Understanding the profile of errors that cause duplicate entries in a patient registry

Scott L DuVall[a], Janice Conrads[b], Alison Fraser MSPH[b], Geraldine Mineau PhD[b,c]

[a]Department of Biomedical Informatics, University of Utah, United States of America

[b]Pedigree and Population Resource, Huntsman Cancer Institute, University of Utah, United States of America

[c]Department of Oncological Sciences, University of Utah, United States of America

Background and Significance

Duplicate records are detrimental to the cost-effective and efficient delivery of health care.[1] Manually identifying and resolving duplicates can cost $60 per case.[2] Patterns have been found in the types of errors that occur in patient registries, suggesting that undetected duplicate records may be similar to those already identified.[3,4]

At the University of Utah, records from all community clinics are merged with hospital records in the Enterprise Data Warehouse (EDW). The Pedigree and Population Resource group at Huntsman Cancer Institute links demographic records from the EDW to the Utah Population Database (UPDB). In last year's linkage, 76,922 duplicate records were identified. The purpose of this study was to compare the differences between clinic and hospital records in the EDW with existing literature.

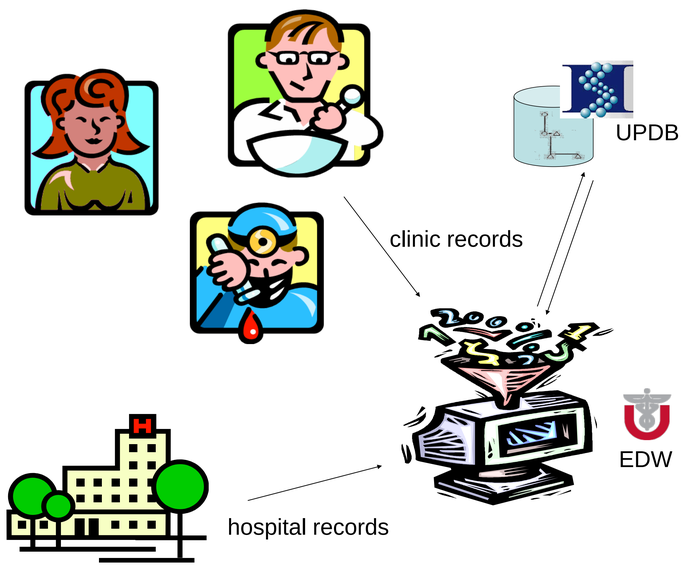

Figure 1. Merging of records in EDW

Methods

Error types described in literature were gathered. A Java program was created to examine and categorize known duplicate records into these error types using state machine templates. The larger string was made into a state machine that consumed the shorter string. If the final state was an "accept" state, the names were categorized as a match.

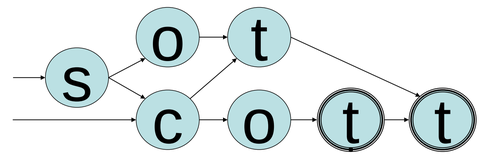

Figure 2. State machine template for Scott allowing one delete

Results

The duplicate records identified in the EDW had approximately the same error types in almost the same order of frequency as those published in literature. Nicknames, different last names for females, and missing social security numbers were much more common in the EDW. Punctuation and spaces and family collisions were less frequent.

	UUHSC	Friedman[3]
Extra names and titles	34.3%	36.9%
Nicknames, spelling variations	21.8%	13.9%
One letter substitutions	13.6%	13.7%
One letter added or deleted	7.6%	12.9%
Punctuation or spaces	1.9%	11.8%
Different last names for females	12.9%	7.8%
Permuted parts of names	3.2%	1.4%
Different first names	2.8%	1.4%
One letter transposed	1.9%	0.8%

Figure 3. Discrepancy in names

	UUHSC	Grannis[4]
Missing SSN	52.4%	35%
Typographical errors	62.7%	35.5%
Spouse (family) collisions	14.8%	47.5%
Unexplained collisions	9.9%	17%
Unexplained mismatch	12.6%	----

Figure 4. Discrepancy in social security numbers

Conclusion

Duplicates in the EDW showed some significant differences from published literature. Duplicate records that did not fit into published error type categories were examined and new categories were defined. By determining which error types exist in patient registries, duplicate records can be better detected and interventions may be introduced to prevent duplicates from happening in the first place.

References

[1]Mays S, Swetnich D, Gorken L. Toward a Unique Patient Identifier. Health Manag Technol. 2002 Mar;23(3):42-4.

[2]Thornton SN, Hood SK. Reducing Duplicate Patient Creation Using a Probabilistic Matching Algorithm in an Open-access Community Data Sharing Environment. Proc AMIA Symp 2005:1135.

[3]Friedman C, Sideli R. Tolerating spelling errors during patient validation. Comput Biomed Res 1992;25:486-509.

[4]Grannis SJ, Overhage JM, McDonald CJ. Analysis of Identifier Performance using a Deterministic Linkage Algorithm. Proc AMIA Symp 2002:305-9.