Supplementary material

Supplementary Appendix 1: UTD restructuring and reorganizing data quality topics described in scoping review articles about EMR data

Data quality topic	Source	Methods	Strengths/Limitations	Use case
Misspelled words	Lai et al	The authors measured spelling errors for clinical and general terms. A clinical term was a word that was not a name or was not present in the Aspells [31] default dictionary. A general term was any other word that was not a clinical term. These methods were used to identify both clinical and general terms: 1) used many dictionaries for misspelling detection 2) used Named Entity Recognition (NER) to avoid misclassification of person names as misspellings 3) used regular expressions to correct emails, URLs, and commonly misspelled words	**Strengths:** A combination of previous methods that worked well. **Limitations:** 1) Since it's a composite of previous methods so its incomparable to previous methods. 2) Authors did not attempt to investigate whether misspellings & misspelling corrections impacted the quality of the information extracted from medical texts.	Applied to: 1) clinical notes from primary care clinics 2) free-text allergy entries 3) free-text medication orders
	Ruch et al	Methods used to improve spelling correction: module 1) string to string edit distance known as Damerau-Levenshtein module 2) syntactic correction – to address word order and agreement problems module 3) processing words with the same parts of speech by applying contextual word sense disambiguation	**Strengths:** NLP tools such as NER and lexical disambiguation can reduce spelling corrections, this can allow for automated processes. **Limitations:** Authors did not attempt to investigate whether misspellings & misspelling corrections impacted on the quality of the information extracted from medical texts.	Applied to electronic patient records
Security	Pantazos et al	The methods used for de-identification has to meet the criteria of: 1) Medical correctness - each health record must show a true medical picture of a patient 2) Anonymity - it's not possible to see who the real patient is 3) Readability - the health record has to represent reality 4) Consistency - the patient's identifiers have to be consistent with the entire medical picture Each method utilized a mapping table to replace existing identifiers with new identifiers. To treat ambiguity, when the de-identification program meets an ambiguous word it: 1) gets deleted if it's a rare word (occurs less than 200 times) 2) gets left if it occurs more than 200 times permutation tables - that mapped existing identifiers with new ones - ensured readability and consistency distorted identifier tables - mapped existing civil registration numbers to a scrambled one - ensured readability and correctness	**Strengths:** While de-identifying EMR data, the authors tried to preserve essential components of the note to mimic the original note. **Limitations:** 1) Did not address spelling errors. 2) Overlooked pharmaceutical names. 3) Missed several clinical abbreviations. 4) Cannot do analysis on geography due to scrambled zip codes. 5) For statistical purposes de-identification should not impact readability or consistency.	Applied to de-identify an existing EMR database
Reducing word variability	Assale et al	1) Tokenization, removal of stop words, numbers and non-ASCII symbols 2) Counted frequency of words - words occurring above an 80% percentage were considered correct, less frequent words were checked if they were present in an Italian dictionary or a medical dictionary, whatever is left over was considered typographical. 3) Used distance metric between strings - "Levenshtein distance". Search in the 80% of most frequent words that had a "distance 1" from the typos. Distance 1 signifies that typological words differ from 1 letter insertion/deletion/substitution from the original word.	**Strengths:** Proposed a method for reducing word variability **Limitations:** 1) Number of false positives is high. 2) Cannot guarantee to correct all errors.	Applied to anamnestic summaries of endocrinology and rheumatology

Continued

Supplementary Appendix 1: Continued

Data quality topic	Source	Methods	Strengths/Limitations	Use case
		4) Also take into account "Damerau-Levenshtein distance" metric where it also takes into account inversions between letters - because its common to invert two adjacent letters (distance 2). 5) Manually inspected to verify no association errors. Ambiguous associations were discarded. 6) Multi-associated words (i.e., words with the same meaning but varied in spelling) were replaced with the most frequent one.		
Sources of noise	Berndt et al	1) converting text to lowercase 2) tokenization 3) removal of tokens with less than three characters or no alphabetical characters 4) normalizing terms 5) removal of stop words 6) removing tokens that only occur once	**Strengths:** Authors addressed text noise by introducing preprocessing methods to reduce noise **Limitations:** The study only used one dataset	Applied to clinical progress notes
Quality of annotations	He et al	Annotation methods included: 1) word segmentation 2) part of speech tagging (with shallow and full parsing of parts of speech tags) 3) named entity tagging 4) relational tagging (i.e., finding relationships among named entities) Annotation quality was evaluated using F1 measure, precision, and recall	**Strengths:** Authors have built a concept of data quality into the creation of a corpus by assessing the quality of annotations **Limitations:** Since there was a limit of annotation resources, the corpus created only covered two departments of a hospital, thus lacking medical terminology in other departments	Applied to: 1) discharge summaries 2) clinical progress notes
	Berndt et al	Quality of Annotations include: 1) "Fall" or "Not fall" was given an operational definition 2) Manual annotation evaluated with Cohen's kappa for interrater agreement	**Strengths:** Generally manual annotations are used as a reference standard to be used for machine learning classification algorithms **Limitations:** Manual annotations are time consuming, costly, and can lead to manual error	Applied to clinical progress notes
Ambiguous abbreviations	Joopudi et al	Utilized a convolutional neural network (CNN) that: 1) Was trained on word embeddings (i.e., a representation of words in a vector) located on journal articles in PubMed 2) Was trained on parts of speech tags 3) Used clinical notes meta information such as author	**Strengths:** Using deep learning allows the user to bypass feature engineering tasks which can be time consuming, furthermore the CNN model worked well on disambiguating abbreviations **Limitations:** While the authors have shown that methods used to disambiguate abbreviations works well, there is no investigation if the corrections have had an impact on the quality of information.	Applied to: 1) de-identified longitudinal patient records with clinical notes 3) publicly available dataset created by University of Minnesota
Reducing manual annotation	Liang et al	Utilized KNN classifier (supervised method) to predict the type of documents (i.e., a document is either "diagnostic errors" or "device related complications"). The process was as follows: 1) Noise Removal: removal of punctuations, words were set to lowercase, white space was removed, stop words were removed 2) Document was converted to a document term matrix 3) Documents were pre-annotated as either "diagnostic errors" or "device related complications" to create a gold standard comparison 4) Evaluated using F measure and Accuracy	**Strengths:** Demonstrated a process that enhances automatic annotation processes that include data quality elements. **Limitations:** The sample size in the study was small and this method was not demonstrated on other types of data such as EMR clinical notes.	Applied to publicly available patient safety documents from WebM&M