Machine Learning Approaches to Identify Nicknames from A Statewide Health Information Exchange

Suranga N. Kasthurirathne, BEng[1,2], Shaun J. Grannis, MD, MS[1,2,3]
[1]School of Informatics and Computing, Indiana University, Indianapolis IN, USA; [2]Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, IN, USA; [3]School of Medicine, Indiana University, Indianapolis, In, USA

Abstract

Patient matching is essential to minimize fragmentation of patient data. Existing patient matching efforts often do not account for nickname use. We sought to develop decision models that could identify true nicknames using features representing the phonetical and structural similarity of nickname pairs. We identified potential male and female name pairs from the Indiana Network for Patient Care (INPC), and developed a series of features that represented their phonetical and structural similarities. Next, we used the XGBoost classifier and hyperparameter tuning to build decision models to identify nicknames using these feature sets and a manually reviewed gold standard. Decision models reported high Precision/Positive Predictive Value and Accuracy scores for both male and female name pairs despite the low number of true nickname matches in the datasets under study. Ours is one of the first efforts to identify patient nicknames using machine learning approaches.

Introduction

The siloed implementation of health information systems and legal restrictions preventing the use of a national level patient identifier[1] has led to the fragmentation of patient information across the US[2]. Fragmentation of patient data impacts data collected within a healthcare system, which may report different patient Identification Numbers (ID's) for data collected at various facilities or clinics, as well as data collected across multiple healthcare systems[3, 4]. Fragmented patient data impedes the delivery of quality patient care by preventing providers from accessing complete patient records, causing inefficiencies and delays, hindering public health reporting and leading to enhanced patient risk[5, 6]. Efforts to address patient fragmentation focus on probabilistic and deterministic patient matching efforts[7] driven by various patient demographics such as patient names, gender, date of birth, address, telephone numbers, as well as identification numbers such as Social Security Number (SSN), etc.

Patient matching accuracy is strongly influenced by the quality and accessibility of data required. Certain data elements may be costly to obtain, incomplete or incorrect. Various clinics and facilities may not capture the same elements in the same format, leading to the need for additional data standardization efforts[8]. Further, not all data elements contribute equally towards matching. The discriminative power contributed by various patient demographics may differ significantly by data type and data set. As an example, patient name elements such as first name, middle name, last name, suffix and SSN may be more diverse, and therefore, hold more discriminative power than data elements such as patient gender, zip code or state of residence.

Patient name elements are widely collected and commonly used pieces of identification used within the healthcare system[9]. However, inconsistencies in the usage and reporting of names pose a significant challenge to patient matching. Some inconsistencies may be caused by misspellings, which conventional patient matching tools address using string comparators[10, 11]. However, string comparators may not address inconsistencies resulting from use of nicknames. Consequently, we hypothesize that supplementing existing patient demographic data with imputed nickname information may improve the accuracy of patient matching.

Nicknames are widely used and researchers have documented evidence of phonological and structural patterns in their use.[13] For example, nicknames can be phonologically similar to given name (i.e. 'Kathryn' and 'Kitty')[12]; they may be based on structural variations such as spelling variations (i.e. 'Vicki' and 'Vickie'), diminutive variations (i.e. the diminutive 'Betty' to the more formal 'Elizabeth') and cross the gender divide (i.e. the nickname 'Andy' may be used for both the female 'Andrea' as well as the male 'Andrew'). However, manually creating nickname lookup tables relevant to a specific population requires significant effort. Further, such efforts would be limited by the reviewers' knowledge and perception of nicknames. An alternate approach is to develop decision models to impute nickname

pairs based on phonological and lexical similarity. Approaches for evaluating phonological similarity and/or patterns in names have been developed previously and include string comparators, phonological similarity measures, N-gram distributions that evaluate term similarities, as well as various algorithms that predict race/ethnicity and gender[13]. These methods present significant potential to provide a wide range of information on the structure and phonological similarity of English nicknames.

We describe efforts to leverage such measures to develop decision models capable of identifying human nicknames from a statewide Health Information Exchange (HIE).

Materials and Methods

Data extraction

We extracted patient data from the master person index of the Indiana Network for Patient Care (INPC)[14], one of the longest continuously running HIE's in the US. The INPC covers 23 health systems, 93 hospitals and over 40,000 providers[15]. To date, the INPC contains data on over 15 million patients having more than 25 million registrations (the same patient can be registered at multiple HIE participants). We used the INPC's patient matching service to identify the same patient across multiple institutions. Next, we analyzed first names for all patients with multiple registrations, and created 'name pairs' when first name for the same patient differed for separate registrations. We excluded all name pairs with mismatching or missing genders, occurred 3 times or less, or contained invalid phrases such as MALE, FEMALE, BOY, GIRL or BABY. For name pairs with frequencies ranging between 3 and 20, we also removed any pairs with Jaro-Winkler[16] or Longest Common Subsequence (LCS)[17] scores of 0. The remaining name pairs were split into male and female genders, and serves as our name pair dataset.

Development of a gold standard

Each first name pair was reviewed by two independent reviewers who tagged each name pair as TRUE (is a nickname) or FALSE (not a nickname). In the event of a disagreement, a third reviewer served as a tiebreaker. Reviewers selected diminutive nicknames as well as nicknames based on phonological and lexical similarities. Nicknames based on familial relationships ('Sr.' for father and 'Jr.' or 'Butch' for son) [18], order of birth or occupation ('Doc' or 'Doctor' used for either a 7th child or a physician) as well as those based on external attributes or personality ('Blondie', 'Ginger', 'Brains' etc.) were not considered for this study.

Preparation of feature sets

We calculated a number of features to represent the phonological and lexical similarity of each first name pair under study (Table 1).

Table 1. Features calculated per each first name pair

Feature name	Description
Frequency	Number of times that the name pair under consideration appeared in the INPC dataset
Modified Jaro-Winkler comparator (JWC)	String comparator which computes the number of common characteristics in two strings, and finds the number of transpositions for one string to be modified to the other[16].
Longest common substring (LCS)	String comparator which generates a nearness metric by iteratively locating and deleting the longest common substring between two strings[17].
Levenstein edit distance (LEV)	String comparator which calculates the minimum number of single character edits (insertions, deletions or substitutions) necessary to change one string into the other[19]
Combined Root mean square	The combined root mean square score of JWC, LCS and LEV string comparators[3].

We also calculated a number of features for each individual name in each name pair (Table 2).

Table 2. Features calculated per each name

Feature name	Description

Race/ethnicity	We used the python ethnicolr package[20] to categorize each name into one of the following categories; white, black, Asian or Hispanic.
Gender	We used the python gender-guesser package[21] to categorize each name into one of the following categories; male, female, androgynous (name is used by both male and female genders) and unknown.
Soundex	Phonetic encoding algorithm based on word pronunciation, rather than how they are spelled[22]
Metaphone	Phonetic encoding algorithm which includes special rules for handling spelling inconsistencies as well as looking at combinations of consonants and vowels[23].
The New York State Identification and Intelligence System algorithm (NYSIIS)	Phonetic encoding algorithm with 11 basic rules that replace common pronunciation variations with standardized characters, remove common characters and replace all vowels with the letter 'A' [3, 24]. The NYSIIS algorithm is more advanced than other phonetic algorithms as it is able to handle phonemes that occur in European and Hispanic surnames.
Number of syllables	We developed a java program that counts the number of syllables in each name using existing language rules[25]. The validity of the program was assessed via manual review of test data.
Bi-Gram frequencies	Researchers have calculated bi-gram frequencies of English words[26]. Frequently occurring bi-grams may represent common phonological sounds. Thus, names that contain multiple commonly occurring phonological sounds have a much higher chance of representing nicknames. We calculated a normalized score representing the frequency of bi-gram counts for each name.
Misspelling frequencies	By computing appearance of bi-grams that occur very infrequently[26], we also calculated a measure for potential misspellings.

In addition to the string comparators listed in Table 2, we created a binary feature agreement vector indicating which of these features agreed for each name pair.

For male and female name pairs, we developed name pair vectors consisting of the feature sets described in Tables 1, 2 and the binary feature agreement vector.

Machine learning process

We leveraged python and the scikit-learn machine learning library[27] to build XGBoost[28] classification models to identify nicknames across male and female name vectors. The XGBoost algorithm is an implementation of gradient boosted ensemble of decision trees[29] designed for speed and performance. XGBoost classification was selected as (a) our own research suggests that ensemble decision trees performed compatibly, or better than other classification algorithms[30, 31], and (b) the algorithm has demonstrated superior performance to other classification algorithms in machine learning competitions organized by Kaggle[32].

In building these models, we sought to address data imbalance present in both name vectors, as well as model overfitting. We split each data vector into random groups of 90% (training and validation dataset) and 10% (holdout test set). Previously, researchers have leveraged both oversampling[33] and under sampling methods[34] to address this challenge. After an exploratory analysis, we adopted the Synthetic Minority Over-sampling Technique (SMOTE)[33] to boost the imbalanced class (nicknames match) . However, various levels of boosting may have different impact on model performance. Similarly, the XGBoost algorithm consisted of multiple parameters which could each impact model performance. Thus, we decided to perform hyperparameter tuning using multiple versions of the training dataset that had been balanced using different boosting levels. Hyperparameter tuning was performed using randomized search and 10-fold cross validation. Features that were modified as part of the hyperparameter tuning process are listed in Appendix A.

The best performing models identified by hyperparameter tuning were applied to the holdout test datasets, which were not artificially balanced via boosting. This ensured that the best decision model would be evaluated against a holdout dataset with the original prevalence of nickname pairs, ensuring that the model was suitable for implementation. Figure 1 presents a flowchart describing our study approach.

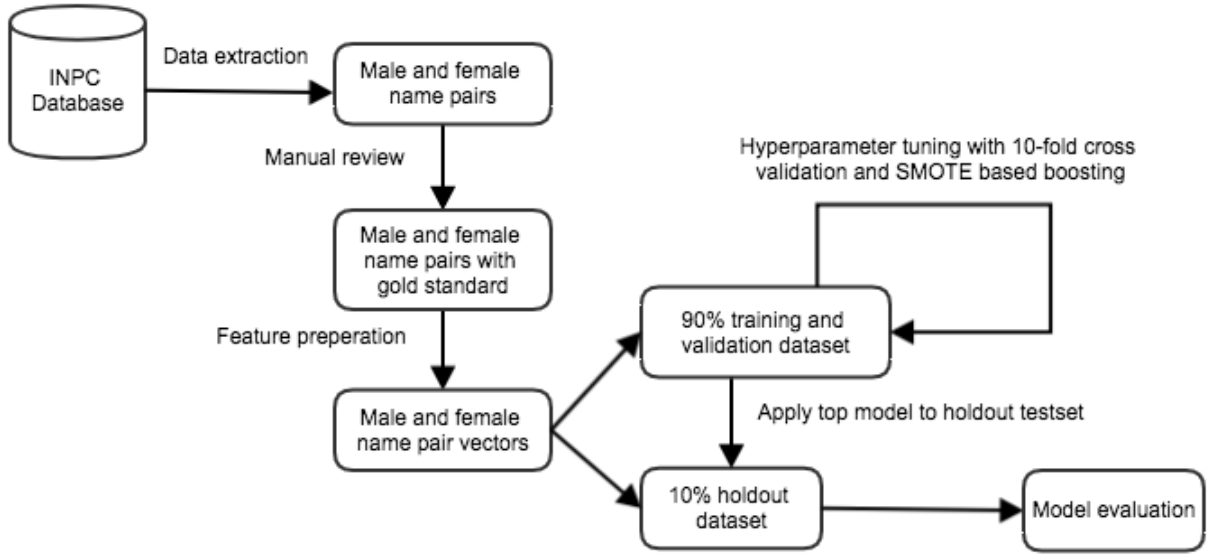

Figure 1. Workflow presenting the complete study approach from data extraction to decision model evaluation.

Analysis

We calculated Positive Predictive Value (PPV) aka precision, sensitivity aka recall, accuracy and f1-score (the harmonic mean between precision and recall) for each decision model under test. Traditionally, area under the ROC curve (AUC) is considered an important performance metric. However, literature suggests that precision-recall curves are more accurate than AUC curves for evaluating unbalanced datasets[35]. Thus, we prepared precision-recall curves for each decision model.

Results

We identified a total of 11,986 male name pairs and 15,252 female name pairs. The manual review of these identified 291 (2.428%) of the male name pairs and 671 (4.4%) of the female name pairs as true nicknames. Kappa scores for male and female nickname reviews, as performed by the two primary reviewers were 0.810 and 0.791 respectively. These scores indicate very high levels of inter-rater reliability in the manual review process.

Figure 2 presents a breakdown of the frequency of true nickname matches as a function of Jaro-Winkler scores for male and female name pairs. The preponderance of male and female nickname match scores for true nicknames ranged from 0.7 to 0.85, with a steep drop as the score approached 1. As presented in Figure 3, frequency of most non-nickname pair scores for male and female datasets ranged between 0-0.05. Pair frequency dropped to 0 from Jaro-Winkler scores between 0.1 and 0.3, after which they rose significantly until Jaro-Winkler scores of 0.5. Frequencies for both male and female datasets fell drastically as Jaro-Winkler scores were increased further.

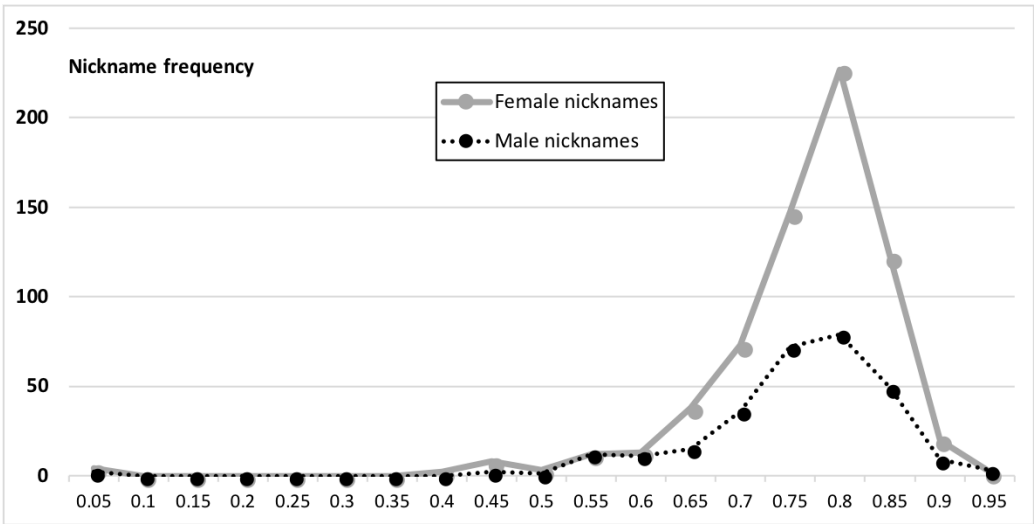

Figure 2. Frequency of nickname matches across Jaro-Winkler scores (0-1)

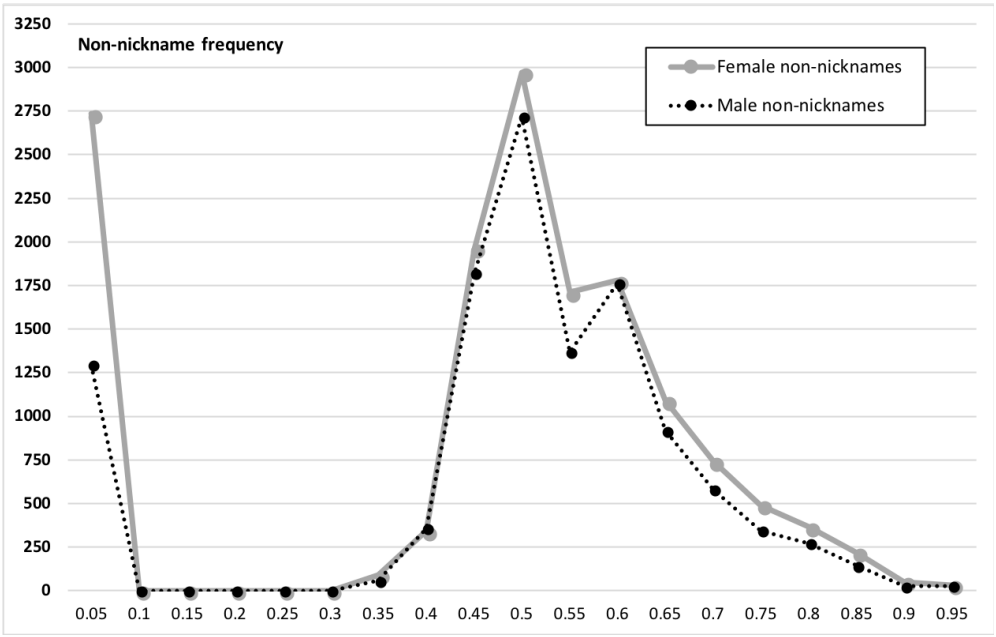

Figure 3. Frequency of non-nicknames across Jaro-Winkler scores (0-1)

Table 3 reports the predictive performance of optimum decision models selected by hyperparameter tuning applied to the holdout test datasets. Figure 4 presents the precision-recall curves reported by these models. Appendix B lists the most important features that contributed to the male and female decision models. Importance was determined by the XGBoost classification algorithm's internal feature selection process which evaluates the number of times a feature is used to split the data across all trees[36].

Table 3. Predictive performance of the machine learning models applied to the holdout test datasets

Performance measure	Male nickname model (%)	Female nickname model (%)
Positive Predictive Value (PPV) aka precision	85.71	70.59
sensitivity aka Recall	42.86	64.29
accuracy	98.50	97.71

f1-score	57.14	67.29

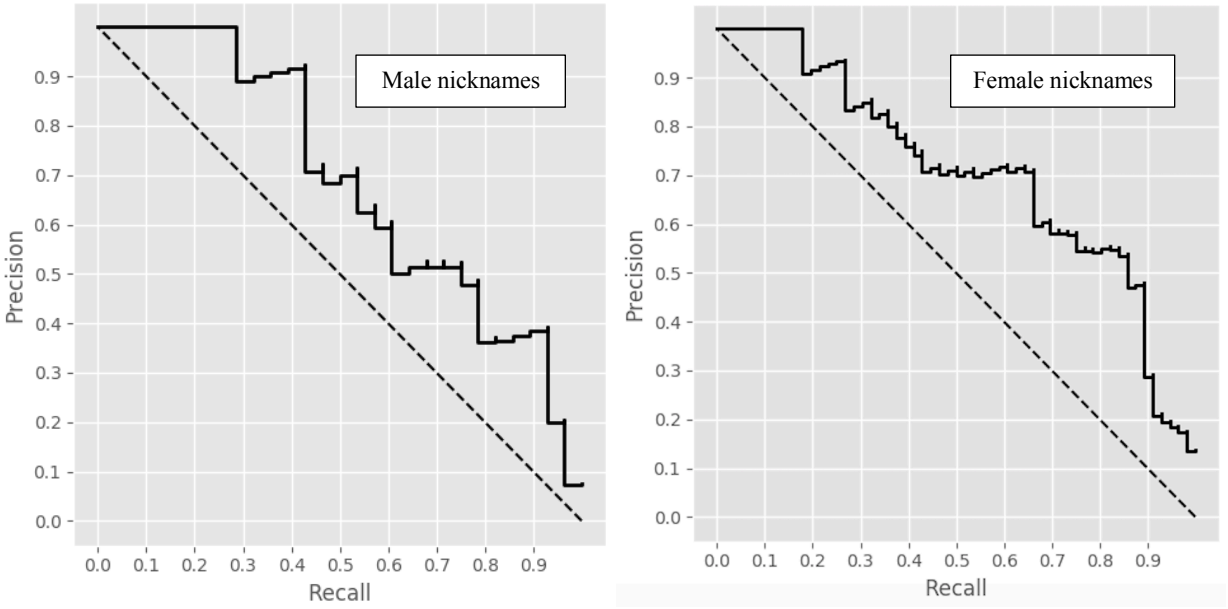

Figure 4. Precision-recall curves reported by male and female nickname prediction models

Decision models performed best when the ratio of true nickname matches to false nickname pairs were boosted to 0.2 for the male nickname model, and 0.3 for the female nickname model. Despite the highly imbalanced nature of the holdout test datasets, decision models performed significantly well with high precision/PPV scores. Both models reported exceptionally high accuracy scores (>97%). However, this is attributed to the unbalanced data of the test data being used[37, 38]. Both models also reported mid-level sensitivity/recall and F1-scores. The weak F1-score is justified on the grounds that it represents a balance between precision and recall. However, our use case requires higher precision/PPV, which makes the models suitable for application in real world scenarios. This also justifies the weak sensitivity/recall scores reported for each model.

Discussion

We present one of the first efforts to impute person nicknames using machine learning approaches. Our analysis revealed that the use of nicknames was higher among females (4.4%) than males (2.4%) within our HIE dataset. Our decision models achieved adequate performance despite the low number of true nickname matches in each of the datasets. The high precision/PPV achieved by each decision model suggests suitability for use in the healthcare domain, where accurately matching patient records is a crucial function. Further, the male nickname model reported significantly high precision/PPV scores despite the male name pair dataset being more imbalanced than the female name pair dataset. However, the sensitivity/recall and F1-scores produced by the male nickname model were lower than the female models. Overall, these results demonstrate the possibility of leveraging existing measures of phonological and lexical similarity to develop decision models to identify true nickname matches. However, these decision models were generated using name pairs from a large scale HIE encompassing 23 health systems, 93 hospitals and over 40,000 providers. It is unclear if smaller data sources with less patient fragmentation would yield similar performance.

We identified a number of challenges. Due to language barriers, our manual review process for nickname identification may have overlooked certain less frequently occurring non-English nicknames used by minority populations. While nicknames used in other languages may also be predicted using machine learning, because our models were trained using predominantly English nicknames, our models may not be valid for non-English datasets. We also note that many persons with non-English names may adopt unofficial English language nicknames. (i.e. an individual named ‘Chen’ may adopt the English nickname ‘Charlie’). Our manual review process did not capture these nicknames as such nickname pairs were negligible given the Indiana population. Also, our manual review was based on first name

pairs only. Thus, nicknames based on an individual's last name (i.e. an individual named James Henderson may adopt the nickname 'Henny') will not be tagged as a nickname as person last names were not available for review.

Our results are not applicable to nicknames used in online communities and forums. Such names may include alphanumeric characters, and may be of different phonetic structure than nicknames used in day to day life [39, 40]. Any effort to identify Internet nicknames falls out of the scope of our efforts. Further, we filtered our dataset to exclude the least frequently occurring name pairs (name pairs with a frequency of 3 or less). Thus, our models cannot be used to evaluate name pairs that occur very infrequently. Further, we built separate decision models for male and female name pairs on the assumption that the gender of each potential name pair would be known. However, gender data may not always be available in real-world datasets. Our effort leveraged patient name data extracted from a statewide HIE serving the people of Indiana. We hypothesize that our models can be applied to patient data in a state with similar demographic characteristics. However, replicating our methods may be challenging to implementers without access to large datasets.

We identified several avenues for future study. We seek to integrate our nickname prediction models into existing patient matching tools to evaluate if the inclusion of nickname information will lead to statistically significant improvements in record linkage performance. Further, our approach required significant human effort for the manual review of name pairs. Research into the feasibly of using readily available nickname lists such as those obtained from various online resources are warranted. If successful, these efforts would significantly reduce human effort required to develop decision models. Additionally, we considered only nicknames based on first name. Research into the use of nicknames based on middle or last names are is warranted. Finally, further effort is necessary to investigate whether our approaches can be used to predict nicknames across other non-English name pairs or Internet-based nicknames.

Conclusions

Supplementing existing patient matching data with nickname information may improve the accuracy of patient matching efforts. However, identifying nickname pairs through manual review requires significant human effort. We leveraged patient demographic data obtained from a statewide Health Information Exchange to develop decision models capable of identifying valid male and female nickname pairs based on their phonological and structural similarities. Our decision models achieved adequate performance despite the low number of true nickname matches in the datasets under study. The high precision/PPV achieved by each decision model suggests its suitability for augmenting existing patient matching tools.

References

1. Hillestad R, Bigelow JH, Chaudhry B, Dreyer P, Greenberg MD, Meili RC, et al. Identity crisis: An examination of the costs and benefits of a unique patient identifier for the US health care system. The RAND Corporation. 2008.
2. Stange KC. The problem of fragmentation and the need for integrative solutions. The Annals of Family Medicine. 2009;7(2):100-3.
3. Grannis SJ, Overhage JM, McDonald CJ, editors. Real world performance of approximate string comparators for use in patient matching. Medinfo; 2004.
4. Zhu VJ, Overhage MJ, Egg J, Downs SM, Grannis SJ. An empiric modification to the probabilistic record linkage algorithm using frequency-based weight scaling. Journal of the American Medical Informatics Association. 2009;16(5):738-45.
5. Mason AR, Barton AJ. The emergence of a learning healthcare system. Clinical Nurse Specialist. 2013;27(1):7-9.
6. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Science translational medicine. 2010;2(57):57cm29-57cm29.
7. Liu S, Wen SW. Development of record linkage of hospital discharge data for the study of neonatal readmission. Chronic Dis Can. 1999;20(2):77-81.

8. The Markle Foundation. Linking Health Care Information: Proposed Methods for Improving Care and Protecting Privacy 2005 [Available from: https://www.markle.org/publications/863-linking-health-care-information-proposed-methods-improving-care-and-protecting-priv.
9. Grannis S, Xu, H, Vest, J, Kasthurirathne, S, Bo, N, Moscovitch, B, Torkzadeh, R, Rising, J. Evaluating the effect of data standardization and validation on patient matching accuracy. Manuscript submitted for publication. 2018.
10. Cohen W, Ravikumar P, Fienberg S, editors. A comparison of string metrics for matching names and records. Kdd workshop on data cleaning and object consolidation; 2003.
11. Walfish M, Hachamovitch D, Fein RA. System and method for automatically correcting a misspelled word. Google Patents; 2000.
12. Van Dam M. On the phonological structure of/i/-suffixed English nicknames. IULC Working Papers. 2003;3.
13. Cassidy KW, Kelly MH, Sharoni LaJ. Inferring gender from name phonology. Journal of Experimental Psychology: General. 1999;128(3):362.
14. McDonald CJ, Overhage JM, Barnes M, Schadow G, Blevins L, Dexter PR, et al. The Indiana network for patient care: a working local health information infrastructure. Health affairs. 2005;24(5):1214-20.
15. Indiana Health Information Exchange Inc. Indiana Health Information Exchange 2012 [Available from: https://www.slideshare.net/mollybutters/overview-indiana-health-information-exchange.
16. Porter EH, Winkler WE, editors. Approximate string comparison and its effect on an advanced record linkage system. Advanced record linkage system US Bureau of the Census, Research Report; 1997: Citeseer.
17. Sideli RV, Friedman C, editors. Validating patient names in an integrated clinical information system. Proceedings of the Annual Symposium on Computer Application in Medical Care; 1991: American Medical Informatics Association.
18. Intellectual Reserve Inc. Traditional Nicknames in Old Documents - A Wiki List 2018 [Available from: https://www.familysearch.org/wiki/en/Traditional_Nicknames_in_Old_Documents_-_A_Wiki_List.
19. Levenshtein VI, editor Binary codes capable of correcting deletions, insertions, and reversals. Soviet physics doklady; 1966.
20. Laohaprapanon S, Sood G. ethnicolr: Predict Race and Ethnicity From Name 2018 [cited 2018. Available from: http://ethnicolr.readthedocs.io/.
21. Python Software Foundation. Gender Guesser 2016 [Available from: https://pypi.org/project/gender-guesser/.
22. Knuth DE. The art of computer programming: sorting and searching: Pearson Education; 1997.
23. Philips L. Hanging on the metaphone. Computer Language. 1990;7(12 (December)).
24. Lynch BT, Arends WL. Selection of a surname coding procedure for the SRS record linkage system. Washington, DC: US Department of Agriculture, Sample Survey Research Branch, Research Division. 1977.

25. Doyle D. Phonics, Syllable and Accent Rules 2010 [Available from: http://english.glendale.cc.ca.us/phonics.rules.html.
26. Norvig P. English letter frequency counts: Mayzner revisited or etaoin srhldcu. Dostopno na http://www norvig com/mayzner html[obiskano 2016-08-07]. 2013.
27. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of machine learning research. 2011;12(Oct):2825-30.
28. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016: ACM.
29. Ye J, Chow J-H, Chen J, Zheng Z, editors. Stochastic gradient boosted distributed decision trees. Proceedings of the 18th ACM conference on Information and knowledge management; 2009: ACM.
30. Kasthurirathne SN, Dixon BE, Gichoya J, Xu H, Xia Y, Mamlin B, et al. Toward better public health reporting using existing off the shelf approaches: A comparison of alternative cancer detection approaches using plaintext medical data and non-dictionary based feature selection. Journal of biomedical informatics. 2016;60:145-52.
31. Kasthurirathne SN, Dixon BE, Gichoya J, Xu H, Xia Y, Mamlin B, et al. Toward better public health reporting using existing off the shelf approaches: The value of medical dictionaries in automated cancer detection using plaintext medical data. Journal of biomedical informatics. 2017;69:160-76.
32. Nielsen D. Tree Boosting With XGBoost-Why Does XGBoost Win" Every" Machine Learning Competition? : NTNU; 2016.
33. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002;16:321-57.
34. Liu XY, Wu J, Zhou ZH. Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2009;39(2):539-50.
35. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one. 2015;10(3):e0118432.
36. Trevor H, Robert T, JH F. The elements of statistical learning: data mining, inference, and prediction: New York, NY: Springer; 2009.
37. Sun Y, Kamel MS, Wong AK, Wang Y. Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition. 2007;40(12):3358-78.
38. Pepe MS. The statistical evaluation of medical tests for classification and prediction: Medicine; 2003.
39. Lindholm L.
The maxims of online nicknames. Pragmatics of computer-mediated communication. 2013;9:437.
40. Ecker R. Creation of Internet Relay Chat Nicknames and Their Usage in English Chatroom Discourse. Linguistik online. 2013;50(6).

Appendix A. Parameters that were evaluated as part of the hyperparameter process.

Hyperparameter	Description
Boosting ratio	Level of boosting performed using SMOTE

Number of estimators	Number of trees
Minimum child weight	Minimum sum of weights of all observations required in a child
Gamma value	the minimum loss reduction required to split a node
Subsample	Fraction of observations to be randomly samples for each tree
Col sample by tree	Fraction of columns to be randomly samples for each tree
Max depth	Maximum depth of each tree

Appendix B: List of top-ranking features that contributed to male and female decision models. (cutoff = 0.5 selected based on variance in feature importance scores).

Male nickname model		Female nickname model	
Feature name	Feature importance (0-1)	Feature name	Feature importance (0-1)
Syllable count comparison	0.997	Soundex comparison	0.995
Soundex comparison	0.995	Syllable count comparison	0.994
Levenstien edit distance	0.9915	Race/Ethnicities match	0.9935
Gender match	0.985	Levenstien edit distance	0.98
Frequency	0.979	Frequency	0.98
Race/Ethnicities match	0.97	Gender match	0.975
Combined Root Mean Square	0.962	Combined Root Mean Square	0.971
NYSIIS comparison	0.935	Bi-gram frequency comparison	0.955
Metphone comparison	0.92	Misspelling frequencies	0.953
Jaro Winkler comparator	0.832	Metphone comparison	0.95
Bi-gram frequency comparison	0.65	Jaro Winkler comparator	0.85
		NYSIIS comparison	0.7