Applying Machine Learning and Application Development to Lower Back Pain and Genetic Medicine

Chethan Jujjavarapu

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Sean D. Mooney, Chair

Patrick J. Heagerty

Jeffrey G. Jarvik

Trevor A. Cohen

Jairam Lingappa

Program Authorized to Offer Degree:

Department of Biomedical Informatics and Medical Education

University of Washington

Abstract

Applying Machine Learning and Application Development to Lower Back Pain and Genetic Medicine

Chethan Jujjavarapu

Chair of the Supervisory Committee:
Sean D. Mooney
Department of Biomedical Informatics and Medical Education

Improvement of the healthcare system is a focal point for academic leaders. In recent years, precision medicine initiatives have gained traction as a solution to improve care by leveraging healthcare analytics and informatic tools to assist clinicians in prescribing individualized treatments based on the patient's health characteristics. This involves data collection, data management and advanced statistical and machine learning methods, and new tools to deliver the promise of data on the outcomes of health and healthcare. To help clinicians, researchers must leverage electronic health record (EHR) data, however these data are complex as they are made up of multiple modalities with an ever increasing volume. While structured EHR data is a popular modality to use for analysis, clinical notes (i.e. unstructured EHR data), for example, provide more granular information about patients that is useful to clinicians. As a result, there is

interest in building cohorts of patients based on unstructured data by using natural language processing (NLP). For analysis, there are recent works that discuss the value of using deep learning to integrate multiple data modalities together to better predict clinical outcomes; however, rigorous testing is needed to fully understand this value. Once data has been collected and analyzed, the final task is understanding how to further patient involvement with this information. In this dissertation, I focus on creating a framework that can build cohorts based on unstructured data, analyze EHR data using the different modalities, and increase patient involvement. The aims are to: 1) compare NLP methods for the classification of lumbar spine imaging findings related to lower back pain, 2) predict decompression surgery by applying machine learning to patients' structured and unstructured health data, and 3) demonstrate patient delivery and sharing of data in a smartphone app to facilitate family communication of genetic results.

Acknowledgements

Throughout the writing of this dissertation I have received a great deal of support and assistance.

I would first like to thank Vikas Rao Pejaver, whose quantitative expertise and mentorship were invaluable throughout my time in the program. He instilled in me the work ethic, confidence, and persistence to lead projects from start to finish. I feel privileged to have worked with him and look forward to his many future accomplishments.

I give thanks to Sean D. Mooney, Gail P. Jarvik, CLEAR Center, and the UW ITHS TL1 program for providing funding for me. I would like to give thanks to my supervisors. I'd like to thank Sean D. Mooney for providing a network of academics to help me achieve my goals. I give thanks to Jeffrey G. Jarvik; his domain expertise pushed me to deepen my knowledge of lower back pain research and his attention to detail influenced my research practices. Patrick Heagerty is the living definition of the ideal leader and quantitative scientist that I aspire to be. His positive attitude when providing feedback always brought me comfort and sharpened my statistics and machine learning skills. I would like to thank Trevor Cohen and his constant curiosity for uncovering the unknown. His feedback always challenged my knowledge and made me want to dig deeper to understand the unknown. Finally, I would like to thank Jairam Lingappa for providing the additional clinical perspective.

Finally, I would like to thank my family: Shilpa, Chandramathi, and Phanindra Jujjavarapu for supporting me during my time in the program.

Table of Contents

Introduction

Background

Precision Medicine (PM) has become a focal point to improve modern medicine[1–3]. This idea gained popularity in 2015 when President Barack Obama announced the United States would start a government-funded initiative to enroll and then collect data from 1 million US citizens[1]. This and other initiatives around the world have propelled scientists to think more deeply about how clinical data can be leveraged for improving healthcare. Under PM, this improved healthcare system will leverage electronic health records (EHR) and other health-related data to tailor treatments and procedures to subpopulations of patients[1]. To achieve this goal, the biomedical informatics community has focused their attention on two tasks: 1) analyzing EHR data and 2) building patient-facing tools to better direct patients' health odysseys[4–9].

Analysis of EHR Data

EHR data is a rich source of information that has given rise to a large volume of diverse types of data ready to be made useful for evidence-based healthcare[7]. In 2009, the Health Information Technology for Economic and Clinical Health (HITECH) Act was signed, which gave financial incentive for hospitals to adopt EHR systems to better monitor patients. In a national survey, only 13% of clinicians reported having a basic EHR system, while in 2012, 72% adopted some type of EHR system[10,11]. As a result, there has been a subsequent production of data that is composed of different types: tabular, free-text, imaging, and patient-generated; however, the human cognition to make sense of these data is limited. The purpose is to create a continually learning healthcare system that can transform data into knowledge that supports clinical decision making[12–17]. Thus, computational-based methods are needed to recognize the patterns in this ever increasing volume of different data types[18].

Machine learning (ML) is a particular computational-based method that builds models to find hidden insights from data[19]. Supervised ML, a subdivision of ML, focuses on developing a model that can accurately predict labels (e.g. malignant vs. benign) based on features (e.g. patient demographics) from the data. Through this process the method learns a relationship between labels and features by training and testing on labelled data, and then predicts on real-world unlabelled data. Post data cleaning, this process automates the discovery of underlying patterns in the data without the need of specific decision rules or to account for the complex interactions between features[20]. As a result, ML has become the preferred method to analyze healthcare data[21,22].

The primary focus of ML has been to build cohorts for analysis using tabular (i.e. structured) EHR data, however free-text (i.e. unstructured) data offers an alternative that has yet to be fully utilized. Unstructured data can contain copious amounts of information that can be used to build patient cohorts; for example, radiology reports are used to record radiologists' observations and impressions of a patient's diagnostic imaging test[23]. This information is vital for assessing the next steps for a patient in their health odyssey, however it cannot be stored in a tabular format. Natural Language Processing (NLP) is the primary method to convert this unstructured data into a format that can be used by ML models for automatic identification[24,25]. To convert data to this format, NLP relies on a sequence of steps that include segmenting text into sentences, tokening the words, and then stemming the words[25,26]. ML models then use these NLP features for prediction. However, NLP is a broad term that encompasses a number of different methods to convert unstructured data. Comparing these different methods that range in complexity from simplistic methods such as n-grams to sophisticated methods such as embeddings is needed to understand under which circumstances one method should be used over another for processing unstructured data[27].

With the addition of NLP, a shift in utilizing EHR data has occurred as there is a growing interest in understanding the value of leveraging both structured and unstructured data for prediction[28–31]. Multimodal deep learning (MDL) has emerged as a way to use deep learning architecture to combine both data types, but more importantly learn the complex relationships between these types of data to improve prediction. For example, Rajkomar et al. built a MDL model to use both data types to predict in-hospital mortality and achieved an area under the curve (AUC) of 0.93-0.94[30]. Zhang et al. developed two MDL models to incorporate structured, unstructured, and temporal data to predict three different clinical outcomes that outperformed models that used only one data type[31]. Miotto et al. developed a MDL architecture to create representations of patients using both data types; when models used these representations to predict a number of diseases, they achieved an AUC of >0.85[32]. With the rise in both the volume and complexity of EHR data, MDL is a possible solution to make sense of these data, but needs further evaluation to assess utility across healthcare systems and comparison to conventional ML methods.

Patient-Facing Tools

Once analyzing EHR data is complete, one of the next steps is to share this information with patients to possibly elicit positive changes in their treatment trajectory. An important clinical practice is genetic testing, which is an essential tool to assist patients and clinicians to better understand the risk of hereditary disease, however lack of patient engagement and education limits its efficacy. Early genetic testing for germline risk variants can promote reproductive autonomy and lead to the recommendation of appropriate medical screening to mitigate risk of developing disease or provide early diagnosis at a more treatable stage. For example, if colorectal cancer (CRC)-associated pathogenic variants are found, colonoscopy with polypectomy can prevent CRC and prophylactic bilateral salpingo-oophorectomy surgery reduces ovarian cancer (OC) risk and are consensus recommendations[33,34]. Thus, early genetic testing is necessary to reduce risk of morbidity and mortality for patients. While genetic testing is important for patients, these test results may also be important for their biological relatives. A patient's positive test result allows for inexpensive and often free direct testing of at-risk family members for that same pathogenic variant. A positive or negative test in a family member is likely to affect their clinical care. However, sharing of genetic results between patients and their biological relatives is infrequent[35–42]. The two most frequently reported communication barriers for sharing are 1) patients have difficulty in clearly communicating the results and meaning, and/or 2) biological relatives have difficulty in interpreting the result[35–37,41].

Family communication tools may improve the dissemination of genetic results among family members. Mobile technology provides a means of communication to improve health behavior for patients[43]. However, mobile health apps' patient privacy is questionable. A recent study found that 81% of diabetes apps do not have privacy policies and would share sensitive patient information to third parties without the patient's permission[44]. Additionally, another study found that 20% (7/35) of health apps would transmit identifiable information over the Internet without encryption[45]. There is an opportunity to leverage mobile technology to increase communication of genetic test results between patients and their family members, while protecting their privacy.

Aims

This work includes three studies to address the following aims:

Aim 1: A Comparison of Natural Language Processing Methods for the Classification of Lumbar Spine Imaging Findings Related to Lower Back Pain

I will assess the 1) performance within and 2) generalizability across the four healthcare systems of different NLP methods: rules, n-grams, controlled vocabulary, and document embeddings, coupled with elastic-net logistic regression (i.e. the ML model) to classify radiology reports for lower back pain-related findings.

Aim 2: Predicting Decompression Surgery by Applying Machine Learning to Patients' Structured and Unstructured Health Data

I aim to predict decompression surgery for patients with lumbar spinal stenosis and lumbar disc herniation by applying MDL to their structured and unstructured data and evaluate performance and generalizability across four healthcare systems. If successful, the ability to identify patients at high risk of surgery could lead clinicians to either try more focused or intensive non-surgical treatments or possibly recommend surgery earlier than they otherwise would. Additionally, if the model predicts patients are unlikely to receive surgery, this may help patients accept their conservative treatment plan.

Aim 3: Demonstrate Patient Delivery and Sharing of Data in a Smartphone App to Facilitate Family Communication of Genetic Results

I aim to build a free secure smartphone app, to 1) lower the barrier to sharing genetic test results with family members by sharing test results with anyone contactable by text or email, 2) provide links to educational material, text to explain how to understand the results, and proper next steps for recipients, and 3) increase security of patient data by allowing for encrypted transmission and minimizing the amount of data needed to register for the app.

Overview

This work advances 1) the need for evaluating clinical ML methods across multiple healthcare systems to assess reliability and 2) the ability to securely share sensitive patient information. The former is applied to the lower back pain domain, while the latter is applied to the genetic testing domain.

Aim 1

Introduction

Lower back pain (LBP) is a common condition, in which patients typically exhibit heterogeneous anatomic phenotypes and undergo a variety of treatments[46–48]. LBP patients frequently receive spinal imaging, and findings identified in the resulting radiology reports are expected to help with phenotyping and decision-making[48]. However, the association between many findings and LBP is uncertain, because findings can be present in both symptomatic and asymptomatic patients[49]. As a result, patients with common aging-related findings may be recommended for LBP-related treatments that are not etiologically linked to their pain. To address this uncertainty, cohort studies and pragmatic trials have investigated patterns of care among patients with LBP and sought to explore subgroups of patients based on the presence of potentially clinically important findings[49–51]. To address the interpretation of radiology findings, the Lumbar Imaging with Reporting of Epidemiology (LIRE) study assessed the effectiveness of including benchmark prevalences in the asymptomatic population for findings found in radiology reports for patients who received a diagnostic imaging test of the lumbar spine to reduce subsequent spine-related interventions at four healthcare systems: Kaiser Permanente of Washington, Kaiser Permanente of Northern California, Henry Ford Health System, and Mayo Clinic Health System[52]. To further assist research investigating the relationship between findings and LBP, the accurate extraction of findings from large patient groups is needed. However, manual annotation is time-consuming. Natural Language Processing (NLP)-based classification pipelines offer an automated alternative to identify key findings in radiology reports[48].

An NLP-based classification pipeline is composed of two parts: 1) NLP methods that extract features from free-text data and convert them to a structured format (or representation) and 2) the machine learning (ML) model that uses these representations for classification. Text conversion or feature generation can be performed using methods that range from relatively simple domain-dependent and highly manual, to sophisticated data-driven scalable strategies[24,53–55]. Task-specific rule-based methods identify terms in the free-text that are typically defined by domain experts for a specific outcome of interest. Word or phrase counting methods (n-grams) convert free-text to grouped consecutive words[54]. Controlled vocabulary methods convert free-text into a standardized language, using resources such as the Unified

Medical Language System (UMLS)'s Metathesaurus, a large vocabulary database that contains information about biomedical and health related concepts, their various names, and the relationships among them[53,56,57]. Document embedding methods use neural networks to represent the semantics of documents as vectors of continuous numerical values[55]. Each of these methods produces different representations that can influence ML performance. Previous studies have investigated the classification performance of these types of NLP methods[27,58,59]. However, to the best of our knowledge only one study assessed generalizability as well[60]. With this study, they investigated the performance of their embeddings on a single external dataset, however without extensive validation on multiple external datasets, there is a risk of overestimating both NLP strategies and ML models' performance[24,60,61].

We hypothesize that NLP methods will have more heterogeneous performance characteristics on external data compared with internal data. The LIRE data provide an opportunity to conduct a systematic evaluation of the utility of different representational methods for identification of image findings in radiology reports drawn from multiple healthcare systems. To assess the reproducibility and reliability of NLP methods, we need to test our methods on multiple external datasets. The purpose of our study is to assess the 1) performance within and 2) generalizability across the four LIRE healthcare systems of different NLP-based feature extraction methods: rules, n-grams, controlled vocabulary, and document embeddings, coupled with elastic-net logistic regression (i.e. the ML model) for classifying radiology reports for LBP-related findings.

Methods

Annotated Dataset

This was a retrospective study that utilized the same annotated dataset from a previous study that showed that ML-based models were superior to rule-based classification[48]. Our work is an extension of this as we 1) expanded our NLP methods to include controlled vocabulary and document embeddings and 2) explicitly assessed generalizability across healthcare systems. All participating IRBs agreed that the LIRE study was minimal risk and granted waivers of both consent and Health Insurance Portability and Accountability Act authorization (IRB approval number is 476829). We used limited dataset consisting of a subsample of the LIRE cohort which consisted of approximately 250,000 patients from four healthcare systems who received a thoracic or lumbar spine plain X-ray, magnetic resonance imaging (MRI), or computed tomography (CT) between October 1, 2013 and September 30, 2016[52]. The LIRE study was a multicenter intervention study that investigated whether inserting text about finding prevalence into lumbar spine imaging reports reduced subsequent spine-related treatments[52]. Once in the study, patients were followed for two years and their data was regularly collected. We randomly sampled 871 index radiology reports, the first radiology report for each patient, and stratified by system and image modality[48]. The sample size was determined based on prior NLP classification tasks[62]. Each report was annotated for the presence of 26 LBP-related findings (Table 1-1) by a team of clinical experts composed of two neuroradiologists, a physiatrist, and a physical therapist. A single report can be annotated for multiple findings. These findings were based on prior research consisting of a review[63], prospective cohort study[49], and randomized control trial[64] that characterized LBP based on its causes and treatments. Out of these 26, eight were considered to be potentially clinically important: *any stenosis, central stenosis, lateral stenosis, foraminal stenosis, disc extrusion, nerve root displacement compression, endplate edema, and listhesis grade 2[49,50,52]*. Further details of this sampling and annotation process are presented in a previous study[48].

Classification Pipeline Overview

Our classification pipeline analyzed the 871 LIRE radiology reports with the goal of learning patterns that are predictive of each of the 26 findings (Figure 1-1). The pipeline can be separated into three steps: preprocessing, featurization, and ML.

Preprocessing and Featurization

For preprocessing, we developed regular expressions to help isolate the finding and impression sections of the 871 radiology reports. For featurization, rules, n-grams, controlled vocabulary mapping, and document embedding methods were used to extract features from the finding and impression sections. Rules require domain experts to identify terms that are related to an outcome of interest. This method is time-consuming, but since the rules were developed by clinician experts, they can be considered a proxy for clinicians' judgement for annotations. In our implementation, we developed regular expressions based on the terms our team of clinical experts identified for each finding during the annotation process. For each report, we split the text into sentences. For each sentence, we identified the presence of a finding using the regular expression and checked for negation[65]. However, the presence of findings may be uncertain as radiology reports can have terms such as "suggesting" and "not definite". We minimized this uncertainty by coding these and other similar terms as indicating the presence of a finding. We used Java (v4.6.0), using Apache Lucene (v6.1.0), Porter Stemmer, and NegEx[66,67].

N-grams is a simple, but powerful method in NLP that converts free-text across the radiology reports to n-grams (phrases of different lengths) and indicates their presence and absence in each report[68]. In our implementation, we used R (v3.6.1) package Quanteda (v2.0.1) to convert the text into un-, bi-, and trigrams.

Controlled vocabulary is a filtered version of the n-grams approach that leverages only clinically-related features from a standardized medical terminology mapped from the text[53]. In our implementation, we first split the text into its constituent sentences using the maximum entropy model in the Apache OpenNLP toolkit to infer the end of a sentence[69]. We then applied MetaMap Lite and an assertion classifier developed by Bejan et al., to each patient's radiology text report to obtain standard UMLS concepts and assertions of whether they were present, absent, conditional, possible or associated with someone else[70,71]. We used MetaMap Lite because previous literature demonstrated MetaMap Lite's performance was comparable to or exceeded MetaMap and other similar methods[71]. In addition, we also implemented a version of the controlled vocabulary method (*controlled vocabulary filter only*) that outputs recognized concepts as raw text, instead of the mapped UMLS terms to assess how a "many-to-one" mapping affects classification performance[72].

Document embeddings is a sophisticated approach that uses a neural network to convert the semantics of text into a continuous numerical vector[55]. For the document embedding method, we used the Python (v3.7.3) package Gensim (v3.7.1) to implement the Distributed Bag of Words (DBOW)[55,73]. We set the vector length to 600, number of epochs to 500, and allowed the model to initially learn word embeddings using the skip-gram architecture prior to learning document embeddings. We pre-trained our DBOW architecture on the full text using two data sets: (1) 522,283 radiology reports from the third version of Medical Information Mart for Intensive Care (MIMIC-III)[74], and (2) the finding and impression sections from 255,094 unannotated reports from the LIRE study. We refer to these as *document MIMIC* and *document LIRE*, respectively. The former reflects a typical pre-training scenario and the latter assesses how pre-training on a corpus similar to our actual train/test corpus of 871 reports affects classification performance. We used these pre-trained models to derive numerical vector representations of each of the 871 radiology reports. At the end of the featurization step, the textual data from radiology reports are represented as *rules*, *n-grams*, *controlled vocabulary*, *controlled vocabulary filter only*, *document MIMIC*, and *document LIRE*.

Rules- and Machine Learning-Based Models

For the *rules*, we used a rule-based model to classify a report as "positive" for a finding if at least one mention in a report was non-negated and "negative" if there was no mention or all mentions of the finding were negated. We then used the trapezoidal rule approximation to calculate the area under the curve (AUC)[75]. For each non-*rule* representation (i.e. feature set) and the finding labels from the annotation process, we developed an elastic-net logistic regression model to predict the presence of each finding (i.e. 26 binary or "one-vs.-rest" classification models) using the R (v3.5.1) package caret (v6.0-80). Within the training set, ten-fold cross validation was used to adjust the value of our regularization parameter (lambda) to perform feature selection on our predictors by shrinking our nonimportant predictors' coefficients towards 0. For each resulting finding-specific model, we identified the optimal threshold based on the training set's receiver operator characteristic (ROC) plot; the threshold is the point closest to the true positive rate of 1 and false positive rate of 0 (i.e. the point closest to the top left corner of the curve) using Euclidean distance[76].

Performance and Generalizability Assessment

We used R (v3.5.1) to evaluate each representation. We used AUC of an ROC plot as the primary evaluation metric. This is because we envision our pipeline as an efficient "first-pass" screening tool intended to favor the identification of more true positives. For performance assessment, we randomly split our full dataset into 80% (697/871) for training and 20% (174/871) for evaluating our model for each finding. We assessed group-level performance by averaging the evaluation AUC across all finding-specific models and across all potentially clinically important finding-specific models, separately. We repeated this process 25 times with each independent repeat using a different random train/test split of the data, so that we could estimate 95% confidence intervals. For each finding/group, a t-test was used to assess significant performance comparing the 25 repeats of the best representation to the next best representation. We used Bonferroni correction to correct for multiple hypothesis testing; for the two groups, we considered p-value 0.025 (0.05/2 groups) to be significant, and for the 26 findings, we considered p-value 0.0019 (0.05/26 findings) to be significant. To assess generalizability across healthcare systems, we trained our model on reports from three systems and evaluated on the fourth, iteratively, for each finding. For each finding, we calculated the mean and standard deviation of the AUC across the four systems. We calculated group generalizability by averaging the AUC across all findings, and across all potentially clinically important findings for each system and then calculated the mean and standard deviation. We chose mean and standard deviation to quantify generalizability, because the former measures the quality, while the latter measures the consistency of performance across systems. For the generalizability assessment, we included all representations except for *rules* because they were developed using reports from all four systems, eliminating the possibility of evaluation using data unavailable at the point of algorithm development.

Results

Data Summary

In our dataset (n = 871), we sampled reports with similar proportions of image type (i.e. x-ray and MRI) and patients' age and gender across our healthcare systems (Table 1-2). For performance assessment, we've shown that our training set is representative of our test set for 23/26 findings by using a t-test to assess the significant difference in the prevalence between sets across the 25 repeats for each finding (Figure 1-2). For generalizability assessment, we found that each healthcare system was comparable since the finding label prevalence across healthcare systems is overall similar with *any degeneration* having the highest label prevalence (0.896) and *listhesis grade 2* having the lowest label prevalence (0.028) (Figure 1-3).

Comparing the Group and Finding Level Predictive Performance of Individual Representations

To assess the best performing representation, we trained and tested 26 finding-specific models for each representation and calculated finding-level and group-level AUC. On average across all findings, we found that the models generally performed well, with average AUC values above 0.87. *N-grams* and *controlled vocabulary* had the best (AUC = 0.960) and worst average (AUC = 0.879) performance, respectively (Table 1-3). At the finding level, *n-grams* had better performance than the corresponding second-best representation (which differed from finding to finding) for 22 findings, 11 of which were statistically significant. These results suggest that on average, the relatively simple methodology of *n-grams* is sufficient to classify our 26 findings.

In addition to assessing the performance of *n-grams*, we were also interested in characterizing the performance relative to *rules*, a representation requiring domain-expertise, and *document LIRE*, an advanced data-driven representation. This comparison is of interest as each of these three representations reflect different disciplines in featurizing textual data that range from domain-expertise to advanced domain-agnostic implementations. On average across all findings, *rules* (AUC = 0.897) performed worse than *n-grams* and *document LIRE*. At the finding level, *rules* was out performed by *n-grams* for eight out of twelve rare findings (prevalence < 20%). Additionally, while *document LIRE* had better overall performance than *rules*, it was not the best representation for any of the findings. This may be due to the fact that *document LIRE* may not have been the best representation but had stable performance across findings (min AUC = 0.799, max AUC = 0.979) compared to *rules* (min AUC = 0.649, max AUC = 0.999). These results also suggest when considering only rare findings, *n-grams* still perform better than other representations.

Comparing the Group and Finding Level Generalizability Performance of Individual Representation across Healthcare Systems

To assess the best generalizable representation, we trained 26 finding-specific models for each representation on data from three systems and tested on the fourth system, iteratively. For each finding/group, we calculated the mean and standard deviation of the test AUC across the four systems. At the group level, *n-grams* had the best average performance across all findings (mean AUC = 0.902) and at the finding level, it was the best performing representation for 10 findings (Table 1-4). The next best representation was *document LIRE* at the group level (mean AUC = 0.879) and it was the best method for 10 findings as well (Table 1-4). Interestingly, when considering standard deviation at the group level, we found that *document LIRE* and *n-grams* were the most (standard deviation = 0.010) and least consistent (standard deviation = 0.051) representations across all findings, respectively (Table 1-5). We found *n-grams* could not generalize well to system two, particularly resulting in a lower sensitivity and higher specificity compared to other representations (Figure 1-4); we verified this result through complementary analyses (Figure 1-5). At the finding level, *document LIRE* was the most consistent representation for 11 findings. These results suggest that while *n-grams* had relatively the best performance, it had the worst consistency across systems. Instead, document embeddings pre-trained on study-specific data (*document LIRE*) had relatively the most consistent classification performance on average across our systems.

Assessing Performance and Generalizability of Potentially Clinically Important Findings

In our previous sections, we focused on all 26 findings, however we consider eight of these findings to be potentially clinically important. As a result, we believe it's important to present results for this important subset of findings. For performance assessment, *n-grams* had the best performance (AUC = 0.954), and it was significantly better than that of *document LIRE*, the second-best representation (AUC = 0.910) (Table 1-3). At the finding level, *n-grams* also had the best performance for all eight potentially clinically important findings, six of which were statistically significant. For generalizability assessment, *n-grams* had better performance (mean AUC = 0.898) than document LIRE (mean AUC = 0.890) (Table 1-4). At the finding level, *n-*

grams and *document LIRE* had the best performance for seven of these findings. For consistency, *document LIRE* was the most consistent representation with standard deviation of 0.007 compared to *n-grams*' 0.076 (Table 1-5). At the finding level, *document LIRE* and *MIMIC* had the most consistent performance for six and two potentially clinically important findings, respectively, with one tie *endplate edema* (Table 1-5). These results indicate for this subset of findings, we still observe the same trend where *n-grams* have the best performance, but *document LIRE* has the best consistency.

Discussion

Manual extraction of information from radiology reports can be burdensome, making automated NLP methods attractive for such tasks. However, correctly estimating these methods' performance across multiple healthcare systems requires an understanding of their generalizability on external datasets. In this study, we compared and contrasted the performance of different NLP methods coupled with elastic-net logistic regression to classify 26 findings related to LBP through performance and generalizability assessment. Our study suggests that if classifier development and deployment occur at the same system, then *n-grams* may be preferable. However, for deployment at multiple systems outside of the system of development, one should consider n-grams with the caveat that it's consistency can vary across systems, while document embeddings pre-trained on study-specific data (*document LIRE*) or a publicly available dataset (*document MIMIC*) had the most consistent performance.

Overall, based on performance assessment, n-grams, a relatively simple, data-driven, domain-agnostic method, is superior to more sophisticated methods (document embeddings and controlled vocabulary) in extracting known findings from text. These results are in line with prior studies[27,77]. Additionally, for rare findings (prevalence < 20%), n-grams had the highest AUCs, which is consistent with a prior study evaluating n-grams coupled with LASSO logistic regression to classify acute LBP (prevalence of 22%)[78]. However, n-grams did not generalize well across the four healthcare systems when compared to document embeddings. This performance-generalizability duality can be explained as follows: the n-grams method is dependent on the precise phrasing in the training text. When we considered performance assessment, we split the full dataset into 80% (697/871 reports) for the train set and 20% (174/871 reports) for the test set; both sets contained the four healthcare systems, and their text were representative of each other. However, when considering each system independently for the generalizability assessment, n-grams were more susceptible to overfitting, i.e., they may have contained predictors more relevant to the training systems than the test system. When comparing summary statistics of the raw text among systems, we found that system two was indeed different from the other systems (Figure 5) and changing classification thresholds for the models tested on this system did not affect performance. In comparison, document embeddings better captured differently worded but synonymous concepts by transforming the raw text into abstract numerical representations that reflect semantics, leading to less deviation in performance across systems but slightly worse performance overall.

Document embeddings are of particular interest because they represent a sophisticated method of featurization that are pre-trained on large-scale corpora to learn more generalizable representations of text. Here, document embeddings were pre-trained on two different data sources, unannotated LIRE reports (smaller but more relevant to LBP) and MIMIC-III (larger but less specific). While *document LIRE* overall performed better than *document MIMIC*, the difference was modest, suggesting that a lack of task-specific corpus is not a barrier for using document embeddings in clinical tasks. This observation is consistent with other studies[59,79,80].

Controlled vocabulary and *controlled vocabulary filter only* are two representations that can be considered filtered versions of the n-grams approach that leverages only clinically related terms. These representations differ from each other in that the former maps the clinically relevant raw terms to standardized terms to then use as features, while the latter does not map and instead uses the clinically relevant raw terms as features. As a result, *controlled vocabulary* performs a "many-to-one" mapping that can affect performance. When comparing these two representations, we found that *controlled vocabulary* marginally outperformed *controlled vocabulary filter only* in both performance and generalizability. Our study indicates that this "many-to-one" mapping is not detrimental to performance, but does not provide a substantial improvement relative to using only the clinically relevant raw terms as features.

Beyond performance and generalizability, scalability and interpretability are important factors to consider when choosing a NLP-based feature extraction method. Rules are the most interpretable method, because they solely rely on domain experts to provide the synonyms to search for in text. However, this method cannot scale well as expanding the synonyms for a 1) more complete identification of findings and 2) larger number of findings will require more time and domain experts. In contrast, n-grams, controlled vocabulary, and document embeddings are domain-agnostic computational methods, and as a result they can scale well to a large number of radiology reports and findings. These methods differ in their interpretability. Controlled vocabulary and n-grams are the most interpretable methods as the former provides an ML model clinically relevant terms as features, while the latter provides the raw text as features. It is relatively easy for a researcher to examine a model's features and coefficients based on either of these two methods and understand what aspects of a radiology report are predictive of the outcome of interest. Document embeddings is the least interpretable method as it uses a neural network to represent a document's semantics as a vector of continuous values. These values are no longer interpretable as they are a result of the interactions of different word embeddings from the radiology reports in the neural network. When considering subsequent implementation, it is important to consider factors such as scalability and interpretability in addition to performance and generalizability.

There are several limitations to this study. First, our pipeline required binary annotations for findings, however the presence of findings may be uncertain as radiology reports can have terms such as "suggesting" and "not definite". We minimized this uncertainty by coding these and other similar terms as indicating the presence of a finding. Second, while our sample size was in line with recommendations for classification tasks, larger training and testing sets could have led to less variable performances across our different NLP methods[62]. Third, we evaluated the algorithms but not the entire pipeline involving the querying and transfer of data; there may be discrepancies in our performance estimates when compared to those at actual deployment. Fourth, we could not assess our rules' generalizability, since the search terms were developed from reports from all four systems. Finally, in the case of document embeddings, because of our limited computational resources, we had to sequentially adjust hyperparameter values in the pre-training step, rather than conducting a grid search. With a more extensive hyperparameter search, we may have been able to improve performance.

Diagnostic imaging is often an early step for LBP patients that eventually leads to interventions, however the association between findings and LBP is uncertain[49]. Jarvik et al. investigated this association and identified eight (potentially clinically important) findings that may be associated with a history of LBP and of these eight, *nerve root contact, disc extrusion*, and *central stenosis* may be associated with a new onset of pain[49,50,52]. We've shown that our pipeline can automate classifying reports for these potentially clinically important findings using n-grams, and can generalize across healthcare systems using document embeddings. Our automated pipeline can assist similar studies by developing large cohorts quickly and inexpensively to investigate the association between findings and a clinical outcome within and across healthcare systems using text-based data.

Aim 2

Introduction

In the United States, low back pain (LBP) is the 5th most common reason for a hospital visit and annually the prevalence is 10-30%[63]. As a result, LBP incurs an annual cost of $100 billion and is the leading contributor to disability and workdays lost[81–83]. Despite numerous available interventions for LBP, it remains difficult to diagnose and treat effectively, in part because LBP has many anatomic and clinical subtypes[84]. Lumbar disc herniation (LDH) and lumbar spinal stenosis (LSS) are two specific spine-related clinical syndromes that are highly associated with LBP[49,81,85]. Patients with LDH experience pain caused by extension of the intervertebral disk material beyond the disk space, which may compress adjacent spinal nerves [85,86]. Patients with LSS experience pain associated with narrowing of the spaces within the spine due to changes in the intervertebral disks and facet joints, which may also compress the spinal nerves[87,88]. These syndromes have overlap as 1) patients with one entity can develop the other and 2) both involve neuropathic lower extremity pain.

Patients with LDH/LSS are often started with non-surgical treatments and if those are not effective then go on to have decompression surgery to relieve the compressed spinal nerves[88–90]. However, decompression has both potential benefits and risks. Recent studies indicate a possible improvement in early health outcomes[91–94]. Randomized controlled trials (RCTs) indicate that decompression surgery offers benefits over non-surgical treatment in the short term, but benefits decrease over time[91,92]. Another study found that LDH patients who underwent surgery had better short-term improvement in function and pain relief compared to non-surgical treatments[94]. A RCT found that LSS patients who received decompression surgery instead of non-surgical treatments had better initial improvement in back pain, but this benefit diminished over time[93]. On the other hand, decompression surgery has potential risks, with 18% of LSS patients experiencing adverse events [95], and up to 9% having clinical worsening within 1 year[96]. Another study found that 3.1% of LDH patients experienced clinical worsening within 1 year as well[97]. Continuation of non-surgical treatment is the default treatment option for patients with LDH/LSS, as many will improve over time without surgery[98]. Therefore, patients with LDH/LSS may be observed for long periods of time- even years- before surgery is considered. Overall, recommendation of decompression surgery is complicated as the outcome can be positive or negative depending on the patient. Early identification of patients at high risk of eventual surgical decompression (i.e. failure of non-surgical treatments) could allow for discussion between a patient and their clinician on the benefits and risks of pursuing surgery informed by the prediction of surgery for a given patient.

A promising method to assist patients and healthcare providers to understand a patient's predicted risk of eventual decompression surgery is machine learning (ML)[99–101]. ML is used to develop predictive models based on learning the relationships from data[102,103]. In recent years, deep learning (DL) has emerged as a popular method to learn the complex representation of raw input data by learning a lower dimensional representation[104]. Several works have applied DL to predict clinical outcomes. Norgeot et al. developed a DL to predict rheumatoid arthritis using structured data[105]. Choi et al. used a recurrent neural network to predict heart failure[106]. Both of these and other similar approaches used structured electronic health record (EHR) data, however with the growing volume and complexity of EHR data, there is interest in utilizing the full data available (i.e. both structured and unstructured data). As a result, multimodal deep learning (MDL) has emerged as a possible way to better model a patient's full characteristics to the outcome of interest[30–32]. However, a recent study indicated that depending on the underlying relationship of the features and outcome of interest, conventional ML methods may provide simpler, cheaper, and more useful data modeling that can achieve comparable, if not better performance than DL-based methods[107]. With the rise in both the volume and complexity of EHR data, MDL is a possible solution to make sense of this data for clinical use. However, rigorously testing this approach against a conventional ML method is needed to truly assess the value of this costly approach.

In the current study, we aim to predict early (within 2 months) and late decompression surgery (within 12 months) for patients with LSS/LDH by applying MDL to their structured and unstructured data and comparing the performance of MDL to LASSO logistic regression. If successful, the ability to identify patients at high risk of surgery could lead clinicians to either try more focused or intensive non-surgical treatments or possibly recommend surgery earlier than they otherwise would. Additionally, if the model predicts patients are unlikely to receive surgery, this may impact their non-surgical treatment plan.

Methods

Data Source

This was a retrospective study that utilized the Lumbar Imaging with Reporting of Epidemiology (LIRE) study dataset which consisted of approximately 250,000 patients from four healthcare systems (Group Health, Kaiser Permanente Northern California, Henry Ford, and Mayo Clinic) who received a thoracic or lumbar spine plain X-ray, magnetic resonance imaging (MRI), or computed tomography (CT) between October 1, 2013 and September 30, 2016[52]. The LIRE study was a multicenter intervention study that investigated if inserting text about finding prevalence into lumbar spine imaging reports reduced subsequent spine-related interventions[52]. Once enrolled in the study, patients were followed for two years and their EHR data was regularly collected along with their information one year prior to enrollment.

Patient Selection

We selected patients who had at least two occurrences of International Classification of Diseases (ICD)-9 and ICD-10 codes related to LSS or LDH (Table 2-1). This criteria was agreed upon by our clinical experts (PS, JF, and JGJ), since it ensured more confidence in identifying patients with these syndromes[108,109]. We based our ICD codes on two previous studies[110,111]. Martin et al. selected ICD-9 codes that were commonly used to describe spine-related problems. These codes were identified by searching the annual updates published by the World Health Organization and referencing the Conversion Tables of new ICD-9 codes published by the National Center for Health Statistics to help identify newly added or modified codes[110]. They then validated their process to group patients based on these codes by comparing it to clinician judgement using sensitivity and specificity analysis. Deyo et al. further grouped their patients with back pain into back and leg pain or herniated disc and lumbar stenosis groups based on ICD-9 codes[111]. We updated the code lists of Martin et al. and Deyo et al. to also include ICD-10[112].

Outcome

From the group of patients with LSS and LDH, we further split them into two prediction tasks: early and late surgery (Figure 2-1). Early and late surgery were separated as two different outcomes of interest based on the clinical rationale that early surgery for LSS/LDH is more likely driven by severe or progressive neurologic deficits, and is therefore fundamentally different from late surgery, which is more likely to be driven by chronic pain[86]. For early surgery, we limited the patients included to those that had at least two LSS/LDH diagnosis codes within the first year prior to LIRE enrollment and then searched two months ahead for the presence (positive) or absence (negative) of their first decompression surgery code. For late surgery, we limited patients to those that had at least two LSS/LDH diagnosis codes within the first year prior to LIRE enrollment and/or the first two months post enrollment and then searched one year ahead for the presence or absence of their first decompression surgery code. The decompression phenotype was developed by using existing Current Procedural Terminology (CPT) code algorithms and manually reviewed lists of each of these types of codes (CPT, ICD-9-PCS) potentially associated with surgery by at least one non-clinician reviewer (Table 2-1)[52,113,114]. Any uncertain codes were also reviewed by two clinician reviewers (PS and JF) and discussed until consensus was achieved by both reviewers. For early surgery, we had a total of 8,387 patients with 198 (2.4%) patients in the positive group. For late surgery, we had a total of 31,210 patients with 1,365 (4.4%) patients in the positive group.

Features

We considered patient demographics, diagnoses, procedures, prescription information, and radiology reports as predictors for the model (Figure 2-1). For demographics, we considered patients' race, age, healthcare system, and ethnicity. For the primary care provider for each patient, we considered their gender, type of clinician, and speciality. For diagnosis, we considered patients' ICD-9 and ICD-10 codes and the day they received the diagnosis. For procedures, we considered patients' CPT and Healthcare Common Procedure Coding System (HCPCS) Level II codes (i.e. procedure codes) and the day they received their procedure code. For prescriptions, we considered the drug name and prescription day. For radiology reports, we considered the finding and impression sections from the first imaging report (i.e. index image report) in the LIRE study along with the type of image (i.e. X-ray, CT, or MRI).

Preprocessing/Featurization

Demographics

This information is composed of patient and provider demographics along with the type of index image. To convert the data into a format for ML, we created dummy variables for the categorical features and normalized the discrete numerical feature (i.e. age) at the patient level.

Diagnosis, Procedures, and Prescriptions

We limited temporal data (diagnosis, prescriptions, and procedures) to the last three months of information for both prediction tasks, so that across the patients we 1) ensure that the time period is consistent and 2) minimize the variability in the amount of available data. The purpose was to minimize any influence from the heterogeneity of these factors on the prediction tasks. For diagnosis codes, we mapped ICD-10 to equivalent ICD-9 codes to minimize redundancy and then assigned all ICD-9 codes to depth level 3 on the ICD hierarchy using crosswalk files from cms.gov. We chose depth level 3 (i.e. the first three digits of ICD codes) to reduce the feature space, but also maintain a level of granularity[115]. ICD codes are organized into a hierarchy based on shared clinical characteristics. The further down in this hierarchy we go, the more specific the disease based on anatomic site, etiology, and manifestations.

Featurization for Classical Machine Learning

We created dummy variables for the features (i.e. diagnosis codes, procedure codes, and drug names) at the patient-level. Further, we excluded extremely rare (<=0.1%) or common (>=99%) features to reduce the feature space to only the most relevant.

Featurization for Deep Learning

We binned the data into one month intervals to reduce the sparsity of the eventual temporal feature matrix. We then created dummy variables for the features (i.e. diagnosis codes, procedure codes, and drug names) at the bin-level for each patient. To maintain the same number of bins (i.e. three), we padded for patients with less than three bins. Finally, we converted the dataframe into a 3D tensor where the depth corresponds to the number of the patients, the height to the number of bins, and the width to the number of unique features.

Index Imaging Reports

For these reports, we isolated and combined the finding and impression sections together. The purpose was to limit the text to only information that pertained to the actual diagnostic image. We then cleaned the text by converting it to lowercase, removing punctuation, removing extra whitespace, removing stopwords, and then isolated the stem of each word using a PorterStemmer from the python package *nltk[116]*.

Featurization for Classical Machine Learning

We converted the cleaned text into uni-, bi-, and trigrams using the python package *scikit-learn[117]*. Further, we excluded extremely rare (<=0.1%) or common (>=99%) n-grams to reduce the feature space to only the most relevant features.

Featurization for Deep Learning

To convert the index reports into a format for the DL architecture, we used word2vec from the python package *gensim[73]*. We first collected reports (n = 123,461) post LIRE enrollment and preprocessed them the same way as the index reports. We pre-trained a word2vec model on these reports using specific parameter settings (skip-gram version and vector length set to 300) from a recent study that investigated the value of using word2vec on radiology reports[60]. We then extracted the vocabulary and the associated embeddings from this pre-trained word2vec model. To maintain the same length for each document, we padded reports to the max length across index reports, 559 for early surgery and 573 for late surgery. We chose this approach to ensure the impression section was included as it summarizes the key findings from the image[118].

Machine Learning

Baseline Model

We used the LASSO logistic regression built using the python package *scikit-learn[119]*.

Because the data naturally has multicollinearity among different features (i.e. diagnosis codes, procedure codes, and prescriptions), this can lead to over- and underestimating relationships between the features and outcome. As a result, we chose LASSO since it performs feature selection through penalization to minimize these redundant features. To identify the optimal regularization parameter (lambda), we performed 5-fold cross validation. We chose the lambda value that led to the highest average F1-score across the folds to shrink the coefficients of the features. We chose the F1-score since it's a popular performance metric for imbalanced datasets, which takes into consideration how well the model can capture the positive group (i.e. minority group), but also the reliability of these positive predictions. Because LASSO's lambda value and its subsequent performance can be affected by how the data is split, we repeated the process of 5-fold cross validation 50 times, each process with a different split of the data into the folds, then chose the prevalent lambda value across repeats[120]. Additionally, to assess the value of each modality, we repeated this process for each data type by itself (i.e. codes, demographics, and textual) (Supplemental Materials).

Multimodal Deep Learning Model

The MDL architecture was built using the python package *PyTorch* and is composed of three entities: 1-layer Convolutional Neural Network (CNN), 1-layer Gated Recurrent Unit (GRU), and two 1-layer Fully-Connected (FC) (Figure 2-1) [121]. This architecture is based on work done by Zhang et al., in which they compared two different MDL architectures that differed in the use of either a CNN or Long Short-Term Memory (LSTM) for both sequences of clinical notes and structured data[31]. Since in our approach we do not have sequences of clinical notes, this comparison is out of scope. Additionally, we decided to use a GRU instead of an LSTM since the former is a simpler architecture, but can lead to similar performance[122,123].

We passed the featurized index reports and the pre-trained word2vec embeddings and vocabulary into a CNN, the featurized temporal data into a GRU, and then concatenated the output from these individual networks with the featurized demographics and then passed it to the FC to make predictions. We included a FC layer to convert the temporal input into embeddings before passing into the GRU as previous studies of this approach showed improvement in prediction performance[124–126]. The MDL was trained using the Adam optimizer with a weight decay and ReLU as the activation function. We used Cross Entropy Loss as the loss function and weighted the positive group and negative group inversely proportional to their prevalence to address the imbalance in our dataset[127]. We minimized subsets of weights from co-adapting (i.e. overfitting to the noise in the training data) by adding a dropout to the hidden layer of the FC to allow all weights to participate in the prediction task[128]. To optimize the hyperparameters (i.e. number of filters, learning rate, dropout rate, GRU hidden size, and weight decay), we 1) split the training data into 80% for training and 20% for validation, 2) used previous works as a starting point for values[31,129], then 3) grid searched to identify the combination of values that was associated to the lowest validation loss (Table 2-3). We trained our model for 30 epochs using a learning rate scheduler to decrease the learning rate value when the validation loss increased to avoid overfitting. Unlike the LASSO optimization, we did not perform 5-fold cross validation as it would have been computationally expensive. Additionally, we repeated this entire process for each individual network (i.e. 1-layer FC, 1-layer GRU with 1-layer FC, and 1-layer CNN with 1-layer FC) in the MDL architecture by itself and its associated data: demographics, temporal, and textual, respectively.

Evaluation

Classical

For each prediction task's dataset, we split it into a training (80%) and test set (20%). Hyperparameter values were optimized using the training set for both model types. The LASSO models were retrained on the full training set using optimized lambda values, while the DL models were retrained using the same training and validation set using the optimized hyperparameter values. The reason for this is because the learning rate scheduler for the DL models needs to monitor the validation loss, so that it can properly update the training process. We then evaluated the models' performance on the test set using the performance metrics: recall, specificity, balanced accuracy, precision, F1-score, area under the curve (AUC), and area under the precision-recall curve (AUPRC). While we calculated these different performance metrics, we prioritized AUC over AUPRC in the analysis and interpretation since 1) both are global metrics that assess overall performance across different thresholds and 2) AUC is a more popular metric in the ML field that is agnostic to positive label prevalence and as a result can be compared across studies. We estimated significant performance between models by performing a t-test on 1,000 bootstrapped test samples[30,105]. We used a Bonferroni correction to correct for multiple hypothesis testing when comparing MDL to the three individual networks (0.05/3 = 0.0167).

Generalizability

For generalizability, we split the data based on the healthcare system. We trained the models on Kaiser Permanente Northern California and tested on the remaining systems. We chose Kaiser Permanente Northern California as the training set, since it made up roughly 80%

of our entire dataset. For the test set, we excluded the Mayo Clinic since it contained a substantially smaller number of patients compared to Henry Ford and Group Health (Table 2-2). For each test system, we then evaluated the models' performance using the performance metrics: recall, specificity, balanced accuracy, precision, F1-score, AUC, and AUPRC. As before, while we calculated these different performance metrics, we prioritized AUC over AUPRC when interpreting results. We estimated significance performance between models by bootstrapping 1,000 samples for each healthcare system in the test set. For each pair of samples (i.e. one sample from each healthcare system in the test set), we calculated the different performance metrics for each sample then averaged. We performed a t-test for each performance metric using each model's resulting 1,000 average values. We used a Bonferroni correction to correct for multiple hypothesis testing when comparing MDL to the three individual networks (0.05/3 = 0.0167).

Results

Data Characteristics

For early surgery, we identified 8,387 patients with a prevalence of 2.4% for decompression surgery (Table 2-2). For late surgery, we identified 31,210 patients with a prevalence of 4.4% for decompression surgery. For early surgery, the average age was 57 years, while for late surgery it was 57.7 years. Both groups were balanced for gender with females representing 56.2% and 56.0%, respectively. The majority of patients from both prediction tasks were 1) white, 63.4% and 65.0%, respectively; and 2) from Kaiser Permanente Northern California, 84.3% and 86.1%, respectively. We found that the majority of early surgery patients had an MRI (69.3%), while late surgery patients had an X-Ray (61.5%).

Classical Performance Assessment

To assess the best performing model for each prediction task, we trained and tested each model, then calculated performance metrics on the test set, and then used a t-test to assess significant performance. For early surgery, we found that MDL had a significantly higher AUC (0.725) and AUPRC (0.061) compared to the baseline model (0.597, 0.047) (Table 2-4). For late surgery, we found that the baseline model had a significantly higher AUC (0.840) and AUPRC (0.266) compared to MDL (0.833, 0.241) (Table 2-4). For early surgery, we found that textual data (i.e. index image reports) was the contributing factor for both the baseline and MDL's performance, while for late surgery diagnosis and procedure codes and drug names in the form of aggregate binary 0/1 and temporal representation was the contributing factor for the baseline and MDL model's performance, respectively (Table 2-5, Supplement Table 2-1, 2-2).

Generalizability Performance Assessment

To assess the most generalizable model for each prediction task, we trained on Kaiser Permanente Northern California data and tested on the remaining healthcare systems. We excluded Mayo Clinic from the test set since it contained a substantially smaller set of patients compared to Group Health and Henry Ford (Table 2-2). MDL had a statistically higher AUC (0.763) compared to the baseline model (0.685), but the baseline model had a higher AUPRC (0.119) than MDL (0.116) for early surgery (Table 2-6). For late surgery, the MDL had a statistically higher AUC (0.760) compared to the baseline model (0.748), but the baseline model had a statistically higher AUPRC (0.177) than MDL (0.175) (Table 2-6). Similar to classical performance, we found that textual data contributed to MDL's generalizability performance for early surgery, while for late surgery, temporal data contributed to MDL's generalizability performance based on AUPRC (Table 2-7). This observation was also present in the baseline predictors, in which the early surgery baseline model's top predictors were composed of mainly textual features, while for late surgery procedure codes were the most important features for prediction (Supplement Table 2-3, 2-4).

Discussion

Early identification of LSS/LDH patients at high risk of eventual surgical decompression (i.e. failure of non-surgical treatments) could allow healthcare providers and patients to discuss the benefits and risks of pursuing surgery or seeking more non-surgical options using individualized information specific to each patient. In our study, we developed a MDL model that leveraged textual, temporal, and demographic information to predict decompression surgery for LSS/LDH patients and then evaluated classical and generalizability performance against a baseline model. For early surgery, MDL was preferred for both assessments. For late surgery, the baseline model was the preferred method for classical performance, while MDL was preferred for generalizability. However, while the difference in performance between MDL and LASSO for predicting late surgery was statistically significant, it was of small magnitude when compared to the difference between the two methods for predicting early surgery (Supplemental Table 2-7). Our study suggests depending on the prediction task, MDL and the baseline model, a conventional ML method can have similar performance. As a result, thorough assessment is needed to quantify the value of DL, a computationally expensive and time-consuming method.

For classical performance evaluation, the MDL models achieved a mean AUC of 0.725 for early surgery and 0.833 for late surgery. These results are similar to prior studies that used DL to predict aspects of lumbar surgeries[130,131]. André et al. assessed the feasibility of training a DL model on synthetic patients generated from EHR data to predict the positive and negative outcomes from decompression surgery resulting in an AUC of 0.78, while Azimi et al. investigated using DL to predict the outcome of surgery for 203 LDH patients resulting in an AUC of 0.82. The difference in our results can be attributed to 1) our larger dataset and 2) Andre et al. using synthetic patients, rather than real patients. As a result, these studies' results are limited in their generalizability, but they are important to acknowledge, so that we can understand our models' performances in the context of similar studies. Interestingly, a previous study by Keeney et al. used logistic regression to predict which Washington State workers with disability claims for back injuries would receive lumbar spine surgery (i.e. decompression, fusion, and/or both) or not, which resulted in an AUC of 0.93[132]. This AUC value outperformed our baseline and DL models for both early and late surgery. Keeney et al. found that the driving feature for this performance was if a patient's injury was first seen by a surgeon or not, and speculated that this may indicate that "who you see is what you get"[132]. If our dataset had this type of information, then our models' performance might have improved.

To the best of our knowledge, we are one of the few studies that evaluated the generalizability of our surgery prediction models across different healthcare systems. A recent study explains that external validation of predictive models in spine surgery are rare[133]. As stated earlier, MDL was the most generalizable model for both prediction tasks. Our rigorous evaluation shows DL-based models can learn a generalizable representation from the training data that can be applied to other healthcare systems' datasets. As Azad et al. stated, if we want to bring ML models into the clinical space, more external validation is needed to prove that performance is not specific to the internal datasets used for training and testing[133].

We observed an interesting trend with the MDL models in which for early surgery, textual data was the contributing data type, while for late surgery temporal data was the contributing data type for both classical and generalizability performance. This same observation was seen in the baseline models' top 10 predictors as well. The drivers for early surgery are likely impending neurologic deficits and anatomic findings related to those deficits; the former can't be known from the EHR elements present, but the latter may be reflected in the spine imaging reports (Supplemental Table 1). The drivers for late surgery are likely ongoing pain and disability, which are proxied by procedure codes reflecting spine-related procedures to relieve pain and indicators of more visits for clinical care (Supplemental Table 2-2).

There are several limitations to this study. First, expanding our hyperparameter value search space could have improved our DL-based models' performances, however we used prior studies to focus our grid search on the most important hyperparameters and their ranges of values. Second, the dataset contained only spine-related diagnosis and procedure codes and pain-relieving drugs, which may limit the generalizability of our results to only the lumbar spine domain. Third, we only used DL and logistic regression for our ML models and did not consider other methods. Including more conventional ML methods might have provided better performance than logistic regression and even DL. However, our objective was to specifically use DL to predict surgery and benchmark this costly method against the most popular and accessible method for researchers: logistic regression. Fourth, a bias in medicine is that sicker patients generally have more data points than healthier patients. We sought to address this by limiting the patients' data to the last 3 months and then binned into one month intervals, so that across the patients we 1) ensure that the time period is consistent and 2) minimize the variability in the amount of available data.

In summary, we built a MDL architecture to predict early and late decompression surgery for LSS/LDH patients. For each prediction task, we compared this architecture's performance within and across different healthcare systems against LASSO logistic regression, a conventional ML method. Through our rigorous testing, we've shown that depending on the prediction task, a conventional ML method can have similar performance to a DL model. This shows that thorough assessment is needed to validate the need for DL over using a conventional ML method. Finally, we believe our MDL architecture can be used as early screening tools to assist clinicians by allowing for early discussions with their patients about possible treatments depending on the prediction.

Aim 3

Introduction

Genetic testing is an essential tool to assist patients and clinicians to better understand the risk of hereditary disease. The cost of genetic testing has decreased and the number of genes routinely evaluated has increased in recent years, due to massively parallel sequencing methods and new discoveries[134,135]. Patients now have increased access to genetic information that can be important for their and their family's health.

Early genetic testing for germline risk variants can promote reproductive autonomy and lead to the recommendation of appropriate medical screening to mitigate risk of developing disease or provide early diagnosis at a more treatable stage. For example, the most common hereditary disease that elicits a genetic clinic visit and testing in adults is cancer, specifically colorectal cancer (CRC), breast cancer (BC), and ovarian cancer (OC). Approximately 5% of CRC and BC, and 10-20% of OC, is due to high penetrance Mendelian conditions[136–138].

CRC accounts for 9.5% of all new cases of cancer[139]. BC is the second leading cause of cancer death in women; 3% of women in the U.S. will die of BC[140]. OC affects 1-2% of women, most of whom will die from it. To mitigate the cancer-related death rate, early detection of Mendelian (germline) cancer predisposition is of grave importance. If CRC-associated pathogenic variants are found, colonoscopy with polypectomy can prevent CRC and prophylactic bilateral salpingo-oophorectomy surgery reduces OC risk and are consensus recommendations[33,34]. Similarly, for BC/OC associated genes, prophylactic mastectomy reduces risk of BC[141]. Thus, early genetic testing is necessary to reduce risk of morbidity and mortality for patients.

While genetic testing is important for patients, these test results may also be important for their biological relatives. A patient's positive test result allows for inexpensive and often free direct testing of at-risk family members for that same pathogenic variant. A positive or negative test in a family member is likely to affect their clinical care. For positive test results, relatives' treatment plans may change to reduce disease risk, while for negative test results, relatives may not be at increased risk and additional testing may not be necessary[35]. However, sharing of genetic results between patients and their biological relatives is infrequent[35–42]. The two most frequently reported communication barriers for sharing are 1) patients have difficulty in clearly communicating the results and meaning, and/or 2) biological relatives have difficulty in interpreting the result[35–37,41]. Nieuwenhoff et al. found that patients had limited knowledge of their test results and this influenced whether or not they would share[36]. For example, terms in the test result like "hereditary" implied danger and motivated patients to share, while terms like 'sensitivity', 'tendency', and 'it runs in the family' made patients perceive the results as normal and did not share[36]. Additionally, if patients shared then there was a risk of arousing fear in their relatives, as they couldn't clearly explain the benefit and threat reduction from getting their own genetic test results[36]. Another recent study reported when patients shared their test results with their biological relatives, over 20% didn't fully understand the results and were unsure if they were at risk for cancer[35]. This 20% was mainly for non-informative test results, indeterminate results or variants of uncertain significance[35]. As a result, patients' explanations had a combination of filtering information and lack of understanding[35].

Family communication tools may improve the dissemination of genetic results among family members. With increasing access to mobile phones and devices, mobile technology, such as apps, have become popular methods to share information[43,142,143]. Studies investigated the value of this technology specifically in families and found that it was a valuable tool for parents and their children to communicate sensitive topics that they didn't feel comfortable discussing in-person[144]. Additionally, this technology facilitates family members being in contact when they are not geographically close[145]. Similarly, in the healthcare space, mobile technology provides a means of communication to improve health behavior for patients[43]. However, mobile health apps' patient privacy and interpretation of results are questionable. A recent study found that 81% of diabetes apps do not have privacy policies and would share sensitive patient information to third parties without the patient's permission[44]. Additionally, another study found that 20% (7/35) of health apps would transmit identifiable information over the Internet without encryption[45]. Finally, there are apps that can analyze genetic test results and provide risk scores, however it's important to encourage patients to seek their healthcare provider's advice on these results[42]. We believe there is an opportunity to leverage mobile technology to increase communication genetic test results between patients and their family members, while protecting their privacy and encouraging the value of healthcare providers' expertise.

We built *ShareDNA*, a free secure smartphone app, to 1) lower the barrier to sharing genetic test results with family members by sharing test results with anyone contactable by text or email, 2) provide links to educational material and text to explain how to understand the results and proper next steps for recipients, and 3) increase security of patient data by allowing for encrypted transmission and minimizing the amount of data needed to register for the app. Here we describe the development of the *ShareDNA* app, and the results of user testing to inform usability and acceptance.

Implementation

Application

Overview

ShareDNA is a smartphone application that allows users to share DNA results with family members in a secure way (Figure 3-1). The application is divided into two parts: the app and server (Figure 3-2).

App

The app faces the user and allows them to upload their documents either by selecting a file from their smartphone or using their smartphone's camera to take a picture of their result. The code itself is encrypted on the user's smartphone. The app is built using HyperText Markup Language (HTML), JavaScript (JS), Cascading Style Sheets (CSS). Users are required to enter their password every time a file is uploaded or downloaded. This ensures only the user and their recipients can access the files. Once uploaded, user’s files are encrypted using ‘AES-256-CBC’ encryption that cannot be decrypted unless a user enters their password again. All communication from the app to the server is encrypted using Secure Socket Layer (SSL) with a 256-bit Certificate. The app interface was designed using Cordova and is available in the Apple App and Google Play store. With this approach, we came across an obstacle in which Apple app guidelines frequently changed which required refactoring the app, especially before the first production build.

Server

User’s encrypted information (i.e. email, password, and test result file) is stored on a Security-Enhanced Linux server with an encrypted file system hosted by University of Washington. The application programming interface (API) is a web application written in Hypertext Preprocessing (PHP) 7+ running on the Apache web server with a MariaDB database for data storage.

Participants

Participant recruitment was conducted in two phases. We first sent invitations to 49 participants who were enrolled in the Electronic Medical Records and Genomics (eMERGE) network clinical study at Kaiser Permanente Washington who had received positive (pathogenic or likely pathogenic) genetic test results. The eMERGE network is a consortium that develops methods to use electronic health record information for genomic research[146]. In the second phase, we sent a batch of 100 invitations in the mail to eMERGE study participants who had received negative test results. In total, we recruited 14 participants, however one dropped out early in the study for unknown reasons. As a result, we only considered the 13 active participants for this study. For each of the 13 participants, we performed an app testing session. Institute Review Board approval was provided by the Kaiser Permanente Health Research Institute Human Subjects Board. Each participant received a $50 incentive.

User-Testing

To evaluate the usability and acceptance of *ShareDNA*, we used the Technology Acceptance Model (TAM) as a framework[147]. We assessed the perceived ease of use (usability) of the tool through observing the participants walk through the procedures and functions for sending relatives their genetic test results and recorded whether the participant successfully accomplished a list of tasks. Testing of the app was done using an iPad. The app was already installed on the device, as well as a picture of a blank genetic test result and contact information for two fabricated relatives. Each participant was also provided a unique email and password to create an account for the app, along with the contact information for two hypothetical at-risk relatives and a hard copy of blank genetic test results. After being consented, we provided a generic scenario with the task to deliver the test results to the two relatives using the ShareDNA app. We had users fill out a usability testing document that indicated how to perform tasks in the app and allowed for user feedback on what could be improved. In addition, we used the 16-item version 3 of the Post Study System Usability Questionnaire (PSSUQ), which is a validated instrument designed for usability evaluations, to assess attitudes towards the app after use[148]. The scale is out of five, with one indicating “strongly disagree” and five indicating “strongly agree”. Finally, we asked participants to vocalize their thoughts and impressions while interacting with the app and then recorded their responses. Testing was done in-person.

Results

Patient Demographics

Our thirteen participants consisted of nine males and four females with an average age of 67.5, minimum age of 60, maximum age of 74, and a standard deviation of 4.8. Our participants were primarily white (10 out of 13).

User-Testing of App

We found on average, the PSSUQ questions with scores above four indicated that users felt comfortable with using this app and could easily learn each app function, however, the lowest scoring question indicated that when users came across a problem, our error messages were not informative enough to help (Table 2-1). These results indicate that reformatting our error messages is needed to better assist users that may have some difficulties with our app. Additionally, participants vocalized their thoughts about sharing via email. Participants expressed a natural inclination to email, as one participant explained: “because it's just what they've done all their lives.”

Issues with App

Users had a number of concerns along with recommendations based on the vocalized impressions of the app (Table 3-2). We found three main themes: 1) certain aspects of our user interface were not intuitive, such as how to select multiple contacts to send a result to, 2) there was a lack of understanding of our security measures, which is why users were confused as to why they needed to enter their password multiple times, and 3) users were confused with modern icons for buttons, such as *Share* and *Downloading*. The second theme is of particular importance, because there seems to be a lack of understanding of the implications of an information leak.

Discussion

User-Interface

Creating an Account and Educational Material

Once the app is opened, *ShareDNA* describes the purpose of the app with a mission statement along with visual slides (Figure 3-1). The user needs to create an account by providing an email address and password to log in to the app. Both the password and email are one-way encrypted, so that they are not stored on the server. Additionally, if the user requires further assistance on how to use the app, they can tap the *Need Help?* Icon on the bottom left of the screen to contact the *ShareDNA* team. Finally, *ShareDNA* provides links to websites that provide educational material. Two of the provided links direct users to our local medical genetics clinic at the University of Washington and to genetic counseling resources across the United States (the National Society of Genetic Counselors 'find a provider' page), particularly for individuals with questions and/or who have a positive result and need follow up care. Additionally, we provide links to two websites with reliable, general genetic condition (Medline) and hereditary cancer specific (National Cancer Institute) information written for the general population for users who may want to research a given condition on their own. The users tap the *Learn More* icon on the bottom right of the screen to access these links.

Upload a Genetic Test Result

When a user uploads a file either from their smartphone local storage or taking a picture with their smartphone camera, the server uses a randomly generated key to encrypt the file and save to the filesystem; the server then erases the key making the file only accessible for sharing and downloading if the user enters their password again.

Sharing a File through Text and/or Email

The user can view a list of their uploaded files and select to share a file. The user must enter their password again which allows the server to create message keys that can be later used by the server to decrypt the files sent to the recipient. Once the device receives the temporary message key for the file, they can select to send a message to a single contact or multiple contacts through the smartphone's native email or text messaging application. Once selected, they can send a message containing a link to the *ShareDNA* web application allowing the recipients to register and access the file. These message keys are only available for 24 hours and can only be used once per recipient; this is to avoid leaving them hanging in emails and texts for the wrong people to read and prevent brute force attacks on our server.

Recipient Viewing the File

After sharing a file, the recipient will receive a link either through email or text. The link takes the recipient to the ShareDNA website, where the recipient will need to create an account and log in to access the file. Once logged in, the recipient can tap the "Testing" button next to the file to learn about the next steps after viewing the file. These steps include 1) how to interpret the results and 2) taking the file to their clinician to discuss if genetic counseling is necessary or not. In order to download a file to their device, recipients must enter their password again. The password is sent to the server to decrypt the file and creates a new temporary file encrypted with a new key that is sent to the device and not stored on the server. The user then uses this key along with their logged in API key to decrypt and download the file to their device. Once the file is on their device, it is stored to a location of their choice in an unencrypted state.

New Features to Add

In a future implementation, we will need to address our app testers' concerns by 1) allowing future users to guide us in implementing the logical steps needed to execute our functions, 2) include documentation in the app that explains why security is needed for clinical data, and 3) using text rather than icons to describe our buttons.

Comparison to Other Existing Software

ShareDNA is similar to the app *FamGenix[42]*. Both apps 1) allow patients to share their genetic information with anyone of their choosing through text or email and 2) store their information on a secure Health Insurance Portability and Accountability Act (HIPAA) compliant server with encryption at rest and in transit. The difference is that *ShareDNA* is a free service focused on sharing genetic test results, while *FamGenix* is a paid sharing service with data analytics. *FamGenix* employs genetic risk algorithms to autogenerate pedigrees and calculate the hereditary cancer risk for a patient; both can be shared with patients' family members. While useful, we believe sharing algorithm-derived risk scores could lead to 1) misinterpretation as these risk scores should only be considered as aids for diagnosis and/or 2) incorrect results. We believe our approach of encouraging patients' family members to share their information with their healthcare provider is a safer option as it 1) minimizes the possibility of misunderstanding and 2) emphasizes healthcare providers' valuable expertise and experience.

Conclusion

ShareDNA is a free secure smartphone app that allows patients to share their genetic test results with others, with emphasis on their family members who may benefit from this information. The main benefit of the app is to provide a secure environment for sharing the genetic test report by requiring minimal user information and encrypting the storage and transportation of data. Our app addresses patients' difficulty in communicating and their relatives' difficulty in interpretation by providing links to educational websites to learn more about genetic testing and text to explain how to interpret these results and next steps for their relatives to get their own testing if needed. Our user-testing indicated that participants felt comfortable with our app, however improvements were suggested to better support potential users, specifically understanding the importance of our security measures (i.e. entering password twice). Our next immediate step will be to implement our participants' recommendations (Table 3-2). A future step would be to perform another usability test to further explore the TAM framework, specifically usefulness and intention to use, by expanding our questions for participants to include ones that directly ask about the usefulness of the app and its educational material and the intention to use the app. A limitation of our study is the sample size. We were able to recruit only 13 participants. A larger sample size may have provided more feedback on how to improve our app. Another limitation is our results are limited to an age group favoring elderly individuals (atleast 60 years old). Another limitation is that our participants came from the eMERGE consortium only. These individuals are familiar with genetic testing, so they may not represent the general population. As a result, our findings are limited in their generalizability. To address these three limitations, a future step would be to perform another usability test, as stated before, but with a larger cohort of participants that come from both the eMERGE consortium and general population with a wider age range. A final limitation is that participants had some familiarity with using a smartphone. While a limitation, this is a necessary prerequisite to use our app. We believe *ShareDNA* will become a useful tool to promote timely communication of genetic risk information to ensure family members are able to make informed decisions about whether or not to access genetic testing. Important features enabled by *ShareDNA*, including file sharing, data encryption, and links to resources, could reduce the barriers to successful cascade screening programs.

Availability and Requirements

Project name: ShareDNA

Project home page: http://sharedna.org/

Project source code page: https://github.com/uwrit/AppShareDNA

Operating system(s): iOS and Android

Programming language: JS, HTML, CSS, PHP

Other requirements: None

License: MIT

Any restrictions to use by non-academics: License required

Conclusion

To conclude this dissertation, I summarize the contributions in fulfillment of the dissertation aims by reviewing the advances in knowledge and acknowledging the limitations and opportunities for future work.

Summary of Contributions

The increase in the volume and complexity of EHR data has given rise to the need for advanced methods to distill these data into information and knowledge that can be acted upon. One of the most common methods is to apply ML methods for clinical prediction tasks, however the generalizability of these predictions are questionable as there is a lack of rigorous testing across independent datasets. Additionally, the ability to send this and other types of sensitive clinical information to patients is limited.

In this work, I address the issue of generalizability in the realm of structured and unstructured EHR data. The first aim assesses the generalizability of information extraction from textual data to build cohorts independent of the healthcare system, while the second aim assesses the generalizability of a ML model that can harness both structured and unstructured data for a prediction task. Finally, I address the issue of sending sensitive clinical information to patients by building a smartphone app that focuses on the security of patients and their information.

Aim 1

In Aim 1, we collected 871 reports from the LIRE study. These reports were labelled for 26 different imaging findings related to LBP. We built a NLP-based ML pipeline to classify these reports for these different findings using four different NLP techniques: rules, n-grams, controlled vocabulary, and document embeddings to represent the text. We performed two types of assessments: classical and generalizability assessment. For classical assessment, we split our dataset into 80% for training and 20% for testing for evaluating each finding-specific ML model (i.e. elastic-net logistic regression). Within the training set, ten-fold cross validation was used to adjust the value of our regularization parameter (lambda) to perform feature selection. We repeated this process 25 times with each independent repeat using a different random train/test split of the data, so that we could estimate 95% confidence intervals. For each finding, a t-test was used to assess significant performance comparing the 25 repeats of the best representation to the next best representation. We used Bonferroni correction to correct for multiple hypothesis testing. For generalizability assessment, we trained our model on reports from three systems and evaluated on the fourth, iteratively, for each finding. For each finding, we calculated the mean and standard deviation of the AUC across the four systems.

For classical assessment, we found that n-grams was the best performing method based on AUC. Interestingly, for generalizability assessment, n-grams had the worst performance and document embeddings had the best performance based on standard deviation. These results indicate that if classifier development and deployment occur at the same system, then n-grams may be preferable. However, for deployment at multiple systems outside of the system of development, one should consider n-grams with the caveat that it's consistency can vary across systems, while document embeddings pre-trained on study-specific data or a publicly available dataset had the most consistent performance.

In support of Aim 1, my contributions in this study are to assess the generalizability of NLP-based feature extraction methods for use in the clinical space. My contributions are:

1. Classical assessment alone is not enough to fully characterize these methods for prediction tasks. Generalizability across healthcare systems helps to rigorously assess the consistency of performance.
2. Reinforcement of the growing evidence that relatively simplistic methods such as n-grams can outperform sophisticated methods like document embeddings.

Aim 2

In Aim 2, we built a MDL architecture to predict early and late decompression surgery for LSS/LDH patients. For each prediction task, we compared this architecture's performance within and across different healthcare systems against LASSO logistic regression, a conventional ML method. We performed two types of assessments: classical and generalizability. For classical assessment, we split each prediction task's dataset into 80% for training and 20% for testing. For generalizability, we split the data based on the healthcare system. We trained the models on Kaiser Permanente Northern California and tested on the remaining systems. We chose Kaiser Permanente Northern California as the training set, since it made up roughly 80% of our entire dataset. For each evaluation, we calculated the test set's area under the curve (AUC) and area under the precision-recall curve (AUPRC). For generalizability, we then calculated these metrics for each healthcare system's dataset in the test set and then averaged. To assess significant performance between our models, we repeated this evaluation for bootstrapped 1,000 samples from the test set and then performed a t-test.

For early surgery, MDL was preferred for both assessments. For late surgery, the baseline model was the preferred method for classical performance, while MDL was preferred for generalizability. However, while the difference in performance between MDL and LASSO for predicting late surgery was statistically significant, it was of small magnitude when compared to the difference between the two methods for predicting early surgery.

In support of Aim 2, my contributions are:

1. Finding that depending on the prediction task, a computationally expensive deep learning-based model is not always the best method as conventional ML methods can perform as well.

2. Emphasizing the need for generalizability assessments of ML models as we've shown that deep learning-based methods based on statistical significance are the preferred method to be applied to other healthcare systems' datasets.

Aim 3

In Aim 3, we built a smartphone app, *ShareDNA*, to securely share patients' genetic test results with their family members. The app allows users to upload their genetic test results as a file or picture to our secured UW server. The user then indicates which of their contacts they'd like to share the test result with, which then prompts the app to provide a link to those recipients either through text or email to download the test result. Additionally, we provide instructions on what to do next when the test result is received and educational material on how to learn more about genetic testing. To assess usability, we recruited 13 participants to test the app. Participants were asked to send a blank test result to two fake recipients and during the process fill out a usability report. Once the test was complete, participants were then asked to fill out a PSSUQ. The PSSUQ scale is out of five, with one indicating "strongly disagree" and five indicating "strongly agree". Finally, we asked participants to vocalize their thoughts and impressions while interacting with the app and then recorded their responses.

Based on PSSUQ scores, we found that overall participants felt comfortable with using this app and could easily learn each app function, but when faced with a problem our error messages were not useful. Additionally, we learned that participants favored sending their information over email instead of text, as one participant explained: "because it's just what they've done all their lives." Finally, our participants vocalized their issues with the app. One particular issue stood out, in which participants did not understand the security measures in the app, specifically the need to enter a password multiple times.

In support of Aim 3, my contributions are:

1. Successfully building a vehicle to allow patients to securely share their sensitive clinical information with anyone of their choosing.
2. Discovering that when it comes to sharing information, patients may prefer a specific mode of transport that should be taken into consideration for future vehicles.

Limitations and Future Works

Aim 1

In our study, we had several limitations. First, our pipeline required binary annotations for findings, however the presence of findings may be uncertain as radiology reports can have terms such as "suggesting" and "not definite". We minimized this uncertainty by coding these and other similar terms as indicating the presence of a finding. Second, a larger training and testing set could have led to less variable performance across our NLP methods. Third, we evaluated the algorithms but not the entire pipeline involving the querying and transfer of data; there may be discrepancies in our performance estimates when compared to those at actual deployment. Fourth, we could not assess our rules' generalizability, since the search terms were developed from reports from all four systems. Finally, in the case of document embeddings, because of our limited computational resources, we had to sequentially adjust hyperparameter values in the pre-training step, rather than conducting a grid search. With a more extensive hyperparameter search, we may have been able to improve performance.

Future efforts could include a larger dataset that encompasses more reports from each healthcare system, so that we can account for more distinct ways clinicians document spinal images to further generalize our results. In addition, our team of clinical experts could be expanded so we can label more different findings beyond our initial 26. Finally, we could include more modern NLP-based methods such as Bidirectional Encoder Representations from Transformers (BERT) in our analysis.

Aim 2

There are several limitations to this study. First, expanding our hyperparameter value search space could have improved our DL-based models' performances, however we used prior studies to focus our grid search on the most important hyperparameters and their ranges of values. Second, the dataset contained only spine-related diagnosis and procedure codes and pain-relieving drugs, which may limit the generalizability of our results to only the lumbar spine domain. Third, we only used DL and logistic regression for our ML models and did not consider other methods. Including more conventional ML methods might have provided better performance than logistic regression and even DL. However, our objective was to specifically use DL to predict surgery and benchmark this costly method against the most popular and accessible method for researchers: logistic regression. Fourth, a bias in medicine is that sicker patients generally have more data points than healthier patients. We sought to address this by limiting the patients' data to the last 3 months and then binned into one month intervals, so that across the patients we 1) ensure that the time period is consistent and 2) minimize the variability in the amount of available data.

In a future effort, we will repeat the process of predicting late surgery, only this time we'd limit identification of LSS/LDH patients to the 1 year prior to LIRE enrollment and this time period would be the data our models would use for prediction. With this approach, we can more easily compare performance between the two prediction tasks.

Aim 3

A limitation of our study is the sample size. We were able to recruit only 13 participants. A larger sample size may have provided more feedback on how to improve our app. Another limitation is our results are limited to an age group favoring elderly individuals (atleast 60 years old). Another limitation is that our participants came from the eMERGE consortium only. These individuals are familiar with genetic testing, so they may not represent the general population. As a result, our findings are limited in their generalizability.

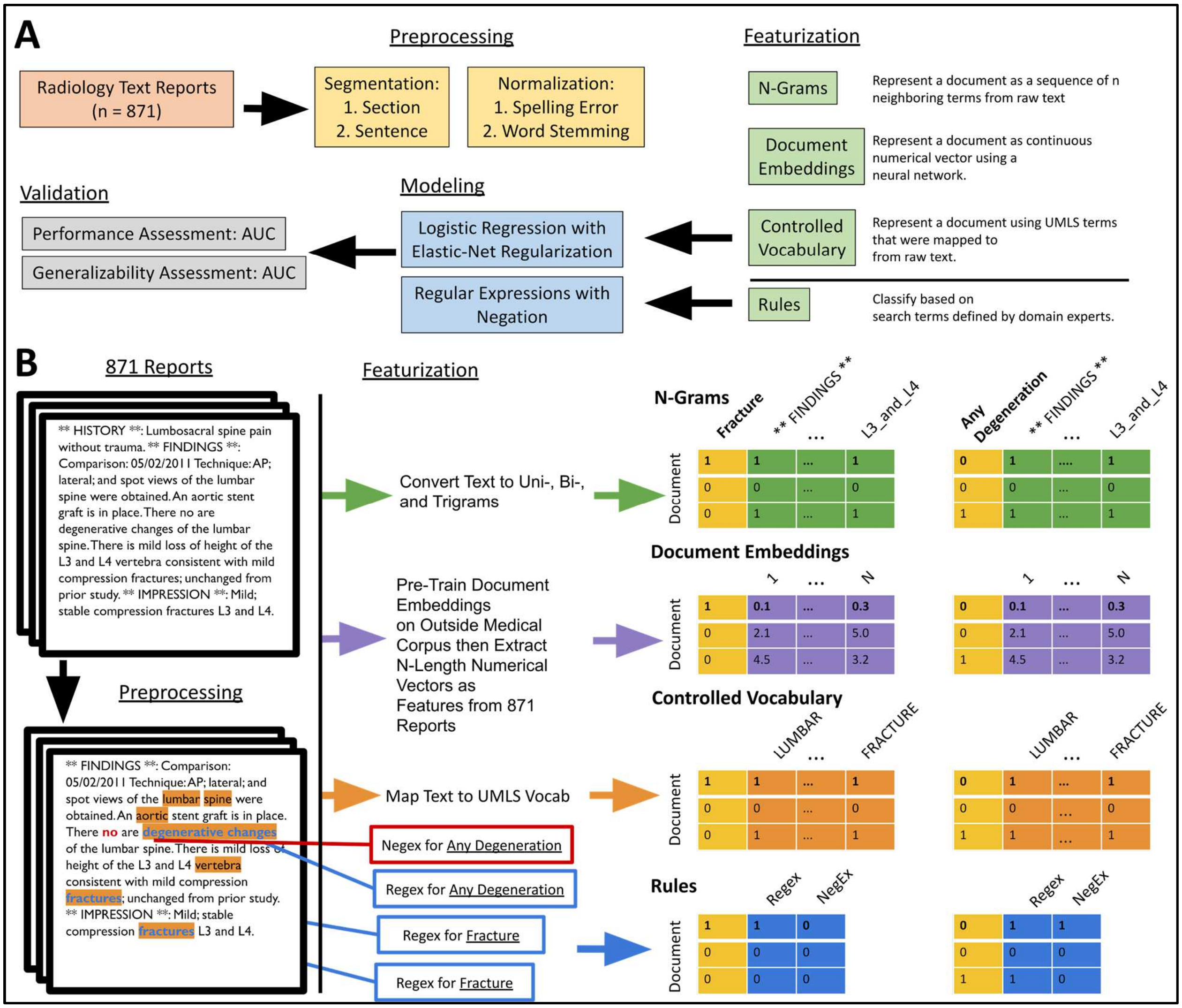
In a future effort, we would perform another usability test, but with a larger cohort of participants that come from both the eMERGE consortium and general population with a wider age range, to further explore the TAM framework, specifically usefulness and intention to use, by expanding our questions for participants to include ones that directly ask about the usefulness of the app and its educational material and the intention to use the app.

Conclusion Overview

This work serves to advance evaluation of prediction pipelines to transform the data in the EHR into knowledge that can then be given to patients and their family members to inform their clinical decisions. In Aim 1, I evaluated the performance and generalizability of different NLP-based feature extraction methods coupled with an ML model to build patient cohorts by classifying reports for different imaging findings. In Aim 2, I built on Aim 1 by then leveraging both free-text and tabular data to predict a clinical outcome and then rigorously assessing the performance within and across different healthcare systems. Finally, in Aim 3, I developed a smartphone app to securely share patients' clinical information with their family members. These three aims serve to bring to vision an evidence-based healthcare system that transforms data into knowledge that can help clinicians and patients' decision making. The contributions of this work will aid in convincing researchers that further evaluation of ML methods needs to be considered before deploying in the clinic and how to share clinical information with patients to keep them and their family involved in the decision making.

Figures/Tables

Figures



A
Preprocessing
Featurization
Radiology Text Reports (n = 871)
Segmentation: 1. Section 2. Sentence
Normalization: 1. Spelling Error 2. Word Stemming
N-Grams
Represent a document as a sequence of n neighboring terms from raw text
Document Embeddings
Represent a document as continuous numerical vector using a neural network.
Validation
Modeling
Performance Assessment: AUC
Generalizability Assessment: AUC
Logistic Regression with Elastic-Net Regularization
Regular Expressions with Negation
Controlled Vocabulary
Represent a document using UMLS terms that were mapped to from raw text.
Rules
Classify based on search terms defined by domain experts.
B
871 Reports
Featurization
Preprocessing
Convert Text to Uni-, Bi-, and Trigrams
Pre-Train Document Embeddings on Outside Medical Corpus then Extract N-Length Numerical Vectors as Features from 871 Reports
Map Text to UMLS Vocab
Negex for Any Degeneration
Regex for Any Degeneration
Regex for Fracture
Regex for Fracture
N-Grams
Document Embeddings
Controlled Vocabulary
Rules
Document

** HISTORY **: Lumbosacral spine pain without trauma. ** FINDINGS **: Comparison: 05/02/2011 Technique: AP; lateral; and spot views of the lumbar spine were obtained. An aortic stent graft is in place. There no are degenerative changes of the lumbar spine. There is mild loss of height of the L3 and L4 vertebra consistent with mild compression fractures; unchanged from prior study. ** IMPRESSION **: Mild; stable compression fractures L3 and L4.

** FINDINGS **: Comparison: 05/02/2011 Technique: AP; lateral; and spot views of the lumbar spine were obtained. An aortic stent graft is in place. There no are degenerative changes of the lumbar spine. There is mild loss of height of the L3 and L4 vertebra consistent with mild compression fractures; unchanged from prior study. ** IMPRESSION **: Mild; stable compression fractures L3 and L4.

Fracture	** FINDINGS **	...	L3_and_L4
1	1	...	1
0	0	...	0
0	1	...	1

Any Degeneration	** FINDINGS **	...	L3_and_L4
0	1	....	1
0	0	...	0
1	1	...	1

	1	...	N
1	0.1	...	0.3
0	2.1	...	5.0
0	4.5	...	3.2

	1	...	N
0	0.1	...	0.3
0	2.1	...	5.0
1	4.5	...	3.2

	LUMBAR	...	FRACTURE
1	1	...	1
0	0	...	0
0	1	...	1

	LUMBAR	...	FRACTURE
0	1	...	1
0	0	...	0
1	1	...	1

	Regex	NegEx
1	1	0
0	0	0
0	0	0

	Regex	NegEx
0	1	1
0	0	0
1	1	0

Figure 1-1. **Overview of the Pipeline**. (A) This visualization shows the different steps of our pipeline

where we collect 871 radiology reports from our four systems, perform preprocessing to clean the text data, perform feature extraction using our four different methods. We load the *n-grams*, *controlled vocabulary*, and *document embeddings* feature matrices into a logistic regression model to predict the presence of these findings. For *rules*, we instead use a rule-based model that classifies a report as "positive" if atleast one mention was non-negated and "negative" if there was no mention or all mentions of the finding were negated. We perform two types of assessments: generalizability and performance based on AUC. (B) A visual representation using the four different NLP methods to featurize the text for two example findings: *fracture* and *any degeneration*. The resulting finding-specific feature matrices are then used for the machine learning model, which uses the first column as the labels and remaining columns as features to predict the presence of these findings. UMLS = Unified Medical Language System, AUC = Area Under the Curve.

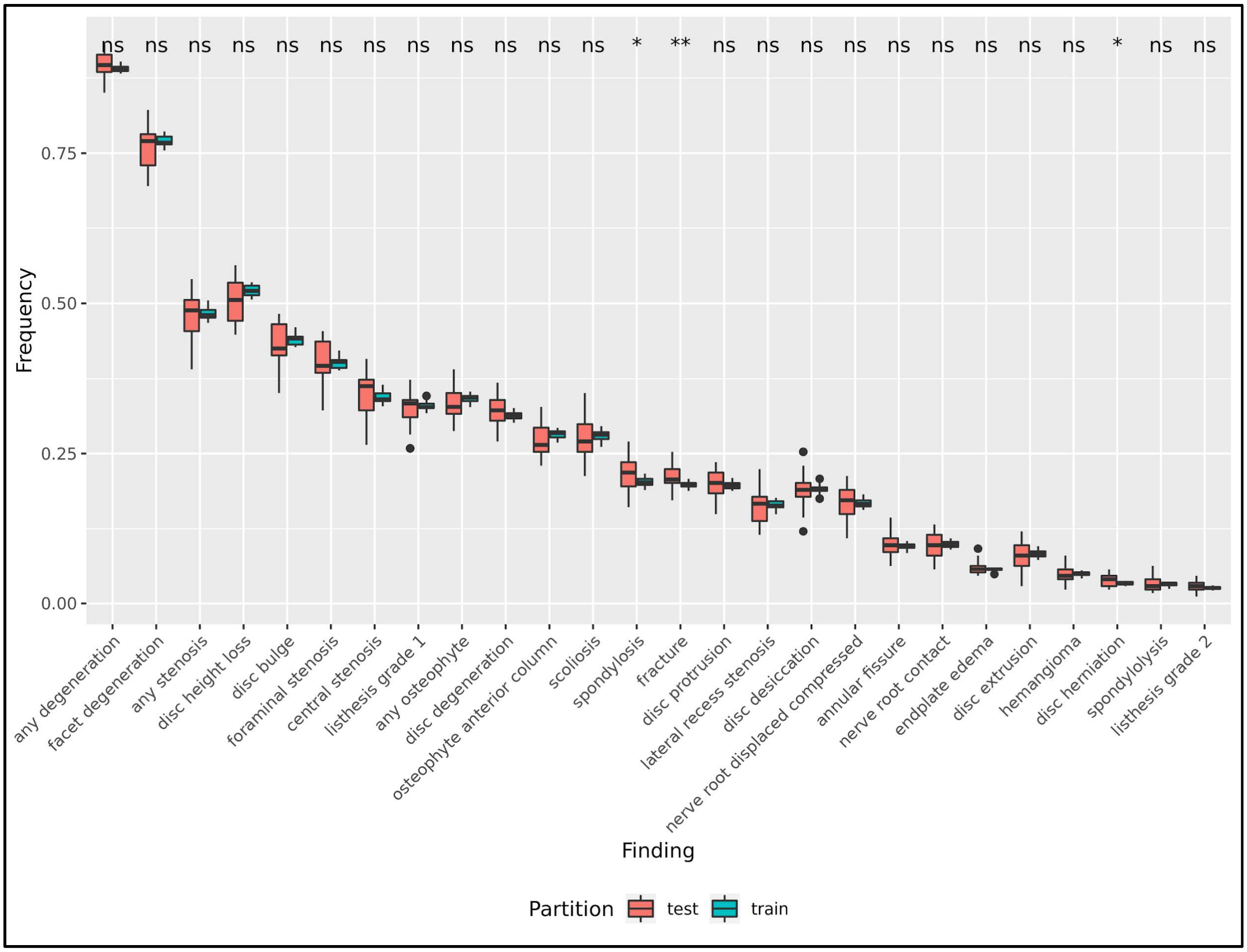


Figure 1-2. **Comparison of the Finding Label Prevalence Between the Training and Test Set**. We compared the finding label prevalence between the train and test sets across the 25 repeats. To assess a significant difference, we performed a t-test between the two sets for each finding. An asterisk indicates a significant difference, while “ns” indicates no significant difference.

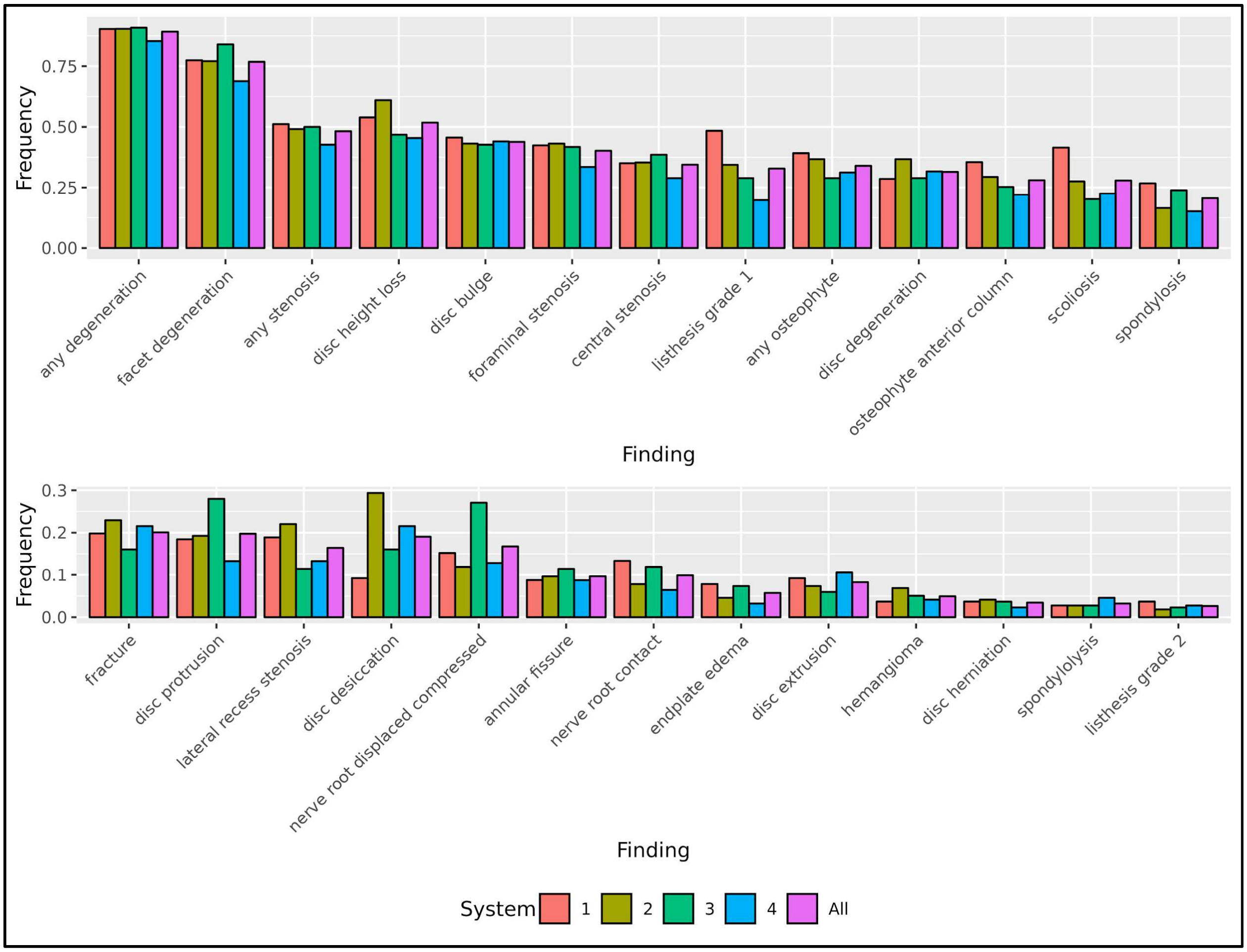


Figure 1-3. **Comparison of the Finding Label Prevalence Across the Healthcare Systems.** We compared the finding label prevalence across the four healthcare systems. 1 = Kaiser Permanente of Washington, 2 = Kaiser Permanente of Northern California, 3 = Henry Ford Health System, 4 = Mayo Clinic Health System, All = all four systems.

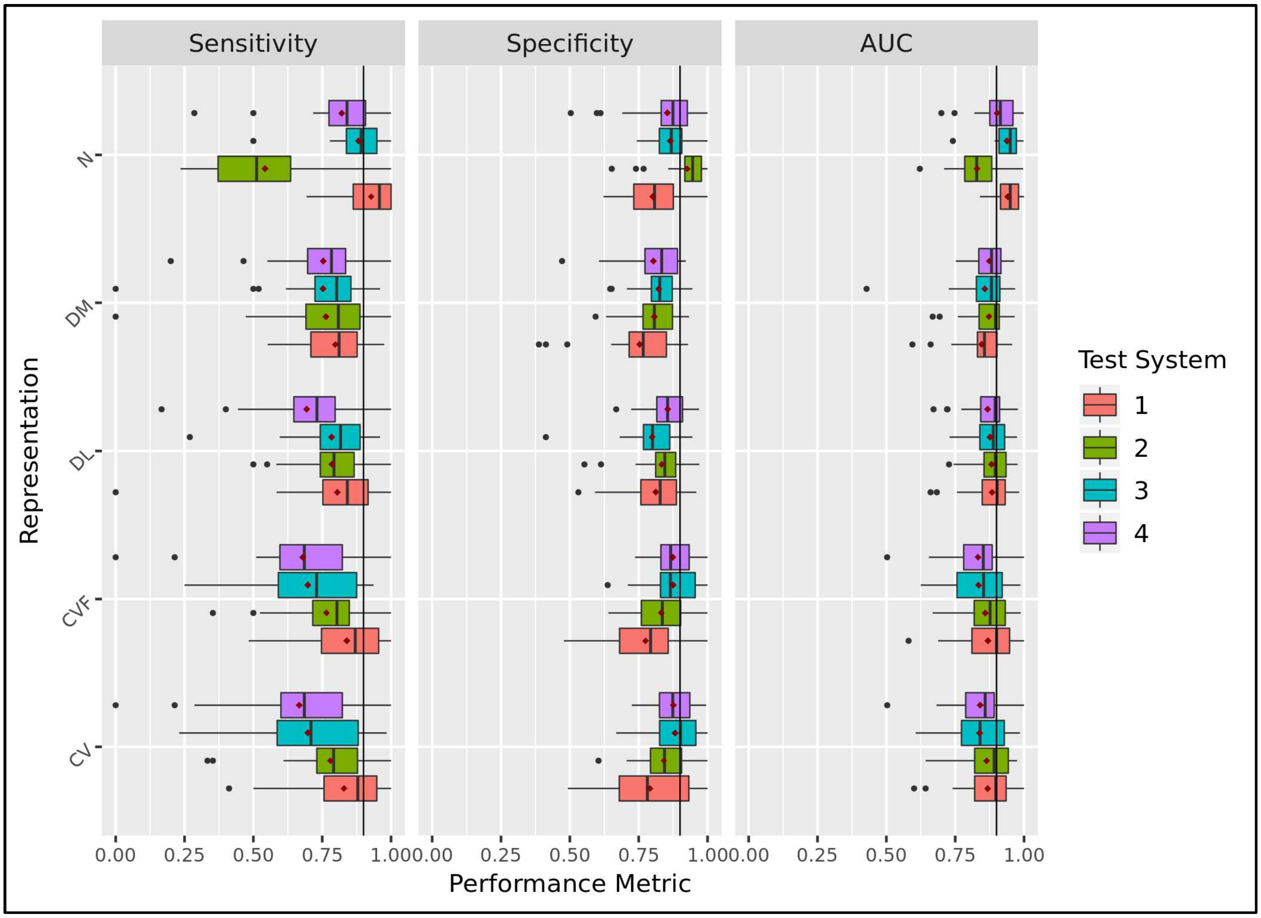


Figure 1-4. **Assessing Generalizability of Individual Representations**. We compared the generalizability of each of our representations and assessed performance using sensitivity, specificity, and AUC. For each representation, we plotted a boxplot to represent the distribution of the 26 findings for each test performance metric across healthcare systems. N = N-grams, DM = Document MIMIC, DL = Document LIRE, CVF = Controlled Vocabulary Filter Only, and CV = Controlled Vocabulary. 1 = Kaiser Permanente of Washington, 2 = Kaiser Permanente of Northern California, 3 = Henry Ford Health System, and 4 = Mayo Clinic Health System.

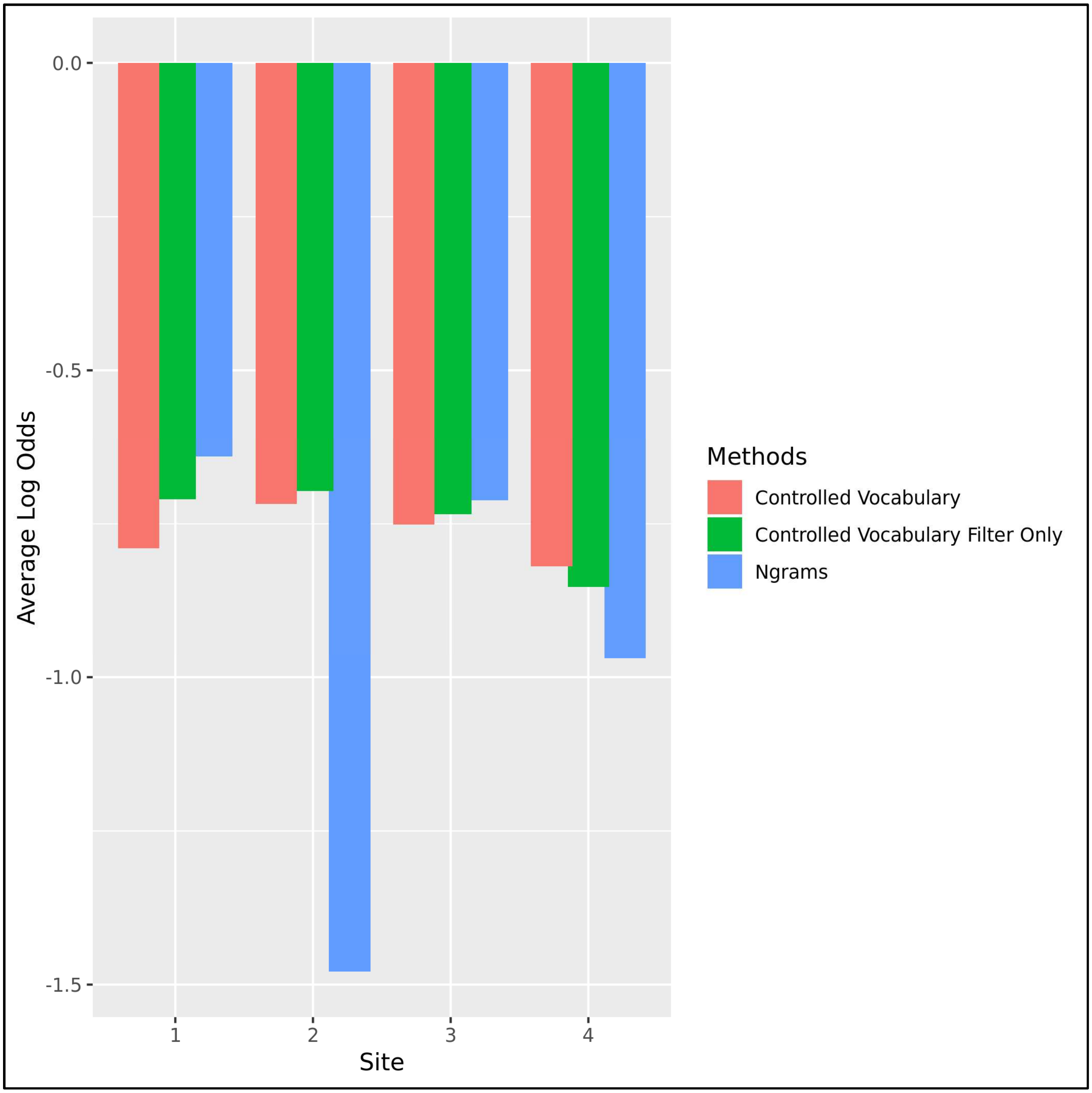


Figure 1-5. **Assessment of Textual Differences Between Systems**. Bar graph shows the average log odds ratio for *controlled vocabulary, controlled vocabulary filter only*, and *ngrams* for each system. For each representation, we calculated the frequency of a feature in three systems and frequency for the fourth system from the feature matrix. We calculated the log odds ratio by dividing the frequency of the fourth system by the three systems and took the log; we then averaged across all features shared between the three systems and fourth system. We repeated this process for each system.

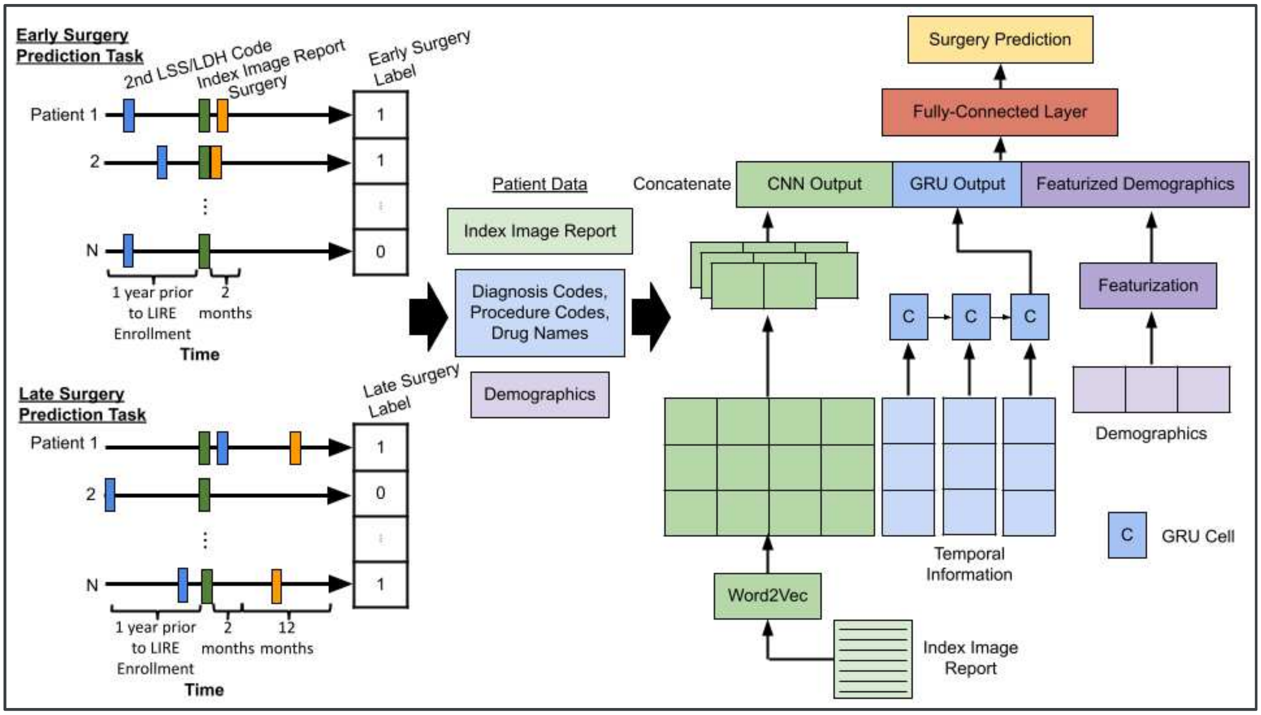
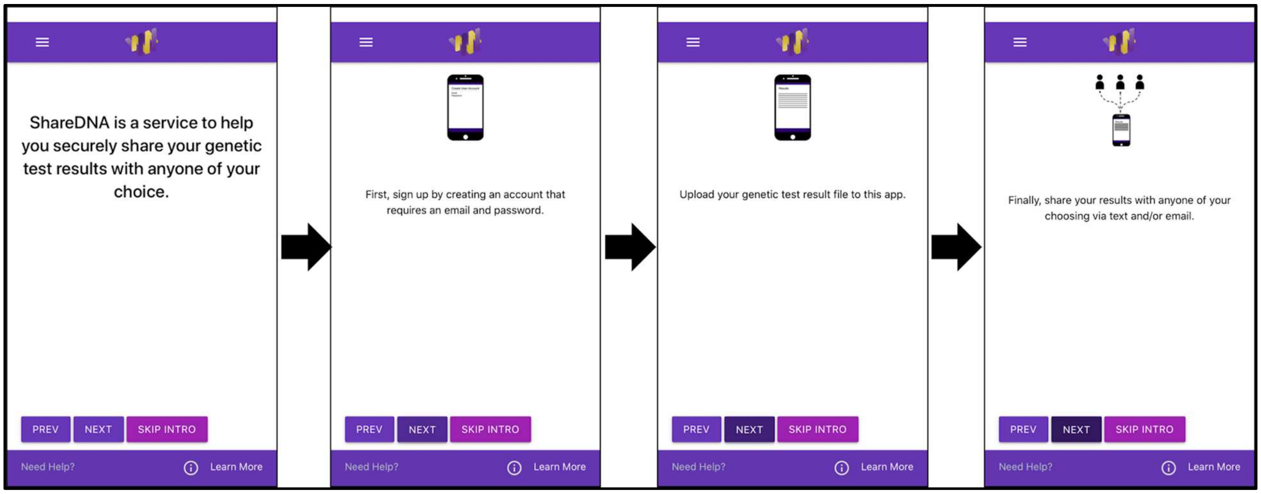


Figure 2-1. **Overview of the Prediction Pipeline**. For early surgery, we identified LSS/LDH patients if they have at least 2 diagnosis codes one year prior to LIRE enrollment and then identified out of these patients as having surgery if they had at least 1 decompression code within 2 months ahead. For late surgery, we identified LSS/LDH patients if they have at least 2 diagnosis codes one year prior to LIRE enrollment and/or 2 months after enrollment and then identified out of these patients as having surgery if they had at least 1 decompression code within 12 months ahead. For each prediction task, we collected patients' demographics, diagnosis codes, procedure codes, drug names, and index image reports. For the multimodal deep learning architecture, the index image reports are passed into a CNN, the diagnosis and procedure codes and drug names are passed into a GRU, and the demographics are featurized. The output from each network are concatenated together along with the featurized demographics and then passed into a fully-connected layer and then to an output layer to make predictions. CNN = Convolutional Neural Network, GRU = Gated Recurrent Unit, LSS = Lumbar Spinal Stenosis, LDH = Lumbar Disc Herniation, LIRE = Lumbar Imaging With Reporting Of Epidemiology



ShareDNA is a service to help you securely share your genetic test results with anyone of your choice.

First, sign up by creating an account that requires an email and password.

Upload your genetic test result file to this app.

Finally, share your results with anyone of your choosing via text and/or email.

Figure 3-1. **Overview of the purpose of *ShareDNA***. *ShareDNA* provides a service to allow users to create an account that only requires their email and password and then they can upload their genetic test results and share with anyone from their contact list.

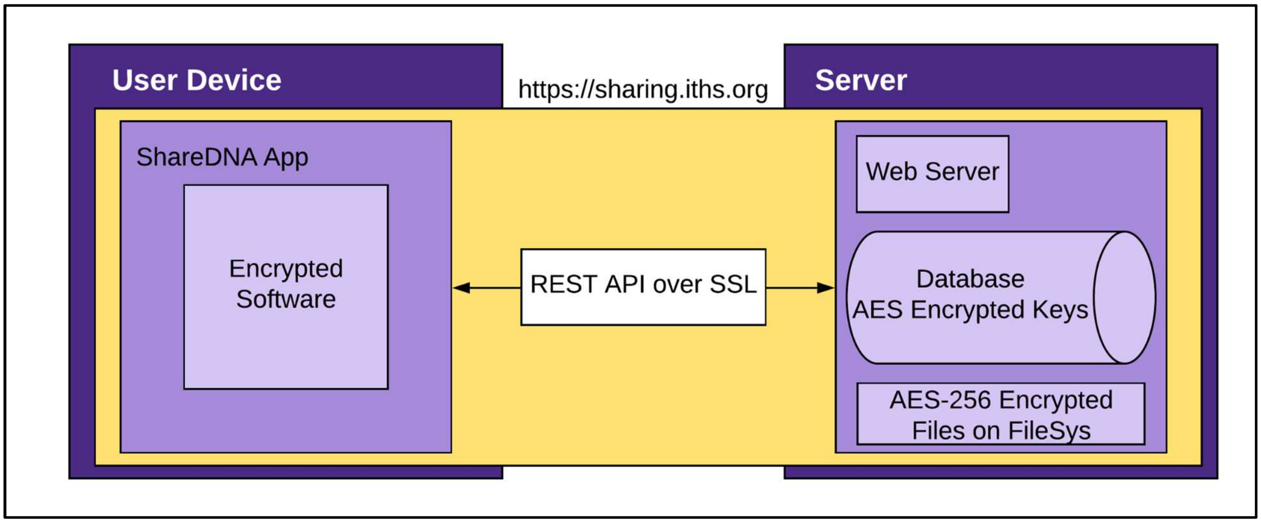


Figure 3-2. **Overview of *ShareDNA's* communication between the client and server-side**. The client side of *ShareDNA* faces the users and when users upload their genetic test results, the information is securely sent to a server maintained by University of Washington for storing.

Tables

Type of Finding	**Imaging Finding**
Deformities	Listhesis-Grade 1
	Listhesis-Grade 2 or higher
	Scoliosis
Fracture	Fracture
	Spondylosis
Anterior Column Degeneration	Annular Fissure
	Disc Bulge
	Disc Degeneration
	Disc Desiccation
	Disc Extrusion
	Disc Height Loss
	Disc Herniation
	Disc Protrusion
	Endplate Edema or Type 1 Modic
	Osteophyte-anterior column
Posterior Column Degeneration	**Any Stenosis**
	Facet Degeneration
Associated with Leg Pain	**Central Stenosis**
	Foraminal Stenosis
	Nerve Root Contact
	Nerve Root Displaced/Compressed
	Lateral Recess Stenosis
Nonspecific Findings and Other	Any Degeneration
	Hemangioma
	Spondylolysis
	Any Osteophyte

Table 1-1. The 26 imaging findings of our study. Any stenosis refers to any of central, foraminal, lateral recess, or not otherwise specified. Any degeneration refers to any of disc degeneration, facet degeneration, or degeneration not otherwise specified. Bold indicates the potentially clinically important findings.

System	**Image Type**	**N in Dataset**	**Average Text Length**	**Average Age**	**Female %**

Kaiser Permanente of Washington	X-Ray	102	132 +/- 34	70.35 +/- 13.84	0.60
	MR	115	267 +/- 106	58.90 +/- 14.35	0.49
	Total	217	203 +/- 105	64.28 +/- 15.20	0.54
Kaiser Permanente of Northern California	X-Ray	104	143 +/- 38	67.51 +/- 16.82	0.61
	MR	114	270 +/- 95	57.1 +/- 14.96	0.53
	Total	218	210 +/- 97	62.06 +/- 16.68	0.56
Henry Ford Health System	X-Ray	103	121 +/- 57	67.15 +/- 16.04	0.72
	MR	115	268 +/- 152	58.95 +/- 15.77	0.5
	Total	218	199 +/- 137	62.96 +/- 16.44	0.61
Mayo Clinic Health System	X-Ray	103	141 +/- 39	69.35 +/- 16.15	0.61
	MR	115	222 +/- 104	55.13 +/- 15.44	0.58
	Total	218	184 +/- 90	61.85 +/- 17.28	0.60
All	X-Ray	413	134 +/- 44	68.58 +/- 15.76	0.63
	MR	458	257 +/- 118	57.52 +/- 15.17	0.52
	Total	871	199 +/- 109	62.79 +/- 16.42	0.58

Table 1-2. We calculated the average text length for the finding and impression sections in each report, the average age of patients, and the proportion of female patients for each healthcare system and each type of report. For average text length, we calculated the average text length for both the finding and impression sections, since these sections were required for our pipeline. For both average text length and age, we included standard deviation.

Finding	**Proportion**	**P-Value**	**N-Grams**	**Document LIRE**	**Rules**	**Document MIMIC**	**Controlled Vocabulary**	**Controlled Vocabulary Filter Only**
All Findings	-	**1.06E-24**	**0.960 (0.949, 0.972)**	0.910 (0.892, 0.929)	0.897 (0.882, 0.911)	0.894 (0.872, 0.916)	0.882 (0.862, 0.901)	0.879 (0.857, 0.902)
Potentially Clinically Important Findings	-	**2.06E-13**	**0.954 (0.925, 0.983)**	0.910 (0.878, 0.942)	0.821 (0.789, 0.852)	0.888 (0.856, 0.920)	0.857 (0.821, 0.894)	0.854 (0.813, 0.895)
any degeneration	0.896	0.03289	**0.947 (0.906, 0.989)**	0.896 (0.820, 0.972)	0.850 (0.770, 0.931)	0.874 (0.789, 0.958)	0.936 (0.911, 0.961)	0.913 (0.882, 0.943)
facet degeneration	0.762	0.13484	**0.970 (0.940, 0.999)**	0.963 (0.935, 0.991)	0.873 (0.832, 0.914)	0.949 (0.919, 0.979)	0.923 (0.888, 0.959)	0.922 (0.883, 0.961)
disc height loss	0.507	**5.46E-10**	**0.931 (0.891, 0.970)**	0.833 (0.791, 0.875)	0.877 (0.829, 0.925)	0.830 (0.774, 0.886)	0.874 (0.812, 0.935)	0.878 (0.826, 0.930)
any stenosis	**0.480**	0.09604	**0.972 (0.950, 0.994)**	0.957 (0.930, 0.983)	0.893 (0.856, 0.930)	0.967 (0.945, 0.988)	0.955 (0.926, 0.984)	0.961 (0.935, 0.987)
disc bulge	0.435	0.00276	**0.986 (0.972, 1.000)**	0.978 (0.956, 1.000)	0.976 (0.953, 1.000)	0.967 (0.939, 0.994)	0.955 (0.923, 0.987)	0.952 (0.920, 0.984)
foraminal stenosis	**0.400**	**0.00153**	**0.950 (0.922, 0.978)**	0.935 (0.899, 0.971)	0.914 (0.876, 0.952)	0.936 (0.906, 0.965)	0.932 (0.898, 0.966)	0.930 (0.896, 0.964)
central stenosis	**0.351**	**6.37E-10**	**0.950 (0.919, 0.981)**	0.903 (0.876, 0.930)	0.773 (0.731, 0.815)	0.915 (0.884, 0.947)	0.907 (0.861, 0.953)	0.901 (0.852, 0.950)
any osteophyte	0.332	**3.44E-07**	**0.955 (0.916, 0.994)**	0.888 (0.845, 0.932)	0.925 (0.894, 0.955)	0.886 (0.835, 0.937)	0.875 (0.818, 0.933)	0.880 (0.819, 0.94)
listhesis grade 1	0.324	0.03888	**0.967 (0.941, 0.994)**	0.927 (0.880, 0.974)	0.958 (0.928, 0.989)	0.910 (0.858, 0.961)	0.945 (0.903, 0.987)	0.947 (0.904, 0.989)

disc degeneration	0.322	0.00203	**0.935 (0.898, 0.973)**	0.830 (0.792, 0.869)	0.904 (0.865, 0.944)	0.784 (0.726, 0.842)	0.909 (0.842, 0.976)	0.909 (0.863, 0.954)
scoliosis	0.274	0.03726	**0.969 (0.943, 0.994)**	0.924 (0.888, 0.961)	0.959 (0.924, 0.994)	0.929 (0.887, 0.971)	0.900 (0.850, 0.949)	0.901 (0.854, 0.948)
osteophyte anterior column	0.271	**5.13E-10**	**0.953 (0.930, 0.976)**	0.867 (0.820, 0.914)	0.913 (0.872, 0.954)	0.846 (0.785, 0.907)	0.882 (0.831, 0.933)	0.874 (0.825, 0.923)
spondylosis	0.217	0.39974	**0.992 (0.977, 1.010)**	0.936 (0.881, 0.99)	0.990 (0.974, 1.010)	0.901 (0.846, 0.955)	0.900 (0.836, 0.964)	0.883 (0.817, 0.949)
fracture	0.212	0.62647	**0.949 (0.912, 0.987)**	0.947 (0.921, 0.973)	0.883 (0.816, 0.95)	0.896 (0.839, 0.953)	0.914 (0.868, 0.960)	0.925 (0.878, 0.972)
disc protrusion	0.197	0.69348	0.977 (0.953, 1.000)	0.940 (0.906, 0.973)	0.927 (0.879, 0.974)	0.910 (0.866, 0.954)	**0.982 (0.952, 1.010)**	0.980 (0.948, 1.010)
disc desiccation	0.189	**1.39E-05**	**0.981 (0.953, 1.010)**	0.957 (0.923, 0.990)	0.958 (0.921, 0.994)	0.923 (0.884, 0.962)	0.817 (0.745, 0.888)	0.822 (0.747, 0.898)
nerve root displaced/compressed	**0.169**	**1.08E-06**	**0.955 (0.913, 0.996)**	0.913 (0.854, 0.972)	0.785 (0.698, 0.872)	0.907 (0.848, 0.966)	0.870 (0.819, 0.921)	0.864 (0.817, 0.911)
lateral recess stenosis	**0.163**	0.00235	**0.966 (0.921, 1.010)**	0.941 (0.909, 0.973)	0.649 (0.567, 0.731)	0.948 (0.918, 0.978)	0.843 (0.771, 0.914)	0.844 (0.773, 0.915)
annular fissure	0.099	0.38358	0.950 (0.888, 1.010)	0.944 (0.886, 1.000)	**0.957 (0.905, 1.010)**	0.922 (0.848, 0.996)	0.763 (0.667, 0.860)	0.755 (0.650, 0.860)
nerve root contact	0.097	**1.13E-08**	**0.972 (0.949, 0.996)**	0.921 (0.859, 0.982)	0.910 (0.827, 0.993)	0.914 (0.856, 0.973)	0.797 (0.716, 0.877)	0.818 (0.741, 0.894)
disc extrusion	**0.079**	**0.00089**	**0.994 (0.963, 1.020)**	0.979 (0.954, 1.000)	0.886 (0.775, 0.997)	0.948 (0.901, 0.995)	0.969 (0.914, 1.020)	0.962 (0.897, 1.030)

endplate edema	**0.059**	**0.00088**	**0.916 (0.831, 1.000)**	0.868 (0.765, 0.970)	0.854 (0.732, 0.976)	0.838 (0.713, 0.963)	0.789 (0.683, 0.896)	0.778 (0.626, 0.930)
hemangioma	0.049	0.0736	0.991 (0.950, 1.030)	0.899 (0.807, 0.991)	**0.999 (0.995, 1.000)**	0.875 (0.705, 1.050)	0.963 (0.867, 1.060)	0.951 (0.845, 1.060)
disc herniation	0.038	0.04629	0.956 (0.878, 1.003)	0.890 (0.715, 1.060)	**0.975 (0.927, 1.020)**	0.912 (0.740, 1.080)	0.760 (0.526, 0.994)	0.786 (0.520, 1.050)
spondylolysis	0.032	0.00567	**0.981 (0.940, 1.020)**	0.832 (0.654, 1.010)	0.936 (0.795, 1.080)	0.866 (0.747, 0.986)	0.754 (0.507, 1.000)	0.785 (0.561, 1.010)
listhesis grade 2	**0.028**	**0.00018**	**0.905 (0.741, 1.070)**	0.799 (0.604, 0.994)	0.767 (0.580, 0.955)	0.683 (0.486, 0.88)	0.735 (0.559, 0.91)	0.706 (0.510, 0.902)

Table 1-3. For each representation, we trained and tested 26 models (one for each finding) on 80% and 20% of the dataset, respectively. For group level, we averaged the AUC across all findings and across all potentially clinically important findings for each representation. We repeated this process 25 times with different splits of the data to calculate 95% confidence intervals (in parentheses). Table shows the best performing representation ordered left to right based on the *All Findings* row (1st row). The first column indicates the finding/group. For the findings, the second column indicates prevalence in the test set represented as a proportion. For each row, we bolded the best performing individual representation based on the average AUC and underlined the second-best representation. We show the 95% confidence interval in the parentheses. Finally, for each finding and group, we performed a t-test comparing the best representation's distribution of AUC values for the 25 repeats to the second-best representation. We bolded the significant comparisons with Bonferroni correction (p-value significance for the groups: 0.025 = 0.05/2 groups, p-value significance for the findings: 0.0019 = 0.05/26 findings). Finally, we bolded the findings that were potentially clinically important. AUC = Area Under the Curve.

Finding	**Proportion**	**N-Grams**	**Document LIRE**	**Document MIMIC**	**Controlled Vocabulary**	**Controlled Vocabulary**

						Filter Only
All Findings	-	**0.902**	0.879	0.868	0.857	0.853
Potentially Clinically Important Findings	-	**0.898**	0.890	0.887	0.834	0.833
any degeneration	0.896	0.905	0.881	0.848	**0.927**	0.874
facet degeneration	0.762	0.919	**0.952**	0.927	0.914	0.898
disc height loss	0.507	**0.845**	0.754	0.748	**0.845**	0.837
any stenosis	**0.480**	0.907	**0.957**	0.955	0.946	0.943
disc bulge	0.435	0.954	**0.963**	0.953	0.930	0.929
foraminal stenosis	**0.400**	0.885	**0.922**	0.913	0.904	0.888
central stenosis	**0.351**	**0.916**	0.881	0.897	0.891	0.878
any osteophyte	0.332	**0.874**	**0.874**	0.860	0.831	0.833
listhesis grade 1	0.324	0.905	0.899	0.891	0.930	**0.936**
disc degeneration	0.322	0.873	0.710	0.716	0.861	**0.908**
scoliosis	0.274	**0.911**	0.892	0.891	0.865	0.870
osteophyte anterior column	0.271	0.832	0.825	0.818	**0.845**	0.825
spondylosis	0.217	**0.985**	0.801	0.764	0.823	0.781
fracture	0.212	0.910	**0.926**	0.889	0.910	0.906
disc protrusion	0.197	0.948	0.935	0.904	0.977	**0.978**
disc desiccation	0.189	**0.929**	0.909	0.850	0.796	0.797
nerve root displaced/compressed	**0.169**	**0.918**	0.906	0.894	0.845	0.855

lateral recess stenosis	**0.163**	0.915	0.932	**0.934**	0.831	0.841
annular fissure	0.099	0.908	**0.921**	0.886	0.766	0.763
nerve root contact	0.097	**0.917**	0.885	0.854	0.736	0.746
disc extrusion	**0.079**	0.947	**0.972**	0.925	0.968	0.959
endplate edema	**0.059**	0.844	**0.865**	0.841	0.725	0.762
hemangioma	0.049	0.917	0.869	0.908	**0.963**	**0.963**
disc herniation	0.038	0.820	**0.826**	0.798	0.725	0.700
spondylolysis	0.032	**0.938**	0.814	0.849	0.775	0.804
listhesis grade 2	**0.028**	**0.838**	0.740	0.681	0.657	0.614

Table 1-4. For each representation, we trained our model on reports from three systems and evaluated on the fourth, iteratively, for each finding. For each finding, we calculated the mean of the AUC across the four systems. We calculated group-level performance by averaging the AUC across all findings, and across all potentially clinically important findings for each system and then calculated the mean across the systems. Table shows the best performing representation ordered left to right based on the *All Findings* row (1st row). The first column indicates the finding/group. For the findings, the second column indicates prevalence in the test set represented as a proportion. Bold value indicates the best performing representation for that finding and group. Finally, we bolded the findings that were potentially clinically important. AUC = Area Under the Curve.

Finding	**Proportion**	**Document LIRE**	**Controlled Vocabulary**	**Document MIMIC**	**Controlled Vocabulary Filter Only**	**N-Grams**
All Findings	**-**	**0.010**	0.012	0.013	0.014	0.051
Potentially Clinically Important Findings	**-**	**0.007**	0.035	0.024	0.031	0.076

any degeneration	0.896	0.021	0.031	0.024	0.041	**0.02**
facet degeneration	0.762	**0.018**	0.039	0.022	0.038	0.064
disc height loss	0.507	0.033	0.050	**0.010**	0.057	0.103
any stenosis	**0.480**	0.023	0.032	**0.013**	0.035	0.051
disc bulge	0.435	0.019	0.022	**0.009**	0.015	0.04
foraminal stenosis	**0.400**	**0.016**	0.058	0.035	0.077	0.042
central stenosis	**0.351**	**0.030**	0.046	0.038	0.056	0.043
any osteophyte	0.332	0.034	0.099	**0.031**	0.094	0.108
listhesis grade 1	0.324	0.032	0.014	**0.009**	0.014	0.076
disc degeneration	0.322	0.033	0.052	0.092	0.049	**0.026**
scoliosis	0.274	**0.017**	0.051	0.025	0.043	0.068
osteophyte anterior column	0.271	0.025	0.084	**0.020**	0.083	0.088
spondylosis	0.217	0.040	0.156	0.036	0.164	**0.015**
fracture	0.212	**0.019**	0.028	0.029	0.039	0.029
disc protrusion	0.197	0.021	**0.011**	0.018	0.016	0.053
disc desiccation	0.189	0.024	0.052	**0.020**	0.051	0.09
nerve root displaced/compressed	**0.169**	0.020	0.018	0.015	**0.011**	0.021
lateral recess stenosis	**0.163**	**0.017**	0.038	0.019	0.034	0.063
annular fissure	0.099	**0.015**	0.064	0.057	0.065	0.063
nerve root contact	0.097	0.071	0.035	0.044	**0.014**	0.09
disc extrusion	**0.079**	**0.007**	0.024	0.015	0.023	0.076
endplate edema	**0.059**	**0.059**	0.164	**0.059**	0.091	0.126
hemangioma	0.049	0.073	0.048	**0.007**	0.048	0.105
disc herniation	0.038	**0.046**	0.079	0.092	0.065	0.094
spondylolysis	0.032	0.107	0.113	0.087	0.108	**0.044**
listhesis grade 2	**0.028**	**0.047**	0.128	0.184	0.099	0.184

Table 1-5. For each representation, we trained our model on reports from three systems and evaluated on the fourth, iteratively, for each finding. For each finding, we calculated the standard deviation of the AUC across the four systems. We calculated group-level consistency by averaging the AUC across all findings, and across all potentially clinically important findings for each system as a test set and then calculated the standard deviation across the systems. Table shows the most consistent representation ordered left to right based on the *All Findings* row (1st row). The first column indicates the finding/group. For the findings, the second column indicates prevalence in the test set represented as a proportion. Bold value indicates the

most consistent representation for that finding and group. Finally, we bolded the findings that were potentially clinically important. AUC = Area Under the Curve.

Group	**Codes**
Lumbar Stenosis	*ICD-9*
	344.6, 344.60, 344.61, 721.4, 721.42, 724, 724.02, 724.03, 724.09
	ICD-10
	G83.4, M47.15, M47.16, M48.05, M48.06, M48.061, M48.062, M48.07, M48.08
Lumbar Disc Herniation	*ICD-9*
	344.6, 344.60, 344.61, 353.4, 355.0, 721.4, 721.42, 722.1, 722.10, 724.3, 724.4
	ICD-10
	G54.4, G57.0, G57.00, G57.01, G57.02, G83.4, M47.15, M47.16, M47.25, M47.26, M47.27, M47.28, M51.15, M51.16, M51.17, M54.18, M51.25, M51.26, M51.27, M54.10, M54.15, M54.16, M54.17, M54.18, M54.30, M54.31, M54.32, M54.4, M54.40, M54.41, M54.42
Decompression	*CPT*
	0274T, 0275T, 22818, 22819, 63003, 63005, 63010, 63011, 63012, 63016, 63017, 63030, 63035, 63042, 63044, 63046, 63047, 63048, 63050, 63051, 63055, 63056, 63057, 63064, 63066, 63077, 63078, 63085, 63086, 63087, 63088, 63090, 63091, 63101, 63102, 63103, 63170, 63172, 63173, 63185, 63190, 63191, 63195, 63197, 63199, 63200, 63266, 63267, 63268, 63271, 63272, 63273, 63276, 63277, 63278, 63281, 63282, 63283, 63286, 63287, 63290, 63301, 63302, 63303, 63305, 63306, 63307, 63308
	HCPCS Level II
	S2350, S2351, S9090
	ICD-9-PCS
	03.02, 03.09, 03.6, 80.50, 80.51, 80.53, 80.54, 80.59
	ICD-10-PCS
	00NX0ZZ, 00NY0ZZ, 01N80ZZ, 01N83ZZ, 01NB0ZZ, 01NB4ZZ, 01NR0ZZ, 0PB40ZZ,

	0PB43ZX, 0QB00ZZ, 0QB03ZX, 0QB03ZZ, 0QB10ZZ, 0QB13ZX, 0QS004Z, 0QS134Z, 0QU007Z, 0QU00JZ, 0QU03JZ, 0QW004Z, 0QW034Z, 0QW104Z, 0RB90ZZ, 0RB93ZX, 0RBB3ZX, 0SB00ZX, 0SB00ZZ, 0SB03ZX, 0SB20ZZ, 0SB23ZX, 0SB23ZZ, 0SB40ZZ, 0SB43ZX, 0SC00ZZ, 0ST20ZZ, 0ST40ZZ
	Kaiser-Specific
	213730, 222494, 222572, 222573, 222590, 222865, 223880, 223899, 223900, 223901, 224085, 224086, 224087, 224088, 224089, 224131, 224170, 224238, 224928, 224929, 226803, 226929, 227538, 227539, 227553, 231207, 231208, 245922, 245923, 245925, 245926, 245927, 245929, 245930, 245931, 245935, 245936, 245937, 245938, 245939, 245940, 245941, 245942, 245944, 245945, 245946, 245947, 245948, 245949, 245963, 245964, 245965, 245977, 245978, 245980, 245983, 245986, 245987, 245988, 245989, 245991, 245993, 245994, 245996, 245997, 245998, 245999, 246000, 246033, 246034, 246035, 246789, 246790, 246791, 246792, 246793, 246794, 246795, 246796, 246797, 251410, 251411, 253030, 707346, 707347, 756636

Table 2-1. List of Codes for Lumbar Stenosis, Lumbar Disc Herniation, and Decompression

Characteristics	**Early Surgery**	**Late Surgery**
N	8,387	31,210
Negative	8,189 (97.6%)	29,845 (95.6%)
Positive	198 (2.4%)	1,365 (4.4%)
Average Days Between LIRE Enrollment and Decompression Surgery	34.3	180.7
Average Age	57	57.7
Gender		
Female	4,713 (56.2%)	17,466 (56.0%)
Race		
White	5,317 (63.4%)	20,287 (65.0%)
Black	991 (11.8%)	2,992 (9.6%)
Unknown	990 (11.8%)	3,871 (12.4%)
Asian	928 (11.1%)	3,516 (11.3%)

Native American	67 (0.8%)	226 (0.7%)
Pacific Islander	50 (0.6%)	199 (0.6%)
Other	27 (0.3%)	65 (0.2%)
Multiracial	17 (0.2%)	54 (0.2%)
Ethnicity		
Not Available	5,945 (70.9%)	22,420 (71.8%)
Not Hispanic	1,233 (14.7%)	4,081 (13.1%)
Hispanic	1,209 (14.4%)	4,709 (15.1%)
Image Type		
MRI	5,810 (69.3%)	11,852 (38.0%)
X-Ray	2,517 (30.0%)	19,189 (61.5%)
CT	60 (0.7%)	169 (0.5%)
System		
Kaiser Permanente	7,071 (84.3%)	26,870 (86.1%)
Henry Ford	654 (7.8%)	1,581 (5.1%)
Group Health	486 (5.8%)	1,755 (5.6%)
Mayo Clinic	176 (2.1%)	1,004 (3.2%)

Table 2-2. Data Characteristics

Hyperparamter	**Values**
Learning Rate	0.001, 0.0001
GRU Hidden Size	200, 300, 400
# of Filters	200, 300, 400
Dropout Rate	0.0, 0.2, 0.5, 0.9
Weight Decay	0.1, 0.01, 0.001

Table 2-3. Hyperparameter Search Space

Target	**Prev.**	**N**	**Model**	**Recall**	**Precision**	**Balanced Accuracy**	**F1**	**AUC**	**AUPRC**
Early Surgery	0.024	824	MDL	0.3 +/- 0.077*	0.086 +/- 0.021*	0.61 +/- 0.039*	0.133 +/- 0.033*	0.725 +/- 0.04*	0.061 +/- 0.014*
			Baseline	0.375 +/- 0.076	0.069 +/- 0.014	0.624 +/- 0.038	0.116 +/- 0.023	0.597 +/- 0.05	0.047 +/- 0.011
Late Surgery	0.044	3,121	MDL	0.725 +/- 0.026*	0.145 +/- 0.006*	0.765 +/- 0.013*	0.242 +/- 0.009*	0.833 +/- 0.012*	0.241 +/- 0.023*

			Baseline	0.663 +/- 0.028	0.156 +/- 0.007	0.75 +/- 0.014	0.253 +/- 0.011	0.84 +/- 0.012	0.266 +/- 0.026

Table 2-4. We compared the performance of the MDL architecture against the baseline (i.e. LASSO). We calculated 1,000 bootstrap samples from the test set. For each sample, we calculated the performance metrics: recall, specificity, balanced accuracy, precision, F1-score, AUC, and AUPRC. We then calculated the average and standard deviation across the samples. For each prediction task, we underline the model that had the best performance for each metric. Finally, we performed a t-test to assess significance between each model's performance metrics for each prediction task; we indicate significance with an asterisk. AUC = Area Under the Curve, AUPRC = Area Under the Precision-Recall Curve, MDL = Multimodal Deep Learning, Prev. = Prevalence.

Target	**Prev.**	N	**Data Type**	**Recall**	**Precision**	**Balanced Accuracy**	**F1**	**AUC**	**AUPRC**
Early Surgery	0.024	824	All	0.3 +/- 0.077	0.086 +/- 0.021	0.61 +/- 0.039	0.133 +/- 0.033	0.725 +/- 0.04	0.061 +/- 0.014
			Demographics	0.475 +/- 0.084*	0.043 +/- 0.008*	0.608 +/- 0.042	0.08 +/- 0.014*	0.593 +/- 0.055*	0.043 +/- 0.01*
			Temporal	1.0 +/- 0.0*	0.024 +/- 0.0*	0.5 +/- 0.0*	0.047 +/- 0.0*	0.5 +/- 0.042*	0.023 +/- 0.002*
			Textual	0.4 +/- 0.083*	0.087 +/- 0.018	0.648 +/- 0.042*	0.143 +/- 0.029*	0.72 +/- 0.043*	0.06 +/- 0.013
Late Surgery	0.044	3,121	All	0.725 +/- 0.026	0.145 +/- 0.006	0.765 +/- 0.013	0.242 +/- 0.009	0.833 +/- 0.012	0.241 +/- 0.023
			Demographics	0.59 +/- 0.031*	0.065 +/- 0.003*	0.6 +/- 0.016*	0.117 +/- 0.006*	0.637 +/- 0.017*	0.076 +/- 0.008*
			Temporal	0.696 +/- 0.027*	0.144 +/- 0.006*	0.753 +/- 0.014*	0.239 +/- 0.01*	0.824 +/- 0.013*	0.255 +/- 0.024*
			Textual	0.396 +/- 0.029*	0.084 +/- 0.006*	0.599 +/- 0.015*	0.138 +/- 0.01*	0.656 +/- 0.017*	0.088 +/- 0.009*

Table 2-5. We compared the performance of the MDL architecture against each individual network (i.e. temporal, textual, and demographics). We calculated 1,000 bootstrap samples from the test set. For each sample, we calculated the performance metrics: recall, specificity, balanced accuracy, precision, F1-score, AUC, and AUPRC. We then calculated the average and standard deviation across the samples. For each

prediction task, we underline the model that had the best performance for each metric. Finally, we performed a t-test to assess significance between each model's performance metrics for each prediction task; we indicate significance with an asterisk. We used a Bonferroni correction to correct for multiple hypothesis testing when comparing MDL to the three individual networks (0.05/3 = 0.0167). AUC = Area Under the Curve, AUPRC = Area Under the Precision-Recall Curve, Prev. = Prevalence

Target	Model	System	Prev.	N	Recall	Precision	Balanced Accuracy	F1	AUC	AUPRC
Early Surgery	MDL	Group Health	0.021	239	0.6 +/- 0.161	0.075 +/- 0.02	0.72 +/- 0.081	0.132 +/- 0.036	0.731 +/- 0.109	0.105 +/- 0.05
		Henry Ford	0.039	324	0.64 +/- 0.097	0.127 +/- 0.021	0.732 +/- 0.05	0.212 +/- 0.033	0.795 +/- 0.047	0.128 +/- 0.031
		Average	0.03	281	0.62 +/- 0.091*	0.101 +/- 0.014*	0.726 +/- 0.046*	0.172 +/- 0.024*	0.763 +/- 0.059*	0.116 +/- 0.029
	Baseline	Group Health	0.021	239	0.3 +/- 0.152	0.056 +/- 0.028	0.595 +/- 0.076	0.094 +/- 0.047	0.656 +/- 0.113	0.149 +/- 0.114
		Henry Ford	0.039	324	0.2 +/- 0.079	0.087 +/- 0.034	0.557 +/- 0.04	0.12 +/- 0.047	0.714 +/- 0.05	0.088 +/- 0.023
		Average	0.03	281	0.25 +/- 0.085	0.071 +/- 0.022	0.576 +/- 0.042	0.107 +/- 0.033	0.685 +/- 0.061	0.119 +/- 0.058
Late Surgery	MDL	Group Health	0.066	878	0.557 +/- 0.05	0.157 +/- 0.013	0.673 +/- 0.025	0.244 +/- 0.021	0.745 +/- 0.025	0.181 +/- 0.026
		Henry Ford	0.05	791	0.443 +/- 0.057	0.169 +/- 0.021	0.664 +/- 0.029	0.244 +/- 0.03	0.776 +/- 0.029	0.168 +/- 0.032
		Average	0.058	834	0.5 +/- 0.039*	0.163 +/- 0.013*	0.669 +/- 0.02*	0.244 +/- 0.019*	0.76 +/- 0.019*	0.175 +/- 0.021*
	Baseline	Group Health	0.066	878	0.843 +/- 0.035	0.096 +/- 0.004	0.644 +/- 0.019	0.173 +/- 0.007	0.728 +/- 0.025	0.162 +/- 0.022
		Henry Ford	0.05	791	0.532 +/- 0.055	0.145 +/- 0.015	0.683 +/- 0.028	0.228 +/- 0.023	0.767 +/- 0.028	0.192 +/- 0.038
		Average	0.058	834	0.688 +/- 0.033	0.121 +/- 0.008	0.664 +/- 0.017	0.2 +/- 0.012	0.748 +/- 0.019	0.177 +/- 0.022

Table 2-6. We compared the generalizability performance of the MDL architecture against the baseline (i.e. LASSO). For each test system, we evaluated models' performance using the performance metrics.

We estimated significance performance between models by bootstrapping 1,000 samples for each test system. For each pair of samples (i.e. one sample from each healthcare system), we calculated different performance metrics for each sample then averaged. We performed a t-test for each performance metric using each model's resulting 1,000 average values; we indicate significance with an asterisk. For each prediction task, we underline the model that had the best average performance metric. AUC = Area Under the Curve, AUPRC = Area Under the Precision-Recall Curve, MDL = Multimodal Deep Learning, Prev. = Prevalence

Target	**Data Type**	**System**	**Prev.**	**Size**	**Recall**	**Prec.**	**Balanced Accuracy**	**F1**	**AUC**	**AUPRC**
Early Surgery	All	Group Health	0.021	239	0.6 +/- 0.161	0.075 +/- 0.02	0.72 +/- 0.081	0.132 +/- 0.036	0.731 +/- 0.109	0.105 +/- 0.05
		Henry Ford	0.039	324	0.64 +/- 0.097	0.127 +/- 0.021	0.732 +/- 0.05	0.212 +/- 0.033	0.795 +/- 0.047	0.128 +/- 0.031
		Average	0.03	281	0.62 +/- 0.091	0.101 +/- 0.014	0.726 +/- 0.046	0.172 +/- 0.024	0.763 +/- 0.059	0.116 +/- 0.029
	Demo.	Group Health	0.021	239	0.3 +/- 0.157	0.068 +/- 0.035	0.606 +/- 0.078	0.11 +/- 0.057	0.656 +/- 0.111	0.058 +/- 0.025
		Henry Ford	0.039	324	0.4 +/- 0.101	0.125 +/- 0.031	0.644 +/- 0.051	0.19 +/- 0.047	0.668 +/- 0.067	0.085 +/- 0.021
		Average	0.03	281	0.35 +/- 0.093*	0.097 +/- 0.024*	0.625 +/- 0.047*	0.15 +/- 0.037*	0.662 +/- 0.066*	0.072 +/- 0.017*
	Temp.	Group Health	0.021	239	0.3 +/- 0.156	0.031 +/- 0.016	0.548 +/- 0.079	0.055 +/- 0.029	0.624 +/- 0.068	0.027 +/- 0.005
		Henry Ford	0.039	324	0.2 +/- 0.086	0.049 +/- 0.021	0.522 +/- 0.044	0.079 +/- 0.034	0.563 +/- 0.057	0.046 +/- 0.008
		Average	0.03	281	0.25 +/- 0.09*	0.04 +/- 0.013*	0.535 +/- 0.045*	0.067 +/- 0.022*	0.593 +/- 0.044*	0.037 +/- 0.005*
	Textual	Group Health	0.021	239	0.7 +/- 0.151	0.102 +/- 0.024	0.784 +/- 0.076	0.178 +/- 0.04	0.815 +/- 0.089	0.121 +/- 0.065
		Henry Ford	0.039	324	0.08 +/- 0.055	0.071 +/- 0.048	0.519 +/- 0.027	0.075 +/- 0.05	0.793 +/- 0.045	0.107 +/- 0.02

		Average	0.03	281	0.39 +/- 0.079*	0.087 +/- 0.027*	0.651 +/- 0.04*	0.126 +/- 0.032*	0.804 +/- 0.05*	0.114 +/- 0.034
Late Surgery	All	Average	0.058	834	0.5 +/- 0.039	0.163 +/- 0.013	0.669 +/- 0.02	0.244 +/- 0.019	0.76 +/- 0.019	0.175 +/- 0.021
	Demo.	Average	0.058	834	0.597 +/- 0.037*	0.094 +/- 0.006*	0.623 +/- 0.019*	0.163 +/- 0.01*	0.655 +/- 0.023*	0.114 +/- 0.013*
	Temp.	Average	0.058	834	0.476 +/- 0.038*	0.152 +/- 0.012*	0.652 +/- 0.019*	0.228 +/- 0.018*	0.72 +/- 0.023*	0.159 +/- 0.019*
	Textual	Average	0.058	834	0.438 +/- 0.036*	0.082 +/- 0.006*	0.569 +/- 0.018*	0.137 +/- 0.011*	0.614 +/- 0.021*	0.085 +/- 0.008*

Table 2-7. We compared the generalizability performance of the MDL architecture against the individual networks (i.e. temporal, textual, and demographics). For each test system, we evaluated models' performance using the performance metrics. We estimated significance performance between models by bootstrapping 1,000 samples for each test system. For each pair of samples (i.e. one sample from each healthcare system), we calculated different performance metrics for each sample then averaged. We performed a t-test for each performance metric using each model's resulting 1,000 average values; we indicate significance with an asterisk. We used a Bonferroni correction to correct for multiple hypothesis testing when comparing MDL to the three individual networks (0.05/3 = 0.0167). For each prediction task, we underline the model that had the best average performance metric. AUC = Area Under the Curve, AUPRC = Area Under the Precision-Recall Curve, Prev. = Prevalence, Prec. = Precision, Demo. = Demographics, Temp. = Temporal

Question	**Min**	**Q1**	**Mean**	**Median**	**Q3**	**Max**	**Standard Deviation**
Overall, I am staisfied with how easy it is to use this app.	2	3	3.62	4	4	5	0.96
It was simple to use this app.	2	3	3.77	4	4	5	0.83

I could effectively complete the tasks and scenarios quickly using this app.	2	3	3.92	4	5	5	1.12
I was able to complete the tasks and scenarios quickly using this app.	2	2.75	3.5	4	4	5	1.09
I was able to efficiently complete the tasks and scenarios using this app.	2	3	3.77	4	4	5	0.83
I felt comfortable using this app.	3	3	4.08	4	5	5	0.9
It was easy to learn to use this app.	2	4	4	4	5	5	0.91
I believe I could become productive quickly using this app.	3	4	4.31	4	5	5	0.63
The app gave error messages that clearly told me how to fix problems.	1	2	2.64	2	3	5	1.12
Whenever I made a mistake using the app, I could recover easily and quickly.	2	2	3.08	3	4	5	1
The information provided with this app was clear.	2	2.75	3	3	3.25	4	0.74
It was easy to find the information I needed	2	3	3.45	4	4	4	0.69

The information provided for the app was easy to understand.	2	4	3.83	4	4	5	0.72
The information was effective in helping me complete the tasks and scenarios.	3	4	3.92	4	4	5	0.64
The organization of information on the app screens was clear.	2	3	3.54	3	4	5	1.05
The interface of this app was pleasant.	3	4	3.85	4	4	5	0.55
I liked using the inferface of this app.	2	3	3.67	4	4	5	0.89
The app has all the functions and capabilities I expect it to have.	2	3	3.75	4	5	5	1.14
Overall I am staistfied with this app.	2	3	3.77	4	4	5	0.93

Table 3-1. **T**he scale is out of five, with one indicating “strongly disagree” and five indicating “strongly agree”. PSSUQ = Post-Study System Usability Questionnaire

Issue	**Recommendation**
Default messaging was impersonal/generic	Leave blank with suggested wording above the text box. Most participants felt the wording should be in first person since it would come from their number/email.

Redundancy in requiring password	Remove additional password requirements once the user is logged into their account. Or an option to require the password before sending test results to recipients.
Adding multiple recipients wasn’t intuitive	Add a feature to “save” recipient contact info and the “+” to add more recipients.
Light greys were difficult to see	Darken grey or change color to indicate the text can be altered.
Confusion from intro screens	Once all screens/dialogue has been rotated through (i.e. pressed “Next” 3 times), enter the login/create an account screen automatically.
“Create an account” was overlooked	If the email entered does not have an account yet, navigate to the “create an account” page with the information already entered.
“Share” icon wasn’t clear	Older users didn’t intuitively know the icon to share and the font was small, a larger button with text would be more clear.

Scrolling function wasn’t shown	Add scrolling sidebar to “More information” section to show additional text is below.
UW branding was confusing to non-UW patients	Consider de-emphasizing UW look and feel if using with external patients.
Some participants didn’t intuitively go to “Files” to send their test results again.	From upper left menu, include a “Share” option.

Table 3-2. Outlines the major issues that participants found along with their recommended improvements

Supplemental

Tables

Feature	**Coefficient**	**Feature**	**Coefficient**
Descend (text)	0.971	Diffus (text)	-0.519
s1 nerv (text)	0.897	Within (text)	-0.533
disc diseas l4 (text)	0.849	Otherwis (text)	-0.537
Larg (text)	0.781	siteID_2 (KP NorCal)	-0.561
stenosi facet (text)	0.771	l4 mild (text)	-0.583
Dx_99282 (proc code)	0.763	Female	-0.637
Equina (text)	0.682	find techniqu multiplanar (text)	-0.649
ligamentum flavum moder (text)	0.581	Sacroiliac (text)	-0.741
14 (text)	0.578	Contact (text)	-0.84
action requir (text)	0.569	Intact (text)	-0.933

Table 2-1. Top 10 and Bottom 10 Predictors for Early Surgery Baseline Model for Classical Performance

Feature	Coefficient	Feature	Coefficient
Dx_72110 (proc code)	1.402	acut bone (text)	-0.376
Dx_99205 (proc code)	1.365	May (text)	-0.388
Dx_64483 (proc code)	1.325	comparison 11 (text)	-0.405
Dx_62311 (proc code)	1.163	disc bulg asymmetr (text)	-0.427
Dx_99204 (proc code)	1.023	stenosi impress (text)	-0.428
sever central (text)	0.878	drugnamerx_DICLOFENAC SODIUM (drug name)	-0.429
Dx_99243 (proc code)	0.842	Degener (text)	-0.43
sever spinal (text)	0.812	foramen l5 (text)	-0.567
Dx_72120 (proc code)	0.795	Dx_805 (dx code)	-0.696
obtain acut fractur (text)	0.789	siteID_2 (KP NorCal)	-0.896

Table 2-2. Top 10 and Bottom 10 Predictors for Late Surgery Baseline Model for Classical Performance

Feature	Coefficient	Feature	Coefficient
view mild (text)	1.13	incident (text)	-0.514
action requir (text)	0.939	canal narrow mild (text)	-0.515
sever central (text)	0.774	upper (text)	-0.515
descend (text)	0.768	multilevel disc (text)	-0.545
also facet (text)	0.736	otherwis (text)	-0.589
s1 diffus disc (text)	0.736	disc bulg facet (text)	-0.696
lumbar spine mri (text)	0.719	l5 mild (text)	-0.738
disc space mild (text)	0.672	intact (text)	-0.757
multiplanar multisequ mr (text)	0.658	raceID_2 (Black)	-0.804
Dx_97001 (proc code)	0.622	sacroiliac joint (text)	-1.027

Table 2-3. Top 10 and Bottom 10 Predictors for Early Surgery Baseline Model for Generalizability Performance

Feature	Coefficient	Feature	Coefficient
Dx_72110 (proc code)	1.71	Degener (text)	-0.449
dx_99205 (proc code)	1.516	Impress (text)	-0.452
dx_72120 (proc code)	1.401	si joint (text)	-0.531
dx_62311 (proc code)	1.365	flexion extens (text)	-0.572
dx_64483 (proc code)	1.341	Dx_805 (dx code)	-0.576
pedicl intact (text)	1.074	foramen l5 (text)	-0.579
dx_99204 (proc code)	0.963	Intercept	-0.584
sever central (text)	0.882	Find (text)	-0.626
sever spinal (text)	0.851	siteID_2 (KP NorCal)	-0.633
sclerosi anterior (text)	0.764	dx_V54 (dx code)	-0.776

Table 2-4. Top 10 and Bottom 10 Predictors for Late Surgery Baseline Model for Generalizability Performance

Target	**Prev.**	**N**	**Data Type**	**Recall**	**Precision**	**Balanced Accuracy**	**F1**	**AUC**	**AUPRC**
Early Surgery	0.024	824	All	0.375 +/- 0.076	0.069 +/- 0.014	0.624 +/- 0.038	0.116 +/- 0.023	0.597 +/- 0.05	0.047 +/- 0.011
			Demographics	0.55 +/- 0.081*	0.037 +/- 0.005*	0.599 +/- 0.041*	0.069 +/- 0.01*	0.599 +/- 0.054	0.046 +/- 0.012
			Codes	0.35 +/- 0.078*	0.032 +/- 0.007*	0.541 +/- 0.039*	0.058 +/- 0.013*	0.597 +/- 0.048	0.039 +/- 0.008*
			Textual	0.35 +/- 0.077*	0.062 +/- 0.013*	0.611 +/- 0.039*	0.106 +/- 0.023*	0.592 +/- 0.052	0.042 +/- 0.009*
Late Surgery	0.044	3,121	All	0.663 +/- 0.028	0.156 +/- 0.007	0.75 +/- 0.014	0.253 +/- 0.011	0.84 +/- 0.012	0.266 +/- 0.026

			Demographics	0.604 +/- 0.029*	0.062 +/- 0.003*	0.593 +/- 0.015*	0.112 +/- 0.005*	0.638 +/- 0.016*	0.072 +/- 0.006*
			Codes	0.722 +/- 0.028*	0.149 +/- 0.006*	0.766 +/- 0.014*	0.246 +/- 0.01*	0.824 +/- 0.014*	0.262 +/- 0.027*
			Textual	0.538 +/- 0.033*	0.07 +/- 0.004*	0.605 +/- 0.016*	0.124 +/- 0.007*	0.655 +/- 0.018*	0.08 +/- 0.007*

Table 2-5. We compared the performance of the baseline using all data types against each individual data type (i.e. codes, textual, and demographics). We calculated 1,000 bootstrap samples from the test set. For each sample, we calculated the performance metrics: recall, specificity, balanced accuracy, precision, F1-score, AUC, and AUPRC. We then calculated the average and standard deviation across the samples. AUC = Area Under the Curve, AUPRC = Area Under the Precision-Recall Curve, Prev. = Prevalence.

Target	Model	System	Prev.	N	Recall	Precision	Balanced Accuracy	F1	AUC	AUPRC
Early Surgery	All	Group Health	0.021	239	0.3 +/- 0.152	0.056 +/- 0.028	0.595 +/- 0.076	0.094 +/- 0.047	0.656 +/- 0.113	0.149 +/- 0.114
		Henry Ford	0.039	324	0.2 +/- 0.079	0.087 +/- 0.034	0.557 +/- 0.04	0.12 +/- 0.047	0.714 +/- 0.05	0.088 +/- 0.023
		Average	0.03	281	0.25 +/- 0.085	0.071 +/- 0.022	0.576 +/- 0.042	0.107 +/- 0.033	0.685 +/- 0.061	0.119 +/- 0.058
	Demo.	Group Health	0.021	239	0.6 +/- 0.165	0.036 +/- 0.01	0.63 +/- 0.083	0.067 +/- 0.018	0.628 +/- 0.107	0.046 +/- 0.02
		Henry Ford	0.039	324	0.64 +/- 0.099	0.109 +/- 0.018	0.716 +/- 0.051	0.186 +/- 0.03	0.715 +/- 0.063	0.146 +/- 0.056
		Average	0.03	281	0.62 +/- 0.096*	0.072 +/- 0.01	0.673 +/- 0.049*	0.127 +/- 0.018*	0.672 +/- 0.062*	0.096 +/- 0.03*
	Codes	Group Health	0.021	239	0.4 +/- 0.158	0.031 +/- 0.012	0.565 +/- 0.08	0.057 +/- 0.023	0.589 +/- 0.078	0.029 +/- 0.009
		Henry Ford	0.039	324	0.12 +/- 0.068	0.035 +/- 0.02	0.493 +/- 0.035	0.054 +/- 0.031	0.611 +/- 0.049	0.092 +/- 0.047
		Average	0.03	281	0.26 +/- 0.088*	0.033 +/- 0.012*	0.529 +/- 0.044*	0.056 +/- 0.019*	0.6 +/- 0.046*	0.06 +/- 0.024*
	Textual	Group Health	0.021	239	0.3 +/- 0.142	0.056 +/- 0.026	0.596 +/- 0.071	0.094 +/- 0.044	0.582 +/- 0.111	0.093 +/- 0.079
		Henry Ford	0.039	324	0.28 +/- 0.095	0.111 +/- 0.036	0.595 +/- 0.048	0.159 +/- 0.052	0.712 +/- 0.057	0.103 +/- 0.033

		Average	0.03	281	0.29 +/- 0.086*	0.083 +/- 0.023*	0.596 +/- 0.043*	0.126 +/- 0.034*	0.647 +/- 0.062*	0.098 +/- 0.043*
Late Surgery	All	Group Health	0.066	878	0.843 +/- 0.035	0.096 +/- 0.004	0.644 +/- 0.019	0.173 +/- 0.007	0.728 +/- 0.025	0.162 +/- 0.022
		Henry Ford	0.05	791	0.532 +/- 0.055	0.145 +/- 0.015	0.683 +/- 0.028	0.228 +/- 0.023	0.767 +/- 0.028	0.192 +/- 0.038
		Average	0.058	834	0.688 +/- 0.033	0.121 +/- 0.008	0.664 +/- 0.017	0.2 +/- 0.012	0.748 +/- 0.019	0.177 +/- 0.022
	Demo.	Group Health	0.066	878	0.696 +/- 0.045	0.095 +/- 0.006	0.616 +/- 0.024	0.167 +/- 0.011	0.645 +/- 0.028	0.127 +/- 0.02
		Henry Ford	0.05	791	0.608 +/- 0.056	0.077 +/- 0.007	0.611 +/- 0.029	0.136 +/- 0.012	0.66 +/- 0.034	0.098 +/- 0.015
		Average	0.058	834	0.652 +/- 0.036*	0.086 +/- 0.005*	0.613 +/- 0.018*	0.152 +/- 0.008*	0.652 +/- 0.022*	0.113 +/- 0.013*
	Codes	Group Health	0.066	878	0.496 +/- 0.048	0.162 +/- 0.015	0.658 +/- 0.024	0.244 +/- 0.022	0.738 +/- 0.023	0.167 +/- 0.021
		Henry Ford	0.05	791	0.418 +/- 0.055	0.167 +/- 0.021	0.654 +/- 0.028	0.238 +/- 0.03	0.77 +/- 0.028	0.18 +/- 0.034
		Average	0.058	834	0.457 +/- 0.036*	0.164 +/- 0.013*	0.656 +/- 0.018*	0.241 +/- 0.019*	0.754 +/- 0.019*	0.173 +/- 0.021*
	Textual	Group Health	0.066	878	0.687 +/- 0.042	0.08 +/- 0.005	0.567 +/- 0.022	0.144 +/- 0.009	0.6 +/- 0.027	0.089 +/- 0.009
		Henry Ford	0.05	791	0.481 +/- 0.057	0.073 +/- 0.008	0.58 +/- 0.029	0.127 +/- 0.015	0.627 +/- 0.033	0.093 +/- 0.016
		Average	0.058	834	0.584 +/- 0.035*	0.077 +/- 0.005*	0.574 +/- 0.018*	0.135 +/- 0.008*	0.614 +/- 0.021*	0.091 +/- 0.009*

Table 2-6. We compared the generalizability performance of the baseline using all data types against each individual data type (i.e. codes, textual, and demographics). For each test system, we evaluated models' performance using the performance metrics. We estimated significance performance between models by bootstrapping 1,000 samples for each test system. For each pair of samples (i.e. one sample from each healthcare system), we calculated different performance metrics for each sample then averaged. We performed a t-test for each performance metric using each model's resulting 1,000 average values; we indicate significance with an asterisk. We used a Bonferroni correction to correct for multiple hypothesis testing when comparing MDL to the three individual networks (0.05/3 = 0.0167). For each prediction task, we underline the model that had the best average performance metric. AUC = Area Under the Curve, AUPRC = Area Under the Precision-Recall Curve

Performance	Target	Prevalence	N	Model	AUC	AUPRC
Classical	Early Surgery	0.024	824	MDL	0.725 +/- 0.040	0.061 +/- 0.014
				Baseline	0.597 +/- 0.050	0.047 +/- 0.011
	Late Surgery	0.044	3,121	MDL	0.833 +/- 0.012	0.241 +/- 0.023
				Baseline	0.840 +/- 0.012	0.266 +/- 0.026
Generalizability	Early Surgery	0.03	281	MDL	0.763 +/- 0.059	0.116 +/- 0.029
				Baseline	0.685 +/- 0.061	0.119 +/- 0.058
	Late Surgery	0.058	834	MDL	0.76 +/- 0.019	0.175 +/- 0.021
				Baseline	0.748 +/- 0.019	0.177 +/- 0.022

Table 2-7. Summary of Major Findings. For each evaluation (i.e. classical and generalizability), we provide the comparison (only using AUC and AUPRC) between MDL and Baseline for each prediction task. AUC = Area Under the Curve, AUPRC = Area Under the Precision-Recall Curve, MDL = Multimodal Deep Learning

References

1. Ginsburg GS, Phillips KA. Precision Medicine: From Science To Value. Health Affairs. 2018;37(5):694–701.

2. Njølstad PR, Andreassen OA, Brunak S, Børglum AD, Dillner J, Esko T, Franks PW, Freimer N, Groop L, Heimer H, Hougaard DM, Hovig E, Hveem K, Jalanko A, Kaprio J, Knudsen GP, Melbye M, Metspalu A, Mortensen PB, Palmgren J, Palotie A, Reed W, Stefánsson H, Stitziel NO, Sullivan PF, Thorsteinsdóttir U, Vaudel M, Vuorio E, Werge T, Stoltenberg C, Stefánsson K. Roadmap for a precision-medicine initiative in the Nordic region. Nat Genet.

2019;51(6):924–30.

3. Wong AH-H, Deng C-X. Precision Medicine for Personalized Cancer Therapy. Int J Biol Sci. 2015;11(12):1410–2.

4. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Information Science and Systems. 2014;2(1):3.

5. Gameiro GR, Sinkunas V, Liguori GR, Auler-Júnior JOC. Precision Medicine: Changing the way we think about healthcare. Clinics. 2018;73:e723.

6. Nayak L, Ray I, De RK. Precision medicine with electronic medical records: from the patients and for the patients. Ann Transl Medicine. 2016;4(1):S61–S61.

7. Kruse CS, Goswamy R, Raval Y, Marawi S. Challenges and Opportunities of Big Data in Health Care: A Systematic Review. Jmir Medical Informatics. 2016;4(4):e38.

8. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. Journal of the American Medical Informatics Association : JAMIA. 2017;24(1):198–208.

9. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assn. 2016;23(5):899–908.

10. DesRoches CM, Campbell EG, Rao SR, Donelan K, Ferris TG, Jha A, Kaushal R, Levy DE, Rosenbaum S, Shields AE, Blumenthal D. Electronic Health Records in Ambulatory Care — A National Survey of Physicians. New Engl J Medicine. 2008;359(1):50–60.

11. Hsiao C-J, Jha AK, King J, Patel V, Furukawa MF, Mostashari F. Office-Based Physicians Are Responding To Incentives And Assistance By Adopting And Using Electronic Health Records. Health Affair. 2017;32(8):1470–7.

12. Harper EM. The Economic Value of Health Care Data. Nurs Administration Q. 2013;37(2):105–8.

13. Liyanage H, Liaw S-T, Lusignan S de. Accelerating the development of an information ecosystem in health care, by stimulating the growth of safe intermediate processing of health information (IPHI). Informatics Prim Care. 2012;20(2):81–6.

14. Bonney S. HIM's role in managing big data: Turning data collected by an EHR into information. J Ahima Am Heal Information Management Assoc. 2013;84(9):62–4.

15. Leventhal R. Trend: big data. Big data analytics: from volume to value. Healthc Informatics Bus Mag Information Commun Syst. 2013;30(2):12, 14.

16. Glaser J, Overhage JM. The role of healthcare IT: becoming a learning organization. Healthc Financial Management J Healthc Financial Management Assoc. 2013;67(2):56–62, 64.

17. Ross MK, Wei W, Ohno-Machado L. ?Big Data? and the Electronic Health Record. Yearb Medical Informatics. 2014;23(01):97–104.

18. Wagholikar KB, Sundararajan V, Deshpande AW. Modeling Paradigms for Medical Diagnostic Decision Support: A Survey and Future Directions. J Med Syst. 2012;36(5):3029–49.

19. Callahan A, Shah NH. Key Advances in Clinical Informatics. 2017;279–91.

20. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. 2018;2(10):719–31.

21. Roberts K, Boland MR, Pruinelli L, Dcruz J, Berry A, Georgsson M, Hazen R, Sarmiento RF, Backonja U, Yu K-H, Jiang Y, Brennan PF. Biomedical informatics advancing the national health agenda: the AMIA 2015 year-in-review in clinical and consumer informatics. J Am Med Inform Assn. 2016;24(e1):e185–90.

22. Deo RC. Machine Learning in Medicine. Circulation. 2015;132(20):1920–30.

23. Hartung MP, Bickle IC, Gaillard F, Kanne JP. How to Create a Great Radiology Report. Radiographics. 2020;40(6):1658–70.

24. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural Language Processing in Radiology: A Systematic Review. Radiology. 2016;279(2):329–43.

25. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assn. 2011;18(5):544–51.

26. Kao A, Poteet SR. Natural Language Processing and Text Mining. 2007;1–7.

27. Zech J, Pain M, Titano J, Badgeley M, Schefflein J, Su A, Costa A, Bederson J, Lehar J, Oermann E. Natural Language–based Machine Learning Models for the Annotation of Clinical Radiology Reports. Radiology. 2018;287(2):171093.

28. Wang Y, Ng K, Byrd RJ, Hu J, Ebadollahi S, Daar Z, deFilippi C, Steinhubl SR, Stewart WF. Early detection of heart failure with varying prediction windows by structured and unstructured data in electronic health records. 2015;2530–3.

29. Scheurwegs E, Luyckx K, Luyten L, Daelemans W, Bulcke TV den. Data integration of structured and unstructured sources for assigning clinical codes to patient stays. J Am Med Inform Assn. 2016;23(e1):e11–9.

30. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, Liu X, Marcus J, Sun M, Sundberg P, Yee H, Zhang K, Zhang Y, Flores G, Duggan GE, Irvine J, Le Q, Litsch K, Mossin A, Tansuwan J, Wang D, Wexler J, Wilson J, Ludwig D, Volchenboum SL, Chou K, Pearson M, Madabushi S, Shah NH, Butte AJ, Howell MD, Cui C, Corrado GS, Dean J. Scalable and accurate deep learning with electronic health records. Npj Digital Medicine. 2018;1(1):18.

31. Zhang D, Yin C, Zeng J, Yuan X, Zhang P. Combining structured and unstructured data for predictive models: a deep learning approach. Bmc Med Inform Decis. 2020;20(1):280.

32. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Scientific Reports. 2016;6(1):26094.

33. Schmeler KM, Lynch HT, Chen L, Munsell MF, Soliman PT, Clark MB, Daniels MS, White KG, Boyd-Rogers SG, Conrad PG, Yang KY, Rubin MM, Sun CC, Slomovitz BM, Gershenson DM, Lu KH. Prophylactic Surgery to Reduce the Risk of Gynecologic Cancers in the Lynch Syndrome. New Engl J Med. 2006;354(3):261–9.

34. Lindor NM, Petersen GM, Hadley DW, Kinney AY, Miesfeldt S, Lu KH, Lynch P, Burke W, Press N. Recommendations for the Care of Individuals With an Inherited Predisposition to Lynch Syndrome: A Systematic Review. Jama. 2006;296(12):1507.

35. Daly MB, Montgomery S, Bingler R, Ruth K. Communicating genetic test results within the family: Is it lost in translation? A survey of relatives in the randomized six-step study. Familial Cancer. 2016;15(4):697–706.

36. Nieuwenhoff HWP van den, Mesters I, Gielen C, Vries NK de. Family communication regarding inherited high cholesterol: Why and how do patients disclose genetic risk? Soc Sci Med. 2007;65(5):1025–37.

37. Vos J, Menko F, Jansen AM, Asperen CJ van, Stiggelbout AM, Tibben A. A whisper-game perspective on the family communication of DNA-test results: a retrospective study on the communication process of BRCA1/2-test results between proband and relatives. Fam Cancer. 2011;10(1):87–96.

38. Koehly LM, Peters JA, Kenen R, Hoskins LM, Ersig AL, Kuhn NR, Loud JT, Greene MH. Characteristics of Health Information Gatherers, Disseminators, and Blockers Within Families at Risk of Hereditary Cancer: Implications for Family Health Communication Interventions. Am J

Public Health. 2009;99(12):2203–9.

39. McGivern B, Everett J, Yager GG, Baumiller RC, Hafertepen A, Saal HM. Family communication about positive BRCA1 and BRCA2 genetic test results. Genetics in Medicine. 2004;6(6):503.

40. Légaré F, Robitaille H, Gane C, Hébert J, Labrecque M, Rousseau F. Improving Decision Making about Genetic Testing in the Clinic: An Overview of Effective Knowledge Translation Interventions. PLOS ONE. 2016;11(3):e0150123.

41. Stoffel EM, Ford B, Mercado RC, Punglia D, Kohlmann W, Conrad P, Blanco A, Shannon KM, Powell M, Gruber SB, Terdiman J, Chung DC, Syngal S. Sharing Genetic Test Results in Lynch Syndrome: Communication With Close and Distant Relatives. Clin Gastroenterol H. 2008;6(3):333–8.

42. Phillips A, Vears DF, Hoyweghen IV, Kuiper J, Borry P. Digital tools for sharing genetic information with family members. Lancet Oncol. 2020;21(7):891–2.

43. Silva BMC, Rodrigues JJPC, Díez I de la T, López-Coronado M, Saleem K. Mobile-health: A review of current state in 2015. J Biomed Inform. 2015;56:265–72.

44. Luo D, Wang P, Lu F, Elias J, Sparks JA, Lee YC. Mobile Apps for Individuals With Rheumatoid Arthritis. Jcr J Clin Rheumatology. 2018;Publish Ahead of Print(NA;):NA;

45. Huckvale K, Prieto JT, Tilney M, Benghozi P-J, Car J. Unaddressed privacy risks in accredited health and wellness apps: a cross-sectional systematic assessment. Bmc Med. 2015;13(1):214.

46. Koes BW, Tulder MW van, Thomas S. Diagnosis and treatment of low back pain. Bmj. 2006;332(7555):1430.

47. Deyo RA, Rainville J, Kent DL. What Can the History and Physical Examination Tell Us

About Low Back Pain? JAMA. 1992;268(6):760–5.

48. Katherine T W, Saeed H, J H Patrick, D R Sean, Pradeep S, T H Hannu, Kathryn J, S C David, P L Curtis, L O Nancy, N M Eric, J S Karen, F K David, H L Patrick, Brent G, R N David, G J Jeffrey. Comparison of Natural Language Processing Rules-based and Machine-learning Systems to Identify Lumbar Spine Imaging Findings Related to Low Back Pain. Academic Radiology [Internet]. 2018;25(11):1422–32. Available from: https://pubmed.ncbi.nlm.nih.gov/29605561

49. Jarvik JJ, Hollingworth W, Heagerty P, Haynor DR, Deyo RA. The Longitudinal Assessment of Imaging and Disability of the Back (LAIDBack) Study: Baseline Data. Spine. 2001;26(10):1158–66.

50. Jarvik JG, Hollingworth W, Heagerty PJ, Haynor DR, Boyko EJ, Deyo RA. Three-Year Incidence of Low Back Pain in an Initially Asymptomatic Cohort: Clinical and Imaging Risk Factors. Spine. 2005;30(13):1541–8.

51. Kim S-Y, Lee I-S, Kim B-R, Lim J-H, Lee J, Koh S-E, Kim SB, Park SL. Magnetic resonance findings of acute severe lower back pain. Ann Rehabilitation Medicine. 2012;36(1):47–54.

52. Jarvik JG, Comstock BA, James KT, Avins AL, Bresnahan BW, Deyo RA, Luetmer PH, Friedly JL, Meier EN, Cherkin DC, Gold LS, Rundell SD, Halabi SS, Kallmes DF, Tan KW, Turner JA, Kessler LG, Lavallee DC, Stephens KA, Heagerty PJ. Lumbar Imaging With Reporting Of Epidemiology (LIRE)—Protocol for a pragmatic cluster randomized trial. Contemporary Clinical Trials. 2015;45(Pt B):157–63.

53. Aronson A. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings / AMIA . Annual Symposium AMIA Symposium. 2001;17–21.

54. Brown PF, deSouza PV, Mercer RL, Pietra VJD, Lai JC. Class-Based n-gram Models of Natural Language. Computational Linguistics. 1992 Jan 1;18:467–79.

55. Mikolov QVL and T. Distributed Representations of Sentences and Documents. Arxiv [Internet]. 2014;abs/1405.4053. Available from: http://arxiv.org/abs/1405.4053

56. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(suppl_1):D267–70.

57. Metathesaurus. In: UMLS® Reference Manual [Internet]. n.d. Available from: https://www.ncbi.nlm.nih.gov/books/NBK9684/

58. Brown AD, Kachura JR. Natural Language Processing of Radiology Reports in Patients With Hepatocellular Carcinoma to Predict Radiology Resource Utilization. J Am Coll Radiol. 2019;16(Hepatology 67 2018):840–4.

59. Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, Kingsbury P, Liu H. A Comparison of Word Embeddings for the Biomedical Natural Language Processing. J Biomed Inform. 2018;87(" Journal of biomedical informatics 2017):12–20.

60. Banerjee I, Chen MC, Lungren MP, Rubin DL. Radiology report annotation using intelligent word embeddings: Applied to multi-institutional chest CT cohort. Journal of Biomedical Informatics. 2018;77:11–20.

61. Cai X, Xie D, Madsen KH, Wang Y, Bögemann SA, Cheung EFC, Møller A, Chan RCK. Generalizability of machine learning for classification of schizophrenia based on resting□state functional MRI data. Hum Brain Mapp. 2019;41(1):172–84.

62. Dreyer KJ, Kalra MK, Maher MM, Hurier AM, Asfaw BA, Schultz T, Halpern EF, Thrall JH. Application of Recently Developed Computer Algorithm for Automatic Classification of Unstructured Radiology Reports: Validation Study1. Radiology. 2005;234(2):323–9.

63. Atlas SJ, Deyo RA. Evaluating and managing acute low back pain in the primary care setting. Journal of General Internal Medicine. 2001;16(2):120–31.

64. Birkmeyer NJO, Weinstein JN, Tosteson ANA, Tosteson TD, Skinner JS, Lurie JD, Deyo R, Wennberg JE. Design of the Spine Patient Outcomes Research Trial (SPORT). Spine. 2002;27(12):1361–72.

65. Thompson K. Programming Techniques: Regular expression search algorithm. Commun Acm. 1968;11(6):419–22.

66. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. J Biomed Inform. 2001;34(5):301–10.

67. M.F. P. An algorithm for suffix stripping. Program [Internet]. 1980;14(3):130–7. Available from: https://doi.org/10.1108/eb046814

68. Harris ZS. Distributional Structure. Word. 2015;10(2–3):146–62.

69. Severance C. The Apache Software Foundation: Brian Behlendorf. Computer. 2012;45(10):8–9.

70. Bejan CA, Vanderwende L, Xia F, Yetisgen-Yildiz M. Assertion modeling and its role in clinical phenotype identification. J Biomed Inform. 2013;46(1):68–74.

71. Dina D-F, J R Willie, R A Alan. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. J Am Med Inform Assn [Internet]. 2017;24(4):841–4. Available from: https://doi.org/10.1093/jamia/ocw177

72. Yetisgen-Yildiz M, Pratt W. The effect of feature representation on MEDLINE document classification. Amia Annu Symposium Proc Amia Symposium Amia Symposium. 2005;849–53.

73. ŘEHŮŘEK R, SOJKA P. Software Framework for Topic Modelling with Large Corpora.

Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks. 2010 May 22;45–50.

74. E.W. J Alistair, J. P Tom, Lu S, H. L Li-wei, Mengling F, Mohammad G, Benjamin M, Peter S, Leo AC, G. M Roger. MIMIC-III, a freely accessible critical care database. Sci Data [Internet]. 2016 May 24;3(1):160035. Available from: https://doi.org/10.1038/sdata.2016.35

75. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J Math Psychol. 1975;12(4):387–415.

76. Perkins NJ, Schisterman EF. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. Am J Epidemiol [Internet]. 2006;163(7):670—675. Available from: https://europepmc.org/articles/PMC1444894

77. Chen P-H, Zafar H, Galperin-Aizenberg M, Cook T. Integrating Natural Language Processing and Machine Learning Algorithms to Categorize Oncologic Response in Radiology Reports. J Digit Imaging. 2018;31(2):178–84.

78. Miotto R, Percha BL, Glicksberg BS, Lee H-C, Cruz L, Dudley JT, Nabeel I. Identifying Acute Low Back Pain Episodes in Primary Care Practice From Clinical Notes: Observational Study. Jmir Medical Informatics. 2020;8(2):e16878.

79. Zhang Y, Li H-J, Wang J, Cohen T, Roberts K, Xu H. Adapting Word Embeddings from Multiple Domains to Symptom Recognition from Psychiatric Notes. Amia Jt Summits Transl Sci Proc Amia Jt Summits Transl Sci. 2018;2017:281–9.

80. V.S. P Serguei, Greg F, Reed M, Yan W, B. M Genevieve. Corpus domain effects on distributional semantic modeling of medical terms. Bioinformatics [Internet]. 2016;32(23):3635–44. Available from: https://doi.org/10.1093/bioinformatics/btw529

81. Martin BI, Deyo RA, Mirza SK, Turner JA, Comstock BA, Hollingworth W, Sullivan SD.

Expenditures and Health Status Among Adults With Back and Neck Problems. Jama. 2008;299(6):656–64.

82. Andersson GB. Epidemiological features of chronic low-back pain. Lancet. 1999;354(9178):581–5.

83. Urits I, Burshtein A, Sharma M, Testa L, Gold PA, Orhurhu V, Viswanath O, Jones MR, Sidransky MA, Spektor B, Kaye AD. Low Back Pain, a Comprehensive Review: Pathophysiology, Diagnosis, and Treatment. Curr Pain Headache R. 2019;23(3):23.

84. Deyo RA, Dworkin SF, Amtmann D, Andersson G, Borenstein D, Carragee E, Carrino J, Chou R, Cook K, DeLitto A, Goertz C, Khalsa P, Loeser J, Mackey S, Panagis J, Rainville J, Tosteson T, Turk D, Korff M, Weiner DK. Report of the NIH Task Force on Research Standards for Chronic Low Back Pain. The Journal of Pain. 2014;15(6):569–85.

85. Amin RM, Andrade NS, Neuman BJ. Lumbar Disc Herniation. Curr Rev Musculoskelet Medicine. 2017;10(4):507–16.

86. Deyo RA, Mirza SK. Herniated Lumbar Intervertebral Disk. New Engl J Medicine. 2016;374(18):1763–72.

87. Genevay S, Atlas SJ. Lumbar Spinal Stenosis. Best Pract Res Clin Rheumatology. 2010;24(2):253–65.

88. Katz JN, Harris MB. Lumbar Spinal Stenosis. New Engl J Med. 2008;358(8):818–25.

89. Mannion AF, Dvorak J, Müntener M, Grob D. A prospective study of the interrelationship between subjective and objective measures of disability before and 2 months after lumbar decompression surgery for disc herniation. Eur Spine J. 2005;14(5):454–65.

90. Machado GC, Ferreira PH, Harris IA, Pinheiro MB, Koes BW, Tulder M van, Rzewuska M, Maher CG, Ferreira ML. Effectiveness of Surgery for Lumbar Spinal Stenosis: A Systematic

Review and Meta-Analysis. Plos One. 2015;10(3):e0122800.

91. Peul WC, Houwelingen HC van, Hout WB van den, Brand R, Eekhof JAH, Tans JTJ, Thomeer RTWM, Koes BW. Surgery versus Prolonged Conservative Treatment for Sciatica. New Engl J Medicine. 2007;356(22):2245–56.

92. Peul WC, Hout WB van den, Brand R, Thomeer RTWM, Koes BW, Group L-THSIPS. Prolonged conservative care versus early surgery in patients with sciatica caused by lumbar disc herniation: two year results of a randomised controlled trial. Bmj. 2008;336(7657):1355–8.

93. Malmivaara A, Slätis P, Heliövaara M, Sainio P, Kinnunen H, Kankare J, Dalin-Hirvonen N, Seitsalo S, Herno A, Kortekangas P, Niinimäki T, Rönty H, Tallroth K, Turunen V, Knekt P, Härkänen T, Hurri H. Surgical or Nonoperative Treatment for Lumbar Spinal Stenosis? Spine. 2007;32(1):1–8.

94. Weinstein JN, Lurie JD, Tosteson TD, Tosteson ANA, Blood EA, Abdu WA, Herkowitz H, Hilibrand A, Albert T, Fischgrund J. Surgical Versus Nonoperative Treatment for Lumbar Disc Herniation. Spine. 2008;33(25):2789–800.

95. Kovacs FM, Urrútia G, Alarcón JD. Surgery Versus Conservative Treatment for Symptomatic Lumbar Spinal Stenosis. Spine. 2011;36(20):E1335–51.

96. Nerland US, Jakola AS, Giannadakis C, Solheim O, Weber C, Nygaard ØP, Solberg TK, Gulati S. The Risk of Getting Worse: Predictors of Deterioration After Decompressive Surgery for Lumbar Spinal Stenosis: A Multicenter Observational Study. World Neurosurg. 2015;84(4):1095–102.

97. Vangen-Lønne V, Madsbu MA, Salvesen Ø, Nygaard ØP, Solberg TK, Gulati S. Microdiscectomy for Lumbar Disc Herniation: A Single-Center Observational Study. World Neurosurg. 2020;137:e577–83.

98. Suri P, Hunter DJ, Jouve C, Hartigan C, Limke J, Pena E, Li L, Luz J, Rainville J. Nonsurgical Treatment of Lumbar Disk Herniation: Are Outcomes Different in Older Adults? J Am Geriatr Soc. 2011;59(3):423–9.

99. Steinmetz MP, Mroz T. Value of Adding Predictive Clinical Decision Tools to Spine Surgery. Jama Surg. 2018;153(7):643.

100. Galbusera F, Casaroli G, Bassani T. Artificial intelligence and machine learning in spine research. Jor Spine. 2019;2(1):e1044.

101. Joshi RS, Lau D, Ames CP. Machine learning in spine surgery: Predictive analytics, imaging applications and next steps. Seminars Spine Surg. 2021;33(2):100878.

102. Wiens J, Shenoy ES. Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. Clin Infect Dis. 2017;66(1):149–53.

103. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2017;19(6):1236–46.

104. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.

105. Norgeot B, Glicksberg BS, Trupin L, Lituiev D, Gianfrancesco M, Oskotsky B, Schmajuk G, Yazdany J, Butte AJ. Assessment of a Deep Learning Model Based on Electronic Health Record Data to Forecast Clinical Outcomes in Patients With Rheumatoid Arthritis. Jama Netw Open. 2019;2(3):e190606.

106. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assn. 2017;24(2):361–70.

107. Chen D, Liu S, Kingsbury P, Sohn S, Storlie CB, Habermann EB, Naessens JM, Larson DW, Liu H. Deep learning and alternative learning strategies for retrospective real-world clinical data. Npj Digital Medicine. 2019;2(1):43.

108. Hebbring SJ. The challenges, advantages and future of phenome□wide association studies. Immunology. 2014;141(2):157–65.

109. Suri P, Stanaway IB, Zhang Y, Freidin MB, Tsepilov YA, Carrell DS, Williams FMK, Aulchenko YS, Hakonarson H, Namjou B, Crosslin DR, Jarvik GP, Lee MT. Genome-wide association studies of low back pain and lumbar spinal disorders using electronic health record data identify a locus associated with lumbar spinal stenosis. Pain. 2021;Publish Ahead of Print(8):2263–72.

110. Martin BI, Lurie JD, Tosteson ANA, Deyo RA, Tosteson TD, Weinstein JN, Mirza SK. Indications for Spine Surgery. Spine. 2014;39(9):769–79.

111. Deyo RA, Bryan M, Comstock BA, Turner JA, Heagerty P, Friedly J, Avins AL, Nedeljkovic SS, Nerenz DR, Jarvik JG. Trajectories of symptoms and function in older adults with low back disorders. Spine. 2015;40(17):1352–62.

112. Kneeman J, Battalio SL, Korpak A, Cherkin DC, Luo G, Rundell SD, Suri P. Predicting Persistent Disabling Low Back Pain in Veterans Affairs Primary Care Using the STarT Back Tool. Pm&amp;r. 2020;

113. Friedly J, Chan L, Deyo R. Increases in Lumbosacral Injections in the Medicare Population. Spine. 2007;32(16):1754–60.

114. Friedly J, Nishio I, Bishop MJ, Maynard C. The Relationship Between Repeated Epidural Steroid Injections and Subsequent Opioid Use and Lumbar Surgery. Arch Phys Med Rehab. 2008;89(6):1011–5.

115. Cartwright DJ. ICD-9-CM to ICD-10-CM Codes: What? Why? How? Adv Wound Care. 2013;2(10):588–92.

116. Bird S, Klein E, Loper E. Natural Language Processing with Python. O’Reilly Media, Inc.;

2009.

117. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Müller A, Nothman J, Louppe G, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: Machine Learning in Python. Arxiv. 2012;

118. Friedman P. Radiologic reporting: structure. Am J Roentgenol. 1983;140(1):171–2.

119. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. J Royal Statistical Soc Ser B Methodol. 1996;58(1):267–88.

120. Bovelstad HM, Nygard S, Storvold HL, Aldrin M, Borgan O, Frigessi A, Lingjaerde OC. Predicting survival from microarray data a comparative study. Bioinformatics. 2007;23(16):2080–7.

121. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Arxiv. 2019;

122. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. Arxiv. 2015;

123. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. Arxiv. 2014;

124. Choi E, Xiao C, Stewart WF, Sun J. MiME: Multilevel Medical Embedding of Electronic Health Records for Predictive Healthcare. Arxiv. 2018;

125. Wang Y, Xu X, Jin T, Li X, Xie G, Wang J. Inpatient2Vec: Medical Representation Learning for Inpatients. 2019 Ieee Int Conf Bioinform Biomed Bibm. 2019;00:1113–7.

126. Steinberg E, Jung K, Fries JA, Corbin CK, Pfohl SR, Shah NH. Language models are an effective representation learning technique for electronic health record data. J Biomed Inform. 2021;113:103637.

127. King G, Zeng L. Logistic Regression in Rare Events Data. Polit Anal. 2001;9(2):137–63.

128. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research [Internet]. 2014; Available from: https://www.cs.toronto.edu/~hinton/absps/JMLRdropout.pdf

129. Zhang Y, Wallace B. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. 2015;

130. André A, Peyrou B, Carpentier A, Vignaux J-J. Feasibility and Assessment of a Machine Learning-Based Predictive Model of Outcome After Lumbar Decompression Surgery. Global Spine J. 2020;219256822096937.

131. Azimi P, Benzel EC, Shahzadi S, Azhari S, Mohammadi HR. The prediction of successful surgery outcome in lumbar disc herniation based on artificial neural networks. J Neurosurg Sci. 2016;60(2):173–7.

132. Keeney BJ, Fulton-Kehoe D, Turner JA, Wickizer TM, Chan KCG, Franklin GM. Early Predictors of Lumbar Spine Surgery After Occupational Back Injury. Spine. 2013;38(11):953–64.

133. Azad TD, Ehresman J, Ahmed AK, Staartjes VE, Lubelski D, Stienen MN, Veeravagu A, Ratliff JK. Fostering reproducibility and generalizability in machine learning for clinical prediction modeling in spine surgery. Spine J. 2021;21(10):1610–6.

134. Marshall DA, MacDonald KV, Robinson JO, Barcellos LF, Gianfrancesco M, Helm M, McGuire A, Green RC, Douglas MP, Goldman MA, Phillips KA. The price of whole-genome

sequencing may be decreasing, but who will be sequenced? Pers Med. 2017;14(3):203–11.

135. Montanez K, Berninger T, Willis M, Harding A, Lutgendorf MA. Genetic testing costs and compliance with clinical best practices. J Genet Couns. 2020;

136. Schwartz GF, Hughes KS, Lynch HT, Fabian CJ, Fentiman IS, Robson ME, Domchek SM, Hartmann LC, Holland R, Winchester DJ. Proceedings of the International Consensus Conference on Breast Cancer Risk, Genetics, & Risk Management, April, 2007. Breast J. 2009;15(1):4–16.

137. Jong MA, Bock GH de, Asperen CJ van, Mourits MJE, Hullu JA de, Kets CM. Germline BRCA1/2 mutation testing is indicated in every patient with epithelial ovarian cancer: A systematic review. Eur J Cancer. 2016;61:137–45.

138. Norquist BM, Harrell MI, Brady MF, Walsh T, Lee MK, Gulsuner S, Bernards SS, Casadei S, Yi Q, Burger RA, Chan JK, Davidson SA, Mannel RS, DiSilvestro PA, Lankes HA, Ramirez NC, King MC, Swisher EM, Birrer MJ. Inherited Mutations in Women With Ovarian Carcinoma. Jama Oncol. 2015;2(4):1–9.

139. Stewart BW, Wild CP. World Cancer Report [Internet]. International Agency for Research on Cancer; n.d. Available from: https://www.drugsandalcohol.ie/28525/1/World%20Cancer%20Report.pdf

140. Society AC. How Common Is Breast Cancer? [Internet]. n.d. Available from: https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html

141. Longo DL, Hartmann LC, Lindor NM. The Role of Risk-Reducing Surgery in Hereditary Breast and Ovarian Cancer. New Engl J Medicine. 2016;374(5):454–68.

142. Istepanian RSH, Lacal JC. Emerging mobile communication technologies for health: some imperative notes on m-health. Proc 25th Annu Int Conf Ieee Eng Medicine Biology Soc Ieee Cat

03ch37439. 2003;2:1414–6.

143. Free C, Phillips G, Galli L, Watson L, Felix L, Edwards P, Patel V, Haines A. The Effectiveness of Mobile-Health Technology-Based Health Behaviour Change or Disease Management Interventions for Health Care Consumers: A Systematic Review. Plos Med. 2013;10(1):e1001362.

144. Devitt K, Roker D. The Role of Mobile Phones in Family Communication. Child Soc. 2009;23(3):189–202.

145. Storch SL, Juarez-Paz AVO. The role of mobile devices in 21st-century family communication. Mob Media Commun. 2018;7(2):248–64.

146. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, Sanderson SC, Kannry J, Zinberg R, Basford MA, Brilliant M, Carey DJ, Chisholm RL, Chute CG, Connolly JJ, Crosslin D, Denny JC, Gallego CJ, Haines JL, Hakonarson H, Harley J, Jarvik GP, Kohane I, Kullo IJ, Larson EB, McCarty C, Ritchie MD, Roden DM, Smith ME, Böttinger EP, Williams MS. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med. 2013;15(10):761–71.

147. Holden RJ, Karsh B-T. The Technology Acceptance Model: Its past and its future in health care. J Biomed Inform. 2010;43(1):159–72.

148. Lewis JR. Psychometric Evaluation of the Post-Study System Usability Questionnaire: The PSSUQ. Proc Hum Factors Soc Annu Meet. 1992;36(16):1259–60.