05H_Summary

Brenna Hanson

2024-03-13

Purpose

In this file, we will detail the data loss on all the iterative multivariate tables we generated. As a reminder, all these tables were generated using the ‘Multivariate’ function of the PCD database. In addition to the specified variables in the function, year was also always set to 2020 as it is our study period of interest. The summary visuals were generated in 05G.

Table

Here is a table that explains the cohorts we are investigating.

Cohort information

cohort	cohortDef	cohortDescription	cohortSize
0	{}	No restrictions	7940
1	{"year": {"operator": "=", "value": "2020"}}	year= 2020	4168
2	{"Race": {"operator": "in", "values": ["Caucasian", "African American", "Asian"]}, "year": {"operator": "=", "value": "2020"}}	year= 2020, restrict Race	4569
3	{"TotalEDInpatientVisits": {"operator": "<=", "value": "9"}, "year": {"operator": "=", "value": "2020"}}	year= 2020, restrict TotalEDInpatientVisits	4341
4	{"Race": {"operator": "in", "values": ["Caucasian", "African American", "Asian"]}, "TotalEDInpatientVisits": {"operator": "<=", "value": "9"}, "year": {"operator": "=", "value": "2020"}}	year= 2020, restrict Race and TotalEDInpatientVisits	4753

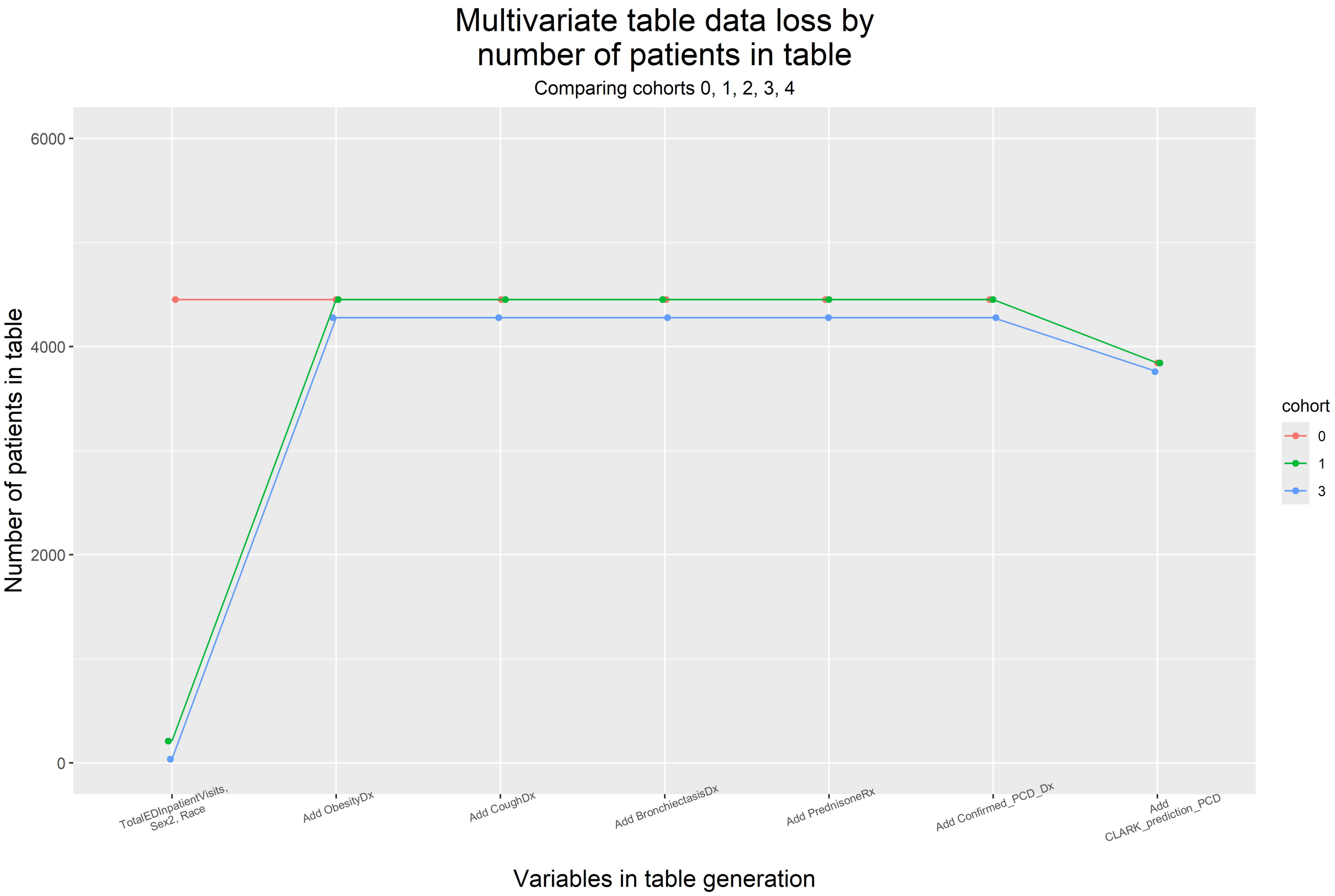
Here is a table that shows how data loss changes by cohort.

Data loss by cohort

description	tableSize using cohort 0	tableSize using cohort 1	tableSize using cohort 3
TotalEDInpatientVisits, Sex2, Race	4453	210	36
Add ObesityDx	4453	4453	4279
Add CoughDx	4453	4453	4279
Add BronchiectasisDx	4453	4453	4279
Add PrednisoneRx	4453	4453	4279
Add Confirmed_PCD_Dx	4453	4453	4279
Add CLARK_prediction_PCD	3843	3843	3760

Plot

Here is a plot that shows how data loss changes by cohort.



Results

- Tables generated with cohort 2 and 4 are purposely excluded. This is because there is an issue with generating these tables, which should be investigated further.
- 3-feature tables generated with cohorts 1 and 3 have suspiciously few patients. The fact that adding a fourth variable makes the tables gain patients indicates an issue. Also, for cohorts 1 and 3, the number of patients jumps with the inclusion of the fourth variable which also indicates an issue.

Note: raw tables were extracted on approximately 3/13.