Data Dictionary

We made every effort to limit the number of variables in the synthetic data to the minimum used in the analysis, both for simplicity and to reduce risk of inadvertent disclosure of PHI, so variables not used in any analyses (e.g., hospitalizations, visit details, etc.) were not included.

The synthetic data file contains 48,784 rows (patient-years) and 160 columns/variables. One note is that, while the original dataset includes multiple observations (years) from the same patient, the synthetic dataset observations do not take this clustering into account; creating adequate time dependencies for multiple years of synthetic data posed significant challenges, and in fact none of our analyses rely on using multiple observations per patient (though of course we do account for this in calculating standard errors in the original paper).

We group the variables into the following categories:

- Variables at time t: A vector of "outcomes" for a given calendar year (t): cost, health, program enrollment, and the commercial risk score. The remaining variables, which are indexed to the year prior to the outcomes (t-1), are used primarily as "predictors" in our experimental algorithms.
- Demographic variables.
- Comorbidity variables at time t-1: A vector of indicators for specific chronic comorbidities (illnesses) that were active in the previous year, and their sum.
- Cost variables at time t-1: Costs claimed from the patients' insurance payer, rounded to the nearest $100 and broken down by type of cost, over the previous year.
- Biomarker/medication variables at time t-1: A set of indicators capturing normal or abnormal values (or missingness) of biomarkers or relevant medications, over the previous year.

Notation:

- `_t`: indicates this is a time dependent variable from year t
- `_tm1`: indicates this is a time dependent variable from year t minus 1 (t-1)

Variables at time t

Variable	**Data Type**	**Description**	**Sample Data**
risk_score_t	Float	Commercial algorithmic risk score prediction for cost in year t, formed using data from year t-1	1.32
program_enrolled_t	Integer	Indicator for whether patient-year was enrolled in program	0
gagne_sum_t	Integer	Total number of active chronic illnesses	3
cost_t	Float	Total medical expenditures, rounded to the nearest 100	1000.00

Variable	Data Type	Description	Sample Data
cost_avoidable_t	Float	Total avoidable (emergency + inpatient) medical expenditures, rounded to nearest 100	100.00
bps_mean_t	Float	Mean systolic blood pressure in year t	120.0
ghba1c_mean_t	Float	Mean HbA1C in year t	5.5
hct_mean_t	Float	Mean hematocrit in year t	40.8
cre_mean_t	Float	Mean creatinine in year t	0.78
ldl_mean_t	Float	Mean low-density lipoprotein in year t	89.0

Total = **10** variables at time t

Demographic variables

Variable	Data Type	Description	Sample Data
race	String	Patient race; values include 'white' or 'black'	white
dem_female	Integer	Indicator for female gender	1
dem_age_band_18-24_tm1	Integer	Indicator for patient age between 18-24	0
dem_age_band_25-34_tm1	Integer	Indicator for patient age between 25-34	0
dem_age_band_35-44_tm1	Integer	Indicator for patient age between 35-44	0
dem_age_band_45-54_tm1	Integer	Indicator for patient age between 45-54	1
dem_age_band_55-64_tm1	Integer	Indicator for patient age between 55-64	0
dem_age_band_65-74_tm1	Integer	Indicator for patient age between 65-74	0
dem_age_band_75+_tm1	Integer	Indicator for patient age 75+	0

Total = **9** demographic variables (including race)

Comorbidity variables at time t-1

Variable	Data Type	Description	Sample Data
gagne_sum_tm1	Integer	Total number of active illnesses	2
alcohol_elixhauser_tm1	Integer	Indicator for alcohol abuse	0

Variable	**Data Type**	**Description**	**Sample Data**
anemia_elixhauser_tm1	Integer	Indicator for deficiency anemia	0
arrhythmia_elixhauser_tm1	Integer	Indicator for arrhythmia	0
arthritis_elixhauser_tm1	Integer	Indicator for arthritis	1
bloodlossanemia_elixhauser_tm1	Integer	Indicator for blood loss anemia	0
coagulopathy_elixhauser_tm1	Integer	Indicator for coagulopathy	0
compdiabetes_elixhauser_tm1	Integer	Indicator for diabetes, complicated	1
depression_elixhauser_tm1	Integer	Indicator for depression	0
drugabuse_elixhauser_tm1	Integer	Indicator for drug abuse	0
electrolytes_elixhauser_tm1	Integer	Indicator for electrolyte disorder	0
hypertension_elixhauser_tm1	Integer	Indicator for hypertension	0
hypothyroid_elixhauser_tm1	Integer	Indicator for hypothyroid	0
liver_elixhauser_tm1	Integer	Indicator for liver disease	0
neurodegen_elixhauser_tm1	Integer	Indicator for neurodegenerative disease	0
obesity_elixhauser_tm1	Integer	Indicator for obesity	0
paralysis_elixhauser_tm1	Integer	Indicator for paralysis	0
psychosis_elixhauser_tm1	Integer	Indicator for psychoses	0
pulmcirc_elixhauser_tm1	Integer	Indicator for pulmonary circulation disorders	0
pvd_elixhauser_tm1	Integer	Indicator for peripheral vascular disorders	0
renal_elixhauser_tm1	Integer	Indicator for renal failure	0
uncompdiabetes_elixhauser_tm1	Integer	Indicator for diabetes, uncomplicated	0
valvulardz_elixhauser_tm1	Integer	Indicator for valvular disease	0
wtloss_elixhauser_tm1	Integer	Indicator for weight loss	0
cerebrovasculardz_romano_tm1	Integer	Indicator for cerebrovascular disease	0
chf_romano_tm1	Integer	Indicator for congestive heart failure	0
dementia_romano_tm1	Integer	Indicator for dementia	0
hemiplegia_romano_tm1	Integer	Indicator for hemiplegia	0
hivaids_romano_tm1	Integer	Indicator for HIV/AIDS	0
metastatic_romano_tm1	Integer	Indicator for metastasis	0
myocardialinfarct_romano_tm1	Integer	Indicator for myocardial infarction	0
pulmonarydz_romano_tm1	Integer	Indicator for pulmonary disease	0
tumor_romano_tm1	Integer	Indicator for tumor	0
ulcer_romano_tm1	Integer	Indicator for ulcer	0

Total = **34** comorbidity variables at time t-1

Cost variables at time t-1

Variable	Data Type	Description	Sample Data
cost_dialysis_tm1	Float	Total costs for dialysis, rounded to nearest 10	990.00
cost_emergency_tm1	Float	Total costs for emergency, rounded to nearest 10	140.00
cost_home_health_tm1	Float	Total costs for home health, rounded to nearest 10	120.00
cost_ip_medical_tm1	Float	Total costs for inpatient medical, rounded to nearest 10	150.00
cost_ip_surgical_tm1	Float	Total costs for inpatient surgical, rounded to nearest 10	200.00
cost_laboratory_tm1	Float	Total costs for laboratory, rounded to nearest 10	90.00
cost_op_primary_care_tm1	Float	Total costs for outpatient primary care, rounded to nearest 10	270.00
cost_op_specialists_tm1	Float	Total costs for outpatient specialists, rounded to nearest 10	180.00
cost_op_surgery_tm1	Float	Total costs for outpatient surgery, rounded to nearest 10	110.00
cost_other_tm1	Float	Total other costs, rounded to nearest 100	300.00
cost_pharmacy_tm1	Float	Total costs for pharmacy, rounded to nearest 10	10.00
cost_physical_therapy_tm1	Float	Total costs for physical therapy, rounded to nearest 10	190.00
cost_radiology_tm1	Float	Total costs for radiology, rounded to nearest 10	120.00

Total = **13** cost variables at time t-1

Biomarker/medication variables at time t-1

Variable	Data Type	Description	Sample Data
lasix_dose_count_tm1	Integer	Number of Lasix doses	0
lasix_min_daily_dose_tm1	Integer	Minimum daily dose of Lasix	20
lasix_mean_daily_dose_tm1	Float	Mean daily dose of Lasix	20
lasix_max_daily_dose_tm1	Integer	Maximum daily dose of Lasix	20
cre_tests_tm1	Integer	Number of creatinine tests	1
crp_tests_tm1	Integer	Number of c-reactive protein tests	0

Variable	Data Type	Description	Sample Data
esr_tests_tm1	Integer	Number of erythrocyte sedimentation rate tests	1
ghba1c_tests_tm1	Integer	Number of GHbA1c tests	1
hct_tests_tm1	Integer	Number of hematocrit tests	1
ldl_tests_tm1	Integer	Number of LDL tests	1
nt_bnp_tests_tm1	Integer	Number of BNP tests	1
sodium_tests_tm1	Integer	Number of sodium tests	1
trig_tests_tm1	Integer	Number of triglycerides tests	1
cre_min-low_tm1	Integer	Indicator for low (< 0.84) minimum creatinine test result	0
cre_min-high_tm1	Integer	Indicator for high (> 1.21) minimum creatinine test result	0
cre_min-normal_tm1	Integer	Indicator for normal minimum creatinine test result	1
cre_mean-low_tm1	Integer	Indicator for low (< 0.84) mean creatinine test result	0
cre_mean-high_tm1	Integer	Indicator for high (> 1.21) mean creatinine test result	0
cre_mean-normal_tm1	Integer	Indicator for normal mean creatinine test result	1
cre_max-low_tm1	Integer	Indicator for low (< 0.84) maximum creatinine test result	0
cre_max-high_tm1	Integer	Indicator for high (> 1.21) maximum creatinine test result	0
cre_max-normal_tm1	Integer	Indicator for normal maximum creatinine test result	1
crp_min-low_tm1	Integer	Indicator for low (< 1) minimum c-reactive protein test result	0
crp_min-high_tm1	Integer	Indicator for high (> 3) minimum c-reactive protein test result	0
crp_min-normal_tm1	Integer	Indicator for normal minimum c-reactive protein test result	1
crp_mean-low_tm1	Integer	Indicator for low (< 1) mean c-reactive protein test result	0
crp_mean-high_tm1	Integer	Indicator for high (> 3) mean c-reactive protein test result	0
crp_mean-normal_tm1	Integer	Indicator for normal mean c-reactive protein test result	1
crp_max-low_tm1	Integer	Indicator for low (< 1) maximum c-reactive	0

Variable	Data Type	Description	Sample Data
		protein test result	
crp_max-high_tm1	Integer	Indicator for high (> 3) maximum c-reactive protein test result	0
crp_max-normal_tm1	Integer	Indicator for normal maximum c-reactive protein test result	1
esr_min-low_tm1	Integer	Indicator for low (< 1) minimum erythrocyte sedimentation rate test result	0
esr_min-high_tm1	Integer	Indicator for high (> 20) minimum erythrocyte sedimentation rate test result	0
esr_min-normal_tm1	Integer	Indicator for normal minimum erythrocyte sedimentation rate test result	1
esr_mean-low_tm1	Integer	Indicator for low (< 1) mean erythrocyte sedimentation rate test result	0
esr_mean-high_tm1	Integer	Indicator for high (> 20) mean erythrocyte sedimentation rate test result	0
esr_mean-normal_tm1	Integer	Indicator for normal mean erythrocyte sedimentation rate test result	1
esr_max-low_tm1	Integer	Indicator for low (< 1) maximum erythrocyte sedimentation rate test result	0
esr_max-high_tm1	Integer	Indicator for high (> 20) maximum erythrocyte sedimentation rate test result	0
esr_max-normal_tm1	Integer	Indicator for normal maximum erythrocyte sedimentation rate test result	1
ghba1c_min-low_tm1	Integer	Indicator for low (< 4) minimum GHbA1c test result	0
ghba1c_min-high_tm1	Integer	Indicator for high (> 5.7) minimum GHbA1c test result	0
ghba1c_min-normal_tm1	Integer	Indicator for normal minimum GHbA1c test result	1
ghba1c_mean-low_tm1	Integer	Indicator for low (< 4) mean GHbA1c test result	0
ghba1c_mean-high_tm1	Integer	Indicator for high (> 5.7) mean GHbA1c test result	0
ghba1c_mean-normal_tm1	Integer	Indicator for normal mean GHbA1c test result	1
ghba1c_max-low_tm1	Integer	Indicator for low (< 4) maximum GHbA1c test result	0
ghba1c_max-high_tm1	Integer	Indicator for high (> 5.7) maximum GHbA1c test result	0
ghba1c_max-normal_tm1	Integer	Indicator for normal maximum GHbA1c test result	1

Variable	Data Type	Description	Sample Data
hct_min-low_tm1	Integer	Indicator for low (< 35.5) minimum hematocrit test result	0
hct_min-high_tm1	Integer	Indicator for high (> 48.6) minimum hematocrit test result	0
hct_min-normal_tm1	Integer	Indicator for normal minimum hematocrit test result	1
hct_mean-low_tm1	Integer	Indicator for low (< 35.5) mean hematocrit test result	0
hct_mean-high_tm1	Integer	Indicator for high (> 48.6) mean hematocrit test result	0
hct_mean-normal_tm1	Integer	Indicator for normal mean hematocrit test result	1
hct_max-low_tm1	Integer	Indicator for low (< 35.5) maximum hematocrit test result	0
hct_max-high_tm1	Integer	Indicator for high (> 48.6) maximum hematocrit test result	0
hct_max-normal_tm1	Integer	Indicator for normal maximum hematocrit test result	1
ldl_min-low_tm1	Integer	Indicator for low (< 50) minimum LDL test result	0
ldl_min-high_tm1	Integer	Indicator for high (> 99) minimum LDL test result	0
ldl_min-normal_tm1	Integer	Indicator for normal minimum LDL test result	1
ldl-mean-low_tm1	Integer	Indicator for low (< 50) mean LDL test result	0
ldl-mean-high_tm1	Integer	Indicator for high (> 99) mean LDL test result	0
ldl-mean-normal_tm1	Integer	Indicator for normal mean LDL test result	1
ldl-max-low_tm1	Integer	Indicator for low (< 50) maximum LDL test result	0
ldl-max-high_tm1	Integer	Indicator for high (> 99) maximum LDL test result	0
ldl-max-normal_tm1	Integer	Indicator for normal maximum LDL test result	1
nt_bnp_min-low_tm1	Integer	Indicator for low (< 100) minimum BNP test result	0
nt_bnp_min-high_tm1	Integer	Indicator for high (> 450) minimum BNP test result	0
nt_bnp_min-normal_tm1	Integer	Indicator for normal minimum BNP test result	1
nt_bnp_mean-low_tm1	Integer	Indicator for low (< 100) mean BNP test result	0
nt_bnp_mean-high_tm1	Integer	Indicator for high (> 450) mean BNP test result	0

Variable	Data Type	Description	Sample Data
nt_bnp_mean-normal_tm1	Integer	Indicator for normal minimum BNP test result	1
nt_bnp_max-low_tm1	Integer	Indicator for low (< 100) maximum BNP test result	0
nt_bnp_max-high_tm1	Integer	Indicator for high (> 450) maximum BNP test result	0
nt_bnp_max-normal_tm1	Integer	Indicator for normal minimum BNP test result	1
sodium_min-low_tm1	Integer	Indicator for low (< 135) minimum sodium test result	0
sodium_min-high	Integer	Indicator for high (> 145) minimum sodium test result	0
sodium_min-normal_tm1	Integer	Indicator for normal minimum sodium test result	1
sodium_mean-low_tm1	Integer	Indicator for low (< 135) mean sodium test result	0
sodium_mean-high_tm1	Integer	Indicator for high (> 145) mean sodium test result	0
sodium_mean-normal_tm1	Integer	Indicator for normal mean sodium test result	1
sodium_max-low_tm1	Integer	Indicator for low (< 135) maximum sodium test result	0
sodium_max-high_tm1	Integer	Indicator for high (> 145) maximum sodium test result	0
sodium_max-normal_tm1	Integer	Indicator for normal maximum sodium test result	1
trig_min-low_tm1	Integer	Indicator for low (< 50) minimum triglycerides test result	0
trig_min-high_tm1	Integer	Indicator for high (> 150) minimum triglycerides test result	0
trig_min-normal_tm1	Integer	Indicator for normal minimum triglycerides test result	1
trig_mean-low_tm1	Integer	Indicator for low (< 50) mean triglycerides test result	0
trig_mean-high_tm1	Integer	Indicator for high (> 150) mean triglycerides test result	0
trig_mean-normal_tm1	Integer	Indicator for normal mean triglycerides test result	1
trig_max-low_tm1	Integer	Indicator for low (< 50) maximum triglycerides test result	0
trig_max-high_tm1	Integer	Indicator for high (> 150) maximum triglycerides test result	0

Variable	**Data Type**	**Description**	**Sample Data**
trig_max-normal_tm1	Integer	Indicator for normal maximum triglycerides test result	1

Total = **94** biomarker/medication variables at time t-1