DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY

Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives

Citation

Gehrmann, Sebastian, Franck Dernoncourt, Yeran Li, Eric T. Carlson, Joy T. Wu, Jonathan Welt, John Foote, et al. 2018. “Comparing Deep Learning and Concept Extraction Based Methods for Patient Phenotyping from Clinical Narratives.” Edited by Jen-Hsiang Chuang. PLOS ONE 13 (2) (February 15): e0192360. doi:10.1371/journal.pone.0192360.

Published version

https://doi.org/10.1371/journal.pone.0192360

Link

http://nrs.harvard.edu/urn-3:HUL.InstRepos:35059946

Terms of use

This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFlOWI2M2EwYTg

Accessibility

https://accessibility.huit.harvard.edu/digital-accessibility-policy

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story

PLOS ONE

RESEARCH ARTICLE

Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives

Sebastian Gehrmann[1,2]*, Franck Dernoncourt[1,3,4], Yeran Li[1,5], Eric T. Carlson[1,6], Joy T. Wu[1,5], Jonathan Welt[1,7], John Foote Jr.[1,8], Edward T. Moseley[1,9], David W. Grant[1,10], Patrick D. Tyler[1,11], Leo A. Celi[1,3]

1 MIT Critical Data, Laboratory for Computational Physiology, Cambridge, MA, United States of America, **2** Harvard SEAS, Harvard University, Cambridge, MA, United States of America, **3** Massachusetts Institute of Technology, Cambridge, MA, United States of America, **4** Adobe Research, San Jose, CA, United States of America, **5** Harvard T.H. Chan School of Public Health, Cambridge, MA, United States of America, **6** Philips Research North America, Cambridge, MA, United States of America, **7** Wellman Center for Photomedicine, Massachusetts General Hospital, Boston, MA, United States of America, **8** Tufts University School of Medicine, Cambridge, MA, United States of America, **9** College of Science and Mathematics, University of Massachusetts, Boston, MA, United States of America, **10** Department of Surgery, Division of Plastic and Reconstructive Surgery, Washington University School of Medicine, St. Louis, MO, United States of America, **11** Department of Internal Medicine, Beth Israel Deaconess Medical Center, Boston, MA, United States of America

* gehrmann@seas.harvard.edu

Check for updates

OPEN ACCESS

Citation: Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, Welt J, et al. (2018) Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PLoS ONE 13(2): e0192360. https://doi.org/10.1371/journal.pone.0192360

Editor: Jen-Hsiang Chuang, Centers for Disease Control, TAIWAN

Received: June 16, 2017

Accepted: January 21, 2018

Published: February 15, 2018

Copyright: © 2018 Gehrmann et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Data for this study are available on physionet under the repository name "Annotated Clinical Texts (ACT) from MIMIC" (URL: https://physionet.org/works/AnnotatedClinicalTextsACTfromMIMIC/). Code and processed data is available on github (URL: https://github.com/sebastianGehrmann/phenotyping). All data besides the annotated texts can be found in the MIMIC-III database and was not modified by the authors. The MIMIC-III database is an openly available, de-identified dataset developed by the MIT Lab for Computational Physiology. Of the

Abstract

In secondary analysis of electronic health records, a crucial task consists in correctly identifying the patient cohort under investigation. In many cases, the most valuable and relevant information for an accurate classification of medical conditions exist only in clinical narratives. Therefore, it is necessary to use natural language processing (NLP) techniques to extract and evaluate these narratives. The most commonly used approach to this problem relies on extracting a number of clinician-defined medical concepts from text and using machine learning techniques to identify whether a particular patient has a certain condition. However, recent advances in deep learning and NLP enable models to learn a rich representation of (medical) language. Convolutional neural networks (CNN) for text classification can augment the existing techniques by leveraging the representation of language to learn which phrases in a text are relevant for a given medical condition. In this work, we compare concept extraction based methods with CNNs and other commonly used models in NLP in ten phenotyping tasks using 1,610 discharge summaries from the MIMIC-III database. We show that CNNs outperform concept extraction based methods in almost all of the tasks, with an improvement in F1-score of up to 26 and up to 7 percentage points in area under the ROC curve (AUC). We additionally assess the interpretability of both approaches by presenting and evaluating methods that calculate and extract the most salient phrases for a prediction. The results indicate that CNNs are a valid alternative to existing approaches in patient phenotyping and cohort identification, and should be further investigated. Moreover, the deep learning approach presented in this paper can be used to assist clinicians during

authors, only LAC was involved in the development of the database. Access can be requested on PhysioNet after completion of a CITI training on "Data or Specimens Only Research". A detailed tutorial how to access and install the data can be found at https://mimic.physionet.org/gettingstarted/access/.

Funding: Franck Dernoncourt is supported by a grant from Philips Research. Leo Anthony Celi is supported by the R01 grant EB017205-01A1 from the National Institute of Biomedical Imaging (http://grantome.com/grant/NIH/R01-EB017205-01A1) and Biomedical Engineering. The content is solely the responsibility of the authors and does not necessarily represent the official views of Philips Research or the National Institute of Bioimaging and Biomedical Engineering. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

chart review or support the extraction of billing codes from text by identifying and highlighting relevant phrases for various medical conditions.

Introduction

The secondary analysis of data from electronic health records (EHRs) is crucial to better understand the heterogeneity of treatment effects and to individualize patient care [1]. With the growing adoption rate of EHRs [2], researchers gain access to rich data sets, such as the Medical Information Mart for Intensive Care (MIMIC) database [3, 4], and the Informatics for Integrating Biology and the Bedside (i2b2) datamarts [5–10]. These data sets can be explored and mined in numerous ways [11]. EHR data comprise both structured data such as International Classification of Diseases (ICD) codes, laboratory results and medications, and unstructured data such as clinician progress notes. While structured data do not require complex processing prior to statistical tests and machine learning tasks, the majority of data exist in unstructured form [12]. Natural language processing methods can extract this valuable data, which in conjunction with analyzing structured data can lead to a better understanding of health and diseases [13] and to a more accurate phenotyping of patients to compare tests and treatments [14–16]. Patient phenotyping is a classification task for determining whether a patient has a medical condition or for pinpointing patients who are at risk for developing one. Further, intelligent applications for patient phenotyping can support clinicians by reducing the time they spend on chart reviews, which takes up a significant fraction of their daily workflow [17, 18].

A popular approach to patient phenotyping using NLP is based on extracting medical phrases from texts and using them as input to build a predictive model [19]. The dictionary of relevant phrases is often task-specific and its development requires significant effort and a deep understanding of the task from domain experts [20]. A different approach is to develop a fully rule-based algorithm for each condition [21]. Due to the laborious task required of clinicians to build a generalizable model for patient phenotyping, models for automated classification using NLP are rarely developed outside of the research area. However, utilizing recent developments in deep learning for phenotyping might prove to be a generalizable approach with less intense domain expert involvement. Applications of deep learning for other tasks in healthcare have shown promising results; examples include mortality prediction [22], patient note de-identification [23], skin cancer detection [24], and diabetic retinopathy detection [25].

A possible drawback to deep learning models is their lack of interpretability. Interpretability means how easily one can understand how a model arrived at a prediction [26]. This is crucial for healthcare applications since results can directly impact decisions about patients health. Furthermore, clinicians have intimate pre-existing knowledge and thus expect applications to support their decision making as opposed to make decisions for them. Therefore, interpretable models are required so that clinicians can trust and control their results [27]. Moreover, the European Union is considering regulations that require algorithms to be interpretable [28]. While much work has been done to understand deep learning NLP models and develop understandable models [29–31], their complex interactions between inputs are inherently less interpretable than linear models that use predefined phrase dictionaries.

In this work, we assess convolutional neural networks (CNNs) as an approach to text-based patient phenotyping. CNNs are designed to identify phrases in text that lead to a positive or negative classification, similar to the phrase dictionary approach, and outperform approaches

to classification problems in other domains [32–34]. We compare CNNs to entity extraction systems using the Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES) [35], and other NLP methods such as logistic regression models using n-gram features. Using a corpus of 1,610 discharge summaries that were annotated for ten different phenotypes, we show that CNNs outperform both extraction-based and n-gram-based methods. Finally, we evaluate the interpretability of the model by assessing the learned phrases that are associated with each phenotype and compare them to the phrase dictionaries developed by clinicians.

Background

Accurate patient phenotyping is required for secondary analysis of EHRs to correctly identify the patient cohort and to better identify the clinical context [36, 37]. Studies employing a manual chart review process for patient phenotyping are naturally limited to a small number of preselected patients. Therefore, NLP is necessary to identify information that is contained in text but may be inconsistently captured in the structured data, such as recurrence in cancer [20, 38], whether a patient smokes [5], classification within the autism spectrum [39], or drug treatment patterns [40]. However, unstructured data in EHRs, for example progress notes or discharge summaries, are not typically amenable to simple text searches because of spelling mistakes, synonyms, and ambiguous terms [41]. To help address these issues, researchers utilize dictionaries and ontologies for medical terminologies such as the unified medical language system (UMLS) [42] and the systematized nomenclature of medicine—clinical terms (SNOMED CT) [43].

Examples of systems that employ such databases and extract concepts from text are the KnowledgeMap Concept Identifier (KMCI) [44], MetaMap [45], Medlee [46], MedEx [47], and the cTAKES. These systems identify phrases within a text that correspond to medical entities [35, 48]. This significantly reduces the work required from researchers, who previously had to develop task-specific extractors [49]. Extracted entities are typically filtered to only include concepts related to the patient phenotype under investigation and either used as features for a model that predicts whether the patient fits the phenotype, or as input for rule-based algorithms [19, 39, 50]. Liao et al. [13] describe the process of extraction, rule-generation and prediction as the general approach to patient phenotyping using the cTAKES [14, 51–53], and test this approach on various data sets [54]. The role of clinicians in this task is both to annotate data and to develop a task-specific dictionary of phrases that are relevant to a patient phenotype. Improving existing approaches often requires significant additional time-investment from the clinicians, for example by developing and combining two separate phrase-dictionaries for pathology documents and clinical documents [20]. The cost and time required to develop these algorithms limit their applicability to large or repeated tasks. While a usable system would offset the development costs, it does not address the problem that a specialized NLP system would have to be developed for every task in a hospital. Moreover, algorithms often do not transfer well between different hospitals, warranting extensive transferability studies [55]. The deep learning based approach we evaluate in this paper does not require any hand-crafted input and can easily be retrained with new data. This can potentially increase the transferability of studies while removing the time required to develop new phrase-dictionaries.

Materials and methods

Data

The notes used in this study are extracted from the MIMIC-III database. MIMIC-III contains de-identified clinical data of over 53,000 hospital admissions for adult patients to the intensive

care units (ICU) at the Beth Israel Deaconess Medical Center from 2001 to 2012. The dataset comprises several types of clinical notes, including discharge summaries (n = 52,746) and nursing notes (n = 812,128). We focus on the discharge summaries since they are the most informative for patient phenotyping [56]. More specifically, we investigate phenotypes that may associate a patient with being a 'frequent flyer' in the ICU (defined as >= 3 ICU visits within 365 days). As many as one third of readmissions have been suggested to be preventable; identifying modifiable risk factors is a crucial step to reducing them [57]. We extracted the discharge summary of the first visit from 415 ICU frequent flyers in MIMIC-III, as well as 313 randomly selected summaries from later visits of the same patients. We additionally selected 882 random summaries from patients who are not frequent flyers, yielding a total of 1,610 notes. The cTAKES output for these notes contains a total of 11,094 unique CUIs.

All 1,610 notes were annotated for the ten phenotypes described in Table 1. The table shows the definitions for each investigated phenotype. To ensure high-quality labels and minimize errors, each note was labeled at least twice for each phenotype. Annotators include two clinical researchers (ETM, JW), two junior medical residents (JF, JTW), two senior medical residents (DWG, PDT), and a practicing intensive care medicine physician (LAC). In the case that the annotators were unsure, one of the senior clinicians (DWG or PDT) decided on the final label. The table further shows the number of occurrences of each phenotype as a measure of dataset imbalance. The frequency varies from 126 to 460 cases, which corresponds to between 7.5% and 28.6% of the dataset. Finally, we provide the Cohen's Kappa measure for inter-rater agreement. While well specified phenotypes such as depression have a very high agreement (0.95), other phenotypes, such as chronic neurologic dystrophies, have a lower agreement 0.71 and required more interventions by senior clinicians.

Table 1. The ten different phenotypes used for this study. The first column shows the name of the phenotype, the second column shows the number of positive examples our of the total 1,610 notes, and the third shows the κ coefficient as inter-rater agreement measure. The last column lists the definition for each phenotype that was used to identify and annotate the phenotype.

Phenotype]	#pos.	κ	Definition
Adv. / Metastatic Cancer	161	0.83	Cancers with very high or imminent mortality (pancreas, esophagus, stomach, cholangiocarcinoma, brain); mention of distant or multi-organ metastasis, where palliative care would be considered (prognosis < 6 months).
Adv. Heart Disease	275	0.82	Any consideration for needing a heart transplant; description of severe aortic stenosis (aortic valve area < 1.0cm^2), severe cardiomyopathy, Left Ventricular Ejection Fraction (LVEF) <= 30%. Not sufficient to have a medical history of congestive heart failure (CHF) or myocardial infarction (MI) with stent or coronary artery bypass graft (CABG) as these are too common.
Adv. Lung Disease	167	0.81	Severe chronic obstructive pulmonary disease (COPD) defined as Gold Stage III-IV, or with a forced expiratory volume during first breath (FEV1) < 50% of normal, or forced vital capacity (FVC) < 70%, or severe interstitial lung disease (ILD), or Idiopathic pulmonary fibrosis (IPF).
Chronic Neurologic Dystrophies	368	0.71	Any chronic central nervous system (CNS) or spinal cord diseases, included/not limited to: Multiple sclerosis (MS), amyotrophic lateral sclerosis (ALS), myasthenia gravis, Parkinson's Disease, epilepsy, history of stroke/cerebrovascular accident (CVA) with residual deficits, and various neuromuscular diseases/dystrophies.
Chronic Pain	321	0.83	Any etiology of chronic pain, including fibromyalgia, requiring long-term opioid/narcotic analgesic medication to control.
Alcohol Abuse	196	0.86	Current/recent alcohol abuse history; still an active problem at time of admission (may or may not be the cause of it).
Substance Abuse	155	0.86	Include any intravenous drug abuse (IVDU), accidental overdose of psychoactive or narcotic medications,(prescribed or not). Admitting to marijuana use in history is not sufficient.
Obesity	126	0.94	Clinical obesity. BMI > 30. Previous history of or being considered for gastric bypass. Insufficient to have abdominal obesity mentioned in physical exam.
Psychiatric disorders	295	0.91	All psychiatric disorders in DSM-5 classification, including schizophrenia, bipolar and anxiety disorders, other than depression.
Depression	460	0.95	Diagnosis of depression; prescription of anti-depressant medication; or any description of intentional drug overdose, suicide or self-harm attempts.

https://doi.org/10.1371/journal.pone.0192360.t001

Concept-extraction based methods

We use cTAKES to extract concepts from each note. In cTAKES, sentences and phases are first split into tokens (individual words). Then, tokens with variations (e.g. plural) are normalized to their base form. The normalized tokens are tagged for their part-of-speech (e.g. noun, verb), and a shallow parse tree is constructed to represent the grammatical structure of a sentence. Finally, a named-entity recognition algorithm uses this information to detect named entities for which a concept unique identifier (CUI) exists in UMLS [43].

Related approaches then use relevant concepts in a note as input to machine learning algorithms to directly learn to predict a phenotype [58, 59]. We specify two different approaches to using the cTAKES output. The first approach uses the complete list of extracted CUIs as input to further processing steps. In the second approach, clinicians specify a dictionary comprising all clinical concepts that are relevant to the desired phenotype (e.g. Alcohol Abuse) as described by Carrell et al. [20].

Our predictive models replicate the process as described by Liao et al. [13]. Each note is represented as a bag-of-CUIs by counting the number of occurrences of each of the CUIs. Due to the fact that cTAKES detects negations, occurrences of negated and non-negated CUIs are counted separately. These features are then transformed using the term frequency–inverse document frequency (TF-IDF). Compared to the bag-of-CUIs, or the bag-of-words of a note as described by Halpern et al. [16], the TF-IDF of the features reflects the importance of a feature to a note. For an accurate comparison to approaches in literature, we train a random forest (RF), a naive Bayes (NB), and a logistic regression (LR) model with these features.

Convolutional neural networks

We use a convolutional neural network (CNN) for text classification to represent deep learning methods, replicating the architecture proposed by Collobert et al. and Kim [33, 60]. The idea behind convolutions in computer vision is to learn filters that transform adjacent pixels into single values [61]. Equivalently, a CNN for NLP learns which combinations of adjacent words are associated with a given concept. An overview of our architecture is shown in Fig 1.

In a CNN, a text is first represented as a sequence of word embeddings in which each word is projected into a distributed representation. A word $\mathbf{x}_i \in \mathbb{R}^k$ is the k-dimensional embedding vector for the i-th word in a text. Consequently, a text of length n is represented the concatenation of its word embeddings $\mathbf{x}_{1:n} = \mathbf{x}_1 \oplus \mathbf{x}_2 \oplus \ldots \oplus \mathbf{x}_n$ (where $\oplus$ is the concatenation operation). Word embeddings have shown to improve performance on other tasks based on EHRs, for example named-entity recognition [62]. Words that occur in similar contexts are trained to have similar word embeddings. Therefore, misspellings, synonyms and abbreviations of an original word learn similar embeddings, which lead to similar results. Consequently, a database of synonyms and common misspellings is not required [20]. Word embeddings can be pre-trained on a larger corpus of texts, which improves results of the NLP system and reduces the amount of data required to train a model [63, 64]. We pre-train our embeddings with word2vec [65] on all discharge notes available in the MIMIC-III database [4].

The embedded text is used as input to the convolutional layer. Convolutions detect a signal from a combination of adjacent inputs. Each convolutional operation applies a filter of trained parameters $w \in \mathbb{R}^{hk}$ to an input-window of width h. A resulting feature c_i is computed as $c_i = f(\mathbf{w} \cdot \mathbf{x}_{i:i+h-1} + b)$. In this equation, $b \in \mathbb{R}$ represents a bias term, and f a non-linear function, in our case a rectified linear unit $f(x) = \max(0, x)$. A filter is applied to every possible word window in the input to produce a feature map $\mathbf{c} = [c_1, c_2, \ldots, c_{n-h+1}]$. This feature map is then reduced to a single value using a pooling operation. More specifically, we use max-over-time-pooling to extract the most predictive value $\hat{c} = \max(\mathbf{c})$ [60]. We combine multiple

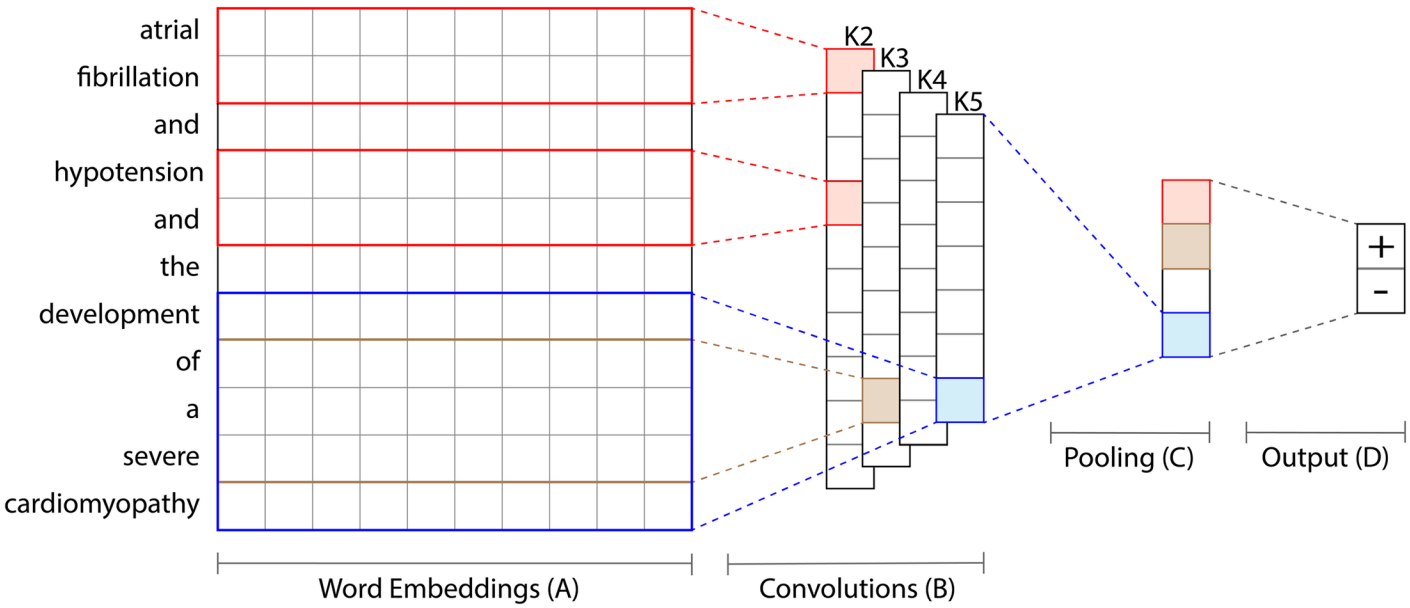

Fig 1. Overview of the basic CNN architecture. (A) Each word within a discharge note is represented as its word embedding. In this example, both instances of the word "and" will have the same embedding. (B) Convolutions of different widths are used to learn filters that are applied to word sequences of the corresponding length. The convolution K2 with width 2 in the example looks at all 10 combinations of neighboring two words and output one value each. There can be multiple feature maps for each convolution width. (C) The multiple resulting vectors are reduced to only the highest value (the one with the most signaling power) for each of the different convolutions. (D) The final prediction ("Does the phenotype apply to the patient?") is made by computing a weighted combination of the pooled values and applying a sigmoid function, similar to a logistic regression. This figure is adapted with permission from Kim [33].

https://doi.org/10.1371/journal.pone.0192360.g001

convolutions per length and of different lengths to evaluate phrases from one to five words long, as illustrated in Fig 1. All convolutions use the same word embeddings as input and only differ in the filters they learn. The combination of many filters of varying length results in multiple outputs $\mathbf{z} = [\hat{c}_1, \ldots, \hat{c}_m]$. A final probability whether the text refers to a patient with a certain condition is computed as $y = \sigma(\mathbf{w} \cdot \mathbf{z} + b)$, with two trainable parameters $\mathbf{w}$ and b and the sigmoid function σ. We train a separate CNN for each phenotype to minimize the negative conditional log-likelihood of training data $\mathcal{L}(\theta) = -\sum_{i=1}^{n} \log p(y_i|\mathbf{x}_i; \theta)$ for a set of parameters θ.

Other baselines

We further investigate a number of baselines to compare with the more complex approaches. We start with a bag-of-words based logistic regression and gradually increase the complexity of the baselines until we reach the CNN. This provides an overview of how adding more complex features impact the performance in this task. Moreover, this investigation shows what factors contribute most to the CNN performance.

Bag of words. The simplest possible representation for a note is a bag of words (BoW) which counts phrases of length 1. Let $\mathcal{F}$ denote the vocabulary of all words, and f_i the word at position i. Let further $\delta(f_i) \in \mathbb{R}^{1\times|\mathcal{F}|}$ be a one-hot vector with a one at position f_i. Then, a note $\mathbf{x}$ is represented as $\mathbf{x} = \sum_i \delta(f_i)$. A prediction is made by computing $y = \sigma(\mathbf{w} \cdot \mathbf{x} + b)$, where $\mathbf{w}$ and b are trainable parameters.

n-grams. The bag-of-words approach can be extended to include representations for longer phrases, also called n-grams, as well. Consider a note "The sick patient". While a bag-of-

words approach considers each word separately, a two-gram model additionally considers all possible phrases of length two. Thus, the model would also consider the phrases "The sick" and "sick patient". Since the number of possible phrases grows exponentially in the size of the vocabulary, the data for longer phrases becomes very sparse and the n-gram size can't be increased too much. In this study, we show results for models with phrase lengths of up to 5.

Embedding-based logistic regression. A typical CNN differs from n-gram based models in both feature representation and model architecture. To study whether models simpler than a CNN have the same expressive power with the same feature representation, we modify the LR to use the same word embeddings as the CNN. Here, a note is represented as a sequence of its word embeddings $\mathbf{x}_{1:n} = \mathbf{x}_1 \oplus \mathbf{x}_2 \oplus \ldots \oplus \mathbf{x}_n$. However, using this representation as input to a LR means that every word is sensitive to its location in a text. An occurrence of "heart" at position 5 would use different parameters in w than the same word at position 10. Therefore, we change the model to use the same weight vector w for each word embedding in a text, computing the score as $\sum_{i=1}^{n} \mathbf{w} \cdot \mathbf{x}_i$. The representation also poses the challenge that inputs vary in length. Thus, we normalize the score by dividing by the length n. Unfortunately, the capacity of this adjusted logistic regression model is insufficient. The AUC of this model is close to 0.5 for all phenotypes, which means that it is not better than chance. Since a typical note can be longer than 4,000 words, all the expressive terms are smoothed out. Using max-pooling instead of averaging as an approach to mitigating this problem has the same result. We thus omit this baseline from further results.

CNN without convolutions. In order to account for the problem that relevant features are smoothed out by the number of features in the model, we train several weight parameters $\mathbf{w}$ instead of only one and combine the pooled results by concatenating them. Then, we take the resulting vector as input to another logistic regression. This architecture is equivalent to a CNN with a convolutional width of one, for which we show results in S1 Table.

Interpretability

The inability of humans to understand predictions of complex machine learning models poses a difficult challenge when such approaches are used in healthcare [26]. Therefore, we consider it crucial that well-performing approaches should be understood and trusted by those who use them. Moreover, bias in the sources of data could lead the model to learn false implications. One such example of bias was in mortality prediction among patients with pneumonia where asthma was found to increase survival probability [27, 66]. This result was due to an institutional practice of admitting all patients with pneumonia and a history of asthma to the ICU regardless of disease severity, so that a history of asthma was strongly correlated with a lower illness severity. To measure the interpretability of each approach, we consider the most frequently used approach in text-classification. In this approach, users are shown a list of phrases or concepts that are most salient for a particular prediction; this list either manifests as an actual list or as highlights of the original text [67, 68].

As a baseline we consider the filtered cTAKES random forest approach (other baselines can be treated equivalently). In the filtered cTAKES approach, clinicians ensure that all features directly pertain to the specific phenotype [22]. We rank the importance of each remaining CUI using the gini importance of the trained model [69]. The resulting ranking is a direct indication of the globally most relevant CUIs. An individual document can be analyzed by ranking only the importance of CUIs that occur in this document.

For the CNN, we propose a modified version of the saliency as defined by Li et al. to compute the most relevant phrases [70]. They define the saliency for a neural network as the norm

of the gradient of the Loss function for a particular prediction with respect to an input x_i as

$$S_1(\mathbf{x}_i) = \left|\frac{\partial L(1, \hat{y}}{\partial \mathbf{x}_i}\right|$$

In the CNN case, an input is only a single word or a single dimension of its embedding. Thus, we extend the saliency definition to instead compute the gradient with respect to the learned feature maps, which we call phrase-saliency. This approximates how much a phrase contributed to a prediction instead of a single word.

$$S_1(\mathbf{x}_{i:i+h-1}) = \left|\frac{\partial L(1, \hat{y}}{\partial \mathbf{c}_i}\right|$$

Since we employ multiple feature maps of different widths, we compute the phrase-saliency across all of them, and use those with maximum value. The saliency is measured on a per-document basis. To arrive at the globally most relevant phrases, we iterate over all documents and measure the phrases that had the highest phrase-saliency while removing duplicate phrases. Alternative methods not considered in this work search the local space around an input [71], or compute a layer-wise backpropagation [72–75].

Evaluation-quantitative performance measures

We evaluate the precision, recall, F1-score, and area under the ROC curve (AUC) of all models as a quantitative measure. The F-score is derived from the confusion matrix for the results on the test set. A confusion matrix contains four counts: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The precision P is the fraction of correct predictions out of all the samples that were predicted to be positive $\frac{\mathrm{TP}}{\mathrm{TP+FP}}$. The recall R is the percentage of true positive predictions in relation to all the predictions that should have been predicted as positive $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$. The F1-score is the harmonic mean of both precision and recall $2 * \frac{P*R}{P+R}$.

For all models, the data is randomly split into a training, validation, and test set. 70% of the labeled data is used as the training set, 10% as validation set and 20% as test set. While splitting, we ensure that patients' notes stay within the set, so that all discharge notes in the test set are from patients not previously seen by the model. The reported numbers across different models for the same phenotype are obtained from testing on the same test set. The validation set is used to choose the hyperparameters for the models. A detailed description of hyperparameters is shown in S1 Text.

All models are trained separately for each of the phenotypes to examine the results in isolation. We present the results for model configurations with the highest F1-score for each model in the main part of the paper. Additional results for different convolution widths for the CNN are shown in S1 Table, and for different models using cTAKES in S2 Table For the other baselines, we present the results of a bag of words representation, and the best performing n-gram length, and additional results in S3 Table.

In summary, we present results for the following approaches:

CNN The convolutional neural network with best performing convolution width

BoW Baseline using a bag of words representation of a note and logistic regression

n-gram Baseline using an n-gram representation of a note and logistic regression

cTAKES full The best performing model that uses the full output from cTAKES

cTAKES filter The best performing model using the filtered CUI-list from cTAKES

Assessing interpretability

In order to evaluate how understandable the predictions of the different approaches are, we conducted a study of the globally most relevant phrases and CUIs. For each phenotype, we computed the five most relevant features, yielding a total of 50 phrases and 50 CUIs. We then asked clinicians to rate the features on a scale from 0 to 3 with the following descriptions for each rating:

0 The phrase/CUI is unrelated to the phenotype.

1 The phrase is associated with the concept subjectively from clinical experience, but is not directly related (e.g. alcohol abuse for psychiatric disorder).

2 The phrase has to do with the concept, but is not a definite indicator of its existence (e.g. a medication).

3 The phrase is a direct indicator of the concept or very relevant (e.g. history of COPD for advanced lung disease).

The features were shown without context other than the name of the phenotype. We additionally provided an option to enter free text comments for each phenotype. We note that all participating clinicians were involved with annotation of the notes and are aware of the definitions for the phenotypes. They were not told about the origin of the phrases before rating them in order to prevent potential bias. In total, we collected 300 ratings, an average of three per feature.

Results

We show an overview of the F1-scores for different models and phenotypes in Fig 2. For almost all phenotypes, the CNN outperforms all other approaches. For some of the phenotypes such as Obesity and Psychiatric Disorders, the CNN outperforms the other models by a large margin. A χ^2 test confirms that the CNN's improvements over both the filtered and the full cTAKES models are statistically significant at a 0.01 level. There is only a minimal improvement when using the filtered cTAKES model, which requires much more effort from clinicians, over the full cTAKES model. The χ^2 test confirms that there is no statistically significant improvement of this method on our data with a p-value of 0.86. We also note that the TF-IDF transformation of the CUIs yielded a small average improvement in AUC of 0.02 ($\sigma = 0.03$) over all the considered models.

In the detailed results, shown in Table 2, we observe that the CNN has the best performance on almost all of the evaluated values. The n-gram and bag-of-words based methods are consistently weaker than the CNN, corroborating the findings in literature that word embeddings

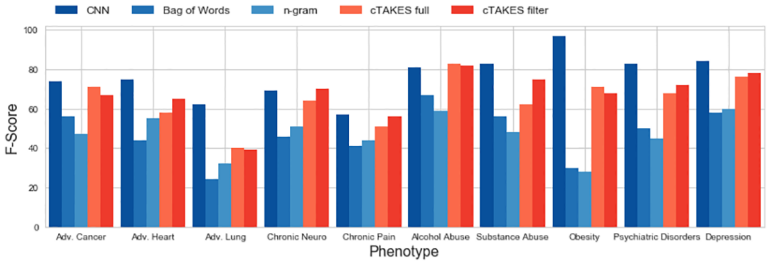

Fig 2. Comparison of achieved F1-scores across all tested phenotypes. The left three models directly classify from text, the right two models are concept-extraction based. The CNN outperforms the other models on most tasks.

https://doi.org/10.1371/journal.pone.0192360.g002

Table 2. This table shows the best performing model for each approach and phenotype. We show precision, recall, F1-Score, and AUC.

		CNN	BoW	n-gram	cTAKES full	cTAKES filter
Adv. Cancer	*P*	**87**	44	41	80	85
	R	65	**77**	55	65	55
	F1	**74**	56	47	71	67
	AUC	**95**	90	88	94	92
Adv. Heart Disease	*P*	74	70	**78**	71	73
	R	**76**	32	42	49	59
	F1	**75**	44	55	58	65
	AUC	**91**	85	85	88	89
Adv. Lung Disease	*P*	**67**	21	27	**67**	43
	R	**57**	29	39	29	36
	F1	**62**	24	32	40	39
	AUC	**89**	76	79	81	87
Chronic Neuro	*P*	69	47	49	75	**80**
	R	**70**	46	54	55	62
	F1	69	46	51	64	**70**
	AUC	84	72	71	87	**86**
Chronic Pain	*P*	**78**	33	42	66	66
	R	45	54	46	41	**48**
	F1	**57**	41	44	51	56
	AUC	73	68	67	78	**85**
Alcohol Abuse	*P*	85	**100**	55	88	91
	R	**79**	50	64	**79**	75
	F1	81	67	59	**83**	82
	AUC	**96**	89	88	95	**96**
Substance Abuse	*P*	83	62	83	**93**	87
	R	**83**	50	33	47	67
	F1	**83**	56	48	62	75
	AUC	**97**	90	86	**97**	**97**
Obesity	*P*	**100**	27	44	64	62
	R	**95**	35	20	80	75
	F1	**97**	30	28	71	68
	AUC	**100**	72	71	99	98
Psychiatric Disorders	*P*	**87**	47	53	74	81
	R	**80**	53	39	63	64
	F1	**83**	50	45	68	72
	AUC	**95**	77	76	88	93
Depression	*P*	**90**	51	51	81	79
	R	**79**	67	73	72	77
	F1	**84**	58	60	76	78
	AUC	93	77	78	**94**	91

https://doi.org/10.1371/journal.pone.0192360.t002

improve performance of clinical NLP tasks [62]. We additionally investigate whether considering longer phrases improves model performance. In Fig 3, we show the difference in F1-score between models with phrases up to a certain length and models that use bag-of-words or bag-of-embeddings. The data used for this figure is shown in S1 and S3 Tables. There is no significant difference in performance for longer phrases in n-gram models. There is, however, a significant improvement for phrases longer than one word for the CNN, showing that the CNN

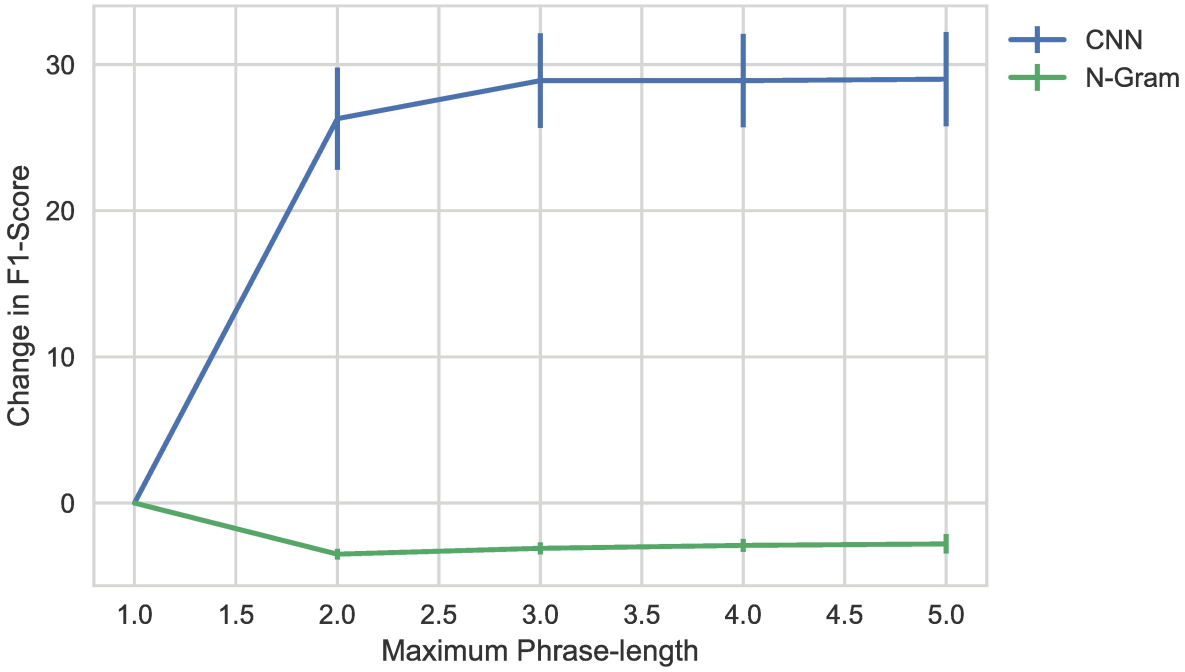

Fig 3. Impact of phrase length on model performance. The figure shows the change in F1-score between a model that considers only single words and a model that phrases up to a length of 5.

https://doi.org/10.1371/journal.pone.0192360.g003

model architecture complements the embedding-based approach and contributes to the result of the model.

Experiments that used both raw text and CUIs as input to a CNN showed no improvement over only using the text as input. This shows that the information encoded in the CUIs is already available in the text and is detected by the CNN. We hypothesize that encoding information available in UMLS beyond the CUI itself can help to improve the phenotype detection in future work.

We show the most salient phrases according to the CNN and the filtered cTAKES LR models for Advanced Heart Disease and for Alcohol Abuse in Table 3. Both tables contain many of the phrases mentioned in the definition shown in Table 1, such as "Cardiomyopathy". We also observe mentions of "CHF" and "CABG" for Advanced Heart Disease for both models, which are common medical conditions associated with advanced heart disease, but are not sufficient requirements according to the annotation scheme. The model still learned to associate those phrases with advanced heart disease, since those phrases also occur in many notes from patients that were labeled positive for advanced heart failure. The phrases for Alcohol Abuse illustrate how the CNN can detect mentions of the condition in many forms. Without human input, the CNN learned that EtOH and alcohol are used synonymously with different spellings and thus detects phrases containing either of them. The filtered cTAKES RF model surprisingly ranks victim of abuse higher than the direct mention of alcohol abuse in a note, and finds that it is very indicative of Alcohol Abuse if an ethanol measurement was taken. While the CUIs extracted by cTAKES can be very generic, such as "Atrium, Heart" or "Heart", the salient CNN phrases are more specific.

In the quantitative study of relevant phrases and CUIs, phrases from the CNN received an average rating of 2.44 ($\sigma = 0.89$), and the cTAKES based approach received an average rating of 1.9 ($\sigma = 0.97$). A t-test for two independent samples showed that there is a statistically

Table 3. The most salient phrases for advanced heart failure and alcohol abuse. The salient cTAKES CUIs are extracted from the filtered RF model.

cTAKES	CNN
Advanced Heart Disease	
Magnesium	Wall Hypokinesis
Cardiomyopathy	Port pacer
Hypokinesia	Ventricular hypokinesis
Heart Failure	p AVR
Acetylsalicylic Acid	post ICD
Atrium, Heart	status post ICD
Coronary Disease	EF 20 30
Atrial Fibrillation	bifurcation aneurysm clipping
Coronary Artery	CHF with EF
Disease	cardiomyopathy, EF 15
Aortocoronary Bypasses	(EF 20 30
Fibrillation	coronary artery bypass graft
Heart	respiratory viral infection by DFA
Catheterization	severe global free wall hypokinesis
Chest	Class II, EF 20
Artery	lateral CHF with EF 30
CAT Scans, X-Ray	anterior and atypical hypokinesis akinesis
Hypertension	severe global left ventricular hypokinesis
Creatinine Measurement	's cardiomyopathy, EF 15
Alcohol Abuse	
Victim of abuse	Consciousness Alert
Ethanol Measurement	Alcohol Abuse
Alcohol Abuse	EtOH abuse
Thiamine	Alcoholic Dilated
Social and personal history	ETOH cirrhosis
Family history	heavy alcohol abuse
Hypertension	evening Alcohol abuse
Injuries risk	Drug Reactions Attending
Pain	alcohol withdrawal compartment syndrome
Sodium	EtOH abuse with multiple
Potassium Measurement	liver secondary to alcohol abuse
Plasma Glucose Measurement	abuse crack cocaine, EtOH

https://doi.org/10.1371/journal.pone.0192360.t003

significant difference between the two with a p-value < 0.00001. This indicates that the five most relevant features from the CNN are more relevant to a phenotype than the five most relevant features from a cTAKES-based model. The free-text comments confirm our descriptive results; the CNN-based phrases are seen as specific and directly relating to a patient's condition while the CUI's are seen as more generic. Moreover, clinicians were impressed to see phrases such as "h o withdrawal" for alcohol abuse (short for "history of", which are typically difficult to interpret by non-experts. Some of the longer phrases from the CNN for the depression phenotype showed the word "depression" amids other diseases, indicating that the phrase is taken from a diagnosis section of the discharge summary. Clinicians commented that this helped them to contextualize the phrase and that it was more helpful than seeing the word "depression" in isolation. This indicates that giving further contextual information can help to increase understanding of predictions.

Discussion

Our results show that CNNs provide a valid alternative approach to the identification of patient conditions from text. However, we notice a strong variation in the results between phenotypes with AUCs between 73 and 100, and F1-scores between 57 and 97, even with consistent annotation schemes. Some concepts such as Chronic Pain are especially challenging to detect, even with 321 positive examples in the data set. This makes it difficult to compare our results to other reported metrics in the literature, since studies typically consider different concepts for detection. This problem is further amplified by the sparsity of available studies that investigate unstructured data [13], and the lack of standardized datasets for this task. We hope that the release of our annotations will support work towards a more comparable performance in text-based phenotyping.

Interpretability

Since bias in data collection and analysis is at times unavoidable, models are required to be interpretable in order for clinicians to be able to detect such biases, and alter the model accordingly [27]. Furthermore, interpretable models lead to an increased trust from the people who use them [26]. The interpretability is typically considered to be a major advantage of rule or concept-extraction based models that are specifically tailored to a given problem. Clinicians have full control over the dictated phrases that are used as input to a model. We demonstrated that CNNs can be interpreted in the same way as concept-extraction based models by computing the saliency of inputs. This even leads to a higher level of interpretability in that the extracted phrases are more relevant to a phenotype than the extracted CUIs. However, a disadvantage of CNNs is that they -by design- consider more different phrases than concept-extraction based methods. Thus, lists of salient phrases will naturally contain more items, making it more difficult to investigate which phrases lead to a prediction. However, restricting the list to a small number of items with the highest saliency coefficients or only including those above a saliency threshold can compensate for the length of the list. The question that developers of such models will ultimately have to answer is whether the trade-off between increased performance is worth the additional effort to extract and show relevant phrases.

Alternative approaches

We further note that both concept-extraction and deep learning are supervised approaches and thus require a labeled dataset. The extraction of all labels for a single discharge summary took clinicians up to 10 minutes, which in our case (with twice-labeled notes) amounts to over 500 hours of annotation work across all annotating clinicians. Therefore, going forward it will be important to combine our approach with methods that alleviate this disadvantage by using semi-supervised methods that do not require fully labeled data-sets [16, 76, 77].

Another approach not considered here that does work without a large annotated data set is to develop a fully rule-based system. While a CNN learns the phrases that lead to a positive label, rule-based approaches require clinicians to define every phrase that is associated with a concept and establish links between them. An example by Mosley et al. looks for certain drug names and the word "cough" within the same line in a description of a patients' allergies [78]. However, due to the heterogeneity of text, clinicians may be unable to consider all of the possible phrases in advance. They also have to consider how to handle negated phrases correctly. Finally, for some clinically important phenotypes such as "Non-Adherence", it is impossible to construct an exhaustive list of associated phrases. Moreover, while rule-based systems require a separate algorithm for each concept, approaches that do not require concept-specific input such as CNNs can be trained for all phenotypes at the same time. This offers an opportunity to

dramatically accelerate the development of scalable phenotyping algorithms for complex clinical concepts in unstructured clinical text that are poorly captured in the structured data. For example, being able to identify patients who are readmitted to hospital due to poor management of problems will have high clinical impact.

Future extensions

While we consider and analyze a simple CNN architecture in this work, future extensions grounded in our findings could explore other deep learning model architectures. Alternatively, one could imagine different feature maps of a CNN for each of the sections in a medical record to capture additional information. Recurrent neural networks could also be applied to directly capture these context specific information instead of using convolutional neural networks. Finally, our unstructured data could be combined with structured input, such as lab results, data from UMLS, or ICD-codes. Augmenting unstructured text with structured data will help to achieve the best possible performance in a given phenotyping task [13].

We anticipate validation of our proposed approach in other types of clinical notes such as social work assessment to identify patients at risk for hospital admissions. Lastly, the CNN creates the opportunity to develop a model that can use phrase saliency to highlight notes and tag patients to support chart review. Future work will explore whether the identification of salient phrases can be used to support chart abstraction and whether models using these phrases represent what clinicians find salient in a medical note. Another opportunity that the learned phrases from a CNN provide is to better understand how different medical conditions are typically described in text. This knowledge can then be turned into improvements to systems like cTAKES and to build better predictive models with a human in the loop [79].

Conclusion

Taking all these points into consideration, we conclude that deep learning provides a valid alternative to concept extraction based methods for patient phenotyping. Using CNNs can significantly improve the accuracy of patient phenotyping without needing any phrase-dictionary as input. We showed that concerns about the interpretability of deep learning can be addressed by computing a gradient-based saliency to identify phrases associated with different phenotypes. We propose that CNNs should be employed alongside concept-extraction based methods to analyze and mine unstructured clinical narratives and augment the structured data in secondary analysis of health records. The deep learning approach presented in this paper could further be used to assist clinicians during chart review by highlighting phrases related to phenotypes of a patient. Moreover, the same methodology could be used to support the identification of billing codes from text.

Supporting information

S1 Table. Overview of CNN results with different convolution widths. Each column name shows the minimum and maximum width of the convolution.
(PDF)

S2 Table. Overview of cTAKES results with different models with all input features and the filtered lists of inputs. While in most cases, the clinician-defined phrase dictionary improves the model performance, the full input performs almost as well and outperforms the filtered model in some.
(PDF)

S3 Table. Results of different n-gram based models. Each column name shows the minimum and maximum length of phrase that has been considered. We observe that in most cases, a simple bag of words (phrase length 1) outperforms all other models.
(PDF)

S1 Text. Additional training information for the models.
(PDF)

Acknowledgments

We thank Alistair Johnson for support with the MIMIC-III database, and Tristan Naumann and Barbara J. Grosz for helpful discussions. We additionally thank the course staff for the course HST.953: Collaborative Data Science in Medicine at Massachusetts Institute of Technology, during which parts of this study were conducted.

Author Contributions

Data curation: Sebastian Gehrmann, Franck Dernoncourt, Yeran Li, Eric T. Carlson, Joy T. Wu, Jonathan Welt, John Foote, Jr., Edward T. Moseley, David W. Grant, Patrick D. Tyler.

Formal analysis: Eric T. Carlson, Edward T. Moseley.

Funding acquisition: Leo A. Celi.

Investigation: Sebastian Gehrmann, Franck Dernoncourt, Eric T. Carlson, Edward T. Moseley.

Methodology: Sebastian Gehrmann, Franck Dernoncourt.

Project administration: Sebastian Gehrmann, Joy T. Wu, Leo A. Celi.

Software: Sebastian Gehrmann, Yeran Li, Edward T. Moseley.

Supervision: Leo A. Celi.

Visualization: Sebastian Gehrmann.

Writing – original draft: Sebastian Gehrmann.

Writing – review & editing: Sebastian Gehrmann, Franck Dernoncourt, Joy T. Wu, Jonathan Welt, Edward T. Moseley, David W. Grant, Patrick D. Tyler, Leo A. Celi.

References

1. Data MC. Secondary Analysis of Electronic Health Records. Springer; 2016.
2. Charles D, Gabriel M, Furukawa MF. Adoption of electronic health record systems among US non-federal acute care hospitals: 2008-2012. ONC data brief. 2013; 9:1–9.
3. Saeed M, Lieu C, Raber G, Mark RG. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. In: Computers in Cardiology, 2002. IEEE; 2002. p. 641–644.
4. Johnson AE, Pollard TJ, Shen L, Lehman LwH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific data. 2016; 3. https://doi.org/10.1038/sdata.2016.35
5. Uzuner Ö, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. Journal of the American Medical Informatics Association. 2008; 15(1):14–24. https://doi.org/10.1197/jamia.M2408 PMID: 17947624
6. Uzuner Ö. Recognizing obesity and comorbidities in sparse data. Journal of the American Medical Informatics Association. 2009; 16(4):561–570. https://doi.org/10.1197/jamia.M3115 PMID: 19390096
7. Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. Journal of the American Medical Informatics Association. 2010; 17(5):514–518. https://doi.org/10.1136/jamia.2010.003947 PMID: 20819854

8. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association. 2011; 18(5):552–556. https://doi.org/10.1136/amiajnl-2011-000203 PMID: 21685143
9. Sun W, Rumshisky A, Uzuner O. Annotating temporal information in clinical narratives. Journal of biomedical informatics. 2013; 46:S5–S12. https://doi.org/10.1016/j.jbi.2013.07.004 PMID: 23872518
10. Stubbs A, Uzuner Ö. Annotating risk factors for heart disease in clinical narratives for diabetic patients. Journal of biomedical informatics. 2015; 58:S78–S91. https://doi.org/10.1016/j.jbi.2015.05.009 PMID: 26004790
11. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics. 2012; 13(6):395–405. https://doi.org/10.1038/nrg3208 PMID: 22549152
12. Murdoch TB, Detsky AS. The inevitable application of big data to health care. Jama. 2013; 309(13): 1351–1352. https://doi.org/10.1001/jama.2013.393 PMID: 23549579
13. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. bmj. 2015; 350:h1885. https://doi.org/10.1136/bmj.h1885 PMID: 25911572
14. Ananthakrishnan AN, Cai T, Savova G, Cheng SC, Chen P, Perez RG, et al. Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. Inflammatory bowel diseases. 2013; 19(7):1411. https://doi.org/10.1097/MIB.0b013e31828133fd PMID: 23567779
15. Pivovarov R, Elhadad N. Automated methods for the summarization of electronic health records. Journal of the American Medical Informatics Association. 2015; 22(5):938–947. https://doi.org/10.1093/jamia/ocv032 PMID: 25882031
16. Halpern Y, Horng S, Choi Y, Sontag D. Electronic medical record phenotyping using the anchor and learn framework. Journal of the American Medical Informatics Association. 2016; p. ocw011. https://doi.org/10.1093/jamia/ocw011 PMID: 27107443
17. Chen L, Guo U, Illipparambil LC, Netherton MD, Sheshadri B, Karu E, et al. Racing Against the Clock: Internal Medicine Residents' Time Spent On Electronic Health Records. Journal of graduate medical education. 2016; 8(1):39–44. https://doi.org/10.4300/JGME-D-15-00240.1 PMID: 26913101
18. Topaz M, Lai K, Dowding D, Lei VJ, Zisberg A, Bowles KH, et al. Automated identification of wound information in clinical notes of patients with heart diseases: Developing and validating a natural language processing application. International Journal of Nursing Studies. 2016; 64:25–31. https://doi.org/10.1016/j.ijnurstu.2016.09.013 PMID: 27668855
19. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association. 2013; 20(1):117–121. https://doi.org/10.1136/amiajnl-2012-001145 PMID: 22955496
20. Carrell DS, Halgrim S, Tran DT, Buist DS, Chubak J, Chapman WW, et al. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. American journal of epidemiology. 2014; p. kwt441. https://doi.org/10.1093/aje/kwt441
21. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. Journal of the American Medical Informatics Association. 2016; 23(6):1046–1052. https://doi.org/10.1093/jamia/ocv202 PMID: 27026615
22. Ranganath R, Perotte A, Elhadad N, Blei D. Deep Survival Analysis. arXiv preprint arXiv:160802158. 2016;.
23. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. Journal of the American Medical Informatics Association. 2017; 24(3):596–606. PMID: 28040687
24. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017; 542(7639):115–118. https://doi.org/10.1038/nature21056 PMID: 28117445
25. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016; 316(22):2402–2410. https://doi.org/10.1001/jama.2016.17216 PMID: 27898976
26. Lipton ZC. The mythos of model interpretability. arXiv preprint arXiv:160603490. 2016;.
27. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2015. p. 1721–1730.

28. Goodman B, Flaxman S. EU regulations on algorithmic decision-making and a "right to explanation". In: ICML Workshop on Human Interpretability in Machine Learning (WHI 2016); 2016.
29. Strobelt H, Gehrmann S, Pfister H, Rush AM. Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. IEEE transactions on visualization and computer graphics. 2018; 24;667–676. https://doi.org/10.1109/TVCG.2017.2744158 PMID: 28866526
30. Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H. Understanding neural networks through deep visualization. arXiv preprint arXiv:150606579. 2015;.
31. Zeiler MD, Krishnan D, Taylor GW, Fergus R. Deconvolutional networks. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE; 2010. p. 2528–2535.
32. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L. Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition; 2014. p. 1725–1732.
33. Kim Y. Convolutional neural networks for sentence classification. arXiv preprint arXiv:14085882. 2014;.
34. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems; 2012. p. 1097–1105.
35. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association. 2010; 17(5):507–513. https://doi.org/10.1136/jamia.2009.001560 PMID: 20819853
36. Ackerman JP, Bartos DC, Kapplinger JD, Tester DJ, Delisle BP, Ackerman MJ. The promise and peril of precision medicine: phenotyping still matters most. In: Mayo Clinic Proceedings. vol. 91. Elsevier; 2016. p. 1606–1616.
37. Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. Big Data. 2015; 3(4):277–287. https://doi.org/10.1089/big.2015.0020 PMID: 27441408
38. Strauss JA, Chao CR, Kwan ML, Ahmed SA, Schottinger JE, Quinn VP. Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm. Journal of the American Medical Informatics Association. 2013; 20(2):349–355. https://doi.org/10.1136/amiajnl-2012-000928 PMID: 22822041
39. Lingren T, Chen P, Bochenek J, Doshi-Velez F, Manning-Courtney P, Bickel J, et al. Electronic Health Record Based Algorithm to Identify Patients with Autism Spectrum Disorder. PloS one. 2016; 11(7): e0159621. https://doi.org/10.1371/journal.pone.0159621 PMID: 27472449
40. Savova GK, Olson JE, Murphy SP, Cafourek VL, Couch FJ, Goetz MP, et al. Automated discovery of drug treatment patterns for endocrine therapy of breast cancer within an electronic medical record. Journal of the American Medical Informatics Association. 2012; 19(e1):e83–e89. https://doi.org/10.1136/amiajnl-2011-000295 PMID: 22140207
41. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. Journal of the American Medical Informatics Association. 2011; 18(5):544–551. https://doi.org/10.1136/amiajnl-2011-000464 PMID: 21846786
42. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research. 2004; 32(suppl 1):D267–D270. https://doi.org/10.1093/nar/gkh061 PMID: 14681409
43. Spackman KA, Campbell KE, Côté RA. SNOMED RT: a reference terminology for health care. In: Proceedings of the AMIA annual fall symposium. American Medical Informatics Association; 1997. p. 640.
44. Denny JC, Irani PR, Wehbe FH, Smithers JD, Spickard III A. The KnowledgeMap project: development of a concept-based medical school curriculum database. In: AMIA; 2003.
45. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association. 2010; 17(3):229–236. https://doi.org/10.1136/jamia.2009.002733 PMID: 20442139
46. Friedman C. Medlee-a medical language extraction and encoding system. Columbia University, and Queens College of CUNY. 1995;.
47. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. Journal of the American Medical Informatics Association. 2010; 17(1):19–24. https://doi.org/10.1197/jamia.M3378
48. Denny JC, Choma NN, Peterson JF, Miller RA, Bastarache L, Li M, et al. Natural language processing improves identification of colorectal cancer testing in the electronic medical record. Medical Decision Making. 2012; 32(1):188–197. https://doi.org/10.1177/0272989X11400418 PMID: 21393557

49. Hripcsak G, Austin JH, Alderson PO, Friedman C. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports 1. Radiology. 2002; 224(1): 157–163. https://doi.org/10.1148/radiol.2241011118 PMID: 12091676

50. Pradhan S, Elhadad N, South BR, Martinez D, Christensen L, Vogel A, et al. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. Journal of the American Medical Informatics Association. 2015; 22(1):143–154. https://doi.org/10.1136/amiajnl-2013-002544 PMID: 25147248

51. Perlis R, Iosifescu D, Castro V, Murphy S, Gainer V, Minnier J, et al. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. Psychological medicine. 2012; 42(01):41–50. https://doi.org/10.1017/S0033291711000997 PMID: 21682950

52. Xia Z, Secor E, Chibnik LB, Bove RM, Cheng S, Chitnis T, et al. Modeling disease severity in multiple sclerosis using electronic health records. PloS one. 2013; 8(11):e78927. https://doi.org/10.1371/journal.pone.0078927 PMID: 24244385

53. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis care & research. 2010; 62(8):1120–1127. https://doi.org/10.1002/acr.20184

54. Carroll RJ, Thompson WK, Eyler AE, Mandelin AM, Cai T, Zink RM, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. Journal of the American Medical Informatics Association. 2012; 19(e1):e162–e169. https://doi.org/10.1136/amiajnl-2011-000583 PMID: 22374935

55. Ye Y, Wagner MM, Cooper GF, Ferraro JP, Su H, Gesteland PH, et al. A study of the transferability of influenza case detection systems between two large healthcare systems. PloS one. 2017; 12(4): e0174970. https://doi.org/10.1371/journal.pone.0174970 PMID: 28380048

56. Sarmiento RF, Dernoncourt F. Improving Patient Cohort Identification Using Natural Language Processing. In: Secondary Analysis of Electronic Health Records. Springer; 2016. p. 405–417.

57. Kocher RP, Adashi EY. Hospital readmissions and the Affordable Care Act: paying for coordinated quality care. Jama. 2011; 306(16):1794–1795. https://doi.org/10.1001/jama.2011.1561 PMID: 22028355

58. Bates J, Fodeh SJ, Brandt CA, Womack JA. Classification of radiology reports for falls in an HIV study cohort. Journal of the American Medical Informatics Association. 2016; 23(e1):e113–e117. https://doi.org/10.1093/jamia/ocv155 PMID: 26567329

59. Kang N, Singh B, Afzal Zubair MEMv and, Kors JA. Using rule-based natural language processing to improve disease normalization in biomedical text. J Am Med Inform Assoc. 2013; 20:876–881. https://doi.org/10.1136/amiajnl-2012-001173 PMID: 23043124

60. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. Journal of Machine Learning Research. 2011; 12(Aug):2493–2537.

61. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998; 86(11):2278–2324. https://doi.org/10.1109/5.726791

62. Wu Y, Xu J, Jiang M, Zhang Y, Xu H. A Study of Neural Word Embeddings for Named Entity Recognition in Clinical Text. In: AMIA Annual Symposium Proceedings. vol. 2015. American Medical Informatics Association; 2015. p. 1326.

63. Erhan D, Bengio Y, Courville A, Manzagol PA, Vincent P, Bengio S. Why does unsupervised pre-training help deep learning? Journal of Machine Learning Research. 2010; 11(Feb):625–660.

64. Luan Y, Watanabe S, Harsham B. Efficient learning for spoken language understanding tasks with word embedding based pre-training. In: Sixteenth Annual Conference of the International Speech Communication Association. Citeseer; 2015.

65. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems; 2013. p. 3111–3119.

66. Cooper GF, Aliferis CF, Ambrosino R, Aronis J, Buchanan BG, Caruana R, et al. An evaluation of machine-learning methods for predicting pneumonia mortality. Artificial intelligence in medicine. 1997; 9(2):107–138. https://doi.org/10.1016/S0933-3657(96)00367-3 PMID: 9040894

67. Lei T, Barzilay R, Jaakkola T. Rationalizing neural predictions. arXiv preprint arXiv:160604155. 2016;.

68. Ribeiro MT, Singh S, Guestrin C. Model-agnostic interpretability of machine learning. arXiv preprint arXiv:160605386. 2016;.

69. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC press; 1984.

70. Li J, Chen X, Hovy E, Jurafsky D. Visualizing and understanding neural models in nlp. arXiv preprint arXiv:150601066. 2015;.

71. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv preprint arXiv:160204938. 2016;.

72. Denil M, Demiraj A, de Freitas N. Extraction of salient sentences from labelled documents. arXiv preprint arXiv:14126815. 2014;.

73. Arras L, Horn F, Montavon G, Müller KR, Samek W. "What is Relevant in a Text Document?": An Interpretable Machine Learning Approach. arXiv preprint arXiv:161207843. 2016;.

74. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one. 2015; 10(7):e0130140. https://doi.org/10.1371/journal.pone.0130140 PMID: 26161953

75. Arras L, Horn F, Montavon G, Müller KR, Samek W. Explaining predictions of non-linear classifiers in NLP. arXiv preprint arXiv:160607298. 2016;.

76. Halpern Y, Choi Y, Horng S, Sontag D. Using anchors to estimate clinical state without labeled data. In: AMIA Annual Symposium Proceedings. vol. 2014. American Medical Informatics Association; 2014. p. 606.

77. Farkas R, Szarvas G, Hegedűs I, Almási A, Vincze V, Ormándi R, et al. Semi-automated construction of decision rules to predict morbidities from clinical texts. Journal of the American Medical Informatics Association. 2009; 16(4):601–605. https://doi.org/10.1197/jamia.M3097 PMID: 19390097

78. Mosley JD, Shaffer CM, Van Driest SL, Weeke PE, Wells QS, Karnes JH, et al. A genome-wide association study identifies variants in KCNIP4 associated with ACE inhibitor-induced cough. The pharmacogenomics journal. 2016; 16(3):231. https://doi.org/10.1038/tpj.2015.51 PMID: 26169577

79. Alba A, Drews C, Gruhl D, Lewis N, Mendes PN, Nagarajan M, et al. Symbiotic Cognitive Computing through Iteratively Supervised Lexicon Induction. In: Workshops at the Thirtieth AAAI Conference on Artificial Intelligence; 2016.