Machine learning and prediction of incident psychiatric hospitalizations using NYC-CDRN data

Project Goal: Prediction of incident psychiatric hospitalization (*bipolar/manic disorders, depression, schizophrenia, psychoses/delusional disorders*), using a 1-year study window

Time period: 01/2013-12/2017

Study Design: Retrospective case-control

Data Source: CDRN [OMOP]

Population:

Potential Cases: 470,242 pts with hx of depression Dx or antidepressant Rx

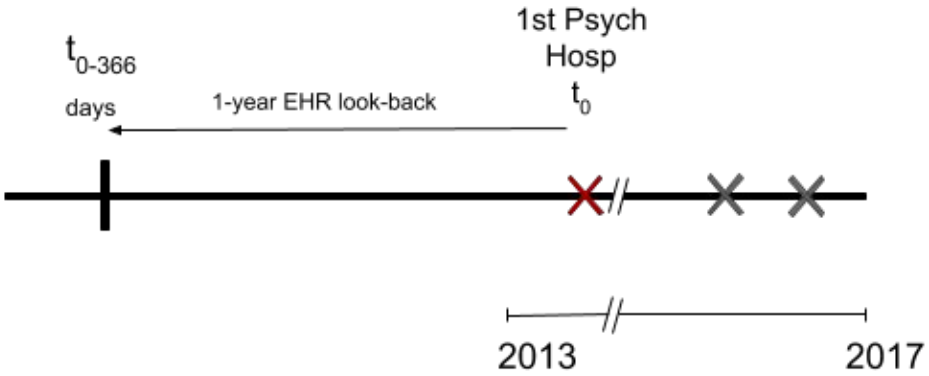
Potential Controls: 1,000,000 pts with no hx of psychiatric Dx or antidepressant Rx

Psychiatric ICD Codes

	ICD-9-CM	ICD-10-CM
All Psychiatric Diagnoses	293.8[1-2], 295, 296, 297, 298, 300.4, 311	F06.[0,2], F20, F21, F22, F23, F24, F25, F28, F29, F30, F31, F32, F33, F34.[1,8], F39
Bipolar / Manic / Other Mood	296.[0,1,4-9]	F30, F31, F34.8, F39
Depression	296.2, 296.3, 300.4, 311	F32, F33, F34.1
Schizophrenia	295	F20, F21, F25
Psychoses (*incl.* Delusional disorders)	293.8[1-2], 297, 298	F06.[0,2], F22, F23, F24, F28, F29

Case Identification

1. **Psych ICD code (primary, secondary) + Inpatient Hospitalization, 2013-2017**
   a. Total, *n* = 8,896
   b. After exclusion, *n* = 7,166
2. **Psych ICD code (primary) + Inpatient Hospitalization, 2013-2017**
   a. Total, *n* = 1,701
   b. After exclusion, *n* = 1,038

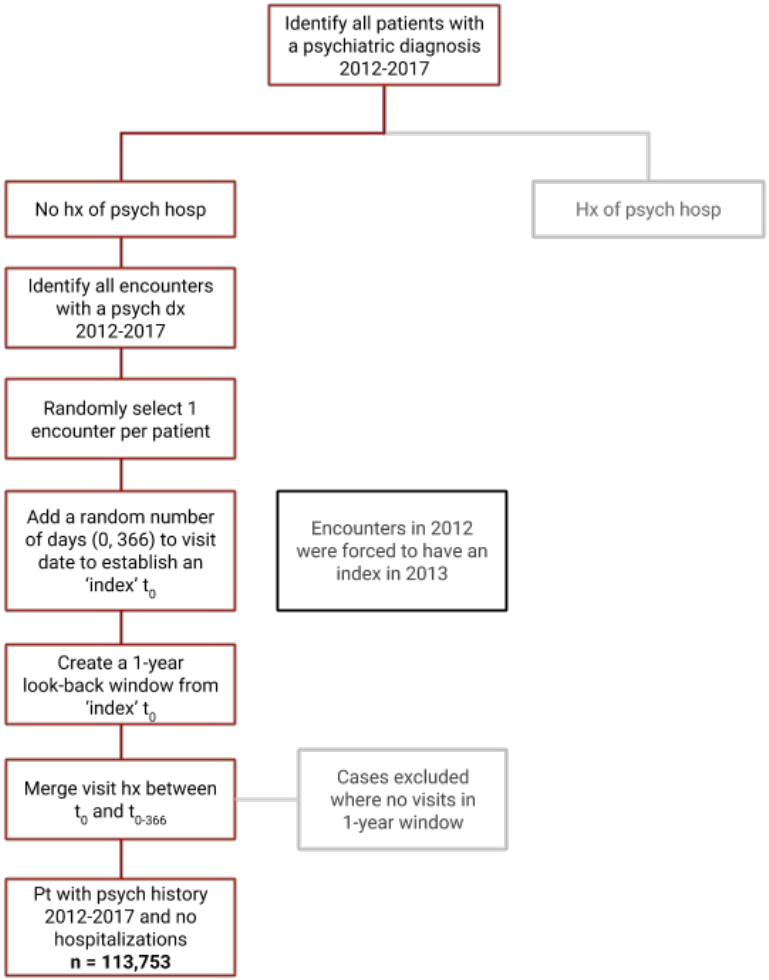


Inclusion criteria:

- Age ≥ 18
- ≥ 2 visits in 1-year window
- ≥ 1 days between first-last visit
- ≥ 1 dx mood | schizophrenia (CCS Categories)

‘Control’ Identification

- All patients have a psychiatric diagnosis in year leading up to ‘index’ t_0.
- Diagnosis encounter and index dates chosen at random



Feature Definitions

Demographics (2):

- **sex, age**
  - **Binary (sex), Count (age)**

Healthcare Utilization (4):

- **ED, Outpatient, Inpatient, Ambulatory**
  - **Counts**

Dx History (283):

- **CCS Categories (ICD9 / 10)**
  - **Counts & Binary (chronic conditions)**

Outpatient Rx History (100):

- **VA Class, top 95% of all outpatient RX in study pop**
  - **Counts**

Procedures (231):

- **CCS Categories (ICD9 / 10)**
  - **Counts**

CCS Algorithm for Diagnoses (ICD-9)

Appendix A - Clinical Classification Software-DIAGNOSES (January 1980 through September 2015)

Revised 03/24/2016

1 Tuberculosis
01000 01001 01002 01003 01004 01005 01006 01010 01011 01012 01013 01014 01015 01016 01080 01081 01082 01083 01084 01085
01086 01090 01091 01092 01093 01094 01095 01096 01100 01101 01102 01103 01104 01105 01106 01110 01111 01112 01113 01114
01115 01116 01120 01121 01122 01123 01124 01125 01126 01130 01131 01132 01133 01134 01135 01136 01140 01141 01142 01143
01144 01145 01146 01150 01151 01152 01153 01154 01155 01156 01160 01161 01162 01163 01164 01165 01166 01170 01171 01172
01173 01174 01175 01176 01180 01181 01182 01183 01184 01185 01186 01190 01191 01192 01193 01194 01195 01196 01200 01201
01202 01203 01204 01205 01206 01210 01211 01212 01213 01214 01215 01216 01220 01221 01222 01223 01224 01225 01226 01230
01231 01232 01233 01234 01235 01236 01280 01281 01282 01283 01284 01285 01286 01300 01301 01302 01303 01304 01305 01306
01310 01311 01312 01313 01314 01315 01316 01320 01321 01322 01323 01324 01325 01326 01330 01331 01332 01333 01334 01335
01336 01340 01341 01342 01343 01344 01345 01346 01350 01351 01352 01353 01354 01355 01356 01360 01361 01362 01363 01364
01365 01366 01380 01381 01382 01383 01384 01385 01386 01390 01391 01392 01393 01394 01395 01396 01400 01401 01402 01403
01404 01405 01406 01480 01481 01482 01483 01484 01485 01486 01500 01501 01502 01503 01504 01505 01506 01510 01511 01512
01513 01514 01515 01516 01520 01521 01522 01523 01524 01525 01526 01550 01551 01552 01553 01554 01555 01556 01560 01561
01562 01563 01564 01565 01566 01570 01571 01572 01573 01574 01575 01576 01580 01581 01582 01583 01584 01585 01586 01590
01591 01592 01593 01594 01595 01596 01600 01601 01602 01603 01604 01605 01606 01610 01611 01612 01613 01614 01615 01616
01620 01621 01622 01623 01624 01625 01626 01630 01631 01632 01633 01634 01635 01636 01640 01641 01642 01643 01644 01645
01646 01650 01651 01652 01653 01654 01655 01656 01660 01661 01662 01663 01664 01665 01666 01670 01671 01672 01673 01674
01675 01676 01690 01691 01692 01693 01694 01695 01696 01700 01701 01702 01703 01704 01705 01706 01710 01711 01712 01713
01714 01715 01716 01720 01721 01722 01723 01724 01725 01726 01730 01731 01732 01733 01734 01735 01736 01740 01741 01742
01743 01744 01745 01746 01750 01751 01752 01753 01754 01755 01756 01760 01761 01762 01763 01764 01765 01766 01770 01771
01772 01773 01774 01775 01776 01780 01781 01782 01783 01784 01785 01786 01790 01791 01792 01793 01794 01795 01796 01800
01801 01802 01803 01804 01805 01806 01880 01881 01882 01883 01884 01885 01886 01890 01891 01892 01893 01894 01895 01896
1370 1371 1372 1373 1374 V1201

2 Septicemia (except in labor)
0031 0202 0223 0362 0380 0381 03810 03811 03812 03819 0382 0383 03840 03841 03842 03843 03844 03849 0388 0389
0545 449 77181 7907 99591 99592

3 Bacterial infection; unspecified site
0200 0208 0209 0218 0219 0228 0229 0230 0231 0232 0233 0238 0239 024 025 0260 0269 0270 0271 0272
0278 0279 0300 0301 0302 0303 0308 0309 0312 0318 0319 03289 0329 0330 0331 0338 0339 0341 0363 03681
03689 0369 037 0392 0393 0394 0398 0399 0400 0401 0402 0403 04042 04081 04082 04089 0410 04100 04101 04102
04103 04104 04105 04109 0411 04110 04111 04112 04119 0412 0413 0414 04141 04142 04143 04149 0415 0416 0417 0418
04181 04182 04183 04184 04185 04186 04189 0419 390 3929 7953 79531 79539 V090 V091 V092 V093 V094 V0950 V0951
V096 V0970 V0971 V0980 V0981 V0990 V0991 V1204

4 Mycoses
1100 1101 1102 1103 1104 1105 1106 1108 1109 1110 1111 1112 1113 1118 1119 1120 1121 1122 1123 1125
11282 11284 11285 11289 1129 1141 1143 1149 11500 11509 11510 11519 11590 11599 1160 1161 1162 1170 1171 1172
1173 1174 1175 1176 1177 1178 1179 118

5 HIV infection
042 0420 0421 0422 0429 0430 0431 0432 0433 0439 0440 0449 07953 27910 27919 79571 7958 V08

6 Hepatitis
0700 0701 0702 07020 07021 07022 07023 0703 07030 07031 07032 07033 0704 07041 07042 07043 07044 07049 0705 07051
07052 07053 07054 07059 0706 07070 07071 0709 07271 57140 57141 57142 57149 5731 5732 5733

Analysis Plan

Construct several predictive models using the below statistical learning methods.

Compare confusion matrix, classification report, AUC score on:

1. Logistic Regression (LR)
    a. L1 regularization
    b. L2 regularization
2. Random Forest (RF)
3. Extreme Gradient Boosting (XGB)
4. Multi-layer Perceptron (MLP)

Model 1:
- Primary, Secondary Dx Hospitalization (outcome)
- All samples (1:X, case:control)

Model 2:
- Primary, Secondary Dx Hospitalization (outcome)
- Random subsample controls (1:1, case:control)

Model 3:
- Primary, Secondary Dx Hospitalization (outcome)
- SMOTE (1:1 ratio, case:control)

Analysis Plan

Construct several predictive models using the below statistical learning methods.

Compare confusion matrix, classification report, AUC score on:

1. Logistic Regression (LR)
    a. L1 regularization
    b. L2 regularization
2. Random Forest (RF)
3. Extreme Gradient Boosting (XGB)
4. Multi-layer Perceptron (MLP)

Model 4:
- Primary Dx Hospitalization (outcome)
- All samples (1:X, case:control)

Model 5:
- Primary Dx Hospitalization (outcome)
- Random subsample controls (1:1, case:control)

Model 6:
- Primary Dx Hospitalization (outcome)
- SMOTE (1:1 ratio, case:control)

Model 1: Results

- Primary, secondary Dx Hospitalization
- All samples (1:X, case:control)
  - n_{case} = 7,166; $n_{control}$ = 89,716

Logistic Regression

Confusion Matrix – Test			
TN	22452	4418	FP
FN	720	1475	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.97	0.84	0.90
1	0.25	0.67	0.36

AUC Score
0.754

Random Forest

Confusion Matrix – Test			
TN	25555	1315	FP
FN	1299	896	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.95	0.95	0.95
1	0.41	0.41	0.41

AUC Score
0.680

XGBoost

Confusion Matrix – Test			
TN	22422	4448	FP
FN	600	1595	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.97	0.83	0.90
1	0.26	0.73	0.39

AUC Score
0.781

Model 2: Results

- Primary, secondary Dx Hospitalization
- Random subsample controls (1:1 ratio, case:control)
  - n_{total} = 14,332

Logistic Regression

Confusion Matrix – Test			
TN	22002	4868	FP
FN	662	1533	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.97	0.82	0.89
1	0.24	0.70	0.36

AUC Score
0.757

Random Forest

Confusion Matrix – Test			
TN	20972	5898	FP
FN	554	1641	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.97	0.78	0.87
1	0.22	0.75	0.34

AUC Score
0.764

XGBoost

Confusion Matrix – Test			
TN	21603	5267	FP
FN	554	1641	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.97	0.80	0.88
1	0.24	0.75	0.36

AUC Score
0.776

Model 3: Results

- Primary, secondary Dx Hospitalization
- SMOTE (1:1 ratio, case:control)
  - n_{total} = 125,692

Logistic Regression

Confusion Matrix – Test			
TN	21979	4891	FP
FN	732	1463	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.97	0.82	0.89
1	0.23	0.67	0.34

AUC Score
0.742

Random Forest

Confusion Matrix – Test			
TN	26563	307	FP
FN	1681	514	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.94	0.99	0.96
1	0.63	0.23	0.34

AUC Score
0.611

XGBoost

Confusion Matrix – Test			
TN	26629	241	FP
FN	1609	586	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.94	0.99	0.97
1	0.71	0.27	0.39

AUC Score
0.629

Model 1,2,3: Multilayer Perceptron Results

- Primary, secondary Dx Hospitalization

MLP All Samples

Confusion Matrix – Test			
TN	25830	1040	FP
FN	1487	708	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.95	0.96	0.95
1	0.41	0.32	0.36

AUC Score
0.641

MLP Subsample

Confusion Matrix – Test			
TN	19909	6961	FP
FN	562	1603	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.97	0.74	0.84
1	0.19	0.73	0.30

AUC Score
0.736

MLP SMOTE

Confusion Matrix – Test			
TN	24377	2493	FP
FN	1247	948	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.95	0.91	0.93
1	0.28	0.43	0.34

AUC Score
0.670

Model 4: Results

- Primary Dx Hospitalization
- All samples (1:X, case:control)
  - n_{case} = 1,038; $n_{control}$ = 89,716

Logistic Regression

Confusion Matrix – Test			
TN	23552	3376	FP
FN	84	215	TP

Classification Report – Test			
	Precision	Recall	F1
0	1.00	0.87	0.93
1	0.06	0.72	0.11

AUC Score
0.797

Random Forest

Confusion Matrix – Test			
TN	26915	13	FP
FN	261	38	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.99	1.00	0.99
1	0.75	0.13	0.22

AUC Score
0.563

XGBoost

Confusion Matrix – Test			
TN	23088	3840	FP
FN	63	236	TP

Classification Report – Test			
	Precision	Recall	F1
0	1.00	0.86	0.92
1	0.06	0.79	0.11

AUC Score
0.823

Model 5: Results

- Primary Dx Hospitalization
- Random subsample controls (1:1 ratio, case:control)
  - n_{total} = 2,076

Logistic Regression

Confusion Matrix – Test			
TN	21750	5178	FP
FN	73	226	TP

Classification Report – Test			
	Precision	Recall	F1
0	1.00	0.81	0.89
1	0.04	0.76	0.08

AUC Score
0.782

Random Forest

Confusion Matrix – Test			
TN	22910	4018	FP
FN	64	235	TP

Classification Report – Test			
	Precision	Recall	F1
0	1.00	0.85	0.92
1	0.06	0.79	0.10

AUC Score
0.818

XGBoost

Confusion Matrix – Test			
TN	22606	4322	FP
FN	49	250	TP

Classification Report – Test			
	Precision	Recall	F1
0	1.00	0.84	0.91
1	0.05	0.84	0.10

AUC Score
0.838

Model 6: Results

- Primary Dx Hospitalization
- SMOTE (1:1 ratio, case:control)
  - n_{total} = 125,576

Logistic Regression

Confusion Matrix – Test			
TN	23548	3380	FP
FN	100	199	TP

Classification Report – Test			
	Precision	Recall	F1
0	1.00	0.87	0.93
1	0.06	0.67	0.10

AUC Score
0.770

Random Forest

Confusion Matrix – Test			
TN	26903	25	FP
FN	273	26	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.99	1.00	0.99
1	0.51	0.09	0.15

AUC Score
0.543

XGBoost

Confusion Matrix – Test			
TN	26899	29	FP
FN	244	55	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.99	1.00	0.99
1	0.65	0.18	0.29

AUC Score
0.591

Analysis Plan

Construct several predictive models using the below statistical learning methods.

Compare confusion matrix, classification report, AUC score on:

1. Logistic Regression (LR)
    a. L1 regularization
    b. L2 regularization
2. Random Forest (RF)
3. Extreme Gradient Boosting (XGB)
4. Multi-layer Perceptron (MLP)

Model 1:
- Primary, secondary Dx Hospitalization (outcome)
- Full data set (1:X, case:control)

Model 2:
- Primary Dx Hospitalization (outcome)
- Full data set (1:X, case:control)

Model 3:
- Primary, secondary Dx Hospitalization (outcome)
- Full Subsample (1:1 ratio, case:control)

Model 4:
- Primary Dx Hospitalization (outcome)
- Full Subsample (1:1 ratio, case:control)

Model 1: Results

- Primary, secondary Dx Hospitalization
- Full data set (1:X, case:control)
  - n_{case} = 8,896; $n_{control}$ = 113,753

Logistic Regression

Confusion Matrix – Test			
TN	28638	5531	FP
FN	795	1831	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.97	0.84	0.90
1	0.25	0.70	0.37

AUC Score
0.768

Random Forest

Confusion Matrix – Test			
TN	33946	223	FP
FN	1879	747	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.95	0.99	0.97
1	0.77	0.28	0.42

AUC Score
0.639

XGBoost

Confusion Matrix – Test			
TN	28397	5772	FP
FN	627	1999	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.98	0.83	0.90
1	0.26	0.76	0.38

AUC Score
0.796

Model 2 Results

- Primary Dx Hospitalization
- Full data set (1:X, case:control)
  - n_{case} = 1,701; $n_{control}$ = 113,753

Logistic Regression

Confusion Matrix – Test			
TN	26656	5187	FP
FN	810	1696	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.97	0.84	0.90
1	0.25	0.68	0.36

AUC Score
0.757

Random Forest

Confusion Matrix – Test			
TN	34121	29	FP
FN	351	136	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.99	1.00	0.99
1	0.82	0.28	0.42

AUC Score
0.639

XGBoost

Confusion Matrix – Test			
TN	29580	4570	FP
FN	86	401	TP

Classification Report – Test			
	Precision	Recall	F1
0	1.00	0.87	0.93
1	0.08	0.82	0.15

AUC Score
0.845

Model 3 Results

- Primary, Secondary Dx Hospitalization
- Full Subsample (1:1, case:control)
  - n_{total} = 17,792

Logistic Regression

Confusion Matrix – Test			
TN	2197	482	FP
FN	786	1873	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.74	0.82	0.78
1	0.80	0.70	0.75

AUC Score
0.762

Random Forest

Confusion Matrix – Test			
TN	2107	572	FP
FN	496	2163	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.81	0.79	0.80
1	0.79	0.81	0.80

AUC Score
0.800

XGBoost

Confusion Matrix – Test			
TN	2138	541	FP
FN	544	2115	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.80	0.80	0.80
1	0.80	0.80	0.80

AUC Score
0.797

Model 4 Results

- Primary Dx Hospitalization
- Full Subsample (1:1, case:control)
  - n_{total} = 3,402

Logistic Regression

Confusion Matrix – Test			
TN	410	86	FP
FN	143	382	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.74	0.83	0.78
1	0.82	0.73	0.77

AUC Score
0.777

Random Forest

Confusion Matrix – Test			
TN	427	69	FP
FN	100	425	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.81	0.86	0.83
1	0.86	0.81	0.83

AUC Score
0.835

XGBoost

Confusion Matrix – Test			
TN	420	76	FP
FN	99	426	TP

Classification Report – Test			
	Precision	Recall	F1
0	0.81	0.85	0.83
1	0.85	0.81	0.83

AUC Score
0.829