ICD-O Fuzzy String Mapping

J. Allen Baron

11/9/2021

Contents

1 PURPOSE

To attempt to map ICD-O names to DO names using approximate ("fuzzy") string matching.

2 Data Setup

The latest version of DO (v2021-10-11) was loaded into a Virtuoso database previously in `notebooks/new_defs-mesh-ncit` and will be reused here. The ICD-O file was provided by Lynn.

The number of DO terms is significantly larger than ICD-O terms, as DO includes all diseases whereas ICD-O focuses on cancer. Instead of using DO in its entirety, I limited the DO terms t only those in the `disease of cellular proliferation` branch. This will help to reduce computation time and spurious matches.

After retrieving only DO cellular proliferation terms, it's interesting to note that the total diseases from each resource are similar (DO: 2,736; ICD-O: 2,955)

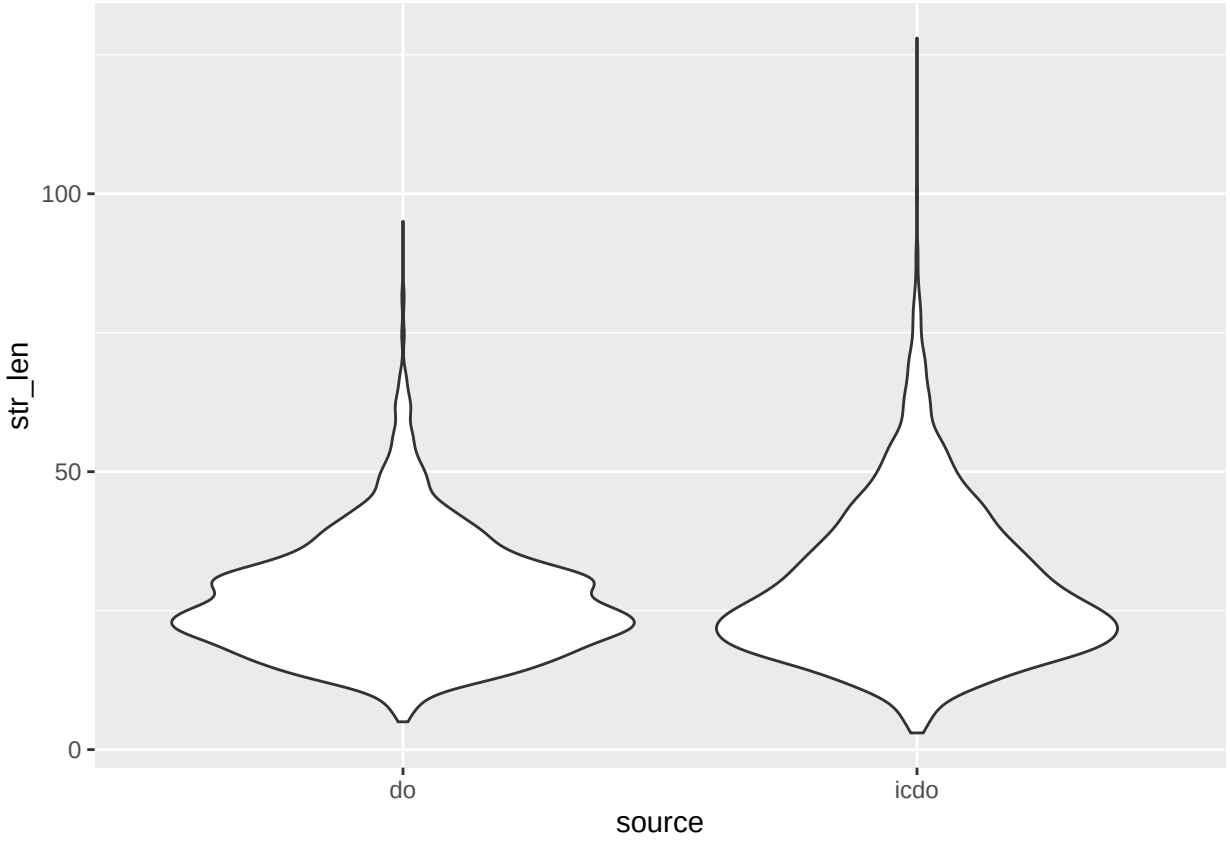
3 Data Preparation

Approximate string matching can take a significant amount of computational resources and can very much be a process of trial and error. To reduce the amount of error, I explored some of the characteristics of the labels in the two datasets.

3.1 Explore String Length

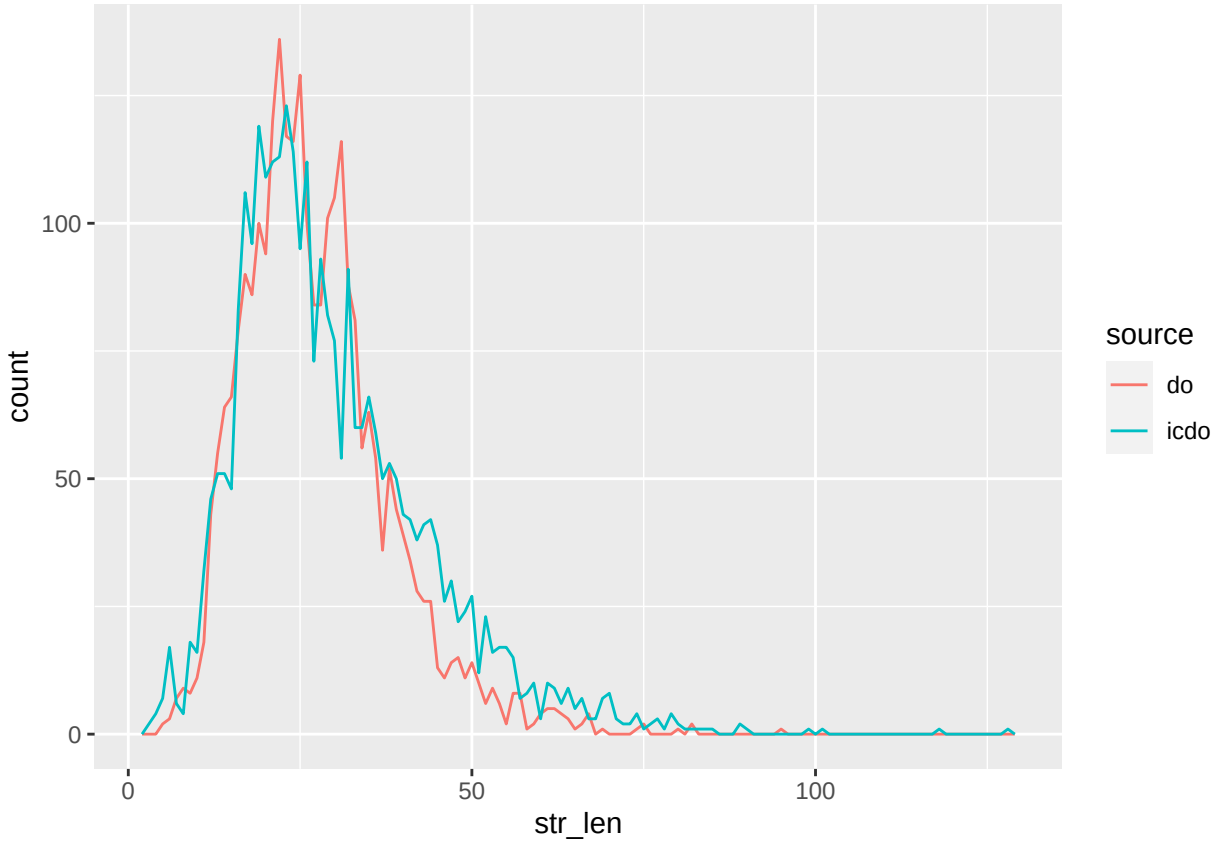
String length influences the number of mismatches that can be allowed before producing completely spurious matches, affecting the balance of false positives and negatives.

Violin plots show that the length of labels in DO and ICD-O are similar, with a maximum around 25 characters and the majority <75. ICD-O tends to have somewhat longer labels with a maximum ~125, compared to DO's ~95.



Previous experience has taught me to limit mismatches to <80% of the maximal length to avoid spurious matches. With a maximal length of 25 that's pretty small for this dataset. It might be possible to break terms up by length but that would depend on if the longer terms in the two sources happen to correspond with one another.

To look at overall counts, I also created a frequency plot and compared 6-stat summaries. No new insights are immediately apparent.



|source | Min.| 1st Qu.| Median| Mean| 3rd Qu.| Max.| |:——|—-:|——-:|——:|———:|——-:|—-:| |do | 5| 20| 26| 27.32091| 33| 95| |icdo | 3| 20| 26| 29.59154| 37| 128|

4 Explore Label Patterns

Both resources have a set pattern for creating labels. If that pattern is readily apparent, I might be able to rearrange one set for better matching results.

Comparing common words found in cancer names across these two resources, the following patterns become apparent:

1. Both use “carcinoma” similarly and primarily at the end of labels
2. DO tends to have “malignant” at the beginning while ICD-O tends to have it at the end after a comma.
3. DO uses “benign” in conjunction with “neoplasm” or “tumor” primarily at the end, while ICD-O uses it in much the same way as malignant
4. Both use “neoplasm” similarly
5. Whereas DO uses “cancer”, ICD-O does NOT. In fact, matches to “cancer” only appear in terms of “precancerous melanosis”

n	do	icdo
1	vulva squamous cell carcinoma	Adenocarcinoma, endocervical type, NOS
2	adenosquamous prostate carcinoma	Intraductal carcinoma, solid type
3	pancreatic signet ring cell adenocarcinoma	Gelatinous adenocarcinoma
4	prostate carcinoma	Infiltrating lobular carcinoma, NOS

n	do	icdo
5	transitional cell carcinoma	Verrucous squamous cell carcinoma
6	gastroesophageal junction adenocarcinoma	Serrated adenocarcinoma
7	subglottis verrucous carcinoma	Lepidic adenocarcinoma
8	fibroepithelial basal cell carcinoma	Squamous cell carcinoma, nonkeratinizing, NOS
9	Bartholin's gland adenocarcinoma	Micropapillary serous carcinoma
10	urethra transitional cell carcinoma	Perihilar cholangiocarcinoma

n	do	icdo
1	malignant type AB thymoma	Granular cell myoblastoma, malignant
2	malignant ovarian germ cell neoplasm	Epithelioid mesothelioma, malignant
3	malignant cystadenoma	Endometrioid cystadenofibroma, malignant
4	nodular malignant melanoma	Perivascular epithelioid tumor, malignant
5	malignant syringoma	Steroid cell tumor, malignant
6	ovarian malignant mesothelioma	Soft tissue tumor, malignant
7	epithelioid malignant peripheral nerve sheath tumor	Thecoma, malignant
8	childhood malignant hemangiopericytoma	Giant cell tumor of tendon sheath, malignant
9	malignant breast melanoma	Mixed tumor, salivary gland type, malignant
10	malignant anus melanoma	Eccrine poroma, malignant

n	do	icdo
1	ovarian benign neoplasm	Mesothelioma, benign
2	anus benign neoplasm	Arrhenoblastoma, benign
3	female reproductive organ benign neoplasm	Unclassified tumor, benign
4	laryngeal benign neoplasm	Skin appendage tumor, benign
5	small intestine benign neoplasm	Adenomyoepithelioma, benign
6	ureteral benign neoplasm	Deep benign fibrous histiocytoma
7	peripheral nervous system benign neoplasm	Hemangiopericytoma, benign
8	intestinal benign neoplasm	Neoplasm, benign
9	Bartholin's gland benign neoplasm	Sweat gland tumor, benign
10	organ system benign neoplasm	Mesenchymoma, benign

n	do	icdo
1	auditory system benign neoplasm	Acinar cell neoplasms
2	parietal lobe neoplasm	Intraductal papillary neoplasm with associated invasive carcinoma
3	jejunal neoplasm	Epithelial neoplasms, NOS
4	myeloid neoplasm	Myeloid or lymphoid neoplasm with PCM1-JAK2
5	hypothalamic neoplasm	Squamous cell neoplasms
6	supraglottis neoplasm	Ductal and lobular neoplasms
7	sensory organ benign neoplasm	Solid and papillary epithelial neoplasm
8	malignant cardiac peripheral nerve sheath neoplasm	Complex mixed and stromal neoplasms
9	ovary neuroendocrine neoplasm	Intraductal papillary mucinous neoplasm with an associated invasive carcinoma
10	vulvar benign neoplasm	Poorly differentiated neuroendocrine neoplasm

n	do	icdo
1	skull base cancer	Precancerous melanosis, NOS
2	retinal cell cancer	Malignant melanoma in precancerous melanosis
3	diffuse gastric cancer	NA
4	vaginal cancer	NA
5	mixed lacrimal gland cancer	NA
6	central nervous system hematologic cancer	NA
7	notochordal cancer	NA
8	lower lip cancer	NA
9	hepatic flexure cancer	NA
10	pericardium cancer	NA

Based on these observations & other known differences, matches might be improved by modifying ICD-O terms to create synonyms as follows:

1. Remove ", NOS".
2. Move "malignant" to the beginning (and dropping the preceding comma).
3. Move "benign" to:
    - the beginning
    - the end
    - the end followed by "neoplasm"
    - the end followed by "tumor"
4. Replace "carcinoma" with "cancer".

5 Improve Labels for Matching

I executed the modifications mentioned in the previous section.

Then, saved the output to repeat grounding with pyobo/GILDA (which I tweaked slightly to track the synonym modifications I made).