Enhancing Electronic Health Records with Tumor Registry Data

Emily R. Pfaff, MS, Ashok Krishnamurthy, PhD, Christopher Baggett, PhD, Timothy S. Carey, MD, MPH

UNC Chapel Hill, Chapel Hill, NC, USA

Introduction

Tumor registries provide high-quality data on tumor type and stage. However, data on follow-up care utilization, testing, outcomes, and non-cancer health data are generally not present in tumor registries. The electronic health record (EHR) may provide this missing detail; however, EHR users are often non-specific or inconsistent in their tumor coding practices. Thus, while both data sources provide valuable information, neither is complete; moreover, linkage may be challenging if regulations prevent the sharing of identifiers. To provide researchers with a fuller picture of patients’ health, we are linking patients in the North Carolina Central Cancer Registry (NCCCR) with patients in the Carolina Data Warehouse for Health (CDWH), UNC Health Care System’s clinical data warehouse. In this paper, we report on the performance of our deterministic, hash-based, privacy-preserving linkage techniques (though identifiers were able to be shared for this study), which we plan to reuse for future projects requiring privacy-preserving linkage.

Methods

We acquired (1) NCCCR data from 2014 (66,928 cancer cases, including duplicates due to multiple cases per patient), and (2) CDWH data, spanning 2009 through 2014 (1,124,107 patients, including duplicates due to changing identifiers over time). Duplicates were purposely left in both datasets in an attempt to match any “version” of a patient in one dataset to any version of that patient in the other. Available identifiers for linkage include social security numbers (SSN), names, dates of birth (DOB), gender, and zip code, with varying levels of completeness in each dataset.

As no gold standard linkage exists between the datasets, we opted to compare the performance of our algorithms with a gold standard algorithm, NCI’s SEER-Medicare algorithm,[1] as described in Dusetzina, et al.[2] Rather than define our denominator as the Cartesian product of all possible links between the datasets, we pre-selected a set of potential matches in a blocking phase using loose criteria (e.g., matching on DOB, first name initial, and gender) also described in Dusetzina, et al. This identified 256,110 potential matches. Using this set of possible matches, we ran three linkage algorithms (Table 1) and compared performance against 10,595 matches identified by the SEER-Medicare algorithm.

Results

Table 1: Sensitivity and specificity of three linkage algorithms, compared with 10,595 SEER-Medicare matches

<table>
<thead>
<tr>
<th>Algorithm</th>
<th># of matches</th>
<th>Sens.</th>
<th>Spec.</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1: Social security number only (hashed)</td>
<td>11,278</td>
<td>0.888</td>
<td>0.992</td>
</tr>
<tr>
<td>#2: First 3 characters first name, first 8 characters last name, date of birth, gender (hashed)</td>
<td>10,310</td>
<td>0.97</td>
<td>1</td>
</tr>
<tr>
<td>#3: First 10 characters first name, last four digits SSN, date of birth, gender (hashed)</td>
<td>8,571</td>
<td>0.809</td>
<td>1</td>
</tr>
</tbody>
</table>

Discussion

A goal of our study is to evaluate identifier subsets that work well for deterministic, hash-based, privacy-preserving linkage, which can be reused for other studies that require linkage but are unable to share identifiers.[3] Though it is a gold standard, SEER-Medicare’s logic would be quite complex to code in a manner to allow for privacy-preserving, hash-based linkage; thus, Algorithm #2’s high performance and simplicity makes it a strong contender. Our data show that algorithms requiring SSN (or last four digits) are at a disadvantage, as SSN may be recorded inaccurately (leading to false positives) or not at all. (SSN was missing in 16.2% of our pre-blocked records.) For this use case, post-linkage, “flags” can be exported and provisioned to the stewards of both NCCCR and CDWH, allowing each side to know which of their patients also exist in the other data source. This information can be used to more accurately identify cohorts of cancer patients, and to allow researchers to access a more holistic picture of patient data in the cancer space.

References