Using Machine Learning on ICD-10 Data to Enhance an Expert Anaphylaxis Case Definition

Kamil Can Kural, Ilya Mazo, Mark Walderhaug, Lei Huang, Luis Santana-Quintero and Ravi Goud (CBER Office of Biostatistics and Epidemiology, CBER HIVE)

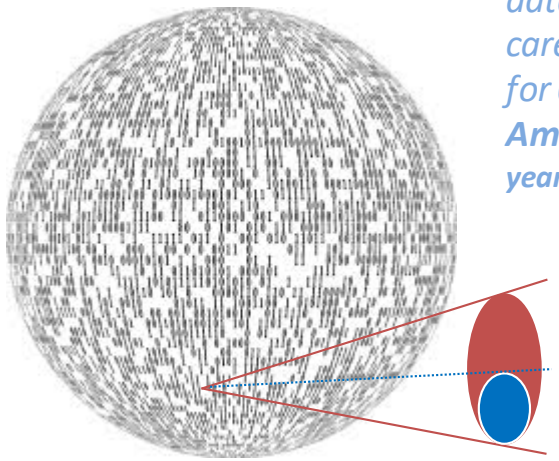
Identification of treatment related adverse events (AEs) is **critical** to FDA's **surveillance** activities.

Case definitions are necessary to identify specific episodes that are potentially AEs from large collections of health care data.

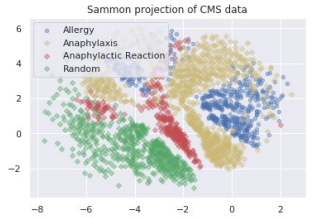
Problem:
Case definitions are time consuming and difficult to construct from administrative data.

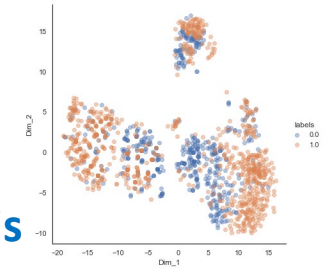
Objective:
Use **machine learning** to enhance a human constructed case definition for detection of true cases of incident **anaphylaxis** in the CMS database.

CMS episodes



*Medicare has data from health care encounters for **60 million Americans** per year*





Allergy
Anaphylaxis

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO
1		Class Label	Index	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Index	Index	Inde	Inde	Inde	Inde	Inde	Inde	Allerg	Aller	Aller	Aller	Allerg	Aller	All
2			Episode ID																																							
3			1	7	11	18	20	28	36	37	41	47	53	55	64	67	69	71	73	88	89	94	96	98	100	101	103	106	107	110	111	115	119	121	124	130	135	143	146	150	154	156
4	ER	1	0	1	1	1	1	1	1	1	1	0	0	1	1	0	1	1	1	1	0	0	1	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	0	1
5	Subsequent hospital care			1	5	28	5	1		1	4	1	1	13	3			4	2	2			5	4					1	2	2			4	1		1	1		7	1		
6	Critical care, first hour			2	3	9	3			1	2	1	2	3	2			4	1	1			5		2		1	3		2			1	5				2	1	3	1	1
7	TOTAL CHARGE	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8	PHARMACY (ALSO SEE 063X, AN EXTENSION OF 025X) - GEN	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
9	LABORATORY - GENERAL CLASSIFICATION	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10	LABORATORY - CHEMISTRY	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
11	EKG/ECG (ELECTROCARDIOGRAM) - GENERAL CLASSIFICATIC	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
12	Anaphylactic shock, unspecified, initial encounter		2	2	1	4	3		4	7		2	4					8	1		6			2	1			1	2	2				9			2		3	2	1	
13	LABORATORY - HEMATOLOGY	1		1		2		1	1	1	1		1		1	1	1	1		1	1	1	1	1	1	1	1	1	1		1		1		1	1		1	1	1	1
14	RESPIRATORY SERVICES - GENERAL CLASSIFICATION	1	1	1	1	2	1	1	1	1	1	1	1	1			1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1		1	1			1	1
15	EMERGENCY ROOM - GENERAL CLASSIFICATION	1		1	1	2	1	1	1	1			1	1		1	1	1	1			1	1	1	1	1	1		1	1		1	1	1	1	1	1	1	1	1		1
16	Angioneurotic edema, initial encounter			2	12		1							6					5								4						1	1			1		4			
17	Anaphyl reaction due to advrs eff drug/med prop admin, ir	2		1	1	2		1		1	1		1	1	1	1	1	6	3	1	1	1	1	1	1	1		3	4	1	2	1	1			1		1	1	1	
18	RADIOLOGY - DIAGNOSTIC - CHEST X-RAY	1		1	1	2		1	1	1				1	1		1	1	1		1	1	1	1	1	1	1	1			1		1	1	1	1	1		1	1	1	1
19	MEDICAL/SURGICAL SUPPLIES AND DEVICES (ALSO SEE 062X	1			1	1		1	1	1	1		1	1	1		1				1	1	1		1	1	1						1				1	1	1			
20	LABORATORY - BACTERIOLOGY & MICROBIOLOGY	1		1		2		1		1			1	1	1	1	1			1	1	1	1	1	1	1	1	1	1		1		1			1	1	1	1	1	1	1
21	Initial hospital care				1	1	2		2				2	1					1				1					1	1			1	1	1		1	1	3	3			
22	MEDICAL/SURGICAL SUPPLIES AND DEVICES (ALSO SEE 062X, AN EXTEN	1		1						1		1	1			1	1	1			1	1	1		1	1				1							1				1
23	Essential (primary) hypertension				1	1			1	1			2	2	1				1	2	1	1		1			3		3	1					1	1	3	5			
24	PHARMACY (ALSO SEE 063X, AN EXTENSION OF 025X) - IV SOLUTIO		1		1				1	1		1		1				1		1		1		1			1					1	1								1
25	Acute respiratory failure with hypoxia	2	4	3		1			1	3	1	4		5		1	1	6		2		3		1	1		1	1	1	1			3	4					5		1
26	Insertion of Endotracheal Airway into Trachea, Via Openinç	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1	1	1
27	RADIOLOGY - DIAGNOSTIC - GENERAL CLASSIFICATION	1	1		1		1	1	1	1	1	1	1		1	1	1	1	1	1		1		1	1	1			1	1		1		1			1	1			
28	INTENSIVE CARE UNIT - GENERAL CLASSIFICATION		1		1	1	1	1	1		1		1		1	1	1		1					1			1		1	1							1	1			
29	Insert emergency airway	1			1	3			1	1			1		1				1			1	1	1			2	1		1			1				2	1	1		1
30	PHYSICAL THERAPY - GENERAL CLASSIFICATION			1					1	1	1		1	1	1	1	1			1	1	1	1	1					1		1	1	1		1	1					1
31	PULMONARY FUNCTION - GENERAL CLASSIFICATION	1		1		1		1		1			1	1	1		1	1	1		1	1		1		1	1				1		1	1	1	1			1		1
32	Acute respiratory failure, unsp w hypoxia or hypercapnia	1			39	3		1	1	2								3						1	1			5		1		1	2	6				1	5	3		
33	PHYSICAL THERAPY - EVALUATION OR RE-EVALUATION			1		1			1						1	1	1	1		1	1	1	1	1		1	1		1		1		1	1	1	1	1	1			
34	LABORATORY - UROLOGY	1		1		2			1				1				1			1	1	1		1	1			1	1				1			1	1	1			
35	Hospital discharge day					1				1			1				1										1		1					1				1			
36	Electrocardiogram report	1	1		1	1			1		1							4	2	2				1			3			2						1		1		
37	LABORATORY - IMMUNOLOGY	1							1	1	1		1			1				1		1		1	1	1		1			1			1		1	1	1		1
38	Respiratory Ventilation, 24-96 Consecutive Hours			1	1			1		1	1		1	1	1	1		1	1	1		1		1	1	1	1		1			1		1			1	1			
39	Acute kidney failure, unspecified			1							1	7						1	1		2				2	1		3	1	1				1	1		1		4	1	
40	OPERATING ROOM SERVICES - GENERAL CLASSIFICATION				1			1	1	1	1			1							1						1											1			
41	Athscl heart disease of native coronary artery w/o ang pctrs				1			1	2		1			1									1						2												

We suggest a **novel way** of using machine learning to **help create case definitions**

and

we apply it to **identify** cases of **anaphylaxis**, a serious, life-threatening allergic reaction.

CBER HIVE

Using Machine Learning on ICD-10 Data to Enhance an Expert Anaphylaxis Case Definition

Kamil Can Kural, Ilya Mazo, Mark Walderhaug, Lei Huang, Luis Santana-Quintero and Ravi Goud (CBER Office of Biostatistics and Epidemiology, CBER HIVE)

Original Workflow Without Chart Confirmed Results

Minimally curate **2500 claims** into groups of 500 likely anaphylaxis, possible anaphylaxis, allergy, and 1000 controls using a subset of ICD-10 codes

Apply **supervised** and **unsupervised machine learning** models to the 2500 cases

Identify **new important codes** and construct **models** to improve the classification of **anaphylaxis cases**

Test machine learning models on chart reviewed data

Data Assumptions

- Anaphylaxis and allergy are similar.
- A classifier that discriminates anaphylaxis from allergy will also discriminate it from other acute health problems.
- Anaphylaxis is rare enough that the random control cases will not contain anaphylaxis episodes.

Methods Used

- Linear discriminant analysis (HIVE-RLDA). (Results not shown)
- T-distributed stochastic nearest neighbor embedding. (t-SNE)
- Decision Tree
- Random Forest

Return Home

Using Machine Learning on ICD-10 Data to Enhance an Expert Anaphylaxis Case Definition

Kamil Can Kural, Ilya Mazo, Mark Walderhaug, Lei Huang, Luis Santana-Quintero and Ravi Goud (CBER Office of Biostatistics and Epidemiology, CBER HIVE)

Latest Workflow with Chart Confirmed Results

Cohort 1

Consists of **2500 samples**

Minimally curated **CMS** claims dataset

High, **intermediate**, and **very low** likelihoods of being Anaphylaxis, plus a **random background** claims dataset

Cohort 2

Consists of **530 samples**

Specifically created to **learn** and **extract** information about **vaccine induced Anaphylaxis**

Claims were identified through an **Anaphylaxis Algorithm** created by experts using **patterns** in **claims data**

Chart Confirmed Results

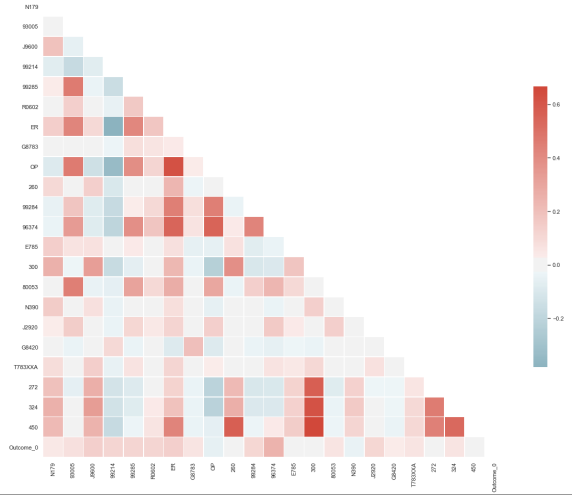
Consists of **159 samples**

A subset of **Cohort 2**

Represents the charts that were **requested**, **received**, and for which chart - **confirmed**, Anaphylaxis **status** is **known**

Feature Selection

- **Logistic Regression with High Regularization**
- **Chi Square Analysis**
- **Random Forest**
- **Light Gradient Boosting**
- **Logistic Regression with Low Regularization**



Feature	Total	Feature
x ray of chest 2 views front and side	5	pneumococcal vaccine for injection into muscle
unspecified osteoarthritis unspecified site	5	hyperlipidemia unspecified
toxic effect of venom of bees accidental (unintentional) initial encounte	5	encounter for immunization
subsequent hospital inpatient care typically 35 minutes per day	5	encounter for general adult medical examination without abnorm
shortness of breath	5	annual wellness visit includes a personalized prevention plan of se
routine electrocardiogram (ekg) with tracing using at least 12 leads	5	administration of pneumococcal vaccine
respiratory ventilation 24 96 consecutive hours	5	vaccine for influenza for injection into muscle
respiratory services general classification	5	vaccine for influenza for administration into muscle 0.5 ml dosage
respiratory inhaled pressure or nonpressure treatment to relieve airwa	5	routine ekg using at least 12 leads including interpretation and re
radiology diagnostic chest x ray	5	op
pb	5	injection diphenhydramine hcl up to 50 mg
unclassified drugs	5	established patient office or other outpatient visit typically 25 mir
other long term (current) drug therapy	5	influenza virus vaccine split virus when administered to individuals
op	5	er
old myocardial infarction	5	current tobacco non user (cad cap copd pv) (dm) (ibd)
injection methylprednisolone sodium succinate up to 125 mg	5	blood test comprehensive group of blood chemicals

Table 1: Coloring scheme of features. Green represents features important for all three cohorts. Blue color represents features important for two out of three cohorts. Red represents features important only for that specific cohort.

Feature Selection

- Use statistical methods as well as machine learning models to identify most salient features between datasets.
- The features are color coded to check robustness between datasets.
- An ideal feature would pass at least 1 test in each dataset.

Methods Used

- Random Forests
- Sammon Mapping
- T-distributed stochastic nearest neighbor embedding (t-SNE)
- Light Gradient Boosting
- Logistic Regression

CBER HIVE

Return Home

Using Machine Learning on ICD-10 Data to Enhance an Expert Anaphylaxis Case Definition

Kamil Can Kural, Ilya Mazo, Mark Walderhaug, Lei Huang, Luis Santana-Quintero and Ravi Goud (CBER Office of Biostatistics and Epidemiology, CBER HIVE)

Results: Feature Selection

90 Features were selected for classification task

Tests Passed	Cohort 1	Cohort 2	Chart Confirmed Results
5	38	9	13
4	33	12	6
3	72	118	93
2	82	60	93
1	127	80	91
0	3876	918	901
Total Features	4228	1197	1197

Table 2: Number of features that satisfies each threshold after running multiple feature selection algorithms. For example, there are 38 features that satisfied all 5 feature selection algorithms for Cohort 1.

Threshold	Cohort 1	Cohort 2	Chart Confirmed Results
5	0.982120 +/- 0.010557	0.883137 +/- 0.067506	0.804325 +/- 0.080723
4	0.984580 +/- 0.007396	0.875336 +/- 0.071413	0.844127 +/- 0.085665
3	0.983860 +/- 0.007900	0.872437 +/- 0.071407	0.812116 +/- 0.090686
2	0.982180 +/- 0.008964	0.872619 +/- 0.075855	0.810952 +/- 0.103367
1	0.983120 +/- 0.008932	0.874118 +/- 0.075392	0.801905 +/- 0.088265
0	0.983740 +/- 0.008431	0.874118 +/- 0.075392	0.800053 +/- 0.081258

Table 3: Success metrics of each trained model in AUC for each individual dataset with different thresholds.

- Fit individual models for each dataset to see how features that satisfied multiple criteria contribute to the models' success.
- Some features were selected even though they did not pass at least 1 test for each Cohort.
- Create the final models using features without expert curated codes. (Remove codes while finding important features)

Return Home

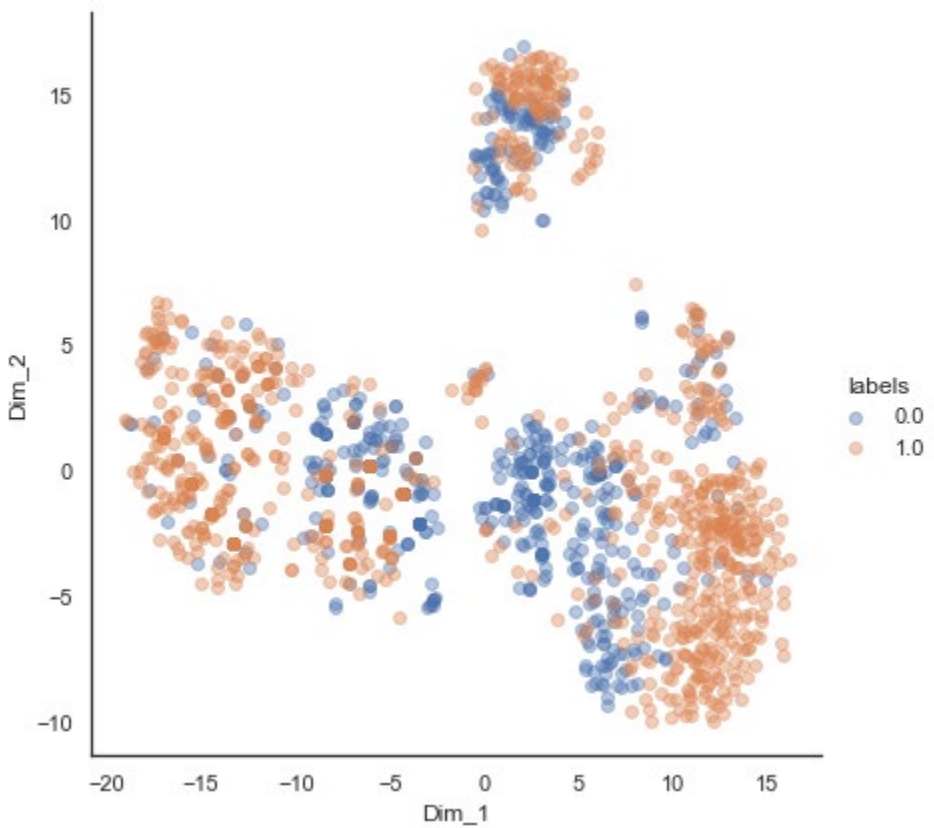
CBER HIVE

Using Machine Learning on ICD-10 Data to Enhance an Expert Anaphylaxis Case Definition

Kamil Can Kural, Ilya Mazo, Mark Walderhaug, Lei Huang, Luis Santana-Quintero and Ravi Goud (CBER Office of Biostatistics and Epidemiology, CBER HIVE)

Results: Unsupervised Machine Learning Models (T-SNE)

- T-SNE analysis overall separates Anaphylaxis episodes (Orange) from Allergy. (blue)
- Further, it shows we might have different influencers for each different dataset based on 3 different clusters.
- 90 features are enough for clear separation.
- Classifications likely had inaccuracies, thus need for improved case definitions.
- After Chart Confirmation, there were multiple changes in Cohort 2 data. (Allergy Cases went up by 52 samples)
- Even the expert definition has less than desired success in identifying Anaphylaxis cases correctly.



Points are colored by group, **anaphylaxis** vs **allergy**

Anaphylaxis – 1

Allergy - 0

CBER HIVE

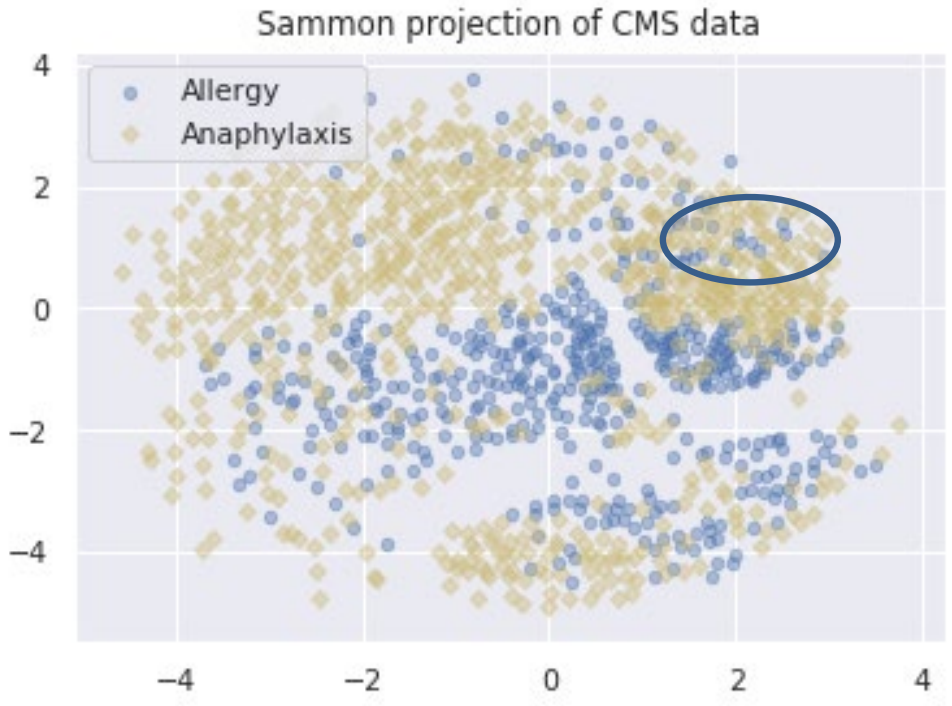
Return Home

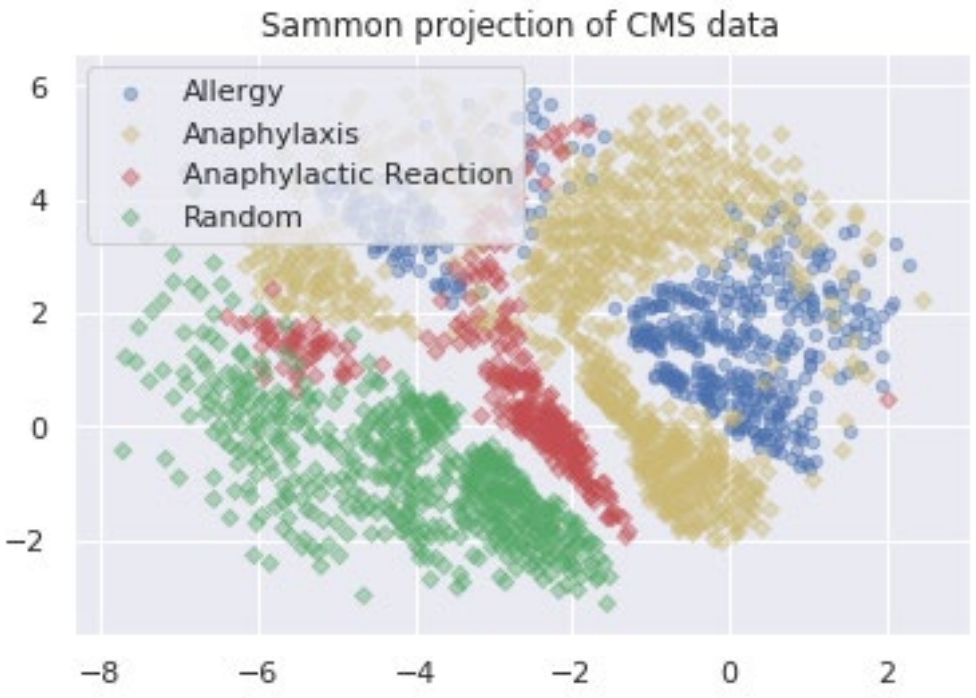
Using Machine Learning on ICD-10 Data to Enhance an Expert Anaphylaxis Case Definition

Kamil Can Kural, Ilya Mazo, Mark Walderhaug, Lei Huang, Luis Santana-Quintero and Ravi Goud (CBER Office of Biostatistics and Epidemiology, CBER HIVE)

Results: Unsupervised Machine Learning Models: Sammon Mapping

- Sammon Mapping performs much better in Classifying CMS Data.
- Again, we can see some of the labels which might be misclassified with expert definitions. (Cohort 2)
- Emphasizes our earlier point which states that the expert definition is not very successful for identifying Anaphylaxis cases.





Points are colored by group, **anaphylaxis** vs **allergy**

CBER HIVE

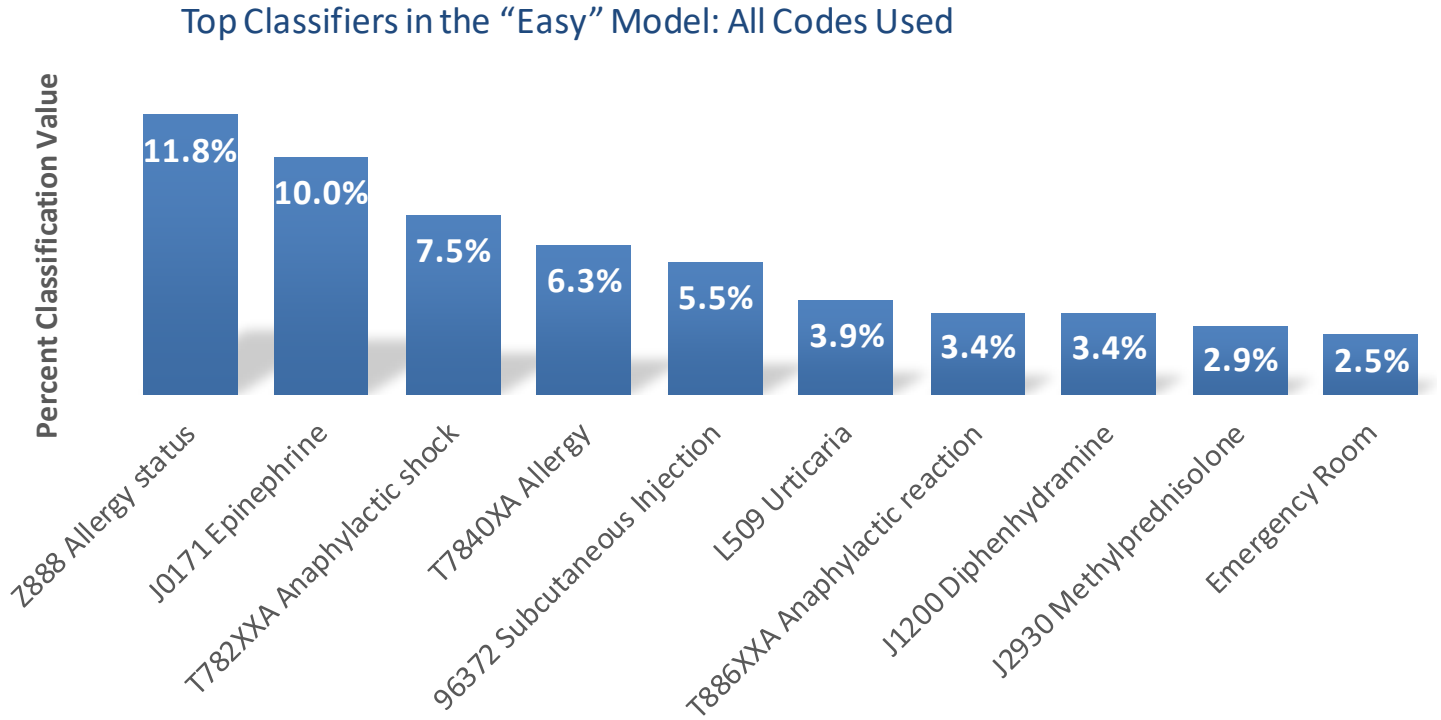
Return Home

Using Machine Learning on ICD-10 Data to Enhance an Expert Anaphylaxis Case Definition

Kamil Can Kural, Ilya Mazo, Mark Walderhaug, Lei Huang, Luis Santana-Quintero and Ravi Goud (CBER Office of Biostatistics and Epidemiology, CBER HIVE)

Results: Supervised Machine Learning Models: Light Gradient Boosting (Easy)

- Supervised classification with Allergy and Anaphylaxis.
- 80% in training set, 20% in testing set stratified and shuffled randomly.
- A separate portion 20% of the data was used for Hyperparameter tuning.
- ***Construct 'Easy' decision trees by including codes used to create the data in the lists.***
- Train decision trees until one with high discrimination is found.
- Extract classifiers and construct final tree.



Machine learning finds the codes experts used to construct the allergy and anaphylaxis groups in the data set. This is expected, since the outcomes and codes the experts used are highly correlated by design of the dataset.

Epinephrine, Allergy Diagnosis, Anaphylaxis Diagnosis, Urticaria are key features.

CBER HIVE

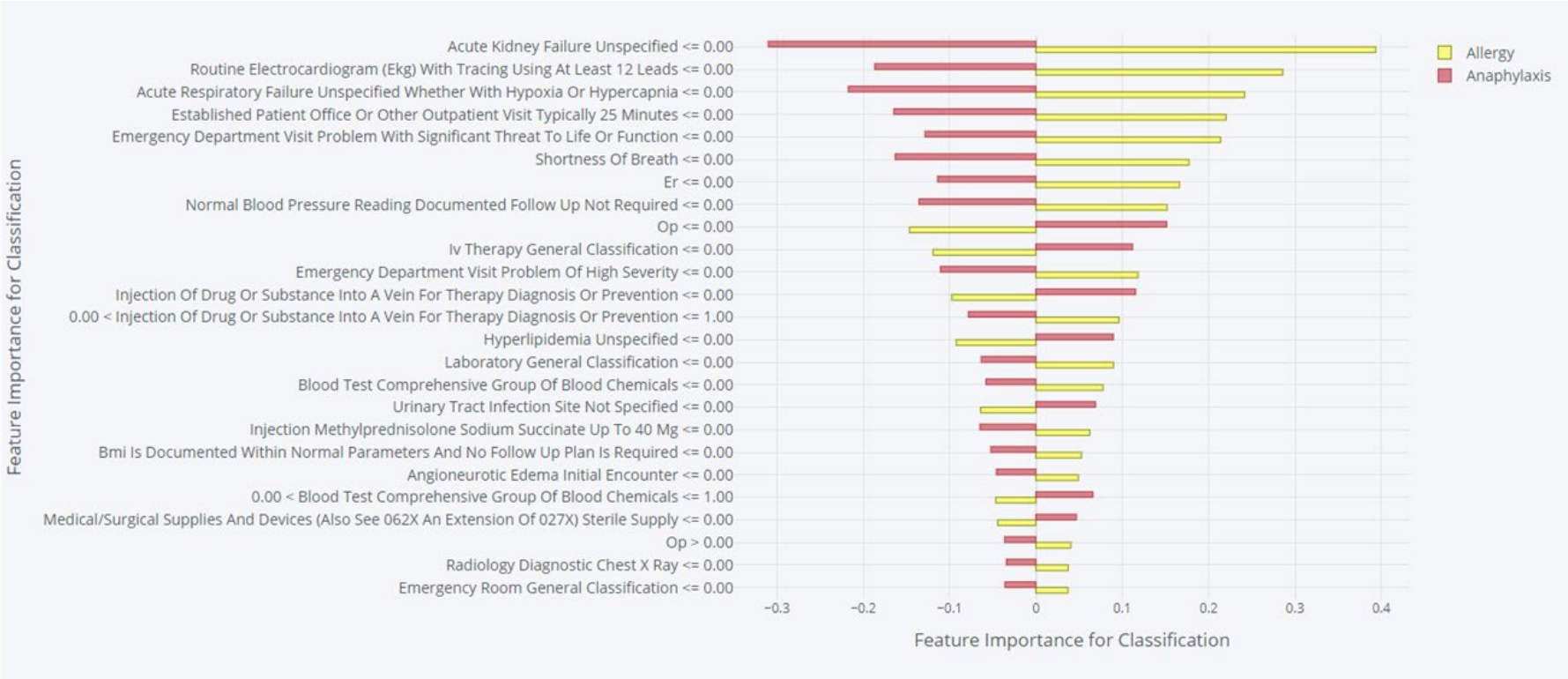
Return Home

Using Machine Learning on ICD-10 Data to Enhance an Expert Anaphylaxis Case Definition

Kamil Can Kural, Ilya Mazo, Mark Walderhaug, Lei Huang, Luis Santana-Quintero and Ravi Goud (CBER Office of Biostatistics and Epidemiology, CBER HIVE)

Results: Supervised Machine Learning Models: Light Gradient Boosting (Hard)

- ***Construct 'Hard' decision trees: remove codes used to manually select the cases and look for other factors embedded in the data.***
- This is necessary to eliminate very strong and highly correlated signals.
- Instead of seeking codes which define cases directly, identifying the patterns this way would be much more likely.
- *Validation:* compare to the codes used in manual case definition built independently by group of experts.



Acute Kidney Failure, Acute Respiratory Failure, Injection Codes and Emergency, Saline administration and Department visit codes were very significant.
Models identified the importance of billing for injections.
Treatment setting is important in identifying Anaphylaxis.

CBER HIVE

Return Home

Using Machine Learning on ICD-10 Data to Enhance an Expert Anaphylaxis Case Definition

Kamil Can Kural, Ilya Mazo, Mark Walderhaug, Lei Huang, Luis Santana-Quintero and Ravi Goud (CBER Office of Biostatistics and Epidemiology, CBER HIVE)

Results: Light Gradient Boosting Model Performance

"Easy"	Predicted as Allergy	Predicted as Anaphylaxis
Allergy	122	1
Anaphylaxis	3	180

"Hard"	Predicted as Allergy	Predicted as Anaphylaxis
Allergy	119	4
Anaphylaxis	6	177

- **Model accuracy** is **96.7%** with considerable information removed.
- "Easy" model setting has around 1200 features while hard setting has only 90 features.
- Information that seemed nonspecific can be used by machine learning models to identify Anaphylaxis cases from ICD-10 codes.

- Confusion matrices for both settings points at a lower model performance for hard model setting.
- Lower model performance for hard model is expected, since the data used 'easy' model codes for determining Anaphylaxis and Allergy cases. As a result, some of the cases are misclassified.
- Our aim is to identify a **pattern** instead of trying to associate a specific code to Anaphylaxis, which might mislead researchers with or without its absence.

CBER HIVE

Return Home

Kamil Can Kural, Ilya Mazo, Mark Walderhaug, Lei Huang, Luis Santana-Quintero and Ravi Goud (CBER Office of Biostatistics and Epidemiology, CBER HIVE)

Conclusions

This work demonstrates that the ***combination of machine learning and a minimally curated*** claims data set could help to quickly identify traits associated with cases of interest, thereby potentially improving the efficiency and accuracy of case definition construction and foster public health innovation.

Validation:
ML methods on the minimally curated dataset succeeded in independently identifying treatment codes (e.g. diphenhydramine, methylprednisolone as well as injection codes) that were used in the manual case definition built by experts.

Python code developed in this project can be made available HIVE for CBER scientists interested in using machine learning.

Kamil Can Kural, Ilya Mazo, Mark Walderhaug, Lei Huang, Luis Santana-Quintero and Ravi Goud (CBER Office of Biostatistics and Epidemiology, CBER HIVE)

Acknowledgements

We would like to thank **Acumen** and **CMS** for providing the data.

"This presentation is an informal communication and represents the best judgment of authors. These comments do not bind or obligate FDA."