An Algorithm to Identify Patient Relocation Events (‘Moves’) from EHR Data
Jessica Meeker¹, MPH, Heather Burris², MD, Mary Regina Boland¹, MA, MPhil, PhD
¹Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania
²Department of Pediatrics, Children’s Hospital of Philadelphia

Abstract. Environmental, social and economic exposures can be extrapolated using address information recorded in Electronic Health Records (EHRs). However, these data often contain administrative errors. In addition, patients can move from one residence to another (i.e., a relocation event). Understanding whether a patient has moved or not is important for appropriate exposure assessment. We developed an algorithm to identify patient relocation events (PREs) from EHRs with 95.7% accuracy, evaluated using manual review of 3,362 patient addresses.

1. Introduction. Accurate information regarding where people live is crucial for the characterization and assignment of exposures based upon a residential address. Most of these approaches focus on location as a static point without incorporating patient relocation events (PREs) or moves. Residential mobility is important to incorporate into exposure assessment otherwise differential misclassification can occur. However, to our knowledge no study has put forth an algorithm that can deal with very large Electronic Health Record (EHR) datasets that contain both administrative errors and true relocation events. In this study we construct an algorithm to adjust for administrative errors and identify PREs. We apply this method to a large hospital cohort.

2. Methods. We designed an algorithm to determine whether a PRE had occurred (Figure 1). We cleaned the data by making all text uppercase, and abbreviating street and avenue. We also discarded all unit/apartment number information. We chose to disregard apartment number changes because in many cases it was absent for one address entry. We used the Damerau–Levenshtein (DL) distance string metric to determine the number of character differences between addresses. The DL metric allows for transpositions, for example “Guardian” and “Gaurdian” would be recognized as being the same street name. We then included the rule that if a first numeric variable was present, and it was the same then it would be counted as a non-move (e.g., "423 Guardian Drive" and "423 Guard Dr" would be considered a non-move even though there are 6 changes between the 2 addresses). Additionally, we determined that even if the character difference were less than or equal to five, but the address number was present and different (“423 Guardian Drive” and “123 Guardian Drive”) it would be counted as a move.

3. Results. We manually reviewed 3,362 addresses to assess the accuracy of our algorithm, which was 95.7% accurate. Our algorithm outperformed a method using only changes in zip codes, which achieved 82.9% accuracy. Two issues occurred for zip code moves: a) moves occurred within zip code and b) data entry errors with the zip codes (e.g., inversion of numbers). Our method was robust to identify moves within zip codes and was not as dependent on inversions.

4. Discussion. Our study addresses an important exposure characterization problem of those factors that are assigned based upon residential addresses. Identifying patients that move will enable assignment of partial exposures based upon the old address from where someone moved from and the new address for a more accurate representation of individual exposures. This will allow for more accurate assessment of the association between exposures and outcomes, giving researchers and policy makers more reliable information. We will make our algorithm shareable for others using EHR data.

References

Figure 1. Algorithm for Identifying PREs from EHRs