DeepSuggest: Using Neural Networks to Suggest Related Keywords for a Comprehensive Search of Clinical Notes

Soheil Moosavinasab[1] Emre Sezgin[1] Huan Sun[2] Jeffrey Hoffman[3,4] Yungui Huang[1] Simon Lin[1,5]

[1] Research Information Solutions and Innovation, The Research Institute at Nationwide Children's Hospital, Columbus, Ohio, United States
[2] Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio, United States
[3] Department of Pediatrics, Nationwide Children's Hospital, Columbus, Ohio, United States
[4] Department of Pediatrics, The Ohio State University College of Medicine, Columbus, Ohio, United States
[5] Department of Biomedical Informatics and Department of Pediatrics, The Ohio State University College of Medicine, Columbus, Ohio, United States

Address for correspondence Simon Lin, MD, MBA, 700 Children's Drive, Columbus, OH 43205, United States (e-mail: simon.lin@nationwidechildrens.org).

ACI Open 2021;5:e1–e12.

Abstract

Objective A large amount of clinical data are stored in clinical notes that frequently contain spelling variations, typos, local practice-generated acronyms, synonyms, and informal words. Instead of relying on established but infrequently updated ontologies with keywords limited to formal language, we developed an artificial intelligence (AI) assistant (named "DeepSuggest") that interactively offers suggestions to expand or pivot queries to help overcome these challenges.

Methods We applied an unsupervised neural network (Word2Vec) to the clinical notes to build keyword contextual similarity matrix. With a user's input query, DeepSuggest generates a list of relevant keywords, including word variations (e.g., formal or informal forms, synonyms, abbreviations, and misspellings) and other relevant words (e.g., related diagnosis, medications, and procedures). Human intelligence is then used to further refine or pivot their query.

Results DeepSuggest learns the semantic and linguistic relationships between the words from a large collection of local notes. Although DeepSuggest is only able to recall 0.54 of Systematized Nomenclature of Medicine (SNOMED) synonyms on average among the top 60 suggested terms, it covers the semantic relationship in our corpus for a larger number of raw concepts (6.3 million) than SNOMED ontology (24,921) and is able to retrieve terms that are not stored in existing ontologies. The precision for the top 60 suggested words averages at 0.72. Usability test resulted that DeepSuggest is able to achieve almost twice the recall on clinical notes compared with Epic (average of 5.6 notes retrieved by DeepSuggest compared with 2.6 by Epic).

Conclusion DeepSuggest showed the ability to improve retrieval of relevant clinical notes when implemented on a local corpus by suggesting spelling variations,

Keywords
- electronic health records
- search engine
- neural networks
- computer
- information storage and retrieval
- user-computer interface
- pediatrics
- electronic data processing

received
February 4, 2020
accepted after revision
March 24, 2021

DOI https://doi.org/10.1055/s-0041-1729982.
ISSN 2566-9346.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

acronyms, and semantically related words. It is a promising tool in helping users to achieve a higher recall rate for clinical note searches and thus boosting productivity in clinical practice and research. DeepSuggest can supplement established ontologies for query expansion.

Introduction

Keyword-driven search of clinical notes greatly expedites retrieval of medical information beyond manual chart review to serve the needs of patient care, quality improvement, and clinical research.[1–3] While searching notes may appear to be simple and quick, determining the optimal set of query keywords is not straightforward. For instance, when searching notes for tonsillectomy patients, using "tonsillectomy" can miss notes containing "tonsilectomy" (common misspelling), "T/A" and "T&A" (nonstandard but commonly used abbreviations), or "adenotonsillectomy" (semantically related concept).

This vocabulary mismatch between the query words and the actual words used in target documents might best be resolved by expanding the query with relevant options.[4,5] Since human recall of synonyms is usually poor, medical ontologies have often been used to assist the expansion of the original query, either interactively[6] or algorithmically.[7] Ontology-driven query expansion strategies have long been adopted by both academic and commercial implementations, such as EMERSE (Michigan University, United States),[8,9] CISearch (Columbia University, United States),[1] SemEHR (King's College London, United Kingdom),[6] EpicCare (Epic, United States), and Cerner PowerChart (Cerner, United States). However, ontology-driven expansion is challenged by the mismatch between the informal, dynamic nature of clinical notes and the formal, static nature of ontologies.[1] Ontology-driven approaches also lack timely updates due to the high curation costs for such efforts.[10]

Unsupervised shallow neural networks, such as Word2Vec[11] and GloVe,[12] have been effective at embedding related concepts from unstructured and unannotated texts when a large corpus is available.[13–18] Ye and Fabbri[19] extensively evaluated different approaches to identify similar terms using semantic embeddings. Although its effectiveness was demonstrated in the TREC Precision Medicine information retrieval tasks by other research teams to expand query Text REtrieval Conference automatically,[20] word embedding is not widely adopted and evaluated as an interactively query optimization tool.

Currrent implementations for medical search engines are limited to ontology-driven methods, and artificial intelligence (AI)-driven methodologies are available to create enhanced search algorithms. Even though scale-up implementation is challenging in terms of potential interoperability issues as well as privacy and security compliance, legislation, and regulations, enhanced medical search engines have potential to improve clinical decision-making. In this paper, we introduce DeepSuggest, an interactive clinical note query platform, by demonstrating its capabilities through use cases, reporting its precision, recall, and effectiveness through quantitative evaluations and usability tests, and discussing its value in research and clinical practice. This work has limited contribution toward the word embedding algorithms or query expansion, but reports findings from a real-world software implementation and its application in clinical research.

We hypothesized that leveraging semantically related words identified through an unsupervised shallow neural network trained on the same corpus of clinical notes to be queried can overcome the vocabulary mismatch problem and produce better search results. Word2Vec can learn semantic (e.g., "myopathy" and "dysferlinopathy") and linguistic (e.g., "discharge," "D/C," and "discharged"; "tonsil" and "tonsill") relationships between the words. In contrast to ontology-based query expansion systems that are purely based on prior knowledge and work without much machine learning, our framework is referred to as AI driven. To test this hypothesis, we developed a unique query expansion tool, which we call "DeepSuggest." DeepSuggest can interactively suggest a list of clinically relevant "keywords/options" (not synonyms only) to a given search term and involve the user to expand or pivot the query (i.e., adapt and update query iteratively to better fit search objectives based on each of search results).

Methods

To create a query expansion tool built upon word-embedding and language-modeling, we have gone through multiple processing and development steps as illustrated by ►Fig. 1. We first tokenized and normalized an institute's corpus of clinical notes. We applied data preprocessing steps including converting notes to lowercase, removing stop words, removing emails, web addresses, dates, and phone numbers. To create the list of n-grams, we used Gensim's phrases and phraser functions. We converted tokens into a sequence of 1 to 4 gram words using Pointwise Mutual Information (PMI) score of 10 or above. PMI defines the strength of association between two words using log ratio between the joint probability of the words occurring together, and the product of probability of each of their occurrences (Eq. 1).

$$\mathrm{PMI}(w1, w2) = \log\left(\frac{P(w1, w2)}{P(w1)P(w2)}\right)$$

Then, we excluded any 1 to 2 gram words and 3 to 4 gram words that appear in our corpus less than 30 and 15 times respectively for two main reasons: (1) to significantly reduce our dictionary size for model to run faster with less memory and (2) to avoid distracting the model by extremely rare words that are clearly extracted due to rare typos or tokenization exceptions. We empirically chose these minimum occurrence

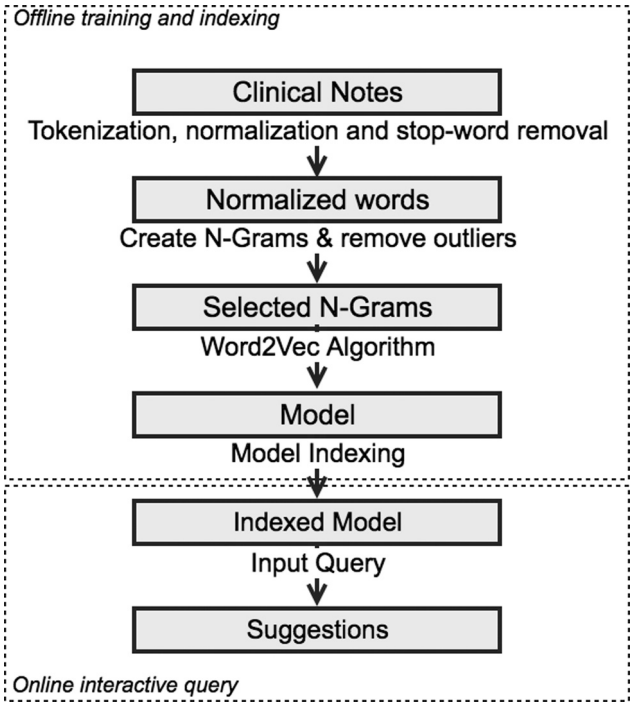

Fig. 1 System architecture of DeepSuggest, an artificial intelligence-driven query expansion system for clinical settings. Word embedding was trained offline by using a local corpus. The search is done on the fly by using Elasticsearch.

cutoffs after looking at a sorted list of unique extracted n-grams and their frequencies. Experts were clinicians who were potential end users of DeepSuggest in clinical practice. After the preprocessing, we employed Word2Vec to create the word embedding vector space, indexed them, and finally designed a user interface to make the interactive query expansion possible (►Fig. 1). Following user-centered design principles and thinking out loud approach during user interaction sessions,[21] we were able to identify needs and expectations, and respond through designing DeepSuggest dashboard. To improve usability and increase adoption of DeepSuggest, we utilized a user-centric approach engaging potential user groups (researchers and clinicians) throughout the development process.[22] The user groups consisted of residents, clinical fellows, and research scientists at Nationwide Children's Hospital (NCH) focusing on epidemiology research including the division of internal medicine, gastroenterology, and biobehavioral health. We compared DeepSuggest against the native note search tool within our electronic medical records (EMR) in Epic. Epic is a widely adopted EMR system which is also used at our institution. It provides a reasonable benchmark to understand the performance of DeepSuggest in contrast to real world implementations. Additionally, comparing to Epic is more convenient than other systems as our research team has access as well as our participants for the study.

Data Source

Our corpus consists of approximately 69 GB of over 66 million clinical notes documenting patient encounters at NCH from 2006 to 2016. After data preprocessing, our vocabulary consists of 6.3 million unique 1 to 4 gram words, representing 5.5 billion total words. The dataset includes additional note and patient information shown in ►Fig. 2 to facilitate patient lookup, chart review, and cohort identification.

Embedding Algorithm

Word2Vec is an unsupervised learning approach that uses neural networks and word embedding to map words into a low dimensional vector space.[11] We used the Continuous Bag-of-Words (CBOW) algorithm in Word2Vec with 400 dimensions and two iterations of optimization to create a statistical model of 1 to 4 gram word similarity across the corpus using word proximity in the notes (16 hours runtime on a physical server with 72 central processing unit (CPU) cores and 512 GB memory). CBOW goes through the corpus and learns word similarity by predicting center word from the surrounding words in a sliding window across each note. The nearest neighbor model ranks the closest semantically relevant N-grams from the trained model using cosine similarity (►Fig. 1). The model created by Word2Vec is indexed by Annoy[23] (a library for nearest neighbors) into a 3.8 GB file to be retrieved quickly and efficiently. We used Word2Vec implementation in the Gensim library using python. The model can be trained incrementally with additional notes over time and in a distributed environment to accommodate larger corpora.

Institutational Review Board Disclosure

The institutational review board (IRB) at NCH reviewed the study and concluded that the project is not qualified as human subjects research, as defined by the United States Department of Health and Human Services and the Food and Drug Administration. This study is exempt from IRB approval.

Use Case Demonstration

We first demonstrate the features and capabilities of DeepSuggest with two clinical use cases and then discuss quantitative and qualitative approaches to measure the success of DeepSuggest. Use cases demonstrates the evaluation of DeepSuggest through comparing search results with SNOMED (use case I) and biomedical ontologies (use case II).

Use Case I: Identification of Tonsillectomy in Telephone Encounter Notes

Tonsillectomy is a relatively common procedure in pediatric care. This use case provides a clinical quality improvement (QI) scenario of postoperative care for tonsillectomy patients through telephone encounters.[24] A first step of this QI study is to identify all the telephone encounter notes with the discussion of tonsillectomy. When "tonsillectomy" was queried, the Systematized Nomenclature of Medicine (SNOMED; SNOMED International, United Kingdom) ontology suggested "excision of tonsil," "adenoid excision," "tonsillectomy with adenoidectomy," "tonsillectomy and adenoidectomy," and "adenotonsillectomy" to expand the query. As expected, the SNOMED-expanded query retrieved more documents

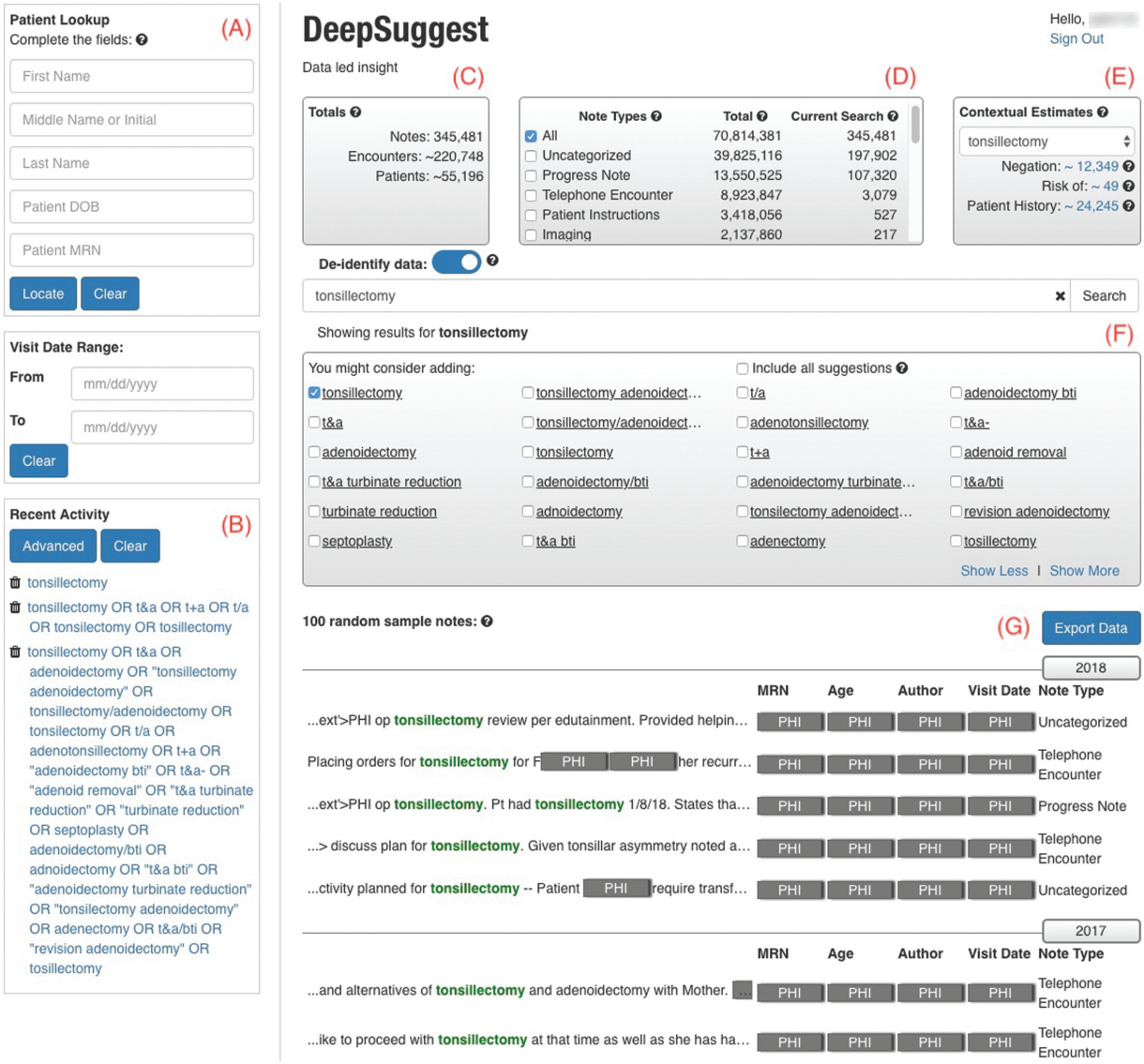

Fig. 2 The design of interactive selection of terms for query expansion by DeepSuggest.

(3,352) than the single keyword "tonsillectomy" (3,079). We also noticed that many synonyms in the ontology, such as "excision of tonsil" or "adenoid excision," were not used by any clinician at NCH in telephone encounter notes. Thus, adding those words offered no real value to query expansion.

By comparison, the AI-driven DeepSuggest learned many informal writings from the local corpus, such as "T&A," "TNA," or common misspellings (►**Table 1**). Using the 38 tonsillectomy-related terms, which were selected by an expert by manually reviewing the list of 60 words suggested by DeepSuggest (DeepSuggest was tuned to provide top 60 words relevant to search terms), the recall of telephone encounter notes that discussing tonsillectomy (14,892 documents returned) was improved by 22.5% compared with documents retrieved by SNOMED-expanded query.

Surprisingly, DeepSuggest also proposed bilateral tube insertion (BTI), a procedure frequently performed together with tonsillectomy and adenoidectomy (T&A/BTI). In addition, the AI-driven query expansion assistant suggested turbinate reduction and other same-day oral or ear procedures. Depending on the scope of the QI study, those additional terms can be either included or excluded, since many QI issues (such as pain management, fluid management, and postoperative bleeding) are shared by these procedures.

This use case also illustrated that the AI feature of DeepSuggest can provoke human thoughts to dynamically change the scope or the aim of the query. To capitalize on this aspect, we designed an interactive feature for users to add a suggested term or not (►**Fig. 2F**). The new pivoting functionality to improve query also challenges the traditional measurement of "false positive," since the judgement of a suggested word as being on-topic versus off-topic can be both dynamic and subjective.

The interface allows to limit the query to a specific patient (►**Fig. 2A**), add boolean constraints (some examples are at ►**Fig. 2B**), show total number of notes, encounters and patients (►**Fig. 2C**), select note types (►**Fig. 2D**), and estimate the context of the matching word (►**Fig. 2E**),[25] suggested

Table 1 Sample keywords recommended by DeepSuggest when querying "tonsillectomy"

Tonsillectomy	Adenoidectomy	Adenotonsillectomy	Together with BTI	Turbinate reduction	Other oral and ear procedures
Tonsillectomy, tonsilectomy, tosillectomy, lingual tonsillectomy	Adenoidectomy, adenectomy, adenoids removed, revision adenoidectomy	T&A, T + A, T/A, TNA, adentonsillectomy, adenotonsillectomy, tonsillectomy/ adenoidectomy, tonsils adenoids removed	BTI, T&A/BTI, BTI/ABR, adenoidectomy/BTI	T&A turbinate reduction, turbinate reduction, turb reduction, turbinate coblation, septoplasty	Wisdom teeth extraction, wisdom teeth removal, tympanoplasty

Abbreviations: BTI, bilateral tube insertion; TA, tonsillectomy and adenoidectomy.
Note: Terms are manually categorized for clarity.

terms (►Fig. 2F), and sample notes with highlighted keywords and note details (►Fig. 2G).

Use Case II: Meal Vouchers (Pivoting Health Equity Research)

This use case further illustrates how DeepSuggest helped pivoting health equity research project with Social Determinants of Health (SDoH) using AI. The researcher was originally interested in studying the documented food insecurity in clinical notes by querying "meal voucher." The AI-driven assistant helped the researcher to think about other health equity issues documented in clinical notes, such as a lack of transportation (e.g., bus tickets, bus passes, cab voucher, taxi voucher, and gas card) and the availability of a charity fund (e.g., compassion fund and gift card) (►Fig. 3). The ontology approach cannot expand "meal voucher," since it is not included in any of the 728 bio-ontologies at BioPortal.[26] Therefore, using biomedical ontologies may fall short in identifying SDoH and health equity research.

Evaluation Methods

The two use cases above illustrate the features and capability of DeepSuggest. Also, these examples allude to the inherent difficulties in assessing the effectiveness of DeepSuggest by the traditional measurements of precision and recall, since the definition of relevance can be highly contextual and subjective. Therefore, we only attempted to use precision and recall to quantitatively probe the limitations DeepSuggest and employed qualitative methods to assess the usability of the DeepSuggest with medical residents.

Evaluation of Precision on Suggested Words

To evaluate the precision of DeepSuggest, three hospital residents manually assessed the relevance of suggested terms generated by DeepSuggest. The residents conducted 11 predefined queries and marked the top suggested 60 terms as relevant or not relevant. This set of 11 queries were primarily inspired from the previous work of Ganesan

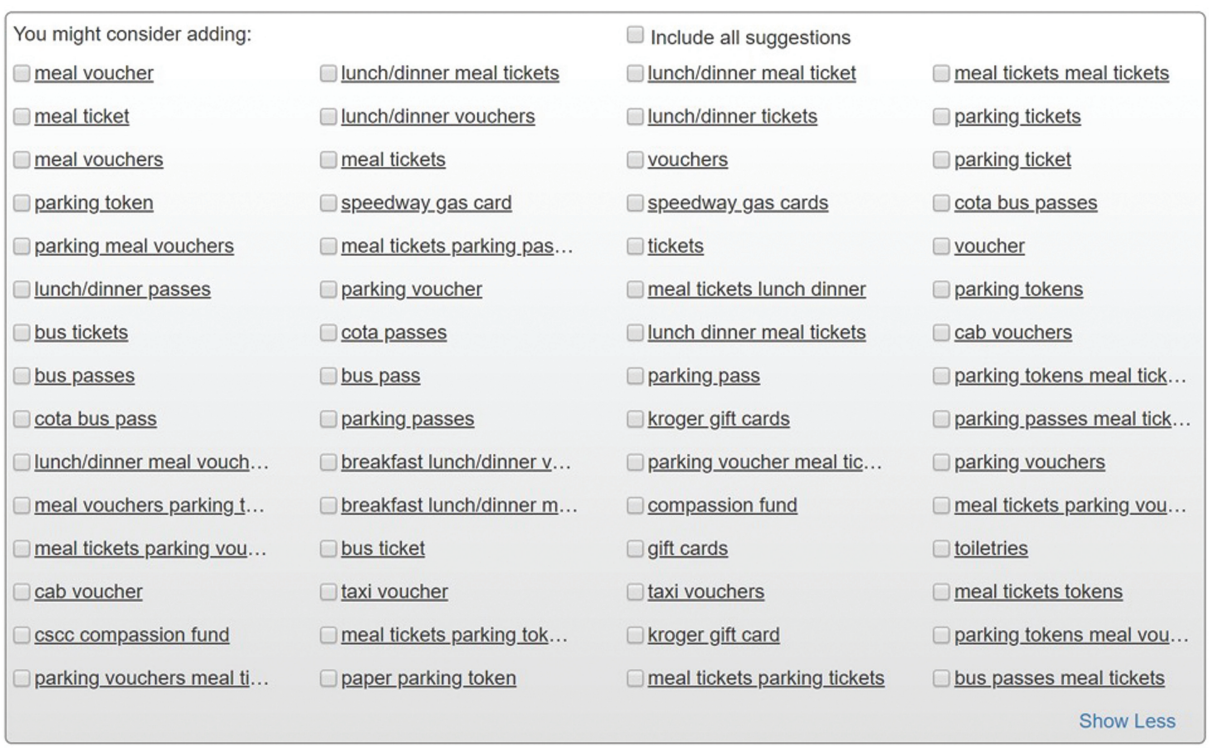

Fig. 3 DeepSuggest result of "meal voucher."

Table 2 Interrater agreement evaluation

Words[a]	Description	Agreement among experts: Fleiss's Kappa score	*p*-Value of the Kappa score[b]	Precision of DeepSuggest using the union score (%)[c]
Asthma	Diagnosis	0.553	0.000[e]	86.44
Fracture	Diagnosis	0.387	0.000[e]	98.30
Tonsillectomy	Procedure	0.387	0.000[e]	71.18
Penicillin	Medication - antibiotic	0.379	0.000[e]	47.45
Pregnancy	Diagnosis	0.369	0.000[e]	83.05
Syncope	Symptom	0.306	0.000[e]	100
Beta blocker	Medication - drug category	0.267	0.000[e]	55.93
Lithium	Medication - drug name	0.201	0.004[d]	55.93
Strata	Brand name of an implanted device treating hydrocephalus	0.098	0.339	42.37
Advair	Medication - brand name inhaler for asthma treatment	−0.012	0.867	81.35
iPhone	Non-medical term	−0.009	0.900	71.18
Average		0.266		72.11

[a]Sorted by Kappa score.
[b]User agreement is significantly different from what would be achieved by chance.
[c]Percent of DeepSuggest results at least verified by one expert.
[d]$p < 0.05$.
[e]$p < 0.001$.

et al,[10] and modified by the authors to accommodate a broad category of medical words in pediatric setting (e.g., drugs, devices, procedures, diseases, and symptoms). For instance, "iPhone" (Apple, United States) was added to test the capability of the system to handle nonmedical terms. ►Table 2 provides the list of test queries. Fleiss's Kappa inter-rater reliability test checked the reliability of responses.[27]

Evaluation of Recall of Known Synonyms

SNOMED, a gold standard medical ontology, was used to assess the recall of known synonyms, similar to Henriksson et al.[28] The SNOMED snapshot of January 2017 contained a list of 994,245 unique terms. This list was filtered to only include the 24,920 unique terms that were also found in the dictionary derived from the NCH corpus, and further filtered down to 6,682 unique terms with at least one synonym reported in SNOMED and existing in DeepSuggest's dictionary. Synonyms were defined as either "synonyms" or "preferred names" in the SNOMED schema.

All synonyms from SNOMED that also appear in the clinical note corpus and the top N (N from 5 to 1000) suggestions from DeepSuggest were extracted for each of the 6,682 SNOMED terms. Recall was calculated as percentage of SNOMED synonyms and preferred names that were seen among the top N suggestions from DeepSuggest. Words among DeepSuggest suggestions but not in SNOMED's vocabulary were disregarded since they might have been relevant but just not listed in SNOMED.

Usability Testing Method on Retrieving Clinical Notes

A group of six residents, representing internal medicine, pediatrics, and neurology, tested DeepSuggest for usability. Each of them was experienced with Epic, the commercial EMR tool implemented at NCH, and Epic search function (average of 3 years of Epic experience). The residents self-reported that they used Epic very frequently, spending on average 12 minutes a day searching clinical notes using Epic.

The residents first used the Epic search function for a given patient to find all previous clinical notes relevant to tonsillectomy. Epic uses SNOMED's ontology to automatically expand the user's search terms. Next, after a brief introduction and training on the purpose, use, and functions of DeepSuggest, the residents were asked to search on the same patient using DeepSuggest and find all previous clinical notes relevant to tonsillectomy again. The number of retrieved relevant clinical notes from each system was used to compare performance of DeepSuggest versus Epic to understand how accurately the notes are retrieved in response to keyword search. The Usability Metric for User Experience (UMUX) usability questionnaire[29] and reaction cards[30] were completed after task completion (see ►Appendix A for details). After the survey, we had an open-ended discussion (~30 minutes) with residents to get a better understanding of their experience. All sessions were audio recorded and log files of DeepSuggest were used in the analysis.

Results

Precision of Suggested Keywords

Three residents evaluated the DeepSuggest results as relevant or nonrelevant, which resulted in divergent expert opinions. Kappa scores, which measure inter-rater agreement, ranged from "moderate" ("asthma," $0.41 < \text{k-score} < 0.60$, $p < 0.001$), "fair" ("fracture, tonsillectomy, penicillin, pregnancy, syncope,

Beta blocker, lithium*," $0.21 <$ k-score < 0.40, $p < 0.001$, *p < 0.05), to almost no agreement ("Strata, Avair, iPhone") (►Table 2). An unfamiliar query ("Strata"; Medtronics, Ireland) and a nonmedical query ("iPhone"; Apple, United States) added to the difference in expert opinions. The Kappa scores reveals that the participants were not in complete agreement on all suggested keywords for each of the words in the list. To err on the liberal definition of relevancy, we used a union of all expert opinions to judge whether a resulting word is relevant or not, meaning a suggestion is considered being relevant if at least one expert marked is as relevant. Highest union scores were achieved with "asthma" (86.44%), "syncope" (100%) and "fracture" (98.30%), and the lowest scores were on "penicillin" (47.45%) and "Strata" (42.37%). The union score reveals that DeepSuggest has good performance on suggesting relevant keywords for diagnosis, symptoms, and procedures as well as nonmedical terms but moredate performance on medications. On average, 72% of the results from DeepSuggest were found to be relevant. Per expert feedback, we did not include any ranking of the suggestions.

Recall of Known Synonyms

Regarding the aforementioned expert feedback and recall rate of the top N words by DeepSuggest (►Fig. 4), we designed the user-interface to suggest up to 60 terms (expanded upon the user's request) to have a balance of providing enough suggestions without overwhelming the users and achieving a recall rate of over 0.5 (►Fig. 4).

Usability Testing on Retrieving Clinical Notes

Assessing on the dimensions of usability (effectiveness, satisfaction, usefulness, and efficiency), we calculated the average score to each question (7-point Likert scale). We found that the UMUX usability score of the prototype implementation of DeepSuggest (average score = 5.3) was similar to the commercial Epic implementation (average score = 5.4; ►Table 3).

Table 3 Usability scores of DeepSuggest versus Epic

Post-test mean scores[a]		
Components	DeepSuggest	Epic
Effectiveness	6.2	5.7
Satisfaction	4.8	5.5
Usefulness	5.3	5.5
Efficiency	5.0	4.8
Average	5.3	5.4

[a]0–7 scale ($n = 6$).

Particularly, effectiveness and efficiency of DeepSuggest was found better than Epic. However, Epic achieved higher score user satisfaction and usefulness of the system. Participants described DeepSuggest as more "appealing," "efficient," and "comprehensive" than their experience with Epic (see ►Appendix A for details). On average, participants retrieved more relevant clinical notes using DeepSuggest than Epic. An average of 5.6 notes retrieved using DeepSuggest versus 2.6 notes from Epic.

Discussion

Query Expansion: Artificial Intelligence Driven versus Ontology Driven

Utilizing ontologies, a clean and standardized set of synonyms and relationships provided by experts, has a clear advantage when expanding query keywords. Thus, ontology-based expansion usually works well when searching a well-written corpus (i.e., articles on PubMed) but suffers when searching a corpus of documents with many informal, nonstandard shorthands, typos, and layman terms (i.e., clinical notes). Accordingly, current search solutions, like the ontology-driven query in Epic, may not fulfill the needs of clinicians and researchers.[31]

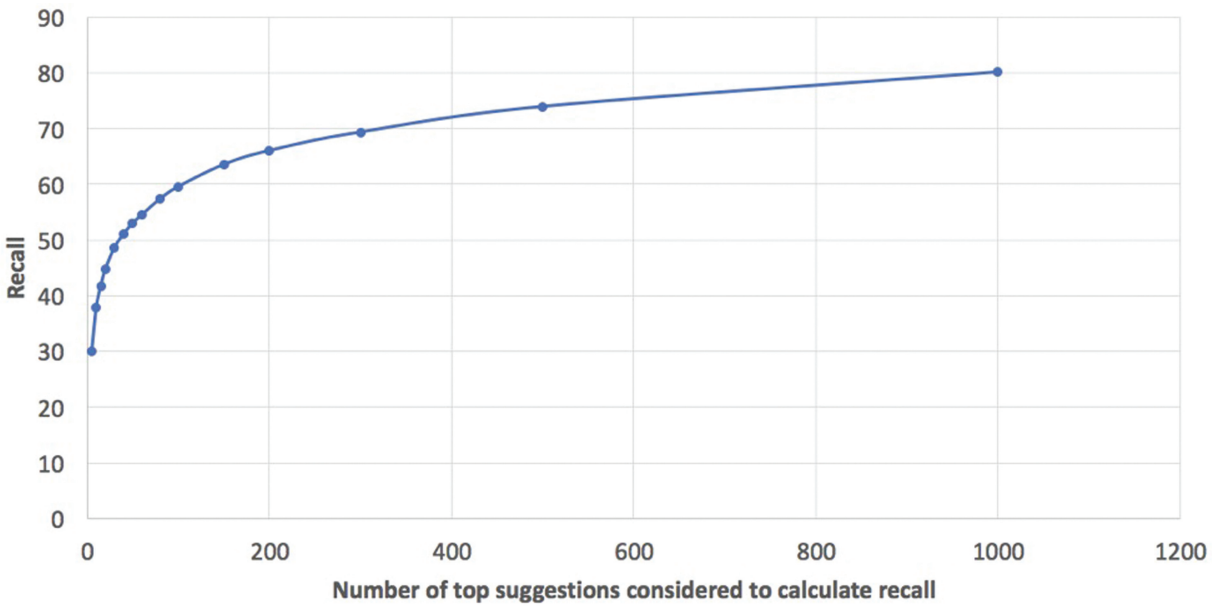

Fig. 4 Recall rate of the top N words by DeepSuggest, considering only Systematized Nomenclature of Medicine synonyms as relevant.

Table 4 A Comparison between two keyword expansion strategies for clinical notes

	Ontology-driven	AI-driven
Example implementations	Epic and Cerner	DeepSuggest
Dictionary size (number of notes)	994,245 in SNOMED (24,921 commonly exist in our local corpus)	6.3 million (extracted from our local corpus)
Creation Technique	Manually created ontology	Automatically learned from a local corpus
Precision on synonym expansion	High due to the use of SNOMED	Lower but acceptable
Recall on synonym expansion	Low, due to out-of-vocabulary terms in clinical notes	Overall high, but lower on known synonyms due to large number of out-of-ontology words in the DeepSuggest corpus.

Abbreviations: AI, artificial intelligence; SNOMED, Systematized Nomenclature of Medicine.

Also, creating and maintaining ontologies up to date is a time-consuming manual process.

The unsupervised neural networks, such as Word2Vec, capture a larger number of related concepts in the clinical note corpus that are beyond the static ontology terms (e.g., generic and user defined abbreviations and common misspellings), which supports the work of Ganesan et al.[10] While the intelligent assistance of DeepSuggest is rudimentary, it is surprisingly effective. DeepSuggest recommends a larger number of relevant words due to its much larger dictionary derived from the local corpus (►Table 4). Participant feedback was positive toward DeepSuggest, with validated usability scores similar to the commercial Epic search system.

However, DeepSuggest cannot replace biomedical ontologies due to its lower recall rate on known synonyms. We have already implemented a new version of DeepSuggest that combined both AI-driven and ontology-driven suggestions. Emerging algorithms can also leverage known ontology relationships in the training process of word embedding.[32,33]

The Lower Recall Rate of DeepSuggest on Known Synonyms

We observed a relatively low recall rate (0.54) on the top 60 suggestions (►Fig. 4) when trying to retrieve synonyms reported in SNOMED. This observation is consistent with the previous report on the difficulty of using unsupervised techniques to outperform ontologies to recall known synonyms.[28] For instance, SNOMED suggests "Excision of tonsil" and "Ts" as top synonyms of tonsillectomy. However, the rank of recall on these two words are pretty low since "T/A," "T&A" and "T + A" are among the top observations in the informal-writing corpus of clinical notes (►Fig. 2F). As such, the evaluation of recall by SNOMED (designed for formal writing) as a gold standard can only act as a proxy.

To study the effect of training size on learning word-similarity in the embedded space, we plotted similarity scores that DeepSuggest returned for 5,000 of SNOMED synonym pairs and compared them with the same set when randomly shuffled. The upward yellow trend in ►Fig. 5 suggests that words with a higher frequency in the NCH corpus tend to be identified as more relevant to their synonym pair reported in the SNOMED database. This only demonstrates a correlation between word frequency and similarity score but not a causation. For example, the lower frequency words might be words with multiple meanings and therefore harder to learn. Although the trend stops when words become highly frequent, the consistency of this stoppage with the distribution of random pairs suggests

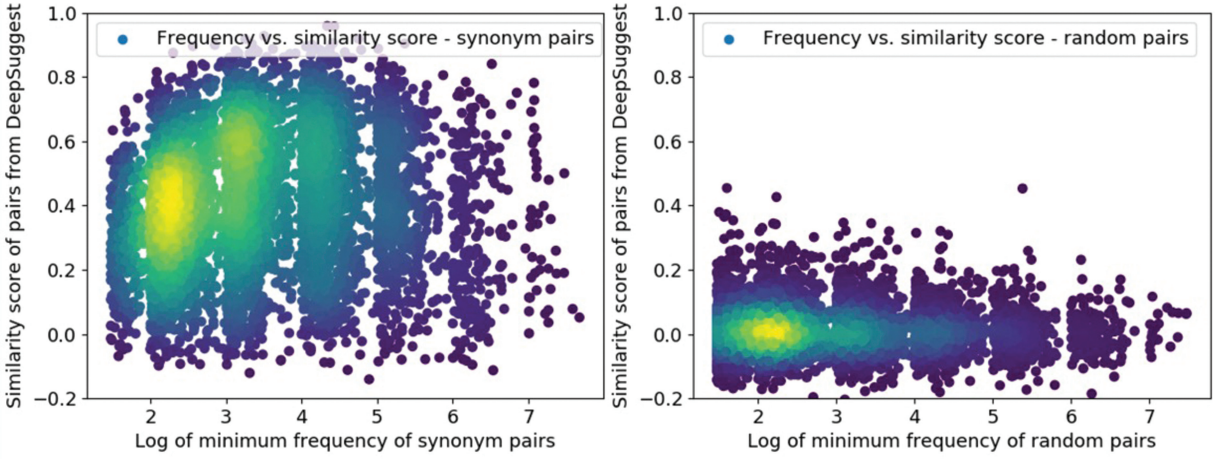

Fig. 5 The effect of frequency of SNOMED words in our training corpus on learning accuracy. The heatmap shows concentration of points, ranging from purple (low) to green (medium) and yellow (high). Left is a distribution of SNOMED synonym pairs, showing an upward trend as the frequency increases, compared with the right graph for random pairs. Synonym pairs have an average similarity score of around 0.4 while random pairs are distributed around the 0.0 similarity score line. SNOMED, Systematized Nomenclature of Medicine.

that the stoppage is due to having less high frequent data points to get plotted.

The actual recall rate remains unknown since SNOMED is highly biased toward formal writing—a comprehensive list of synonyms and relevant words to be used in the evaluation does not exist. As such, although the recall rate on known synonyms is low, the clinical note retrieval amount significantly increased when residents used DeepSuggest over Epic; averaged among six users, DeepSuggest retrieved 5.6 relevant notes versus Epic's 2.6. This increase in retrieving relevant notes demonstrates that DeepSuggest's interactive user interface, shown in ►Fig. 2F, prompts users to include relevant words to their query easily.

In the future, we plan to improve the system by combining AI-driven suggestions with ontology-driven suggestions to have a more comprehensive and accurate list, resulting in 100% precision60 for known synonyms through ontologies. We believe that with the accessibility to the synonyms at SNOMED, we could improve precision by providing additional resource and also a fail-safe method for our model errors on, such as, identifying uncommon words, tokenization.

Reaching Beyond Query Expansion to Knowledge Expansion

We observed during the user evaluations that the proffered query expansion helped users to expand their knowledge of specific words. For example, while participants did not know the term Strata (Medtronics, Ireland) during accuracy testing, the query suggestions hinted at possible relationships and meanings. Interestingly, users stated that the "irrelevant" terms listed by DeepSuggest do not bother them much because the relevance is all contextual. Thus, we speculate that "query expansion" may influence "knowledge expansion" for users in the long term, expanding the user's knowledge of relevant words and any latent relationships. This indirect effect may be investigated further to determine its impact on medical training.

Limitations of DeepSuggest

To balance the time needed to create word embedding models and the need to search for less frequent keywords, DeepSuggest used empirically derived frequency thresholds for N-grams. This will limit search performance on N-grams below the chosen frequency thresholds. Besides the mentioned shortcoming of the algorithm, the major limitation of this study is the lack of ground truth for evaluation. It was not practical to annotate the large corpus for each query. To compensate, we evaluated using data from SNOMED, expert opinion and usability testing. However, precision might be overestimated based on the union score. Although most suggestions from DeepSuggest are relevant to the input query, precision is not perfect (0.72 on average and average Kappa score is "fair" at 0.266), and users' opinions frequently differ regarding relevance and usefulness of the suggestions (►Table 2). In addition, we were not able to check precision of each suggestion due to high volume rather selectively checked by focusing on top results. This may result with false positives and lower selectivity if a selected term or abbreviation have other meanings. The relevance of suggestions may also vary depending on the task and search question of interest. For example, with stomachache as an input, if the goal is to find all patients reported with a stomachache, only variations of stomachache such as tummy ache, belly ache, and stomach pain are relevant, while nausea and vomiting suggested by DeepSuggest will be considered irrelevant. However, if the goal is to find patients with any of the commonly reported gastrointestinal symptoms, nausea and vomiting would be relevant words for expansion. Consequently, we designed the UI to be interactive and allow users to choose keywords of interest for expansion rather than automatically including all suggestions in the new query (►Fig. 2F).

Since DeepSuggest cannot define the meaning of suggested terms, unfamiliar suggested terms are especially challenging. This issue was addressed by an interactive tooltip that displays examples of word usage in the local corpus (►Fig. 6) for each suggested term.

Another limitation of DeepSuggest was the testing environment. We only used the single corpus from a pediatric hospital and evaluated using pediatric relevant keywords. Therefore, the performance of DeepSuggest in adult corpus is unknown. In addition, potential Epic configurations which may affect the vocabulary and terminology of notes entered by providers were not considered as a factor in the evaluation of DeepSuggest.

We carefully designed a user interface (UI), which provides rich information and addresses the aforementioned limitations of the DeepSuggest algorithm regarding precision, recall, and unfamiliar suggestions. The user evaluation in the usability testing suggests that the interface design and implementation were successful.

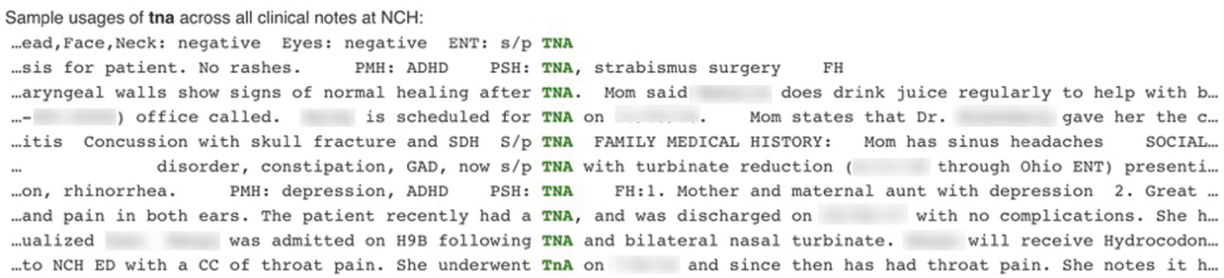

Fig. 6 An interactive tooltip showing the example usage of a word. The tooltip will show up once the user clicks on a suggested word (protected health information redacted).

Other Word Embedding Approaches

Recently, a new generation of word embedding methods emerged: Embeddings from Language Models (ELMo)[34] and Bidirectional Encoder Representations from Transformers (BERT),[35] introduced by AI2 and Google, have gained popularity because of being context aware, as opposed to Word2Vec that provide one embedding vector for each word type regardless of the context the word appeared in. In addition, fastText (a library for text classification and representation) provides a resourceful library for word embedding models.[36] However, when implemented on our corpus, it showed to be ineffective because too many of the suggested words appeared to be heavily morphologically similar, including typos and abnormally large words, in addition to the lack of semantically relevant suggestions. A comprehensive study of different embedding methods is needed in future, but previously studies hint that the performance improvement with different embedding methods may be subtle.[19]

Conclusion

As a proof of concept, DeepSuggest demonstrated its ability to improve retrieval of relevant clinical notes when implemented on our local corpus by suggesting spelling variations, acronyms, and semantically related words. The system shows promise in helping users to achieve a higher recall rate for clinical note searches, thus boosting productivity in clinical care and research. Our novel approach contributed to the science and literature by (1) leveraging neural-network powered unsupervised learning on a locally derived clinical corpus to infer word relatedness, which can be used to enhance the textual search performance for clinical care or research purposes, and (2) retaining human input in the loop to ensure the functionality and effectiveness of the system in real-world scenarios. Future research could also investigate the understanding and perceptions of health care providers with information systems success models[37] and integrate the AI-driven approach with the ontology-driven approach of query expansion. As a future implementation of DeepSuggest, we plan to utilize other embeddings methods utilizing FAISS, and deep learning based models such as BERT and ELMo for training the contextualized embeddings. In addition, we are planning to deploy source code to public repositories for sharing DeepSuggest with other health institutions and researchers.

Protection of Human and Animal Subjects

No human subjects were involved in this project and institutional review board approval was not required.

Clinical Relevance Statement

Implementation of AI in medical records shows promising development in terms of assisting providers better search clinical cases, symptoms, and diagnosis. Intelligent search engines, as DeepSuggest, could improve clinical decision-making through effective search assistance and expansion of queries.

Funding

This study received its financial support from Patient-Centered Outcomes Research Institute (grant number: ME-2017C1-6413).

Conflict of Interest

None declared.

Acknowledgments

The authors would like to acknowledge the contribution of Dan Digby for the user interface implementation, Koushik Ginna for initial programming of the Gensim and annoy library, Robert Strouse for designing the wireframe, and Tran Bourgeois for assisting usability testing. They are also thankful to Melody Davis for constructively copy-editing the manuscript.

The expansion of this study is (partially) supported through a Patient-Centered Outcomes Research Institute (PCORI) Award (ME-2017C1–6413) under the name of "Unlocking Clinical Text in EMR by Query Refinement Using Both Knowledge Bases and Word Embedding."

All statements in this report, including its findings and conclusions are solely those of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee.

References

1 Natarajan K, Stein D, Jain S, Elhadad N. An analysis of clinical queries in an electronic health record search utility. Int J Med Inform 2010;79(07):515–522
2 Terry AL, Chevendra V, Thind A, Stewart M, Marshall JN, Cejic S. Using your electronic medical record for research: a primer for avoiding pitfalls. Fam Pract 2010;27(01):121–126
3 Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care 2013;51(08, Suppl 3):S30–S37
4 Abdulla AAA, Lin H, Xu B, Banbhrani SK. Improving biomedical information retrieval by linear combinations of different query expansion techniques. BMC Bioinformatics 2016;17(Suppl 7):238
5 Rivas AR, Iglesias EL, Borrajo L. Study of query expansion techniques and their application in the biomedical information retrieval. ScientificWorldJournal 2014;2014:132158
6 Wu H, Toti G, Morley KI, et al. SemEHR: a general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. J Am Med Inform Assoc 2018;25(05):530–537
7 Zhu D, Wu S, Carterette B, Liu H. Using large clinical corpora for query expansion in text-based cohort identification. J Biomed Inform 2014;49:275–281
8 Seyfried L, Hanauer DA, Nease D, Albeiruti R, Kavanagh J, Kales HC. Enhanced identification of eligibility for depression research using an electronic medical record search engine. Int J Med Inform 2009;78(12):e13–e18
9 Hanauer DA, Mei Q, Law J, Khanna R, Zheng K. Supporting information retrieval from electronic health records: a report of University of Michigan's nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). J Biomed Inform 2015;55:290–300
10 Ganesan K, Lloyd S, Sarkar V. Discovering related clinical concepts using large amounts of clinical notes: supplementary issue: big data analytics for health. Biomed Eng Comput Biol 2016;7s2: BECB.S36155

11 Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality. NIPS. Accessed 2013 at: https://arxiv.org/abs/1310.4546
12 Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Accessed 2014 at: https://www.aclweb.org/anthology/D14-1162/
13 Minarro-Giménez JA, Marín-Alonso O, Samwald M. Exploring the application of deep learning techniques on medical text corpora. Stud Health Technol Inform 2014;205:584–588
14 Turner CA, Jacobs AD, Marques CK, et al. Word2Vec inversion and traditional text classifiers for phenotyping lupus. BMC Med Inform Decis Mak 2017;17(01):126
15 Glicksberg BS, Miotto R, Johnson KW, et al. Automated disease cohort selection using word embeddings from electronic health records. Pac Symp Biocomput 2018;23:145–156
16 Wang Y, Liu S, Afzal N, et al. A comparison of word embeddings for the biomedical natural language processing. J Biomed Inform 2018;87:12–20
17 Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform 2018;22(05):1589–1604
18 Wang Y, Rastegar-Mojarad M, Komandur-Elayavilli R, Liu H. Leveraging word embeddings and medical entity extraction for biomedical dataset retrieval using unstructured texts. Database (Oxford) 2017;2017:bax091
19 Ye C, Fabbri D. Extracting similar terms from multiple EMR-based semantic embeddings to support chart reviews. J Biomed Inform 2018;83:63–72
20 Roberts K, Demner-Fushman D, Voorhees EM, et al. Overview of the TREC 2017 Precision Medicine Track. In TREC. Accessed 2017 at: https://pubmed.ncbi.nlm.nih.gov/32776021/
21 Galitz WO. The Essential Guide to User Interface Design: An Introduction to GUI Design Principles and Techniques. John Wiley & Sons; 2007
22 Cheung CS, Tong EL, Cheung NT, et al. Factors associated with adoption of the electronic health record system among primary care physicians. JMIR Med Inform 2013;1(01):e1
23 Bernhardsson E. Annoy: Approximate Nearest Neighbors in C++/Python. Python package version 1.13. Accessed 2018 at: https://pypi.org/project/annoy/
24 Norinkavich KM, Howie G, Cariofiles P. Quality improvement study of day surgery for tonsillectomy and adenoidectomy patients. Pediatr Nurs 1995;21(04):341–344
25 Turchin A, Pendergrass ML, Kohane IS. DITTO - a tool for identification of patient cohorts from the text of physician notes in the electronic medical record. AMIA Annu Symp Proc 2005;2005:744–748
26 Salvadores M, Alexander PR, Musen MA, Noy NF. BioPortal as a dataset of linked biomedical ontologies and terminologies in RDF. Semant Web 2013;4(03):277–284
27 McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(03):276–282
28 Henriksson A, Conway M, Duneld M, Chapman WW. Identifying synonymy between SNOMED clinical terms of varying length using distributional analysis of electronic health records. AMIA Annu Symp Proc 2013;2013:600–609
29 Finstad K. The usability metric for user experience. Interact Comput 2010;22:323–327
30 Joey B, Trish M. Measuring Desirability: New methods for evaluating desirability in a usability lab setting. Proceedings of Usability Professionals Association. Accessed 2002 at: https://www.researchgate.net/publication/228721563_Measuring_Desirability_New_-methods_for_evaluating_desirability_in_a_usability_lab_setting
31 Davis Z, Khansa L. Evaluating the epic electronic medical record system: a dichotomy in perspectives and solution recommendations. Health Policy Technol 2016;5:65–73
32 Bian J, Gao B, Liu T-Y. Knowledge-Powered Deep Learning for Word Embedding. In: Lecture Notes in Computer Science. 2014:132–148
33 Grover A, Leskovec J. node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining 2016 August 13; 855–864. Accessed 2016 at: https://dl.acm.org/doi/10.1145/2939672.2939754
34 Peters ME, Neumann M, Iyyer M, et al. Deep contextualized word representations. arXiv preprint. Accessed 2018 at: https://arxiv.-org/abs/1802.05365?ref=hackernoon.com
35 Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. Accessed 2018 at: https://arxiv.org/abs/1810.04805
36 Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Linguist 2017;5:135–146
37 The DeLone and McLean Model of Information Systems Success. A ten-year update. J Manage Inf Syst 2003;19:9–30

Appendix A Usability testing items

UMUX Questionnaire								
Instructions: Please read the questions and select your response from 1 to 7 for the system you have just used (Epic or DeepSuggest). Response scales: (1) strongly disagree, (2) disagree, (3) somewhat disagree, (4) neither agree nor disagree, (5) somewhat agree, (6) agree, and (7) strongly agree.								
Questions		Responses (scale)						
1	The system capabilities meet my requirements	(1)	(2)	(3)	(4)	(5)	(6)	(7)
2	Using this system is a frustrating experience	(1)	(2)	(3)	(4)	(5)	(6)	(7)
3	This system is easy to use	(1)	(2)	(3)	(4)	(5)	(6)	(7)
4	I have spent too much time correcting things with this system	(1)	(2)	(3)	(4)	(5)	(6)	(7)

Product reaction cards				
Instructions: Select/highlight how you feel about the system.				
Accessible	Desirable	Gets in the way	Patronizing	Stressful
Appealing	Easy to use	Hard to use	Personal	Time consuming
Attractive	Efficient	High quality	Predictable	Timesaving
Busy	Empowering	Inconsistent	Relevant	Too technical
Collaborative	Exciting	Intimidating	Reliable	Trustworthy
Complex	Familiar	Inviting	Rigid	Uncontrollable
Comprehensive	Fast	Motivating	Simplistic	Unconventional
Confusing	Flexible	Not valuable	Slow	Unpredictable
Connected	Fresh	Organized	Sophisticated	Usable
Consistent	Frustrating	Overbearing	Stimulating	Useful
Customizable	Fun	Overwhelming	Straight Forward	Valuable

Abbreviation: UMUX, Usability Metric for User Experience.