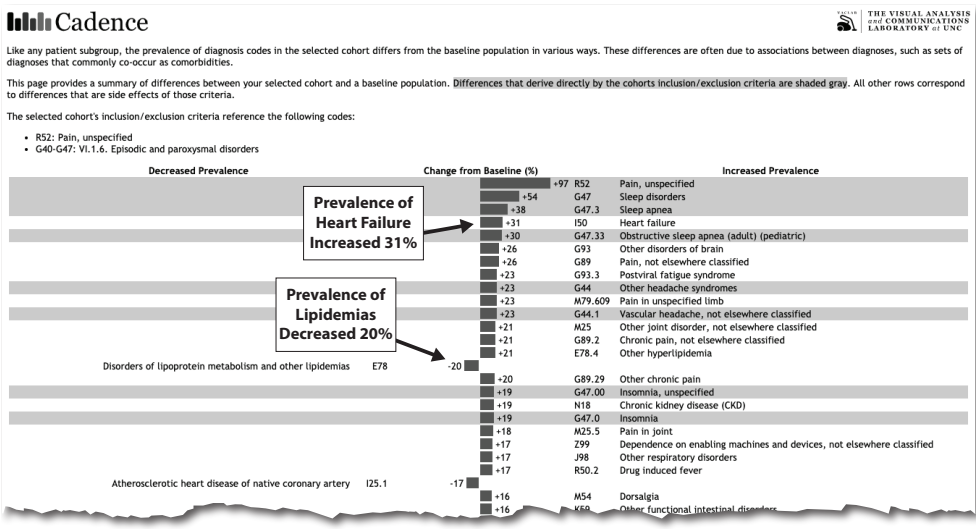

Figure 2: The selection bias report shows a ranked list of differences in prevalence of diagnoses (or families of diagnoses) between the selected cohort and the population baseline. The gray line items represent constrained variables with expected differences in prevalence. The white line items are "side effects" of the inclusion/exclusion criteria.

currently provide statistical significance estimates for the differences, we plan to incorporate bootstrap estimates of the standard errors to account for multiple testing. The report also uses color coding to help analysts distinguish between variables directly changed by the inclusion/exclusion criteria versus those that change as a side effect.

Results

Fig. 1 shows Cadence being used to analyze a cohort of cardiology patients with a pain diagnosis, with an analytical goal of identifying risk factors for opiate abuse/addiction. The analyst has applied additional inclusion criteria to focus on a smaller cohort of patients with "Episodic and Paroxysmal Disorders" because they appear to have a higher rate of opiate disorders. Right clicking on the corresponding cohort icon, the user selects "Compare to global population..." (Fig. 1B) to display the selection bias report (Fig. 2). The report shows that four of the top differences in prevalence are, as expected, related to the inclusion criteria. However, several non-constrained diagnoses (e.g., a 31% higher rate of heart failure) had a higher prevalence than the baseline. In contrast, other diagnoses, such as lipidemias (E78) had a lower prevalence than in the baseline. Initial qualitative feedback from analysts has encouraged a continuing effort to improve our approach.

Conclusion

Integrating baseline prevalence statistics into cohort selection tools enables the creation of selection bias reports that can help contextualize user-defined cohorts and protect against unexpected shifts in distributions which can invalidate findings. User studies and additional interactive reporting capabilities are planned to evaluate and improve these capabilities in the future.[1]

References

1. Shawn N. Murphy, Griffin Weber, Michael Mendis, Vivian Gainer, Henry C. Chueh, Susanne Churchill, and Isaac Kohane. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Jour. of the Amer. Med. Info. Assoc.*, 17(2):124–130, March 2010.
2. David Gotz, Jonathan Zhang, Wenyuan Wang, Joshua Shrestha, and David Borland. Visual Analysis of High-Dimensional Event Sequence Data via Dynamic Hierarchical Aggregation. *IEEE Tran. on Vis. and Comp. Gr.*, 26(1):440–450, January 2020.
3. David Borland, Wenyuan Wang, Jonathan Zhang, Joshua Shrestha, and David Gotz. Selection Bias Tracking and Detailed Subset Comparison for High-Dimensional Data. *IEEE Tran. on Vis. and Comp. Gr.*, 26(1), 2020.

[1]This work is made possible in part by support from the National Science Foundation (Grant #1704018).