Preventing exacerbation in respiratory patients by early prediction of exacerbation

http://www.brighthealing.com/

Final Report!!

Team 3- MD Team

Kun-Lin Tsai ,Yuan-Yu Zhang ,Chris Lo

Business Problem

- Customers: public health care system (public welfare)
- Purpose: decreasing lethal rate and lowering social costs (preventing exacerbation)
- Goals: Establishing a knowledge based model for predicting exacerbators with:

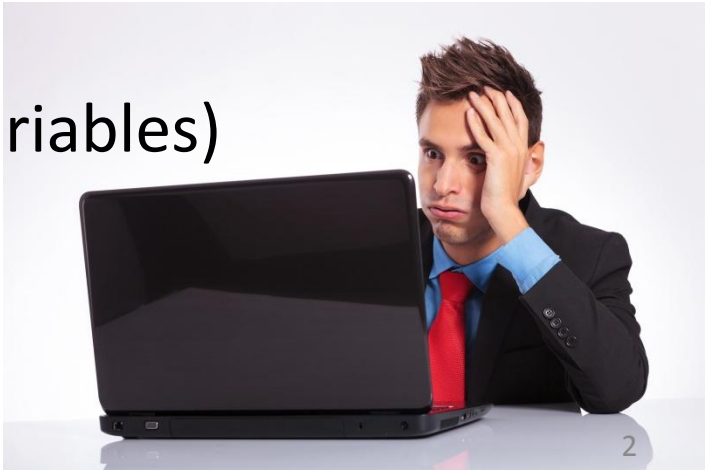
a. Standard rules (current: physical examination)

b. Fast (DNA exam)

c. Low cost (throat swab)

d. Less measurements (predictive variables)

e. High accuracy (better one)



Data mining goal

- Purpose: Finding out predictive variables and creating a model for preventing exacerbation
- Goals: selecting useful variables for creating predictive model by:

a. Dimension Reduction (from many exam)

b. With comparatively few predictors

c. With good AUC (correct prediction)

Data profile

sid	Exacebator	V3	V4	V29	V30	V31	V33	V36	V37	V38	V39
1296	0	0	1	0.49505	0	0.340659	0	0.47619	0.288395	0.522788	0.241652
1725	0	0	0	0.50495	0	0.626374	0	0.428571	0.480195	0.489276	0.439683
2195	0	0	1	0.554455	0	0.549451	0	0.404762	0.573315	0.687668	0.411311
2712	0	0	0	0.346535	0	0.373626	0	0.333333	0.387769	0.542895	0.324455
2913	0	0	0	0.376238	0	0.395604	0	0.345238	0.350938	0.580429	0.274078
1269	0	0	1	0.29703	1	0.274725	1	0.470238	0.213343	0.288204	0.269253
2836	0	0	1	0.356436	0	0.274725	0	0.386905	0.313412	0.601877	0.232388
91	0	0	0	0.514851	0	0.461538	0	0.529762	0.323141	0.449062	0.308242
5571	0	0	1	0.356436	0	0.21978	0	0.369048	0.309937	0.620643	0.222158
2645	0	0	0	0.326733	1	0.21978	0	0.464286	0.295344	0.352547	0.328315
4059	0	0	1	0.336634	1	0.406593	0	0.279762	0.364837	0.379357	0.387956
1724	0	0	0	0.405941	0	0.230769	0	0.494048	0.496178	0.754692	0.318278
5122	0	0	0	0.39604	1	0.175824	0	0.345238	0.314802	0.400804	0.323876
5566	0	0	0	0.376238	0	0.461538	0	0.488095	0.52467	0.829759	0.307663
3626	0	0	1	0.415842	1	0.296703	0	0.345238	0.246699	0.352547	0.275429
5254	0	0	0	0.376238	1	0.373626	0	0.428571	0.259903	0.365952	0.283343

- Resource: a competition from crowdanalytix (MODELING: Predict Exacerbation in patients with Respiratory Diseases)
- Dataset:

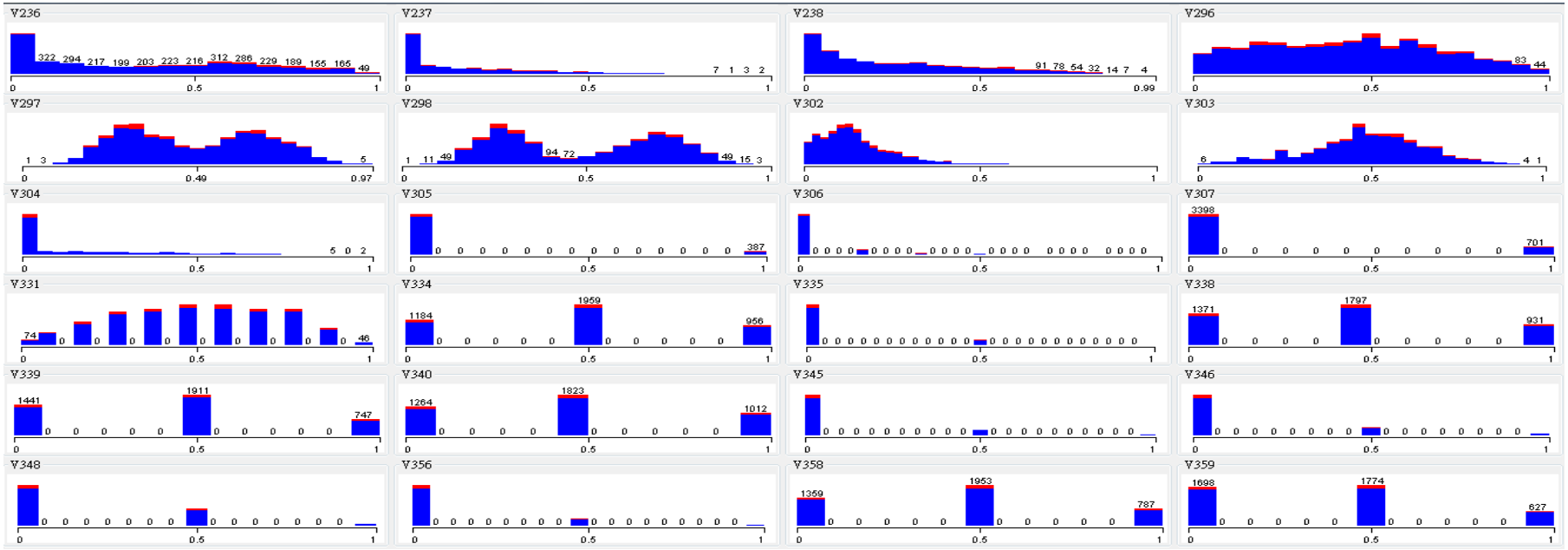
a. 330 columns: medical derived data (numerical)

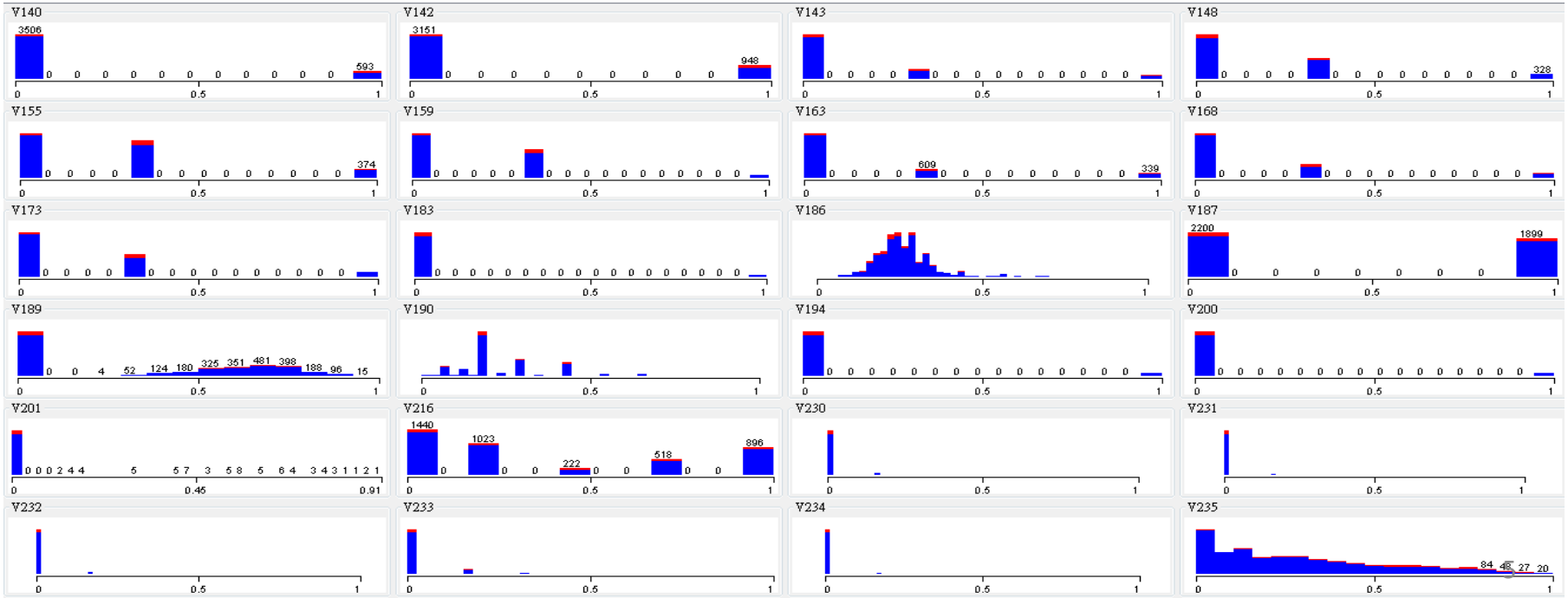
b. 1000 columns: genetic data (nominal)

c. 300 variables after data cleaning

d. 4000 patients (rows)

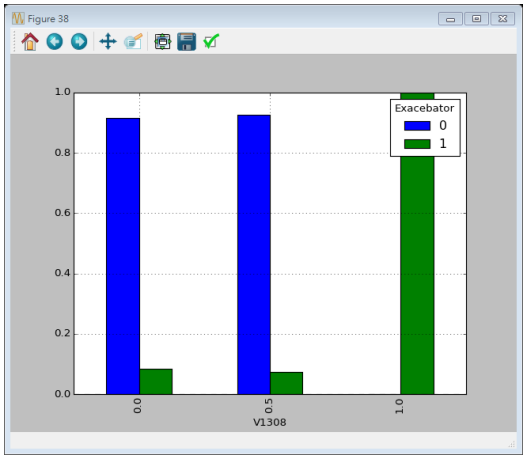
e. 1= excerabator, 0 = non excerabator (supervised)

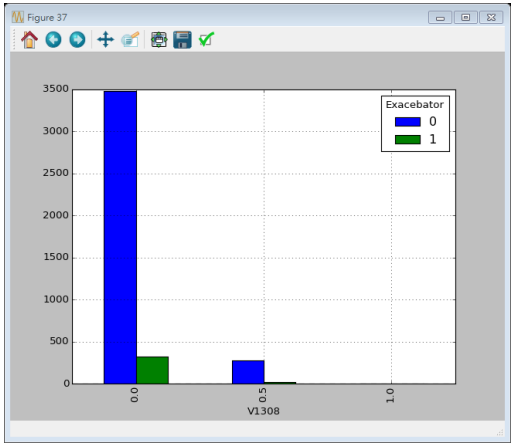

V236
322 294 217 199 203 223 216 312 286 229 189 155 165 49
V237
7 1 3 2
V238
91 78 54 32 14 7 4
V296
83 44
V297
1 3 5
V298
1 11 49 94 72 49 15 3
V302
V303
6 4 1
V304
5 0 2
V305
387
V306
V307
3398 701
V331
74 46
V334
1184 1959 956
V335
V338
1371 1797 931
V339
1441 1911 747
V340
1264 1823 1012
V345
V346
V348
V356
V358
1359 1953 787
V359
1698 1774 627

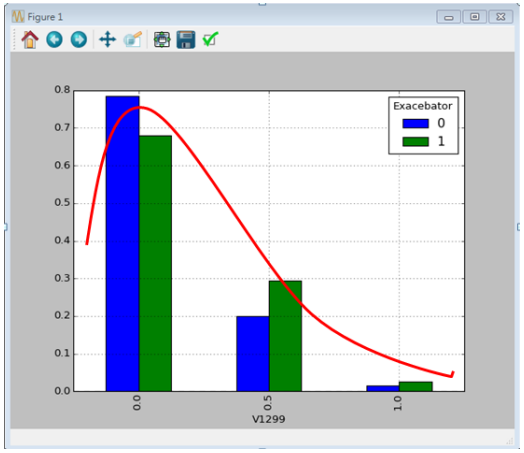

V140
3506 593
V142
3151 948
V143
V148
328
V155
374
V159
V163
609 339
V168
V173
V183
V186
V187
2200 1899
V189
4 52 124 180 325 351 481 398 188 96 15
V190
V194
V200
V201
5 5 7 3 5 8 5 6 4 3 4 3 1 1 2 1
V216
1440 1023 222 518 896
V230
V231
V232
V233
V234
V235
84 48 27 20

Method

New variable representing the effect of 1000 genes







2000 patients

No exacerbation

Has exacerbation

The percentage of patients that have genotype 0

The percentage of patients that have genotype 0

SCORE = [ ] - [ ]

Get thousand scores

sid	Exacebator	gene_score
3934	0	-3.9972
3891	0	-0.47717
5863	0	1.423719
3245	0	0.534213
2894	0	0.390069

Error Report

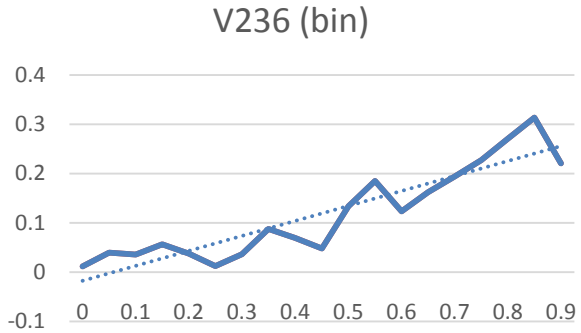
Class	# Cases	# Errors	% Error
0	941	521	55.36663
1	86	19	22.09302
Overall	1027	540	52.58033

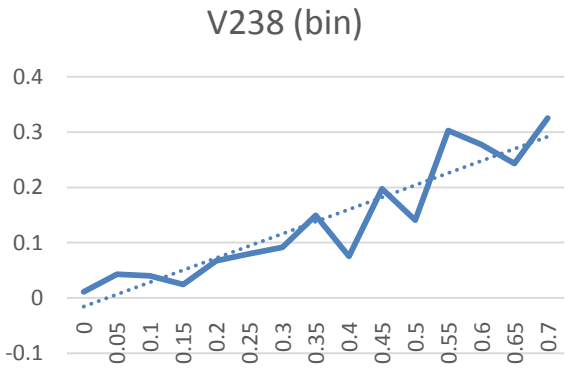
Test data for result of oversampling LR

Method

- Dimension reduction of measure part:

PCA → Explore the variable → See the correlation between variable





	V236	V238
V236	1	0.971942
V238	0.971942	1

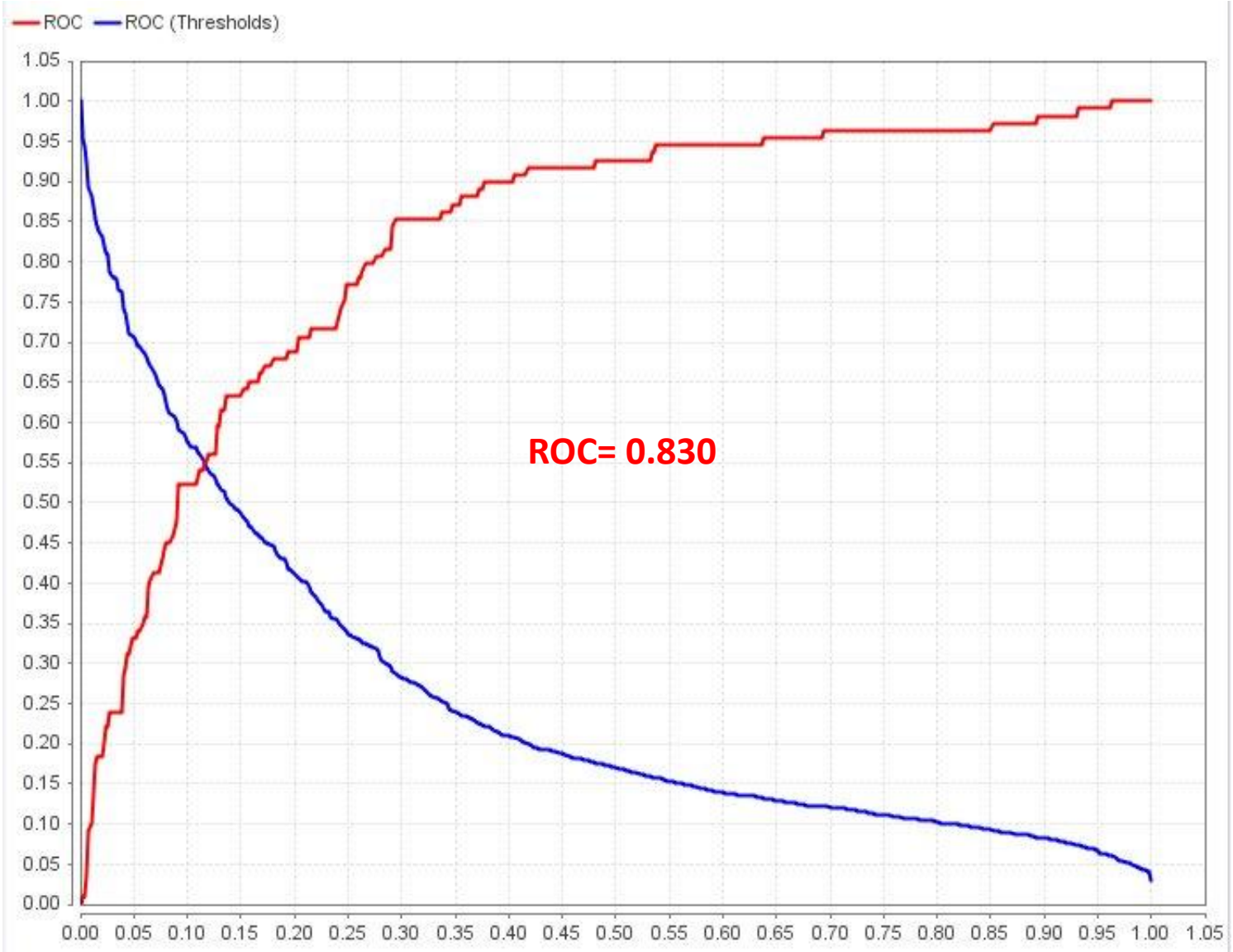
	V236	V238	V235	V304	V306	V72	V44	V210	V95	V33	V52	V59	gene_score
V236	1	0.9719425	0.8763534	0.776434	0.7000467	0.886288	0.8115886	0.7843874	0.838509	0.7928668	0.7682743	0.7694282	0.4842537
V238	0.9719425	1	0.9310565	0.7540425	0.6664926	0.872358	0.7926077	0.7623184	0.8418543	0.7695104	0.7443778	0.7461838	0.4690119
V235	0.8763534	0.9310565	1	0.7721516	0.6898759	0.8265442	0.799447	0.7753721	0.8111597	0.7363577	0.7590846	0.7621039	0.4610386
V304	0.776434	0.7540425	0.7721516	1	0.9715084	0.8153418	0.9918597	0.9955713	0.7886435	0.8451775	0.9828038	0.9883315	0.6155134
V306	0.7000467	0.6664926	0.6898759	0.9715084	1	0.7510618	0.9580819	0.967444	0.7128907	0.7890316	0.9565522	0.9610597	0.6077977
V72	0.886288	0.872358	0.8265442	0.8153418	0.7510618	1	0.8389928	0.8208115	0.8411286	0.8028894	0.8074124	0.8093702	0.4991833
V44	0.8115886	0.7926077	0.799447	0.9918597	0.9580819	0.8389928	1	0.9908399	0.8126925	0.8564577	0.9787427	0.9824421	0.6138727
V210	0.7843874	0.7623184	0.7753721	0.9955713	0.967444	0.8208115	0.9908399	1	0.7944247	0.8472186	0.9805853	0.9855422	0.6202189
V95	0.838509	0.8418543	0.8111597	0.7886435	0.7128907	0.8411286	0.8126925	0.7944247	1	0.8078122	0.7862495	0.7873466	0.4533659
V33	0.7928668	0.7695104	0.7363577	0.8451775	0.7890316	0.8028894	0.8564577	0.8472186	0.8078122	1	0.8387829	0.8425171	0.4830728
V52	0.7682743	0.7443778	0.7590846	0.9828038	0.9565522	0.8074124	0.9787427	0.9805853	0.7862495	0.8387829	1	0.975925	0.591304
V59	0.7694282	0.7461838	0.7621039	0.9883315	0.9610597	0.8093702	0.9824421	0.9855422	0.7873466	0.8425171	0.975925	1	0.5990278
gene_score	0.4842537	0.4690119	0.4610386	0.6155134	0.6077977	0.4991833	0.6138727	0.6202189	0.4533659	0.4830728	0.591304	0.5990278	1

Performance

- Logistic regression:
  - Partition the data 0.5:0.3:0.3 (training : validation: test)
  - Oversampling
  - Using selected variable (V238,V306,V33,V44,V95,gene_score)
- Validating model performance

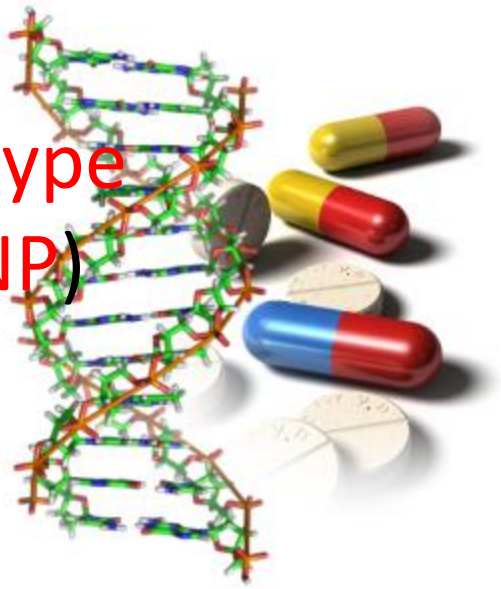
Predictors	Area under curve (AUC)
Only V238 (medical derived)	0.791
All medical derived data (w/o gene score)	0.819
Both all medical derived data and gene score	**0.830**

ROC of Both all medical derived data and gene score



conclusions

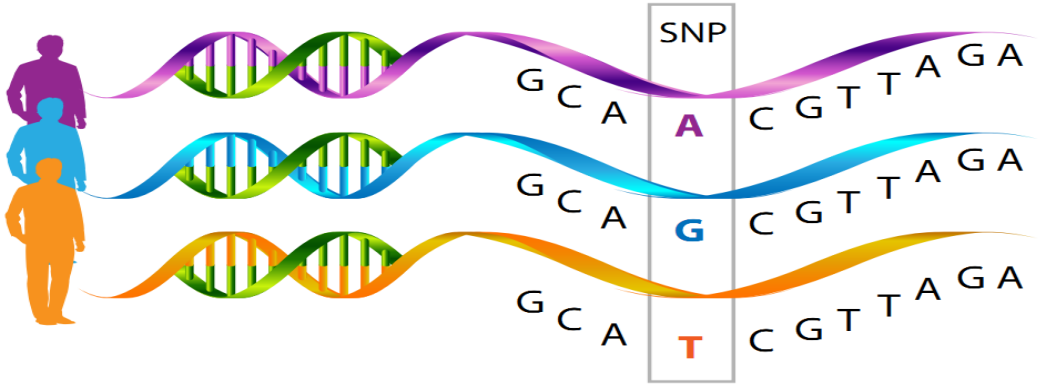
- Medical derived data play more important role in prediction. Also, genetic score can improve predicting results.
- We create a model for exacerbators with good predicting power (AUC = 0.83)
- Diseases are always related to genotype (single nucleotide polymorphism, SNP)



- Our model is:

1. A better predicting model for exacerbators (currently: physical exams ex: cough, sputum)
2. Need many measurements but easy to perform (gene screen is fast)
3. With biological and clinical meaning

- No more discussion on the relations between each variables
- Improvement: good predictors weighting



Thank
You
Merci
Gracias
Spacibo
Danke
Mèsi
Xièxie
Takk
aDank
BarakAllahuFiik
Grazie
Chokrane
Kiitos
Dziakuju
Mauruuru
GratiasAgimus
Aciu
Dankie
Motashakkeram
Tak
Sagolun
TananVäga
GratiasAgo
KopKhunKha
Gracie
CámOn
Chnorakaloutioun
KhobChaiDeu
Gratzias
Multumesc
Saha
Salamat
KamSahHamnida
Obrigado
Tenki
Aabhar
Dhanyavaad
Arigatô
Bayarlalaa
Hvala
Nirringrazzjak
NajisTuke
Grandmerce
aNiKie
Tanemirt
TerimaKasih
Rahmat
Buznyg
ThintKo
Blagodaram
Matondo
Mercé
Taiku
Dakujem
Gràcies
Faleminderit
Efharisto
TapadhLeat
Köszönöm
Murakoze
WaadMahadsantahay
Grazzi
Barkal

Thank you!!