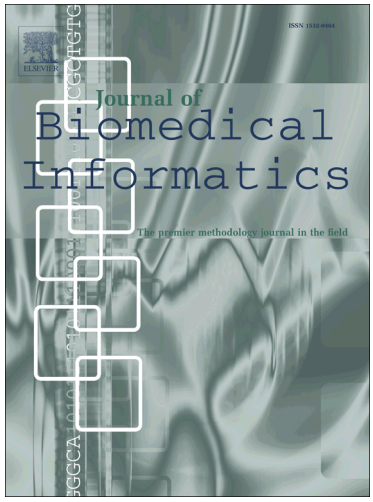


Toward better public health reporting using existing off the shelf approaches: The value of medical dictionaries in automated cancer detection using plaintext medical data

Suranga N. Kasthurirathne, Brian E. Dixon, Judy Gichoya, Huiping Xu, Yuni Xia, Burke Mamlin, Shaun J. Grannis

PII: S1532-0464(17)30078-3
DOI: http://dx.doi.org/10.1016/j.jbi.2017.04.008
Reference: YJBIN 2759

To appear in: *Journal of Biomedical Informatics*

Received Date: 23 August 2016
Revised Date: 9 April 2017
Accepted Date: 10 April 2017

Please cite this article as: Kasthurirathne, S.N., Dixon, B.E., Gichoya, J., Xu, H., Xia, Y., Mamlin, B., Grannis, S.J., 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), doi: http://dx.doi.org/10.1016/j.jbi.2017.04.008

Toward better public health reporting using existing off the shelf approaches: The value of medical dictionaries in automated cancer detection using plaintext medical data

Suranga N. Kasthurirathne, BEng[1], Brian E. Dixon, MPA, PhD[2,3], Judy Gichoya, MD, MS[4], Huiping Xu, PhD[3], Yuni Xia, PhD[5], Burke Mamlin, MD[2,4], Shaun J. Grannis, MD, MS[2,4]

[1]Indiana University School of Informatics and Computing, Indianapolis, IN, USA; [2]Regenstrief Institute, Indianapolis, IN, USA; [3]Indiana University Richard M. Fairbanks School of Public Health, Indianapolis, IN, USA; [4]Indiana University School of Medicine, Indianapolis, IN, USA; [5]Indiana University-Purdue University, Indianapolis, IN, USA

Corresponding author:

Suranga N. Kasthurirathne
Indiana Universty School of Informatics and Computing (SOIC)
535 W. Michigan Street, IT 475
Indianapolis, IN 46202, USA
(317)-278-4636,
snkasthu@iupui.edu

Keywords

Public health reporting, medical dictionaries, decision models, cancer, pathology, feature selection, data preprocessing

Abstract

Objectives

Existing approaches to derive decision models from plaintext clinical data frequently depend on medical dictionaries as the sources of potential features. Prior research suggests that decision models developed using non-dictionary based feature sourcing approaches and "off the shelf" tools could predict cancer with performance metrics between 80%-90%. We sought to compare non-dictionary based models to models built using features derived from medical dictionaries.

Materials and Methods

We evaluated the detection of cancer cases from free text pathology reports using decision models built with combinations of dictionary or non-dictionary based feature sourcing approaches, 4 feature subset sizes, and 5 classification algorithms. Each decision model was evaluated using the following performance metrics: sensitivity, specificity, accuracy, positive predictive value, and area under the receiver operating characteristics (ROC) curve.

Results

Decision models parameterized using dictionary and non-dictionary feature sourcing approaches produced performance metrics between 70-90%. The source of features and feature subset size had no impact on the performance of a decision model.

Conclusion

Our study suggests there is little value in leveraging medical dictionaries for extracting features for decision model building. Decision models built using features extracted from the plaintext reports themselves achieve comparable results to those built using medical dictionaries. Overall, this suggests that existing "off the shelf" approaches can be leveraged to perform accurate cancer detection using less complex Named Entity Recognition (NER) based feature extraction, automated feature selection and modeling approaches.

1. Background and Significance

The widespread adoption of electronic health record (EHR) systems has produced readily available clinical data for a myriad of primary and secondary healthcare needs (Murdoch & Detsky, 2013; Savova et al., 2010). Much of these data are recorded as unstructured clinical reports (Jiang et al., 2011) dictated or typed by clinicians and must therefore be transformed into actionable information to realize their full value.

Analyzing and extracting relevant information from unstructured clinical data has gained significant importance within the healthcare domain (Murdoch & Detsky, 2013), which requires the contextualization of concepts of interest, or "named entities". The identification of named entities is also referred to as "named entity recognition" (NER). Target entities for NER can be obtained from both dictionary and non-dictionary based sources. Dictionary-based approaches for NER depend on medical dictionaries or controlled vocabularies (Imler, Vreeman, & Kannry, 2016) as a source for named entities. Non-dictionary based approaches derive named entities from informal sources such as empirical knowledge, or directly from clinical data being analyzed (Cheng, Wei, & Tseng, 2006).

Dictionary-based approaches have been traditionally used for extracting information from clinical data using NER (Rindflesch, Tanabe, Weinstein, & Hunter, 2000; Song, Yu, & Han, 2015). Dictionaries provide a well curated and comprehensive body of medical terms for use as potential features (feature sourcing) (Wang & Patrick, 2009). However, scientific literature indicates that; (a) the use of short names/terms in dictionaries is associated with an increased number of false positives (Tsuruoka & Tsujii, 2004) and (b) spelling variations in dictionaries contribute to decreased accuracy (Song et al., 2015). These limitations hinder accurate dictionary-based NER in plaintext data using text mining (Krauthammer & Nenadic, 2004). The decision-making accuracy of dictionary-based NER approaches, as evaluated using various performance measures, are well below acceptable levels for use in clinical or research needs (Kang, Afzal, Singh, van Mulligen, & Kors, 2012; Spasić, Livsey, Keane, & Nenadić, 2014). These approaches are more susceptible to over-fitting (Domingos, 1999). Also, given that controlled medical dictionaries are routinely modified, with terms being added/deprecated or expanded

(Bodenreider, 2008; Vreeman, 2007), dictionary-based NER approaches require ongoing time and resource-heavy manual curation to stay up to date with evolving medical terminology (Grannis & Vreeman, 2010).

Another challenge in decision model building is determining which feature subset selection approach is optimal for a given dataset. Researchers typically have used manual or expert-driven feature selection (Cheng et al., 2006). However, these approaches are cumbersome, and require specialized expertise. They also fail to consider contextual aspects of the dataset such as healthcare facility or disease specific behavior.

Given significant advances in the field of machine learning, we previously evaluated whether non-dictionary based feature selection approaches requiring varying levels of human intervention could be used to identify cancer in plaintext pathology reports (Kasthurirathne et al., 2016). In that study we observed that non-dictionary based feature selection can perform equally, or better than feature selection informed by clinicians with specialized expertise.

We extend the prior research in this analysis for the following reasons. First, we did not previously compare the performance of non-dictionary based decision models to dictionary-based decision models, nor are we aware of similar comparative analyses. Second, we note that feature selection approaches (e.g., dictionary and non-dictionary) are a key element in developing free text case detection methodologies, and also that there is a paucity of peer-reviewed evidence-based best practice guidance regarding choice of feature selection approaches. Therefore, this subsequent analysis represents a novel methodological contribution because dictionary and non-dictionary approaches previously have not been directly compared in the context of free-text cancer case detection. Consequently, in this work, we evaluate the performance of automated cancer detection performed using decision models built using features obtained from both non-dictionary based feature sources and dictionary-based feature sources.

2. Materials and Methods

2.1 Sources of Data and Cancer Diagnosis

We obtained a convenience sample of 7,000 plaintext pathology reports extracted from the Indiana Network for Patient Care (INPC), a robust, statewide Health Information Exchange (HIE) serving several large health systems, including more than 100 hospitals, in Indiana (McDonald et al., 2005; Overhage, 2016). These pathology reports were extracted from seven diverse health systems representing over 30 hospitals within the INPC, and recorded between the years 1996 to 2012. The reports were manually reviewed by three clinicians who tagged them as either positive or negative for the presence of cancer. Pathology reports were selected for this study due to their completeness and availability as well as their suitability to be used for cancer diagnosis.

2.2 Selection of a vocabulary for dictionary-based feature selection

Selection of a medical dictionary that contained a comprehensive set of cancer related tokens was a considerable challenge. Several dictionaries including the International Classification of Diseases (ICD) (World Health Organization, 2007), Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) (International Health Terminology Standards Development Organisation, 2016), Logical Observation Identifiers Names and Codes (LOINC) (McDonald et al., 2003) and Medical Subject Headings (MeSH) (US National Library of Medicine, 2015) contain support for cancer related terms. However, none of these vocabularies focus exclusively on cancer. Instead, they contain a wide range of concepts representing other medical conditions. A potential solution to this challenge was to subset various vocabularies by choosing all the descendants of the top-level class/concept on cancer, and combine them to a single list of tokens. However, this approach posed several challenges: (a) it required repeated curation of the selected token list as each vocabulary was updated; (b) considerable manual intervention was necessary to select subsets of cancer related tokens from each vocabulary; and (c) working with multiple vocabularies added considerable complexity to the workflow.

We identified two candidate dictionaries that were specific to cancer: the tumor taxonomy for the developmental lineage classification of neoplasms developed by Jules J. Berman (Berman, 2004) and the International Classification of Diseases for Oncology (ICD-O). We selected the tumor taxonomy as it was the largest nomenclature of

neoplasms currently available, with over twice the number of neoplasm names found in ICD-O or other medical nomenclatures including the UMLS, SNOMED, and the National Cancer Institute's Thesaurus. The tumor taxonomy contains 122,632 different terms encompassing 5,376 neoplasm concepts. On average, each neoplasm concept has an average of 23 different synonyms (Berman, 2004).

2.3 Preparation of feature subsets

We extracted feature subsets using features obtained from non-dictionary based and dictionary-based feature sources.

2.3.1 Dictionary-based approach

A Perl script leveraged the Lingua Stopwords module (Estudillo-Valderrama et al., 2014) to remove stop words in the dictionary and to count occurrences of tokens identified using the Berman taxonomy. Token identification was performed after stemming each word using the Perl Lingua Stem module (cpan.org, 2014) and comparing the root forms. The Negex algorithm (Chapman, Bridewell, Hanbury, Cooper, & Buchanan, 2001) was used to identify positive/negative context of use for each of the identified tokens. For each report we counted the presence of each token in positive and negated context, and transformed this data into an input vector.

2.3.2 Non-Dictionary based approach

The Perl script was reused to remove all stop words from the pathology report set and count the frequency of unique features appearing across entire pathology report set. From these, we removed low prevalence tokens appearing less than three times across all reports. The Negex algorithm was used to identify positive/negative context of use for each remaining token. We counted the presence of each token in positive and negated contexts per each report, and compiled this data into an input vector.

The above approaches yielded two separate input vector sets, each consisting of thousands of features. Given that such a large number of features would lead to increased model complexity and over-fitting, we used information gain, also known as Kullback-

Leibler divergence (Polani, 2013; J. Yang, Qu, & Liu, 2014; Y. Yang & Pedersen, 1997) to identify the most relevant features for decision model building. We ranked all tokens from the pathology reports in descending order using information gain scores.

2.3.4 Feature Subset Sizes

We hypothesized that varying feature subset sizes would affect various performance metrics, including precision, recall, specificity, and accuracy. To test this hypothesis, we chose feature subset sizes of 5, 10, 15, and 20 from each of the two feature sets based on rankings assigned by the information gain algorithm. These feature sizes were chosen because we had used similar feature subset sizes with considerable success in our previous study.

2.4 Decision Models

We randomly selected 700 (10%) of the plaintext reports as hold-out test data. The remaining 6300 reports (90%) were used to train decision models using alternative feature selection approaches, feature subset sizes, and classification algorithms against the gold standard produced by manual review. After training, each decision model was tested using the 10% hold-out test data (figure 1).

The five classification algorithms selected for our study were simple logistic regression (SLR), naïve Bayes (NB), k-nearest neighbor (KNN), random forest (RF), and J48 decision tree (J48). These classification algorithms were selected as a representative subset of the types of algorithms most widely used in the public health field. Each algorithm represents a specific group or type of classification algorithm with its own unique traits. SLR, RF and J48 follow a discriminative learning approach while NB is based on adaptive learning (Dietterich, Becker, & Ghahramani, 2002). Decision trees such as RF and J48 are nonparametric and, therefore, make no assumptions on the distribution of input data, and are flexible and robust with respect to nonlinear and noisy relations among input features and class labels (Friedl & Brodley, 1997). NB assumes conditional independence of features (Lewis, 1998).

The training and testing of various decision models was performed using version 3.6.11 of Weka (Hall et. al., 2009).

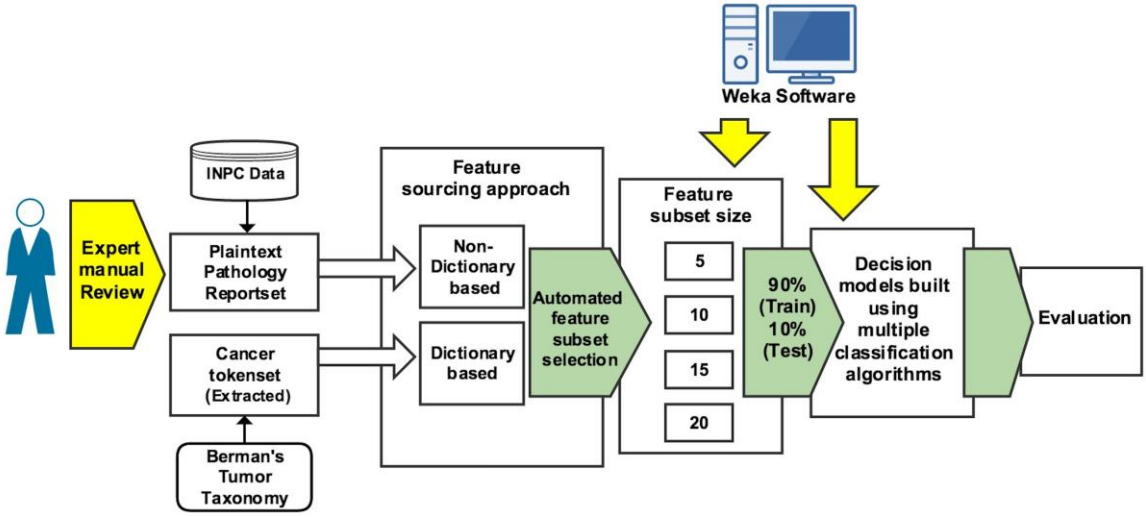


Figure 1. The flowchart presenting our study approach from data selection to the evaluation of decision model performance

2.5 Statistical analysis

By applying the 2 feature sourcing approaches and the 4 feature subset sizes, we extracted 8 (2 x 4) different feature subsets for decision model building and evaluation. Each of these 8 feature subsets was applied to 5 different classification algorithms for a total of 40 (8 x 5) decision models. These decision models were tested using the 10% holdout data, and analyzed using the metrics of sensitivity, specificity, accuracy, PPV and the areas under the curve (ROC). We did not perform any specific optimization to maximize a given performance metric. Rather, we used the default thresholds defined by Weka software for each classification algorithm. We used version 9.4 of the Statistical Analysis System (SAS) software (SAS Institute Inc., Cary, NC) to compare the performance of each of these 40 decision models. We compared the predicted outcomes of each decision model to the gold standard produced by manual review. The performance of each decision model, evaluated using sensitivity, specificity, PPV, and overall accuracy, were estimated using proportions and 95% confidence intervals. To account for the clustering effect of multiple methods applied to the same pathology report set, and assess the effects of multiple feature sourcing approaches, feature subset size,

and classification algorithm on the accuracy of cancer detection, we used a marginal logistic regression based on generalized estimating equations (Leisenring W, Pepe MS & Longton G, 1997; Leisenring W Alono T, Pepe MS, 2000). In evaluating performance, we also included the main effects, 2-way interactions, and 3-way interaction of the 3 factors in the model to allow for differential effects of a factor as other factors were changed. Standard errors of the accuracy measures were calculated using robust sandwich variance estimation methods (Kauermann & Carroll, 1999). Comparison of these accuracy metrics across each decision model was performed using a multiple comparison approach with a Bonferroni adjustment. Accuracy was also evaluated using the receiver operating characteristics (ROC) curve and the area under the ROC curve (AUC). A nonparametric approach was used to estimate and the 95% confidence interval of the AUC, as well as the comparison of multiple AUC values (DeLong, DeLong, & Clarke-Pearson, 1988).

3. Results

Manual review of the 7,000 pathology reports identified 1,950 (27.86%) as cancer positive, and the remaining 5,050 (72.14%) as cancer negative. Among the training set reports (N=6,300), 1,757 (27.89%) were manually labeled as cancer positive. In the test set (N=700), 201 (28.7%) reports were manually labeled as cancer positive.

Parsing of the Berman dictionary to assess the dictionary-based feature sourcing approach produced a total of 7,302 unique tokens. Parsing of the pathology report set for non-dictionary based features produced a total of 17,601 unique tokens. Of these, 8,121 tokens that appeared only once or twice were removed due to low prevalence, resulting in a total of 9480 tokens for evaluation. Tokens identified via the dictionary and non-dictionary based approaches were not limited to clinical terms. They also represented other semantic types such as patient-provider interactions, medical procedures, drugs, medical devices and geographic locations. We used Metamap (Aronson & Lang, 2010), to map each token to concepts from the UMLS Metathesaurus (Bodenreider, 2004) enabling identification of the semantic types to which each token belonged. A summary

of these results is presented in Table 1. A more detailed breakdown of the frequencies for each semantic type is presented in Appendix A.

Table 1: Distribution of tokens mapped to UMLS semantic groups from the UMLS Metathesaurus. Note that while the dictionary and non-dictionary approaches showed proportional differences when all tokens where considered (e.g., Disorders and Geographic Areas), those differences were less pronounced when limited to the top 20 tokens.

Semantic group	Dictionary		Non-dictionary	
	All tokens	Top 20 tokens	All tokens	Top 20 tokens
Activities & Behaviors (ACTI)	51 (0.54%)	0	156 (1.8%)	0
Anatomy (ANAT)	1204 (12.8%)	4 (11.1%)	1108 (13.1%)	6 (16.2%)
Chemicals & Drugs (CHEM)	1670 (17.7%)	1 (2.8%)	1158 (13.7%)	0
Concepts & Ideas (CONC)	2018 (21.5%)	12 (33.3%)	2363 (28.0%)	13 (35.1%)
Devices (DEVI)	65 (0.7%)	1 (2.8%)	144 (1.7%)	2 (5.4%)
Disorders (DISO)	2346 (25%	9 (25%)	904 (10.7%)	11 (29.7%)
Genes & Molecular Sequences (GENE)	873 (9.3%)	1 (2.8%)	597 (7.1%)	1 (2.7%)
Geographic Areas (GEOG)	115 (1.2%)	1 (2.8%)	555 (6.6%)	0
Living Beings (LIVB)	402 (4.3%)	1 (2.8%)	499 (5.9%)	0
Objects (OBJC)	143 (1.5%)	2 (5.6%)	285 (3.4%)	3 (8.1%)
Occupations (OCCU)	4 (0.04%)	0	22 (0.3%)	0
Organizations (ORGA)	16 (0.2%)	0	34 (0.4%)	0
Phenomena (PHEN)	77 (0.8%)	0	97 (0.5%)	0
Physiology (PHYS)	242 (2.6%)	1 (2.8%)	210 (2.5%)	0
Procedures (PROC)	162 (1.7%)	3 (8.3%)	295 (3.5%)	1 (2.7%)

Lists of the feature subsets for the top 5, 10, 15 and 20 tokens for each approach, together with summary statistics can be found in Appendix B.

3.1 Performance metrics

3.1.1 Sensitivity

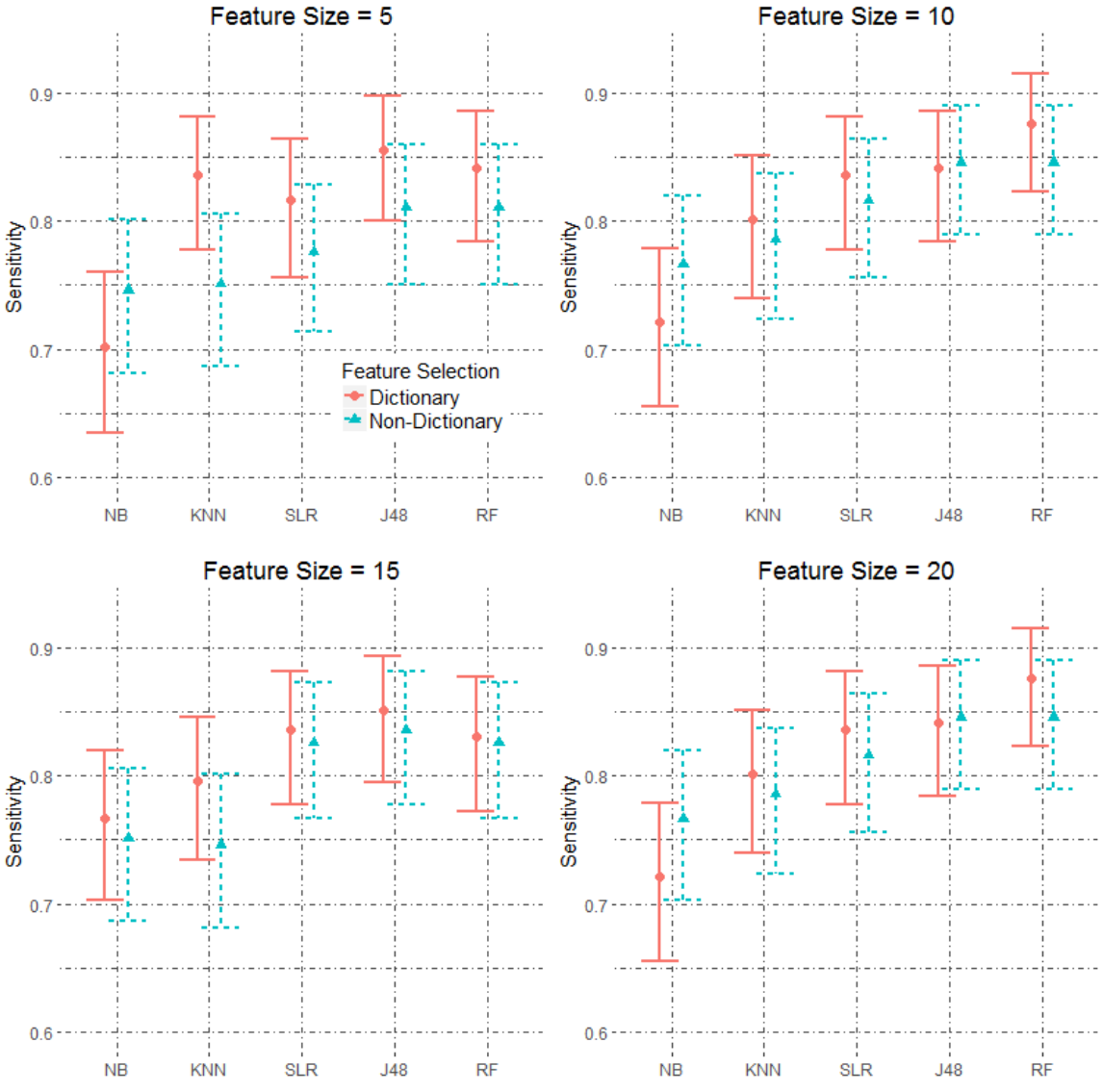


Figure 2. Estimated sensitivity and 95% confidence interval across each (a) feature sourcing approach (b) classification algorithm and (c) feature subset size

The results of the sensitivity analysis are summarized in **Figure 2**. Most decision models produced sensitivity values greater than 70% and no statistical difference between dictionary and non-dictionary approaches were noted.

3.1.2 Specificity

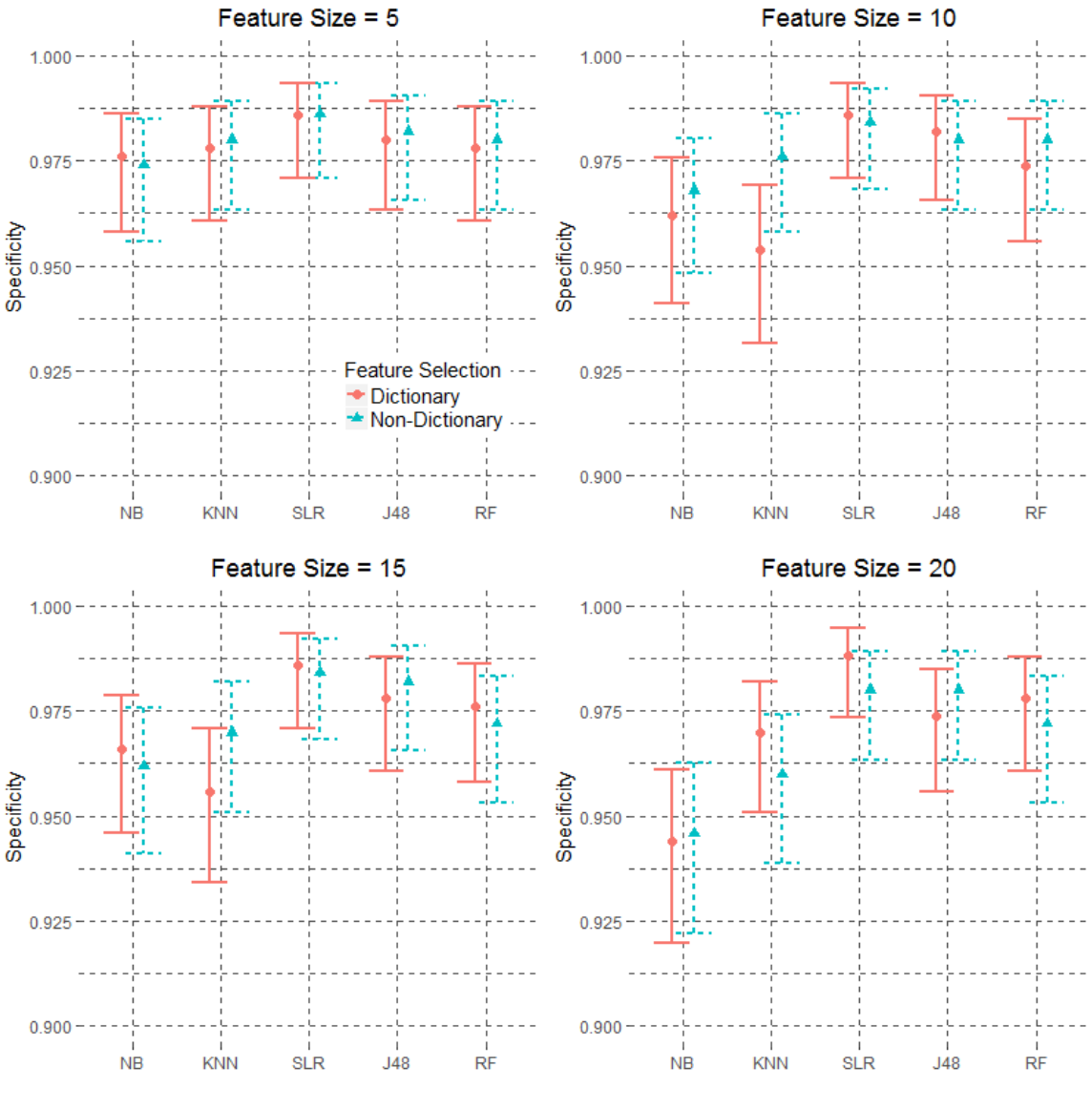


Figure 3. Estimated specificity and 95% confidence interval across each (a) feature sourcing approach (b) classification algorithm and (c) feature subset size

The results of the specificity analysis are summarized in Figure 3. Each decision model yielded specificity values greater than 90%.

3.1.3 Accuracy

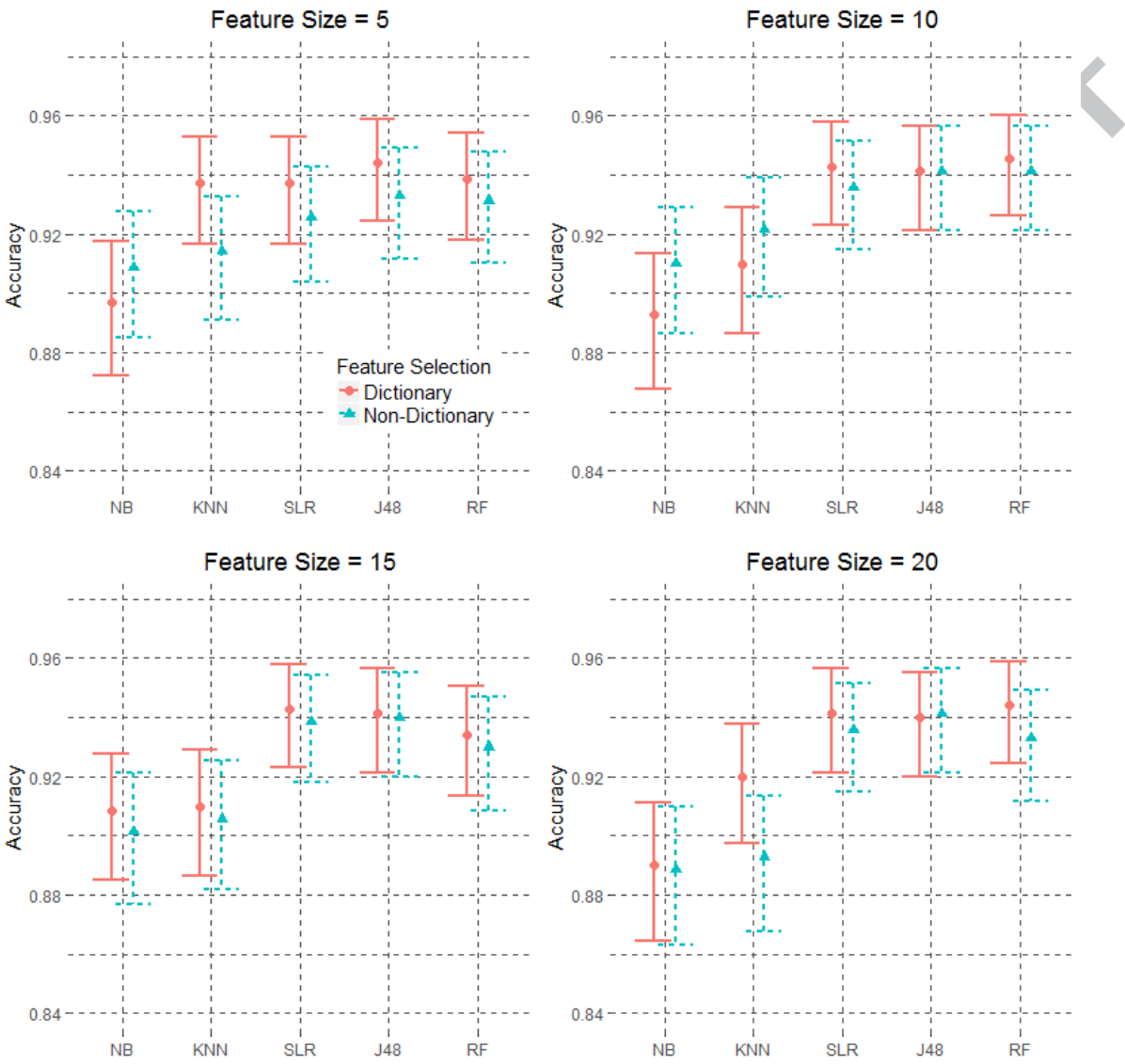


Figure 4. Estimated accuracy and 95% confidence interval across each (a) feature sourcing approach (b) classification algorithm and (c) feature subset size

The results of the accuracy analysis are summarized in Figure 4. Most decision models produced accuracy values greater than 85%.

3.1.4 Positive Predictive Value (PPV)

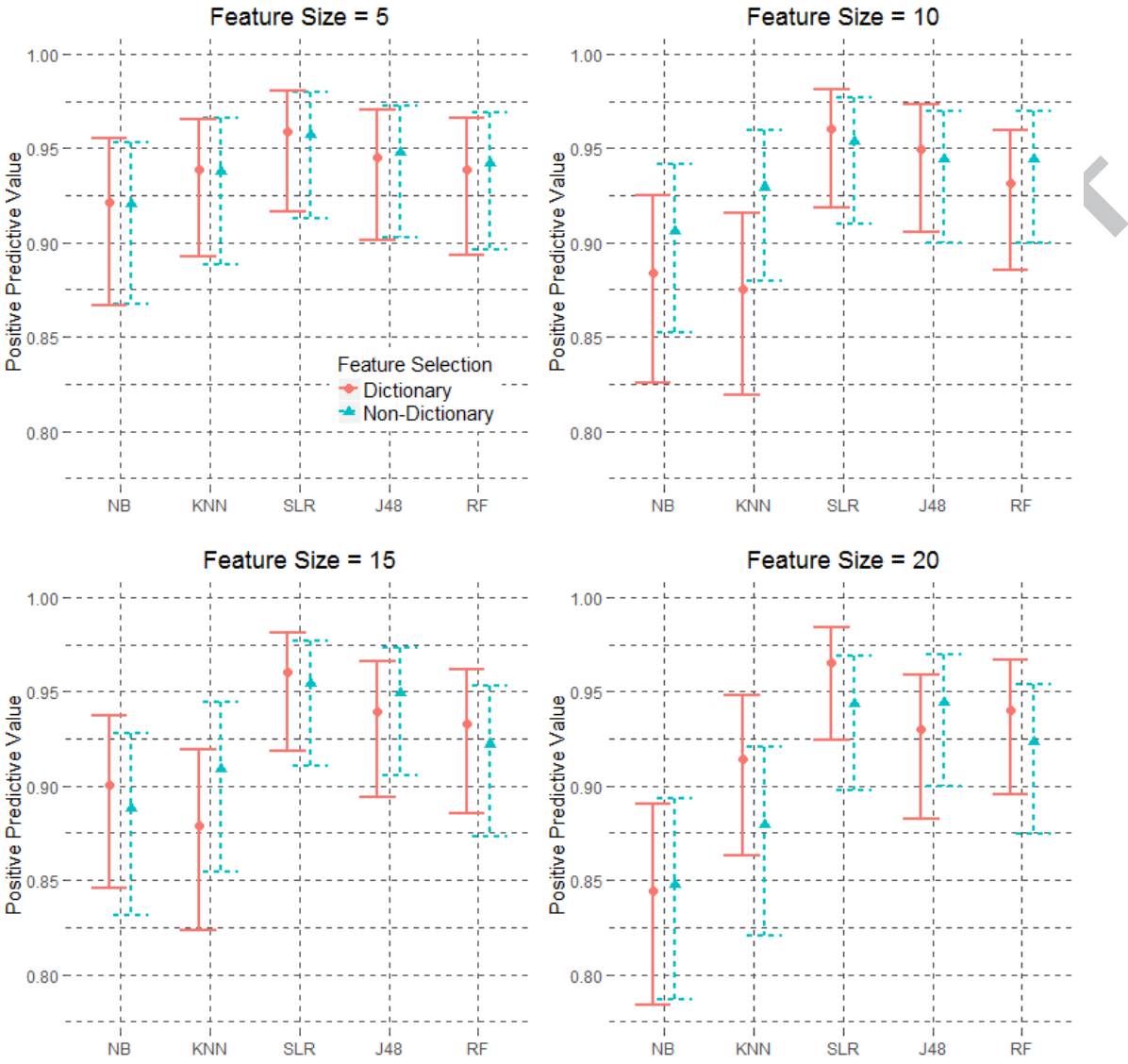


Figure 5. Estimated PPV and 95% confidence interval across each (a) feature sourcing approach (b) classification algorithm and (c) feature subset size

The results of the PPV analysis are summarized in **Figure 5**. A majority of decision models produced PPV values greater than 85%.

3.1.5 Area Under the ROC Curve (AUC)

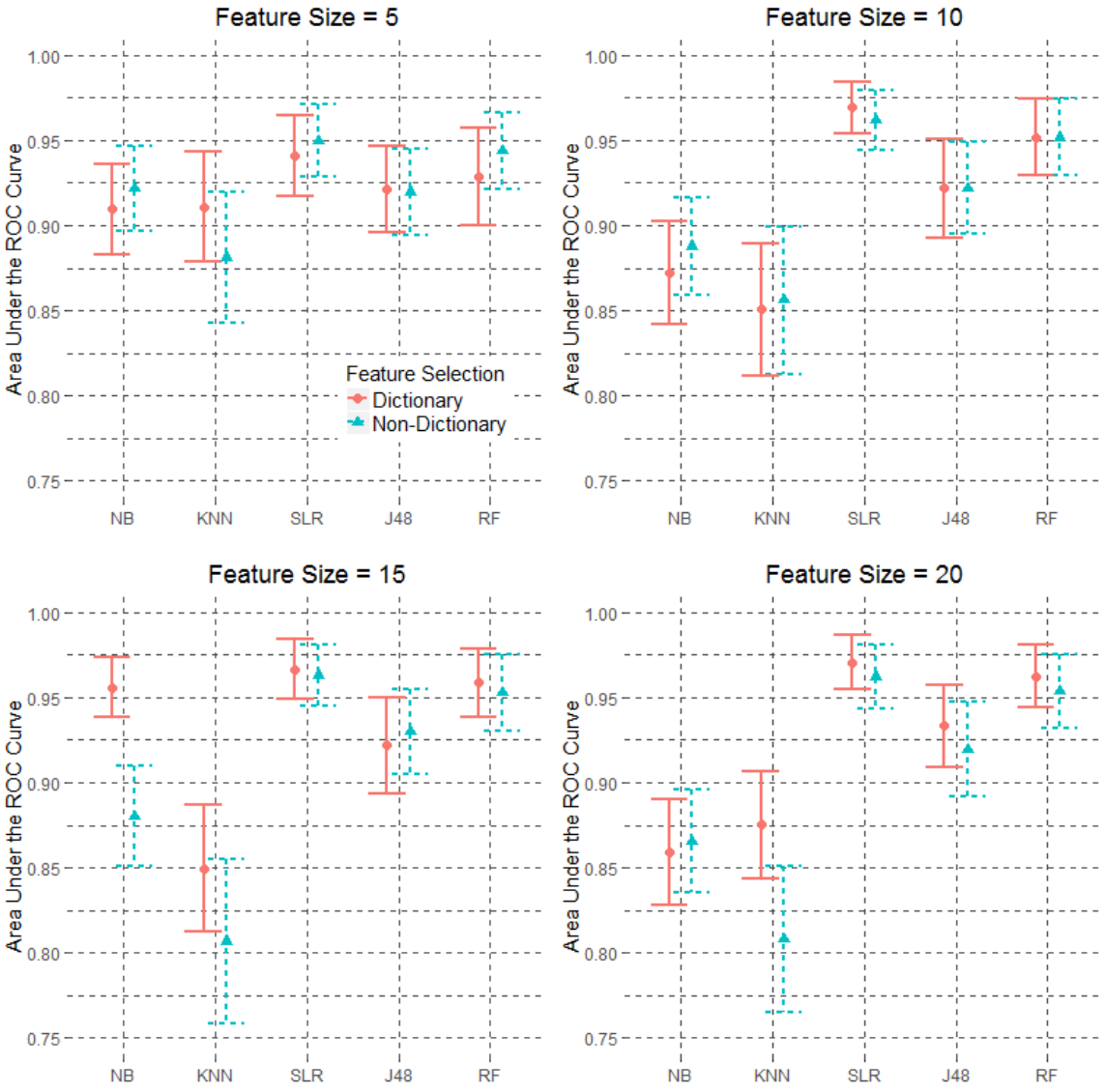


Figure 6. Estimated area under the ROC curve and 95% confidence interval across each (a) feature sourcing approach (b) classification algorithm and (c) feature subset size

The results of the AUC analysis are summarized in **Figure 6**. A majority of decision models produced AUC values greater than 80%.

3.2 Comparison of decision model performance

As indicated by figures 2-6 above, there were variations in performance metrics calculated across decision models built by varying feature sources, feature subset sizes and classification algorithms. When comparing p-values adjusted using the Bonferroni approach , many of these variations were not statistically significant (with statistical significance defined as the adjusted $p<0.05$). Table 2 summarizes the comparisons.

Table 2. A comparison of statistically significant differences in decision models built using varying combinations of feature sourcing approaches (FS), feature subset sizes (SS) and classification algorithms (CA).

Performance metric	**Comparison approaches**		
	(a) Identical FS & CA, varying SS	(b) Identical SS & CA, varying FS	(c) Identical FS & SS, varying CA
Sensitivity	No significant difference	No significant difference	KNN underperformed when used with non-dictionary based FS and SS of 15 or greater. For dictionary based FS with a SS of 10 or smaller, NB underperformed.
Specificity	No significant difference	No significant difference	NB underperformed when used with dictionary-based FS and SS of 20.
Accuracy	No significant difference	No significant difference	SLR, RF and J48 outperformed NB and KNN across most SS and FS.
Positive Predictive Value (PPV)	No significant difference	No significant difference	NB underperformed across dictionary based and non-dictionary based FS and SS of 20
Area Under the ROC Curve	For NB based models using non-	For SS of 15 and NB CA, dictionary-based	KNN and NB based models tended to

(AUC)	dictionary based FS, a SS of 5 yielded best results. For dictionary-based FS, NB yielded the best AUC with a SS of 15	models outperformed non-dictionary based models	underperform when used with non-dictionary based FS

Each decision model was built using the training set (90% reports) and tested using the test set (10% reports). To assess how decision models performed when trained on smaller training sets, we built multiple decision models using variations of training/test data splits ranging from 10% (train on 10%, test on 90%) to 90% (train on 90%, test on 10%), and assessed their performance using Area under the ROC curve values. While the performance measures for most models tends to improve as training/test split proportion increases from 10% to 90%, there are no significant differences between each model (Appendix C). Also of note is that, with the exception of kNN, the non-dictionary based models exhibited superior performance metrics for a given training/test split proportion when compared to their dictionary-based counterpart.

4. Discussion

Decision models built using dictionary and non-dictionary based feature sources can identify positive cases of cancer from plaintext pathology reports with performance measures ranging between 70%-90%. Different feature sourcing approaches and feature subset sizes did not result in significant changes in performance metrics reported across decision models evaluated in this study. However, decision models built using NB and KNN algorithms tended to underperform compared to others. Furthermore, decision model performance did not always improve with feature subset size.

For optimized sensitivity, specificity, accuracy and PPV, decision models built using any feature sourcing approach, classification algorithm and feature subset size generally produced statistically similar results. For optimized ROC, our results indicate that decision models built using any feature sourcing approach, feature subset size and algorithms other than KNN and NB are approximately equivalent options.

Our results suggest that non-dictionary based approaches should be considered for identifying cancer cases in free-text documents. First, our results indicate that there is no statistical performance difference between non-dictionary and dictionary based approaches. Second, developing and maintaining non-dictionary based approaches is less resource intensive than dictionary based methods. Consequently, given similar performance and fewer resource requirements, non-dictionary based approaches may be optimal in many circumstances. Another argument for the use of non-dictionary based approaches are the wider range of semantic types represented by concepts identified by this approach.

However, using non-dictionary based approaches that derive features solely from existing data raises the question of whether such results are generalizable and can be reproduced across pathology reports obtained from multiple healthcare facilities. Given that the pathology reports used in our study were extracted from a statewide HIE system, which connects more than 100 highly heterogeneous healthcare systems ranging from sophisticated multi-institution organizations to small critical access hospitals, we believe that they are sufficiently diversified, and of similar quality and completeness to clinical reports collected at other healthcare systems, and thus, represent an acceptable test dataset to demonstrate generalizable use.

Overall, these results suggest that existing "off the shelf" approaches can be leveraged to support accurate cancer detection using simple NER based data extraction and modeling approaches, and without the additional effort required to manage medical dictionaries. However, the suboptimal performance of NB and KNN based models warrants further investigation. We hypothesize that decreased performance exhibited by NB based models are due to the assumption of conditional independence among features, which is highly unlikely for this dataset (Lewis, 1998). KNN based models assume linear scaling for every additional feature, an assumption that may lead to inaccuracies in calculating distance measures, especially as noisy or less discriminating features are added to the model. Better scaling approaches may enhance the performance of KNN based models (García-Laencina, Sancho-Gómez, Figueiras-Vidal, & Verleysen, 2009).

We were challenged to identify ready-to-use medical dictionaries that could be leveraged to extract tokens indicative of cancer positive status without significant preprocessing and

curation of the dictionary. While Berman's tumor taxonomy was adequate for the purposes of this study, we found little evidence of other disease specific dictionaries that could be used for dictionary-based decision modeling for other illnesses. In comparison, medical dictionaries that are not disease specific cannot be used to build decision models for illnesses without appropriate filtering of relevant tokens, which requires additional effort. This raises questions regarding the potential of leveraging existing medical dictionaries for any disease specific NER based tasks without considerable human intervention.

These results extend the findings of our previous work to obtain actionable information from unstructured clinical documents. They demonstrate the potential of realistic, practical, and low complexity solutions in extracting substantial value from unstructured clinical documents, and can contribute significant value to various public health tasks. Since many public health notifiable conditions are communicated in free text reports, classification of text reports is meaningfully linked to surveillance and many other public health initiatives. We hypothesize that this approach may also contribute to evidence-based best practices to solve similar challenges across different medical domains, including free text microbiology reports, and may also support the identification of positive results within the reportable laboratory results. Further, such work may also contribute to evidence-based best practices to solve similar challenges across different medical domains.

5. Conclusion

Our previous work demonstrated the potential of leveraging existing "off the shelf" approaches to perform automated cancer case detection from plaintext pathology reports solely using non-dictionary based feature-sourcing approaches. The results of the current study extend that previous work by performing one of the first direct comparisons between dictionary and non-dictionary feature selection approaches. Given each methods' approximate statistical equivalency, we conclude that when a sufficiently representative training data set is available, the added effort of using complex medical dictionaries as a source of features for decision model building does not result in

significant performance improvement. Our findings present significant potential for existing public health reporting efforts. They are of considerable value to healthcare professionals who must adhere to various state or nationally mandated communicable disease reporting laws, but lack adequate resources to do so using existing approaches.

References

Aronson, A. R., & Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association, 17*(3), 229-236.

Berman, J. J. (2004). Tumor taxonomy for the developmental lineage classification of neoplasms. *BMC cancer, 4*(1), 88.

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research, 32*(suppl 1), D267-D270.

Bodenreider, O. (2008). Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearbook of medical informatics*, 67.

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., & Buchanan, B. G. (2001). A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics, 34*(5), 301-310. doi:http://dx.doi.org/10.1006/jbin.2001.1029

Cheng, T.-H., Wei, C.-P., & Tseng, V. S. (2006). *Feature selection for medical data mining: comparisons of expert judgment and automatic approaches.* Paper presented at the Computer-Based Medical Systems, 2006. CBMS 2006. 19th IEEE International Symposium on.

cpan.org. (2014). Lingua::Stem. Retrieved from http://search.cpan.org/dist/Lingua-Stem/lib/Lingua/Stem.pod

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837-845.

Dietterich, T. G., Becker, S., & Ghahramani, Z. (2002). *Advances in Neural Information Processing Systems 14: Proceedings of the 2001 Conference* (Vol. 2): MIT Press.

Domingos, P. (1999). The role of Occam's razor in knowledge discovery. *Data mining and knowledge discovery, 3*(4), 409-425.

Estudillo-Valderrama, M. A., Talaminos-Barroso, A., Roa, L. M., Naranjo-Hernandez, D., Reina-Tosina, J., Areste-Fosalba, N., & Milan-Martin, J. A. (2014). A distributed approach to alarm management in chronic kidney disease. *IEEE Journal of Biomedical & Health Informatics, 18*(6), 1796-1803. doi:http://dx.doi.org/10.1109/JBHI.2014.2333880

Friedl, M. A., & Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment, 61*(3), 399-409.

García-Laencina, P. J., Sancho-Gómez, J.-L., Figueiras-Vidal, A. R., & Verleysen, M. (2009). K nearest neighbours with mutual information for simultaneous classification and missing data imputation. *Neurocomputing, 72*(7), 1483-1493.

Grannis, S., & Vreeman, D. (2010). *A vision of the journey ahead: using public health notifiable condition mapping to illustrate the need to maintain value sets.* Paper presented at the AMIA Annual Symposium Proceedings.

Imler, T. D., Vreeman, D. J., & Kannry, J. (2016). Healthcare Data Standards and Exchange *Clinical Informatics Study Guide* (pp. 233-253): Springer.

International Health Terminology Standards Development Organisation. (2016). SNOMED CT, The Global Language of Healthcare. Retrieved from http://www.ihtsdo.org/snomed-ct

Jiang, M., Chen, Y., Liu, M., Rosenbloom, S. T., Mani, S., Denny, J. C., & Xu, H. (2011). A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association, 18*(5), 601-606.

Kang, N., Afzal, Z., Singh, B., van Mulligen, E. M., & Kors, J. A. (2012). Using an ensemble system to improve concept extraction from clinical records. *Journal of Biomedical Informatics, 45*(3), 423-428. doi:http://dx.doi.org/10.1016/j.jbi.2011.12.009

Kasthurirathne, S. N., Dixon, B. E., Gichoya, J., Xu, H., Xia, Y., Mamlin, B., & Grannis, S. J. (2016). 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, 60*, 145-152.

Kauermann, G., & Carroll, R. J. (1999). The Sandwich Variance Estimator: Efficiency Properties and Coverage Probability of Confidence Intervals.

Krauthammer, M., & Nenadic, G. (2004). Term identification in the biomedical literature. *Journal of Biomedical Informatics, 37*(6), 512-526. doi:http://dx.doi.org/10.1016/j.jbi.2004.08.004

Lewis, D. D. (1998). *Naive (Bayes) at forty: The independence assumption in information retrieval.* Paper presented at the European conference on machine learning.

McDonald, C. J., Huff, S. M., Suico, J. G., Hill, G., Leavelle, D., Aller, R., . . . Hook, J. (2003). LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clinical chemistry, 49*(4), 624-633.

McDonald, C. J., Overhage, J. M., Barnes, M., Schadow, G., Blevins, L., Dexter, P. R., . . . Committee, I. M. (2005). The Indiana network for patient care: a working local health information infrastructure. *Health Affairs, 24*(5), 1214-1220.

Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. *JAMA, 309*(13), 1351-1352.

Overhage, J. M. (2016). The Indiana Health Information Exchange. In B. E. Dixon (Ed.), *Health Information Exchange: Navigating and Managing a Network of Health Information Systems* (1 ed., pp. 267-279). Waltham, MA: Academic Press.

Polani, D. (2013). Kullback-Leibler Divergence. *Encyclopedia of Systems Biology*, 1087-1088.

Rindflesch, T. C., Tanabe, L., Weinstein, J. N., & Hunter, L. (2000). *EDGAR: extraction of drugs, genes and relations from the biomedical literature.* Paper presented at the Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing.

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction

System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association, 17*(5), 507-513.

Song, M., Yu, H., & Han, W.-S. (2015). Developing a hybrid dictionary-based bio-entity recognition technique. *BMC medical informatics and decision making, 15*(1), 1.

Spasić, I., Livsey, J., Keane, J. A., & Nenadić, G. (2014). Text mining of cancer-related information: Review of current status and future directions. *International journal of medical informatics, 83*(9), 605-623. doi:http://dx.doi.org/10.1016/j.ijmedinf.2014.06.009

Tsuruoka, Y., & Tsujii, J. i. (2004). Improving the performance of dictionary-based approaches in protein name recognition. *Journal of Biomedical Informatics, 37*(6), 461-470.

US National Library of Medicine. (2015). Medical Subject Headings (MeSH®). Retrieved from https://www.nlm.nih.gov/pubs/factsheets/mesh.html

Vreeman, D. J. (2007). Keeping up with changing source system terms in a local health information infrastructure: running to stand still.

Wang, Y., & Patrick, J. (2009). *Cascading classifiers for named entity recognition in clinical notes.* Paper presented at the Proceedings of the workshop on biomedical information extraction.

World Health Organization. (2007). International Classification of Disease (ICD-10). Available: h ttp.

Yang, J., Qu, Z., & Liu, Z. (2014). Improved Feature-Selection Method Considering the Imbalance Problem in Text Categorization. *The Scientific World Journal, 2014*.

Yang, Y., & Pedersen, J. O. (1997). *A comparative study on feature selection in text categorization.* Paper presented at the ICML.

Appendix A.

A detailed breakdown of the distribution of UMLS semantic types and groupings identified via dictionary and non-dictionary based feature selection. Using the metamap tool, we identified semantic types of the 7,302 dictionary based tokens and 9,480 non-dictionary based tokens extracted from the cancer report set, and categorized their distribution across each semantic type.

Id	barb.	Semantic type name	Dictionary	Non-Dictionary
Activities & Behaviors (ACTI)				
T051	evnt	Event	0	9
T052	acty	Activity	34	68
T053	bhvr	Behavior	8	4
T054	socb	Social Behavior	4	11
T055	inbe	Individual Behavior	2	43
T056	dora	Daily or Recreational Activity	1	6
T057	ocac	Occupational Activity	1	11
T064	gora	Governmental or Regulatory Activity	1	3
T066	mcha	Machine Activity	0	1
			51 (0.54%)	156 (1.85%)
Anatomy (ANAT)				
T017	anst	Anatomical Structure	3	19
T018	emst	Embryonic Structure	26	19
T021	ffas	Fully Formed Anatomical Structure	0	0
T022	bdsy	Body System	39	28
T023	bpoc	Body Part, organ or organ component	648	625
T024	tisu	Tissue	64	69
T025	cell	Cell	212	46
T026	celc	Cell Component	40	99
T029	blor	Body Location or Region	98	118
T030	bsoj	Body Space or Junction	28	32
T031	bdsu	Body Substance	46	53
			1204	1108

			(12.83%)	(13.15%)
Chemicals & Drugs (CHEM)				
T103	nnon	Nucleic Acid, nucleoside or nucleotide	21	17
T104	opco	Organophosphorus Compound	0	0
T109	aapp	Amino Acid, peptide or protien	479	253
T110	carb	Carbohydrate	0	0
T111	bodm	Biomedical or Dental Material	32	49
T114	bacs	Biologically Active Substance	186	123
T115	chvf	Chemical Viewed Functionally	5	7
T116	phsu	Pharmacologic Substance	303	225
T118	chem	Chemical	3	2
T119	chvs	Chemical Viewed Structurally	4	6
T120	strd	Steroid	0	0
T121	eico	Eicosanoid	0	0
T122	nsba	Neuroreactive Substance or Biogenic Amine	0	0
T123	horm	Hormone	122	20
T124	enzy	Enzyme	131	49
T125	rcpt	Receptor	21	12
T126	antb	Antibiotic	2	15
T127	elii	Element, ion, or isotope	27	49
T129	inch	Inorganic Chemical	30	31
T130	orch	Organic Chemical	185	174
T131	clnd	Clinical Drug	0	0
T192	hops	Hazardous or Poisonous Substance	24	18
T195	imft	Immunologic Factor	68	80

T196	irda	Indicator, reagent or diagnostic aid	15	23
T197	vita	Vitamin	12	5
T200	lipd	Lipid	0	0
			1670 (17.79%)	1158 (13.75%)
Concepts & Ideas (CONC)				
T077	cnce	Conceptual Entity	80	139
T078	idcn	Idea or Concept	88	138
T079	tmco	Temporal Concept	72	119
T080	qlco	Qualitative Concept	654	516
T081	qnco	Quantitative Concept	285	403
T082	spco	Spatial Concept	263	275
T089	rnlw	Regulation or Law	1	2
T102	grpa	Group Attribute	2	1
T169	ftcn	Functional Concept	280	235
T170	inpr	Intellectual Product	274	500
T171	lang	Language	2	8
T185	clas	Classification	17	27
			2018 (21.5%)	2363 (28.05%)
Devices (DEVI)				
T074	medd	Medical Device	65	140
T075	resd	Research Device	0	3
T203	drdd	Drug Delivery Device	0	1
			65 (0.7%)	144 (1.71%)
Disorders (DISO)				

T019	cgab	Congenital Abnormality	170	18
T020	acab	Acquired Abnormality	10	14
T033	fndg	Finding	348	228
T037	inpo	Injury or Poisoning	11	16
T046	patf	Pathologic Function	84	90
T047	dsyn	Disease or Syndrome	650	256
T048	mobd	Mental or Behavioral Dysfunction	8	6
T049	comd	Cell or Molecular Dysfunction	19	10
T050	emod	Experimental Model of Disease	1	2
T184	sosy	Sign or Symptom	44	42
T190	anab	Anatomical Abnormality	17	14
T191	neop	Neoplastic Process	984	208
			2346 (25%)	904 (10.73%)
Genes & Molecular Sequences (GENE)				
T085	mosq	Molecular Sequence	0	0
T086	nusq	Nucleotide Sequence	2	1
T087	amas	Amino Acid Sequence	1	0
T088	crbs	Carbohydrate Sequence	0	0
T028	gngm	Gene or Genome	870	596
			873 (9.3%)	597 (7.08%)
Geographic Areas (GEOG)				
T083	geoa	Geographic Area	115	555
			115 (1.23%)	555 (6.59%)
Living Beings (LIVB)				

T001	orgm	Organism	3	5
T002	plnt	Plant	68	93
T004	fngs	Fungus	5	16
T005	virs	Virus	40	13
T007	bact	Bacterium	3	14
T008	anim	Animal	4	5
T010	vtbt	Vertebrate	0	0
T011	amph	Amphibian	0	7
T012	bird	Bird	10	26
T013	fish	Fish	8	20
T014	rept	Reptile	2	5
T015	mamm	Mammal	53	34
T016	humn	Human	10	10
T096	grup	Group	1	3
T097	prog	Professional or Occupational Group	26	60
T098	popg	Population Group	64	69
T099	famg	Family Group	4	12
T100	aggp	Age Group	11	5
T101	podg	Patient or Disabled Group	4	2
T194	arch	Archaeon	0	0
T204	euka	Eukaryote	86	100
			402 (4.29%)	499 (5.92%)
Objects (OBJC)				
T167	sbst	Substance	18	29
T168	food	Food	30	56
T071	enty	Entity	1	2

T072	phob	Physical Object	1	4
T073	mnob	Manufactured Object	93	194
			143 (1.53%)	285 (3.38%)
Occupations (OCCU)				
T090	ocdi	Occupation or Discipline	1	7
T091	bmod	Biomedical Occupation or Discipline	3	15
			4 (0.04%)	22 (0.26%)
Organizations (ORGA)				
T092	orgt	Organization	2	9
T093	hcro	Health Care Related Organization	13	22
T094	pros	Professional Society	1	3
T095	shro	Self-help or Relief Organization	0	0
			16 (0.17%)	34 (0.4%)
Phenomena (PHEN)				
T034	lbtr	Laboratory or Test Result	19	12
T038	biof	Biologic Function	1	1
T067	phpr	Phenomenon or Process	14	38
T068	hcpp	Human-caused Phenomenon or Process	0	3
T069	eehu	Environmental Effect of Humans	0	1
T070	npop	Natural Phenomenon or Process	43	42
			77 (0.82%)	97 (0.5%)
Physiology (PHYS)				
T032	orga	Organism Attribute	37	37
T039	phsf	Physiologic Function	7	9

T040	orgf	Organism Function	34	27
T041	menp	Mental Process	10	23
T042	ortf	Organ or Tissue Function	11	6
T043	celf	Cell Function	30	13
T044	moft	Molecular Function	37	22
T045	genf	Genetic Function	23	3
T065	edac	Educational Activity	0	3
T201	clna	Clinical Attribute	53	67
			242 (2.58%)	210 (2.49%)
Procedures (PROC)				
T058	hlca	Health Care Activity	7	41
T059	lbpr	Laboratory Procedure	59	71
T060	diap	Diagnostic Procedure	13	35
T061	topp	Therapeutic or Preventive Procedure	68	127
T062	resa	Research Activity	15	16
T063	mbrt	Molecular Biology Research Technique	0	2
T065	edac	Educational Activity	0	3
			162 (1.73%)	295 (3.5%)
Total of semantic types identified			9388 (100%)	8427 (100%)

Appendix B. The list of top 5, 10, 15, 20 features selected using dictionary and non-dictionary based sourcing approaches, together with summary statistics describing their frequency of appearance in positive and negative contexts across (a) all reports as well as reports labeled as (b) cancer positive and (c) cancer negative via manual review.

Dictionary based tokens

	All reports				Cancer negative reports				Cancer positive reports			
Token (Context)	Total	Max	Std. dev.	Avg.	Total	Max	Std. dev.	Avg.	Total	Max	Std. dev.	Avg.
Top 5 tokens												
tumor (P)	5197	34	2.139	0.742	194	13	0.374	0.038	5003	34	3.386	2.566
tumor (N)	1447	15	0.824	0.207	122	8	0.235	0.024	1325	15	1.410	0.679
carc (P)	1	1	0.012	0.000	0	0	0.000	0.000	1	1	0.023	0.001
carc (N)	0	0	0.000	0.000	0	0	0.000	0.000	0	0	0.000	0.000
cin (P)	33	4	0.095	0.005	29	4	0.108	0.006	4	1	0.045	0.002
cin (N)	1	1	0.012	0.000	1	1	0.014	0.000	0	0	0.000	0.000
invas (P)	558	9	0.466	0.080	32	2	0.089	0.006	526	9	0.841	0.270
invas (N)	133	6	0.181	0.019	40	2	0.095	0.008	93	6	0.306	0.048
ca (P)	241	6	0.234	0.034	50	3	0.116	0.010	191	6	0.395	0.098
ca (N)	10	2	0.045	0.001	2	1	0.020	0.000	8	2	0.078	0.004
Top 10 tokens												
metastat (P)	1059	10	0.696	0.151	42	4	0.136	0.008	1017	10	1.225	0.522
metastat (N)	179	9	0.255	0.026	16	2	0.063	0.003	163	9	0.467	0.084
neop (P)	1	1	0.012	0.000	0	0	0.000	0.000	1	1	0.023	0.001

neop (N)	0	0	0.000	0.000	0	0	0.000	0.000	0	0	0.000	0.000
lym (P)	4	1	0.024	0.001	1	1	0.014	0.000	3	1	0.039	0.002
lym (N)	0	0	0.000	0.000	0	0	0.000	0.000	0	0	0.000	0.000
nod (P)	4	1	0.024	0.001	1	1	0.014	0.000	3	1	0.039	0.002
nod (N)	0	0	0.000	0.000	0	0	0.000	0.000	0	0	0.000	0.000
cell (P)	5624	34	1.904	0.803	1885	23	1.136	0.373	3739	34	2.820	1.917
cell (N)	695	6	0.400	0.099	351	6	0.322	0.070	344	6	0.545	0.176
Top 15 tokens												
involv (P)	949	18	0.602	0.136	165	3	0.200	0.033	784	18	1.048	0.402
involv (N)	382	8	0.323	0.055	32	3	0.095	0.006	350	8	0.574	0.179
adenoca (P)	5	2	0.032	0.001	0	0	0.000	0.000	5	2	0.060	0.003
adenoca (N)	0	0	0.000	0.000	0	0	0.000	0.000	0	0	0.000	0.000
stat (P)	4	1	0.024	0.001	3	1	0.024	0.001	1	1	0.023	0.001
stat (N)	0	0	0.000	0.000	0	0	0.000	0.000	0	0	0.000	0.000
margin (P)	4310	35	2.205	0.616	1157	15	0.937	0.229	3153	35	3.714	1.617
margin (N)	984	13	0.673	0.141	75	4	0.154	0.015	909	13	1.191	0.466
differ (P)	24	1	0.058	0.003	9	1	0.042	0.002	15	1	0.087	0.008
differ (N)	1	1	0.012	0.000	1	1	0.014	0.000	0	0	0.000	0.000
Top 20 tokens												

mass (P)	2920	48	1.485	0.417	757	14	0.698	0.150	2163	48	2.448	1.109
mass (N)	287	5	0.251	0.041	156	4	0.208	0.031	131	5	0.336	0.067
cassett (P)	7549	33	1.792	1.078	4940	16	1.205	0.978	2609	33	2.771	1.338
cassett (N)	236	16	0.470	0.034	99	7	0.231	0.020	137	16	0.808	0.070
section (P)	11130	44	3.104	1.590	5543	23	1.790	1.098	5587	44	4.903	2.865
section (N)	996	14	0.707	0.142	350	7	0.363	0.069	646	14	1.184	0.331
rad (P)	15	1	0.046	0.002	5	1	0.031	0.001	10	1	0.071	0.005
rad (N)	0	0	0.000	0.000	0	0	0.000	0.000	0	0	0.000	0.000
grade (P)	1427	10	0.742	0.204	330	5	0.350	0.065	1097	10	1.218	0.563
grade (N)	310	5	0.268	0.044	162	3	0.210	0.032	148	5	0.377	0.076
All tokens	1,317,948	1416	208.9	171.1	790,751	1279	153.5	112.1	527,197	1416	289.4	195.3

Non-dictionary based tokens

	All reports				**Cancer negative reports**				**Cancer positive reports**			
Feature (Context)	Total	Max	Std. dev.	Avg.	Total	Max	Std. dev.	Avg.	Total	Max	Std. dev.	Avg.
Top 5 tokens												
tumor (P)	5197	34	2.139	0.742	194	13	0.374	0.038	5003	34	3.386	2.566
tumor (N)	1447	15	0.824	0.207	122	8	0.235	0.024	1325	15	1.410	0.679

carcinoma (P)	2658	25	1.325	0.380	40	6	0.121	0.008	2618	25	2.232	1.343
carcinoma (N)	656	10	0.488	0.094	196	4	0.254	0.039	460	10	0.813	0.236
invasion (P)	784	9	0.489	0.112	20	5	0.099	0.004	764	9	0.851	0.392
invasion (N)	641	11	0.444	0.092	19	2	0.067	0.004	622	11	0.791	0.319
slide (P)	3199	22	1.344	0.457	711	22	0.745	0.141	2488	21	2.028	1.276
slide (N)	88	4	0.152	0.013	26	3	0.082	0.005	62	4	0.254	0.032
cell (P)	5624	34	1.904	0.803	1885	23	1.136	0.373	3739	34	2.820	1.917
cell (N)	695	6	0.400	0.099	351	6	0.322	0.070	344	6	0.545	0.176
Top 10 tokens												
metastat (P)	1059	10	0.696	0.151	42	4	0.136	0.008	1017	10	1.225	0.522
metastat (N)	179	9	0.255	0.026	16	2	0.063	0.003	163	9	0.467	0.084
lymph (P)	5723	55	3.441	0.818	883	26	1.169	0.175	4840	55	5.928	2.482
lymph (N)	824	15	0.650	0.118	90	4	0.166	0.018	734	15	1.163	0.376
node (P)	6524	58	3.881	0.932	920	26	1.277	0.182	5604	58	6.681	2.874
node (N)	946	16	0.741	0.135	136	8	0.231	0.027	810	16	1.314	0.415
return (P)	791	5	0.389	0.113	132	3	0.193	0.026	659	5	0.614	0.338
return (N)	2	1	0.017	0.000	0	0	0.000	0.000	2	1	0.032	0.001
adenocarcin	840	12	0.652	0.120	3	1	0.024	0.001	837	12	1.179	0.429

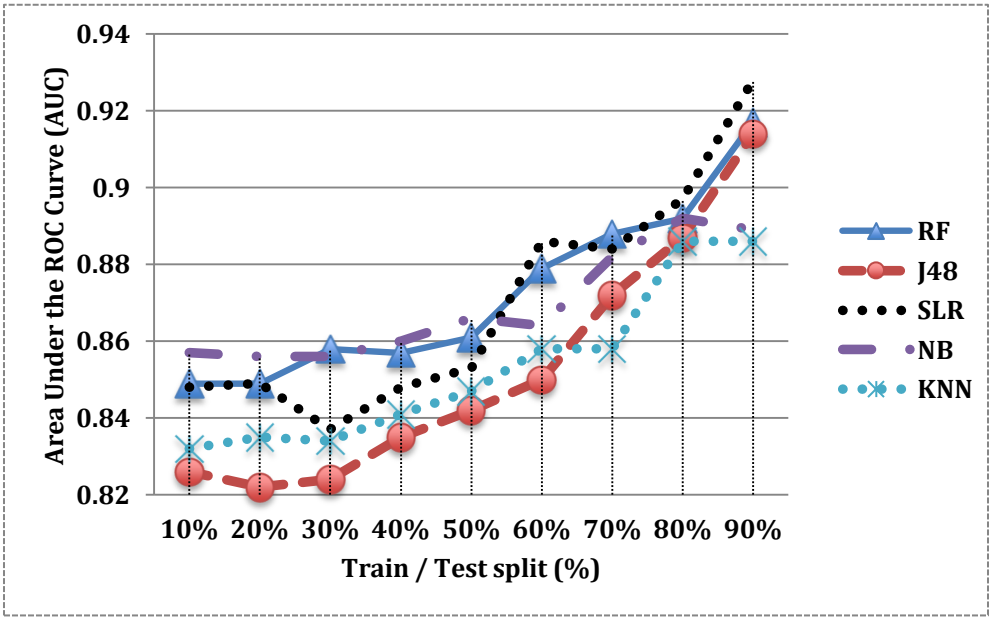
oma (P)												
adenocarcinoma (N)	61	4	0.114	0.009	6	1	0.034	0.001	55	4	0.207	0.028
Top 15 tokens												
margin (P)	4310	35	2.205	0.616	1157	15	0.937	0.229	3153	35	3.714	1.617
margin (N)	984	13	0.673	0.141	75	4	0.154	0.015	909	13	1.191	0.466
involv (P)	949	18	0.602	0.136	165	3	0.200	0.033	784	18	1.048	0.402
involv (N)	382	8	0.323	0.055	32	3	0.095	0.006	350	8	0.574	0.179
consult (P)	3311	15	1.041	0.473	1258	10	0.650	0.249	2053	15	1.528	1.053
consult (N)	79	2	0.107	0.011	16	1	0.056	0.003	63	2	0.180	0.032
differenti (P)	746	9	0.494	0.107	68	2	0.122	0.013	678	9	0.870	0.348
differenti (N)	40	2	0.081	0.006	4	1	0.028	0.001	36	2	0.146	0.018
mass (P)	2920	48	1.485	0.417	757	14	0.698	0.150	2163	48	2.448	1.109
mass (N)	287	5	0.251	0.041	156	4	0.208	0.031	131	5	0.336	0.067
Top 20 tokens												
cassett (P)	7549	33	1.792	1.078	4940	16	1.205	0.978	2609	33	2.771	1.338
cassett (N)	236	16	0.470	0.034	99	7	0.231	0.020	137	16	0.808	0.070
phone (P)	688	5	0.368	0.098	149	4	0.205	0.030	539	5	0.577	0.276
phone (N)	0	0	0.000	0.000	0	0	0.000	0.000	0	0	0.000	0.000

left (P)	7951	40	2.622	1.136	3446	40	1.690	0.682	4505	38	3.921	2.310
left (N)	332	8	0.337	0.047	68	3	0.146	0.013	264	8	0.585	0.135
grade (P)	1427	10	0.742	0.204	330	5	0.350	0.065	1097	10	1.218	0.563
grade (N)	310	5	0.268	0.044	162	3	0.210	0.032	148	5	0.377	0.076
microscop (P)	7876	10	1.014	1.125	5008	6	0.878	0.992	2868	10	1.236	1.471
microscop (N)	128	5	0.168	0.018	54	3	0.119	0.011	74	5	0.253	0.038
All tokens	1,557,188	1663	222.5	162.9	936,867	1539	185.5	113.9	620,321	1663	318.2	221.5

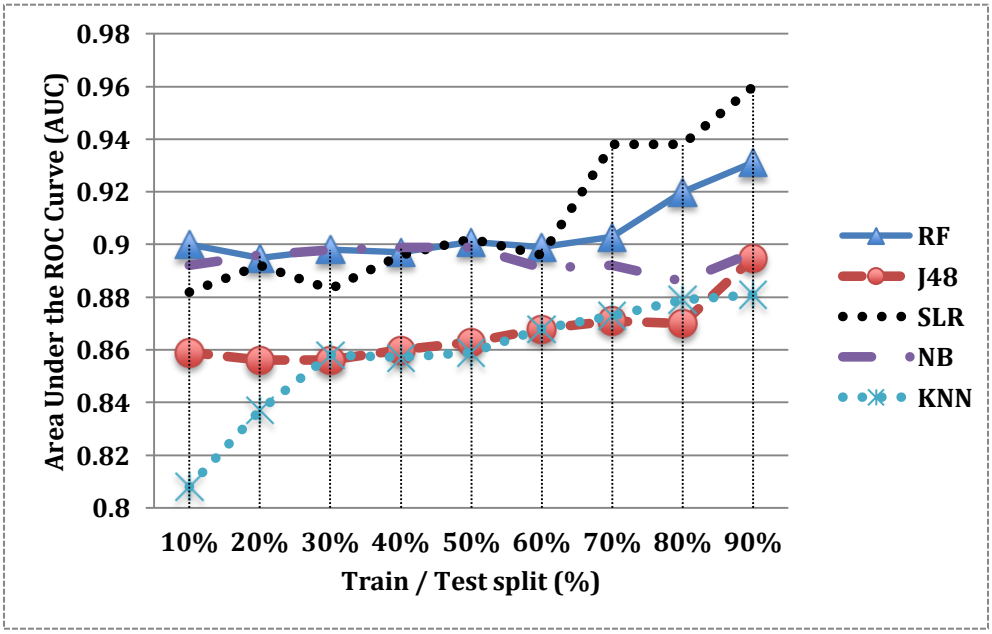
Appendix C. The performance of each model when trained using incremental train/test splits starting from 10% and increasing to 90%.

1) Dictionary based decision models

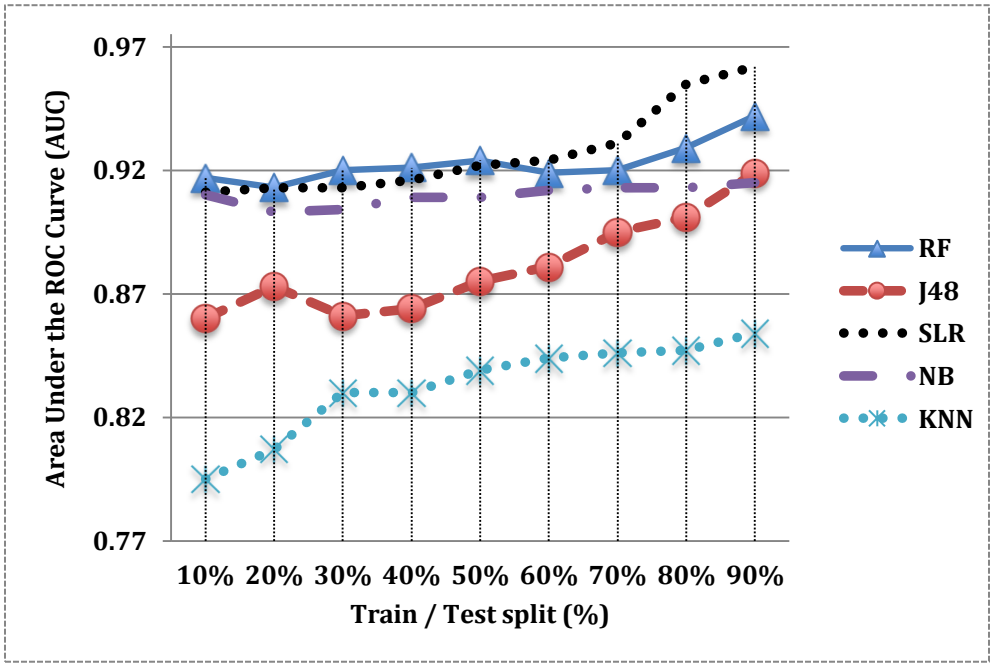
5 tokens



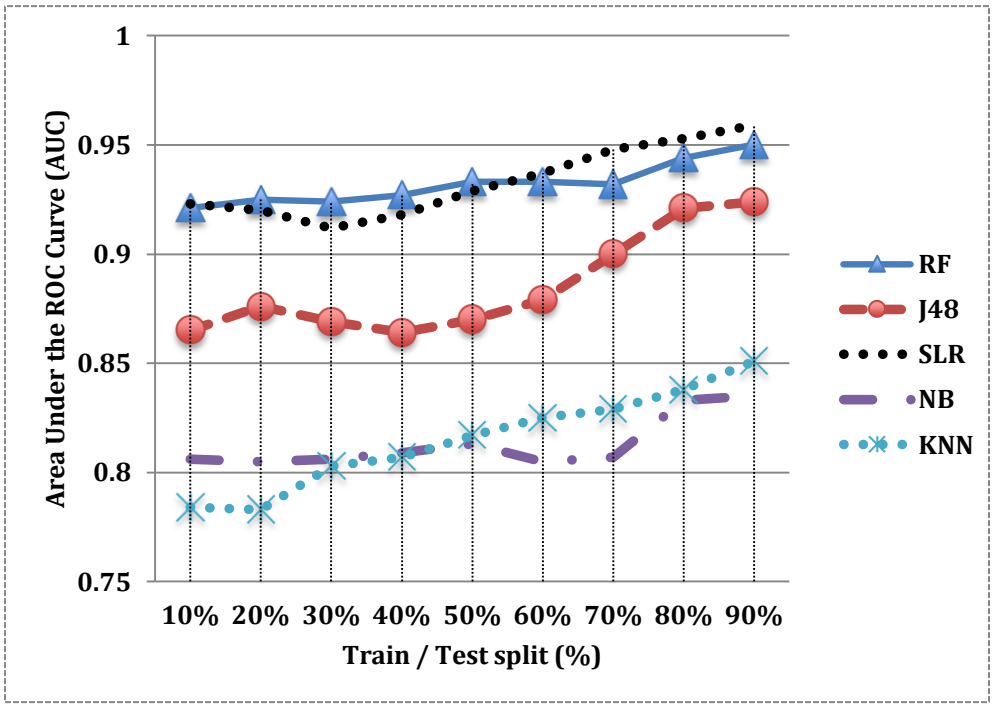
10 tokens



15 tokens

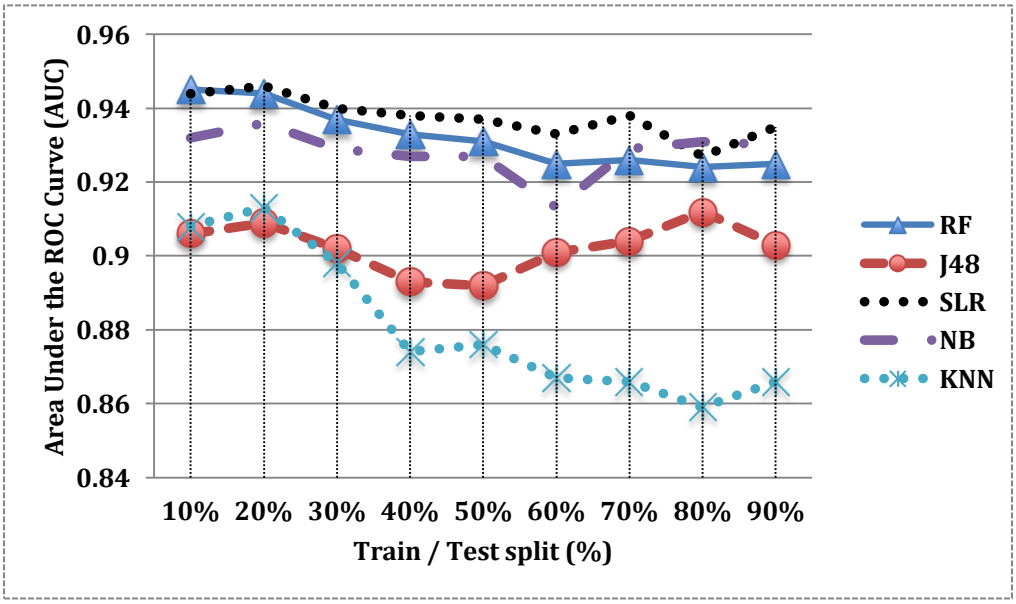


20 tokens

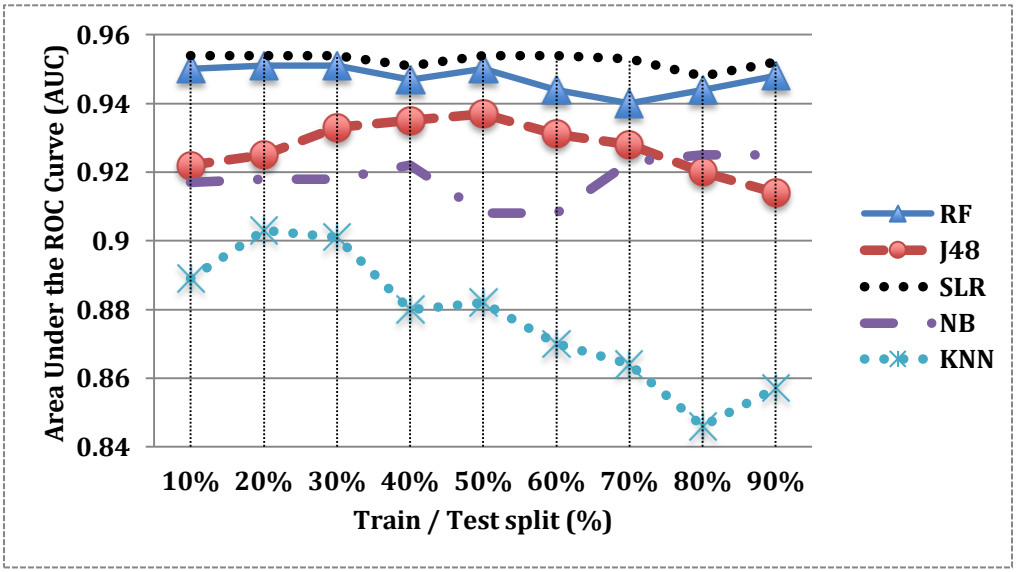


2) Non-dictionary based decision models

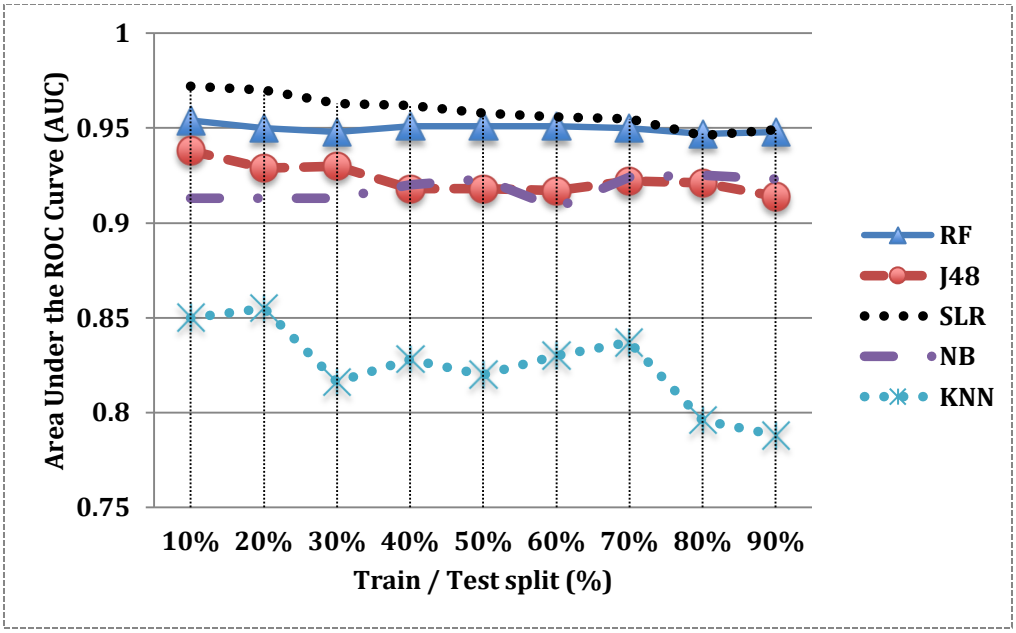
5 tokens



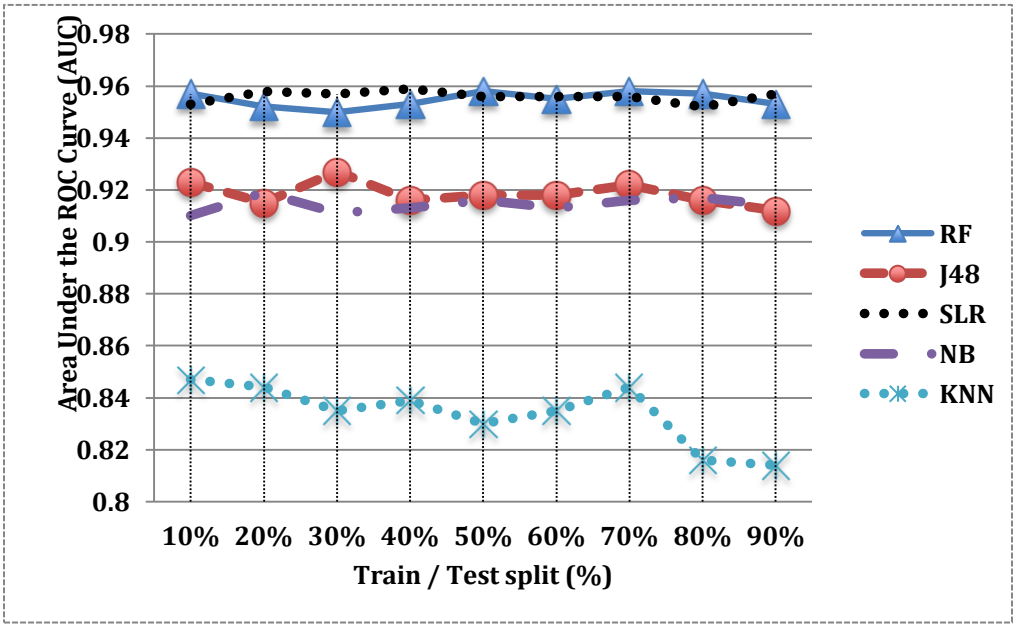
10 tokens



15 tokens

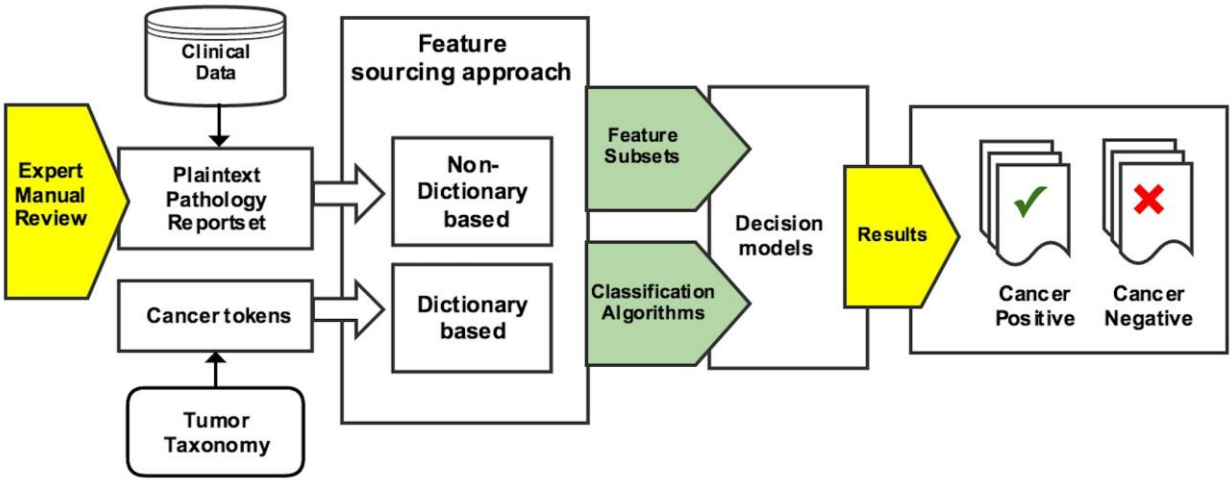


20 tokens



The authors declare no conflicts of interest.

Graphical abstract



Highlights

Dictionary and non-dictionary based feature sources can be used to build decision models to predict cancer using plaintext data.

Decision models parameterized using dictionary and non-dictionary feature sourcing approaches yielded performance metrics between 70-90%.

Feature source and feature subset size had no impact on the performance of a decision model.

Decision models built using features extracted from the plaintext reports themselves achieve comparable results to those built using medical dictionaries.

Non-dictionary based approaches may be generalized for other health analytics applications and healthcare domains.