**I. List Components and Creation Process**

**A. Whitelist Creation**

**1. Whitelist with all original components (whitelist\_061418.json)**

We assembled a whitelist of ‘safe’ words using lists of terms and phrases from various medical term libraries and common word lists. First, medical terms from the NLM UMLS Lexicon (Source 1), the NLM MeSH thesaurus (Source 2), the MeSH scope notes (Source 2), the MeSH tree file for biomedical data (Source 3), the Medline Plus heath topic records (Source 4), and medical terms and definitions from SNOMED CT US (Source 5) were added to the Whitelist. Two lists of common medical abbreviations were added to the Whitelist (Source 6). FDA-approved drug names, dosages, and active ingredients (Source 7) and a list of ICD9 diagnosis names and codes were also added to the Whitelist (Source 8). Next, lists of the 20k most common English words (Source 9) and 1k most common verbs and their tenses (Source 10) were added to the whitelist.

**2. FP addition round 1 (whitelist\_plus\_fps-3\_cleaned.json)**

~1000 false positives generated by running whitelist\_061418.json only on ~500 notes from UCSF training batch 3 and the I2B2 corpus, and then selecting the union of these false positive lists, were added to the Whitelist. All added words were tokenized on whitespace and symbols, and converted to lowercase. All single letters, numeric and alphanumeric words were removed from the Whitelist.

**3. Names subtraction round 1 (whitelist\_090618\_nonames.json)**

All first names from the Social Security National data file (Source 11, see Blacklist Creation for details) and all last names from the 2010 US Census (Source 12), were removed from the whitelist.

**4. FP addition round 2 (whitelist\_090618\_nonames\_addback100\_ucsf\_u\_i2b2.json)**

~6700 false positives generated by running whitelist\_nonames.json only on UCSF training batch 5 and the I2B2 corpus, and then selecting the union of these false positive lists, were added to the Whitelist. All added words were tokenized on whitespace and symbols, and converted to lowercase. All single letters and integers were removed from the Whitelist, but mixed alphanumeric words were not removed from the Whitelist.

**5. Names subtraction round 2 (whitelist\_090618\_nonames\_addback100\_ucsf\_u\_i2b2\_noperiods\_noeponyms.json)**

A list of 638 medical terms named after people (“eponymous terms”) were compiled from a Wikipedia summary of the most common eponymous medical terms (Source 13). 209 of these terms overlapped with whitelist terms in whitelist\_090618\_nonames\_addback100\_ucsf\_u\_i2b2.json and were removed from the Whitelist.

**6. Common words addition (whitelist\_090618\_nonames\_addback100\_ucsf\_u\_i2b2\_noperiods\_noeponyms\_commonwords.json)**

9 common words that contributed the most false positives during whitelist-only tests on UCSF training batch 5 and the I2B2 corpus after subtracting the names in Step 5 were added back to the Whitelist.

**7. Names subtraction round 3 (whitelist\_090618\_nonames\_addback100\_ucsf\_u\_i2b2\_noperiods\_noeponyms\_commonwords\_nonames.json)**

The overlap between whitelist\_090618\_nonames\_addback100\_ucsf\_u\_i2b2\_noperiods\_noeponyms\_commonwords.json and the full list of Social Security and 2010 Census names was manually inspected, and a list of 38 non-ambiguous names (defined as names that were more likely to be a name than a common word) were removed from the Whitelist.

**8. FP addition round 3 (whitelist\_plus\_fps\_091718\_nonames.json)**

A list of false positives, generated by running whitelist\_090618\_nonames\_addback100\_ucsf\_u\_i2b2\_noperiods\_noeponyms\_commonwords.json only on UCSF training batch 5, was sorted manually into four different categories of false positives: medical terms, medical acronyms, UCSF-specific medical terms, and common English words. All 240 words in these four lists were added to the Whitelist.

**B. Blacklist Creation**

**1. Blacklist with all original components (names\_blacklist\_061418-1.json)**

We assembled a Blacklist of names using last names occurring 100 or more times in the 2010 U.S. census (Source 12), and first names occurring 5 or more times for each year of birth from 1879-2017 from the U.S. Social Security website (Source 11). All single letters were removed from the Blacklist, and all names added to the Blacklist were tokenized on whitespace and symbols, and converted to lowercase.

**2. FP subtraction (names\_blacklist\_sub\_100\_fps.json)**

To minimize the number of common words (i.e. new, walks, knee, home, child, etc.) in the Blacklist, we removed 977 words in the Blacklist that contributed to precision errors when names\_blacklist\_061418-1.json was run on the UCSF training batch 3.

**3. Names addition (names\_blacklist\_sub\_100\_fps\_withnames\_x2.json)**

The list of 977 false positives was manually examined for words that were either ambiguous names (i.e. words that could either be a common word or a name) or unambiguous names (i.e. words that are most likely names, and not common words). 122 of these words were added back to the Blacklist.

**4. Blacklist separation (firstnames\_minus\_fps.json and lastnames\_minus\_fps.json)**

A first names blacklist was created by subtracting the list of false positives in Step 2 from the full Social Security first names list, and adding the names in Step 3. A last names blacklist was created by subtracting the list of false positives in Step 2 from the full 2010 Census last names list, and adding the names in Step 3.

**Whitelist and Blacklist Online Sources:**

* Source 1: <https://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lexicon/2017/release/LEX/LEXICON>
* Source 2: <https://www.nlm.nih.gov/mesh/>
* Source 3: <ftp://nlmpubs.nlm.nih.gov/online/mesh/2017/>
* Source 4:  <https://medlineplus.gov/xml.html>
* Source 5: <https://www.nlm.nih.gov/healthit/snomedct/us_edition.html>
* Source 6: <https://nurseslabs.com/medical-terminologies-abbreviations-listcheat-sheet/>
* Source 7: <https://www.fda.gov/Drugs/InformationOnDrugs/ucm079750.htm>
* Source 8: <https://www.cms.gov/Medicare/Coding/ICD9ProviderDiagnosticCodes/codes.html>
* Source 9: <https://raw.githubusercontent.com/first20hours/google-10000-english/master/20k.txt>
* Source 10: <https://www.worldclasslearning.com/english/five-verb-forms.html>
* Source 11: <https://www.ssa.gov/oact/babynames/limits.html>
* Source 12: <https://www.census.gov/topics/population/genealogy/data/2010_surnames.html>
* Source 13: <https://en.wikipedia.org/wiki/List_of_eponymously_named_diseases>