10.1148/ryai.2020190137

Appendix E1

PHI Category Definition

Safe Harbor defines 18 types of identifiers which, when removed, are sufficient to deidentify data, provided the covered entity has no actual knowledge that the remaining data could not be used to identify an individual.

The Safe Harbor method does not explicitly require the removal of information related to the names and addresses of health care workers or health care systems who interact with the subjects of a report. However, we have opted to include these in our annotations and comparisons, as this information is often relatively specific and distinguishable (eg, "patient recently received care at XXX private practice"), particularly when combined with other identifying information. Furthermore, removing this information does not affect the quality of the dataset for downstream applications. Existing public datasets of clinical notes, such as MIMIC-III and the i2b2 2014 De-Identification Challenge Dataset, also remove this information for complete de-identification. We have therefore included two additional categories in our annotations: "clinician names" and "hospital/institution names". In addition, we have added an annotation category for "vendor and tool names", which is designed to capture specific medical device and software vendor names, as well as specific products or platforms (developed in-house or by a third-party vendor), which could serve to help identify the location at which a subject received care.

For the purpose of maintaining consistent date-shifts in downstream applications, we included in the "dates" category spans of text consisting only of a year or month, if that year or month was relevant to the patient's medical history. For instance, in the sentence "Comparison is made to MRI from last August", the text span "August" would be tagged as a PHI element. However, the publication date of a journal article relevant to the patient's case would not be tagged as PHI.

Other similar studies have often labeled *all* names, dates, or ages appearing in the text as PHI to be redacted, likely because it is easier to design systems that identify *all* strings formatted as names or dates than just those which represent PHI. However, as *(a)* we wish to quantify the actual amount of PHI appearing in our corpus, and *(b)* removing or obfuscating non-PHI information may impair the quality of the dataset (eg, a patient's age may be informative about risks for conditions mentioned in the radiology report), we only label names, ages, and dates that represent PHI (for instance, the name of an author of a journal article cited in a radiology report, just as the year of publication above, is not labeled).

Appendix E2

Search Strategy

The authors identified publicly-available de-identification software packages designed to be used out-of-the-box through a keyword search in MEDLINE and Cochrane Library, using the key terms "anonymi*," "data loss prevention," "de-identification," "optical character recognition,"

"regular expression," "parsing," "deidentif*," "de-identif*," "medical data," "pathology report," "radiology report," "electronic medical record," "health record," "personal health data," and "protected health information." In addition, major bioinformatics and clinical informatics journals were searched for similar terms. References from included or related studies and reviews were followed to identify additional articles.

Appendix E3

Parameters and Training Protocols

General.—

Many of the software packages involve their own word tokenization steps as part of their pipeline. While the vast majority of tokens produced by the different systems are aligned, in some cases, this may result in the tested systems labeling only *part* of one of our reference-standard tokens as PHI. We resolve this by saying that if *any* character of one of our (SpaCy-tokenized) tokens is labeled by a system as PHI, the entire token will be considered PHI by that system.

MIST/Carafe.—

For MIST/Carafe, we used Carafe's prepackaged tokenizer to split the full documents into individual tokens, the default sentence boundary detector, and the default BIO (beginning-inside-outside) tagging scheme (19), which labels each token as either the *beginning* of a PHI mention, *inside* a PHI mention, or *outside* of a PHI mention. For instance, in a correct labeling of the sentence, "Information received by Dr. Jane Doe," "Dr." would be labeled as "**B**-HC_NAME," indicating the *beginning* of a span of text corresponding to the name of a health care provider, "Jane" and "Doe" would be labeled "**I**-HC_NAME" as these tokens are *inside* of the same HC_NAME span, and the other words would be labeled "**O**," as they are not part of any PHI text span. The default conditional log likelihood maximization was used as a training objective with L2 parameter regularization to prevent overfitting. Training was run until convergence with a maximum of 200 iterations. The default word feature set was used.

NLM-Scrubber.—

The NLM-Scrubber produces only four output categories: "PERSONALNAME," "ADDRESS," "ALPHANUMERICID," and "DATE." We mapped our labels to these four labels in a many-to-one manner as follows: our "Patient Names," "Healthcare Provider Names," "Vendors/Tools," and "Healthcare Location Names" categories mapped to NLM's "PERSONALNAME" category. Our "Addresses/Geographic Locations" category mapped to NLM's "ADDRESS" category. Our "Phone Numbers," "Medical Record Numbers," and "Other Identification Codes" categories mapped to NLM's "ALPHANUMERICID" category, and our "Dates" category mapped to their "DATE" category. The system's rules were not designed to remove titles such as "Dr.," "MD," "RN," and the like, so we modified our labels to exclude this information before running the classifier. Similarly, commas and periods within date strings, eg, the comma in "May 4, 2018" were not designed to be removed, so we modified our labels in order not to penalize the system for not including these tokens as PHI. Lastly, their system considered *time* strings, eg, "5:15 PM" as PHI, whereas our annotators did not, so we similarly did not penalize the system for including time strings. The system cannot process reports with non–American Standard Code for

Information Interchange (ASCII) characters, so we replaced these characters with whitespace in the reports which have them.

MIT deid.—

We used the default settings and the "deid.pl" perl script to produce predictions for each document. All predicted "Name" string categories were mapped to a single "NAMES" category, which corresponds to the combination of "Patient Names" and "Healthcare Provider Names" in our PHI schema. In addition, MIT deid predictions for "Hospital" and "Address" were mapped to one category, "ADDRESS," as were our "Healthcare Location Names" and "Addresses/Geographic Locations" categories. Otherwise, there was a relatively clean one-to-one correspondence between our categories and the system's. As with the NLM-Scrubber, we modified our labels in order not to penalize the machine for "missing" commas and periods within date strings, as well as titles such as "Dr." or "RN" within name strings.

Emory HIDE.—

We used all of the default parameters for CRFSuite, over the features generated by HIDE's software.

NeuroNER.—

We converted our data into the BIO format expected by NeuroNER (see the previous MIST/Carafe section). We used the default hyperparameters and settings to produce predictions for each document (notably: SpaCy tokenizer, stochastic gradient descent optimizer, character embeddings and hidden states with 25 dimensions, using a conditional random field to produce the final predictions).