L3CatTXST: NIH Long COVID Computational Challenge (L3C) Results

Jelena Tešić on behalf of N3C
Computer Science

Rasim Musal
Information Systems and Analytics

TEXAS STATE UNIVERSITY

N3C Long COVID Challenge

National COVID Cohort Collaborative L3C challenge: determine if the patient who has tested positive for SARS-CoV-2 in an outpatient or hospital setting (ICU or non-ICU) developing PASC/Long COVID.

- The N3C's data consists of existing patient records at 94 participating institutions.
- The data itself can only be accessed through a secure cloud portal hosted by NCATS known as the Enclave
- With collaborative efforts it consists of:
- 20 billion rows, 1,757.1 million clinical observations, 16.4 million patients, and 6,438,192 SARS-CoV-2 cases
- Under the university's DUR we have access to Level 2 data.

Exploratory Data Analysis

- 57,672 patients total
- 9,031 patients were recorded as testing positive for Long COVID after four weeks after the infection

Data Set	Rows X Columns (% missing values) Test	Train
care_site	8,367 x 8 (66%)	26 x 8 (1%)
condition_era	2,484,521 x 8 (0%)	13,639 x 8 0%
condition_occurrence	**6,316,765 x 21 (37%)**	**36,451 x 21 (35%)**
condition_to_macro	**1,286,673 x 8 (6%)**	**8,388 x 8 (5%)**
device_exposure	422,167 x 19 (44%)	2,836 x 19 (45%)
drug_era	**2,090,455 x 9 (0%)**	**12,698 x 9 (0%)**
drug_exposure	**13,611,559 x 28 (42%)**	**66,050 x 28 (39%)**
location	25,142 x 9 (57%)	281 x 9 (60%)
long_COVID	**57,675 x 2 (13%)**	**300 x 2 (0%)**
manifest_safe	69 x 6 (23%)	300 x 13 (23%)
measure	32,569,723 x 29 (33%)	198,151 x 30 (33%)
measure_to_macro	17,839,906 x 8 (0%)	112,243 x 8 (2%)
micro_to_macro	3,524,398 x 28 (54%)	19,430 x 26 (54%)
note	321,151 x 19 (59%)	2,710 x 19 (66%)
note_nlp	7,580,262 x 21 (38%)	60,486 x 21 (45%)
observation	**6,869,266 x 25 (49%)**	**43,355 x 25 (43%)**
observation_period	**45,404 x 7 (0%)**	**234 x 7 (0%)**
payer_plan_period	1,370,746 x 26 (69%)	6,029 x 26 (69%)
person	**57,672 x 26 (29%)**	**300 x 26 (28%)**
procedure_occurrence	**278,981 x 19 (22%)**	**14,645 x 19 (23%)**
procedures_to_macro	991,579 x 8 (5%)	5,247 x 8 (5%)
provider	31,664 x 18 (51%)	477 x 18 (56%)
visit_occurrence	350,934 x 23 (49%)	19,411 x 23 (49%)

TABLE I
L3C CHALLENGE TRAINING AND TESTING DATA FRAME SOURCES AND PERCENTAGE OF MISSING DATA. WE HAVE USED THE **BOLDED** DATA SOURCES.

- condition information for 38,044 patients
- 14,476 conditions that lasted from 1-409 days.
- observation information for 38,340 patients
- 2,744 observations plus 14,159 prescribed drugs.

Attribute Selection

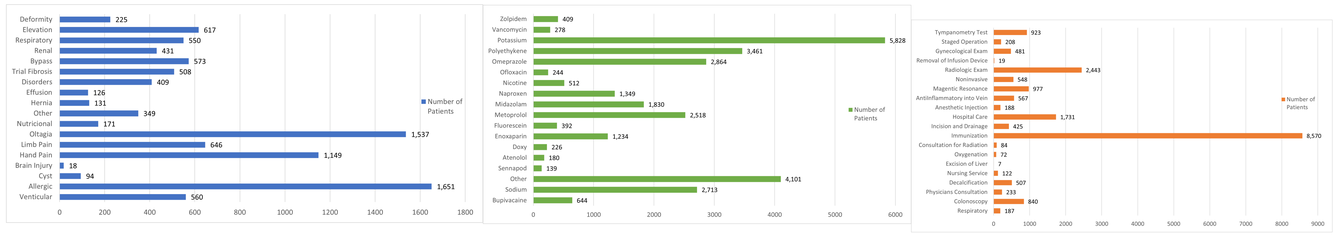


Figure 1. Twenty attributes were selected for the conditions, drugs and procedures based on the uncertainty of the Long COVID label through Lindley entropy.

Fig 2. The most frequent conditions and drugs, per patient in the training data set (left) and their average duration

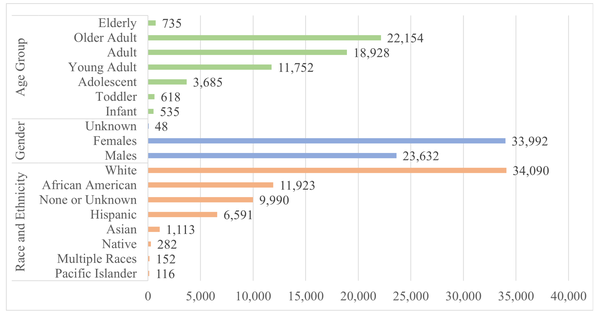
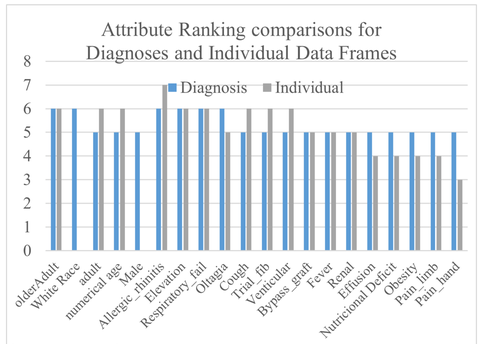


Figure 3. Demographics and conditions dataframe

Feature Importance



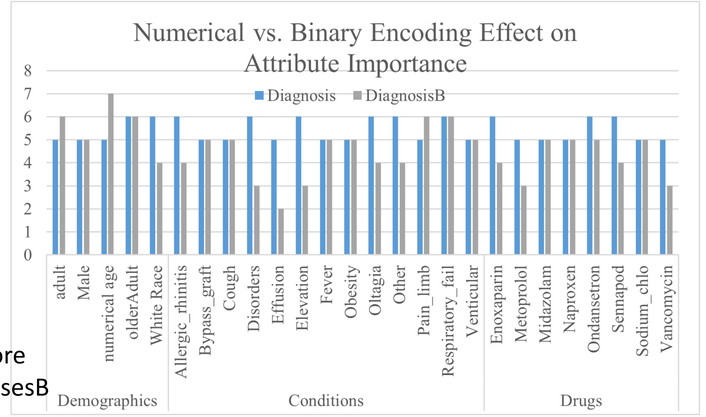


Figure 4. Diagnoses Feature Importance: Five or more selection methods selected to feature in for DiagnosesB (binary) and Diagnoses (cumulative).

Modeling

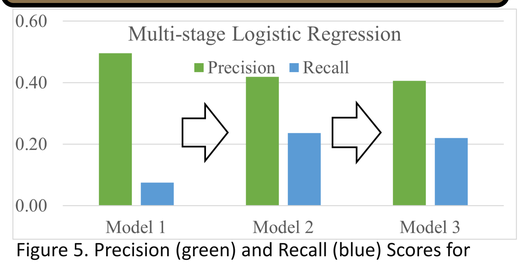


Figure 5. Precision (green) and Recall (blue) Scores for each of the three stages of the logistic regression modeling on the entire data set.

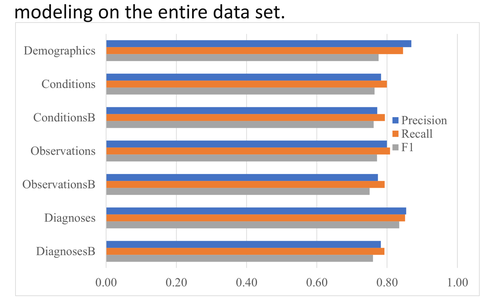


Figure 6. Selecting the most informative data frames for the Random Forest modeling based on precision P, recall R, and F1-measure on the training holdout set

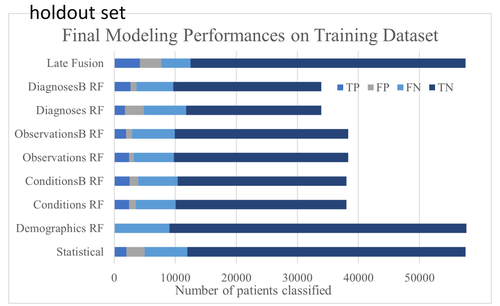


Figure 7. Final model performances from Table VII on entire training dataset. Late fusion resulted in precision .548, recall 0.467, F1 0.504 and **accuracy 0.856** on the training set.

Acknowledgment

Grad students: June Yu and Mirna Elizondo
The analyses described in this poster were conducted with data or tools accessed through the NCATS N3C Data Enclave https://covid.cd2h.org and N3C Attribution & Publication Policy v 1.2-2020-08-25b supported by NCATS U24 TR002306, CHERR, NVIDIA, and CS department. This research was possible because of the patients whose information is included in the data and the organizations and scientists who have contributed to the ongoing development of this community resource.