DiagnoSmart Platform
Path to Unbiased Diagnosis

DiagnoSmart platform

Proof of concept statistical report

Table of content

I. Context

Women are often considered less seriously than men when complaining about symptoms like aches and pains, even if these symptoms can be a warning sign for serious diseases such as heart attack, kidney stones etc. Moreover, men's symptoms are the norm in terms of diagnosis in several diseases despite being different from the symptoms experienced by women.

SymptomSavvy propose a DiagnoSmart platform, a CDS (Clinical Decision Support) tool for quick check of the risk for a selection of diseases at the emergency admission

considering with great attention the difference in terms of symptoms between women and men.
The tool would use only easily available data to avoid heavier work for caregivers at the emergency department. Indeed, a tool time consuming and asking for more data would not be used by caregivers that would otherwise disregard a woman's symptoms. Then the tool needs to be quick and easy to use, and to use only routinely collected data and information given by the patient.

The tool, even if it can be used for men and women, would decrease the risk to disregard a woman's symptom if the patient is not considered seriously or if the woman's symptoms for a disease are unknown by the caregiver. Above all, it can in any case avoid forgetting about a potential diagnosis.

Thus, the objective of the DiagnoSmart is to decrease the risk of wrong diagnosis, avoid medical errancy, and save lives by giving a tool to improve women's diagnosis.

II. Literature and state of the art

One disease stands among misdiagnosed conditions in women: Heart attack.

University of Leeds, conducted studies using the UK national heart attack register MINAP, and found that almost one third of patients had an initial diagnosis which differed from their final diagnosis. And it shows that women were 50% more likely to receive a misdiagnosis of a heart attack compared to men [8].

Receiving a prompt diagnosis and receiving appropriate treatment following a heart attack are crucial for facilitating optimal recovery. A timely diagnosis not only influences immediate treatment but can also have lasting effects on long-term care. Women who received a misdiagnosis faced a roughly 70% higher risk of mortality within 30 days compared to those accurately diagnosed, mirroring similar outcomes observed in men.

Another study published in the Journal of the American Heart Association found that women under the age of 55 were seven times more likely to be misdiagnosed in the emergency department when experiencing a heart attack.

Based on the analysis of the UK's National Heart Attack Registry, it has been predicted that by administering the same standard of care in women as men, at least 8000 deaths could have been prevented in female cardiac arrest patients.

Several studies have shown the potential power of machine learning algorithm for diagnostic prediction in emergency departments. Kareemi et al. [6] have reviewed 7 studies comparing machine learning based diagnoses with usual cares in emergency. All studies showed that the machine learning based diagnosis outperformed usual care for all metrics. This shows that misdiagnosis could be decrease with the use of machine learning algorithm.

Concerning heart attack, Ahsan et al. [7] published a review of Machine learning-based heart disease diagnosis showing that a lot of them have been developed and that if even

challenges exist, the production of clinical decision support tools for heart attack diagnosis is possible.

III. The methodology

In reason of the time limitation, we decided to focus on only one machine learning algorithm that would predict acute myocardial infarction. Other machine learning algorithm would be produced in a second time.

IV. Database and data analysis

1. MIMIC-IV database

The proof-of-concept of DiagnoSmart development rely as a firs instance on MIMIC-IV (Medical Information Mart for Intensive Care) database. This relational database contains health-related data associated with patients who were admitted to critical care units at the Beth Israel Deaconess Medical Center in Boston, Massachusetts between 2011 and 2019 (more than 200,00 patients). It is a publicly available database widely used for research purposes in the healthcare domain.

The MIMIC-IV database includes a wide range of clinical data, including demographics, vital signs, laboratory tests, medications, procedures, and clinical notes. One component of the MIMIC-IV database is the Emergency Department (ED) data. The composition of this sub-dataset is described in Appendix 1.

MIMIC-IV provides a valuable resource for researchers and clinicians to conduct studies related to emergency medicine, critical care, and healthcare quality improvement. However, it's essential to adhere to data usage policies and ethical considerations when accessing and analysing the data. For this purpose, all members of the team that had access to the data, have obtained a certification by completing the CITI Data or Specimens Only Research (training provided by the MIT) beforehand [4].

2. Files description

The files used for this proof of concept are the following:

edstays.csv

It contains “the primary tracking table for emergency department visits. It provides the time the patient entered the emergency department and the time they left the emergency department.”[1].

This table is use for determination of the gender subgroup and descriptive statistics.

Number of rows: 425,087.

Fields: [subject_id, hadm_id, stay_id, intime, outtime, gender, race, arrival_transport, disposition].

triage.csv

It contains vital signs and chief of complaints from the triage step in the mergency department.

"*Patients are assessed at triage by a single care provider and asked a series of questions to assess their current health status. Their vital signs are measured and a level of acuity is assigned. Based on the level of acuity, the patient either waits in the waiting room for later attention, or is prioritized for immediate care." [1]*

This file is used to extract the input variables of the machine learning algorithm. The fields are all free text which make their use challenging.

Number of rows: 425,087.

Fields: [subject_id, stay_id, temperature, heartrate, resprate,o2sat, sbp, dbp, pain, acuity, chiefcomplaint]

diagnosis.csv

It contains "*billed diagnoses for patients. Diagnoses are determined after discharge from the emergency department*"[1]

This file was used to determine the ground-truth of our variable of interest (in our context the infarction status).

Number of rows: 899,050

Fields: [subject_id, stay_id, seq_num, icd_code, icd_version, icd_title]

3. Statistics on the used data

MIMIC-IV-ED database contains :

- 205,504 unique patients
- 202,441 unique admissions
- 425,087 unique stays

The rate of women within the stays is 54.08%.

A set of descriptives statistics are presented in Appendix 2. It shows that there is a significant[1] difference between women and men on dispositions as 39.67% of men are admitted to the hospital against only 35.05% of women. On the opposite, 59.34% of women are sent home while only 53.89% of men (Appendix 2, Figure 1, and Table 2).

V. The machine learning algorithm for acute myocardial infarction risk prediction: DiagnoSmart-AMI

1. The variable of interest (Infarction status) definition

As we decide to predict the risk to be in presence of an acute myocardial infarction, we consider as positive patients with the billed diagnosis registered in the emergency department with the following ICD codes:

When the ICD-9 version is used, the patients with ICD code 410 (Acute myocardial infarction) is used [2].

When considering the ICD-10 version, the patients with ICDs starting with I21, I22, and I23 are used [3].

Limitation: This method to define the ground-truth has some limitations:

- We consider as positive to acute myocardial infarction solely well diagnosed patients. If the patient was sent home with an acute myocardial infarction, he will be considered negative.
- We don't have the time of diagnosis so we can't know if it would have saved time and thus lives to use DiagnoSmart-AMI.

2. The input variables selection and definition

Two types of data are available at triage for all patients.

One value of each vital sign at admission: Body temperature, heart rate, respiratory rate, oxygen saturation, systolic and diastolic blood pressure.

Chief of complaint which are free-text data. It means it contains around 25,000 unique symptoms.

After data analysis and split of the values to obtain one symptom per row, the RegEx method was used for symptoms extraction.

Were extracted from the chief of complaints column:

- The side (left or right when precised) (dict_side) (see Appendix 3 Figure 2)

[1] The p-value from the chi-test is 2.89e-252.

- The location of the symptom on the body (dict_location) (see Appendix 3 Figure 3)
- The symptoms of interest (dict_symptom) (see Appendix 3 Figure 4)
- The symptoms close to a diagnosis that would make the product unnecessary for the patients (dict_diagnosis) (see Appendix 3 Figure 5).

After the classification of the symptoms according to the previous dictionaries, the following symptoms are extracted (45 features):

"jaundice, hyperglycemia, dehydration, Hematemesis, distention, nausea, swelling, tachycardia, bleed, fatigue, fever, cough, itch, paralysis, diarrhea, dizzy, hemorroids, neurologic, lump, numbness, seizure, migraine, sore, smelling urine, hearing loss, rash_redness, hypoglycemia, dyspnea, anemia, throat foreign body sensation, constipation, dysuria, anxiety, hematuria, pain_back, pain_neck, pain_chest, pain_joint, pain_abdominal, pain_head, pain_urinary track, paralysis_face, paralysis_arm, cramps_abdominal, pain_arm_left"

The symptoms are not necessarily selected toward acute myocardial infarction as the purpose was to select among them the most clinically and statistically relevant symptoms.

When combined with the vital signs' parameters, it leads to 51 input features.

Descriptive statistics on the 51 features can be found in Appendix 4 Table 3.

The patients containing one symptom from dict_diagnosis were removed from the training and evaluation sets as if the diagnosis is already made at admission, DiagnoSmart is unnecessary or less relevant. In some cases, it also implies that symptoms are not registered when the diagnosis is registered instead. At this step, 113,410 stays are removed.

Limitation: This method to define the input data has some limitations:

- If symptoms too obvious chief of complaint is directly cardiac arrest which means the symptoms are not included in the list. So training and evaluation dataset, are made of unobvious cardiac arrest.

3. The selected patients and time points

The selected time point is the time of triage. Thus, all input data are data collected during the triage and it corresponds to only one value per feature.

After cleaning of the dataset (remove patients with none of the selected symptoms, remove patients with missing vital signs, remove patients with evident diagnosis at triage) the number of stays is 249,692 among which 362 women and 551 men had acute myocardial infarction. The infarction rates are respectively 0.26% considering only the one not diagnosed at admission before or during triage, and 0.54%.

As the rate of cardiac arrest is low[2] it was decided to keep the ratio of infarction/not infarction to 50/50 for a first version of the algorithm. A version with the ratio closer to real rate will be used in the next version of the model.

For this first version we use a 50% rate of women in order to produce a model that would add as much importance into men and women diagnosis. Indeed, previous experiments showed that keeping a 40% women rate leads to a model with 88% of accuracy for men and 76% of accuracy for women.

Moreover, we can't use only women for the diagnosis as the number of patients for the training would be too small.

4. The machine learning model

Selected data have been spitted through a stratified 6 folds keeping 5 folds for a 5-fold cross validation and one fold for the independent testing.

We use as input of the model a matrix of side (1,51) for each patient.

The ground-truth is a binary value:

- 1 if the patient has a billed diagnosis registered in the emergency department.
- 0 otherwise.

DiagnoSmart-AMI use a XGBoost model and predict a risk of having an acute myocardial infarction.

XGBoost, which stands for eXtreme Gradient Boosting, is an optimized and scalable implementation of gradient boosting machines, a powerful ensemble learning technique.

It combines several well-known methods:

Gradient Boosting

- Gradient boosting is a machine learning technique that builds a predictive model in the form of an ensemble of weak learners, typically decision trees.
- It works by sequentially adding weak learners to a model, with each new learner correcting the errors made by the existing ensemble.
- The final prediction is made by aggregating the predictions of all the weak learners.

Decision Trees

- Decision trees are simple models that make predictions based on a series of hierarchical decisions.

[2] In the publication [5] the national incidence rate 2016-2018 was 1.6 (range 1.60-1.70) cardiac arrests per 10,000 ED visits in the ED reported to the Swedish Registry for Cardio-Pulmonary Resuscitation (SRCR) during 2007-2018)

- Each decision tree is a weak learner, meaning it performs slightly better than random guessing but is not particularly strong on its own.

XGBoost improves upon traditional gradient boosting by introducing several optimizations and regularization techniques to make the model more accurate and less prone to overfitting. It uses a gradient descent algorithm to minimize a loss function while adding new trees to the ensemble. XGBoost incorporates both a linear model and a tree model, allowing it to capture linear and complex patterns in the data. Regularization techniques such as L1 and L2 regularization are used to control the complexity of the model and prevent overfitting. XGBoost also supports parallel processing and distributed computing, making it highly scalable and efficient for large datasets.

XGBoost has several advantages:

- Fast and efficient: XGBoost is designed for speed and can handle large datasets with millions of samples and features.
- Regularization: Built-in regularization techniques help prevent overfitting and improve generalization performance.
- Flexibility: XGBoost can be used for both regression and classification tasks and supports various loss functions.
- Feature importance: XGBoost provides feature importance scores, allowing users to understand which features are most influential in making predictions.

Overall, XGBoost is a versatile and powerful machine learning algorithm that is widely used in practice due to its effectiveness, efficiency, and scalability.

It was decided to create one model per disease instead of using a classification model to predict the disease for several reasons. The first one is in reason of the short timing, creating a multi-class model would imply more cleaning work. Secondly, the creation of multi-classes models in this context, is challenging for several reasons including for example imbalanced-classes or cases of multi-diagnosis.

Limitation: This method as some limitations:

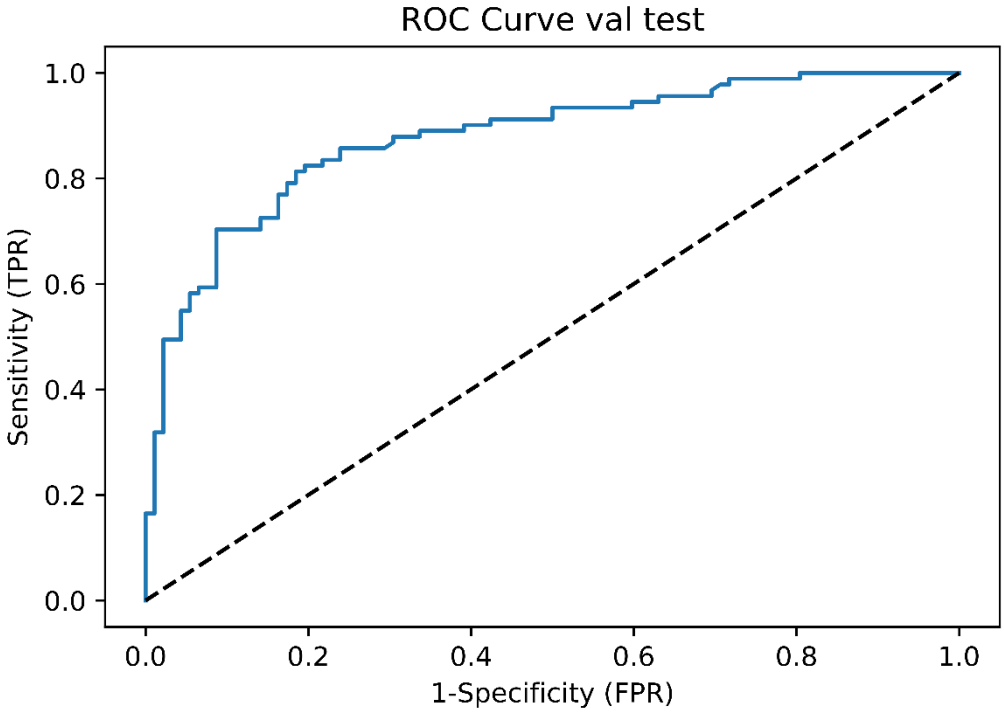
- In the future a better method to handle hot-vector symptoms variable can be seen. Especially as a negative value for a symptom can be either negative as non-present or missing (Missing At Random (MAR) type of missing data).
- In the future, a multi-class method to predict all diseases at a time can replace the one model – one disease.

DiagnoSmart Platform
Path to Unbiased Diagnosis

VI. Results of the first version of DiagnoSmart-AMI

Here is the performance of our first version of DiagnoSmart:

specificity	sensitivity	auprc	auc	accuracy	precision
0.82	0.77	0.83	0.85	0.80	0.81



When using 50% of women rate during training and testing, the performances for men and women subset are similar.

VII. Conclusion

This proof-of-concept shows that there is a great potential for machine learning based diagnosis in emergency department, especially for heart-attack diseases.

However, during the development we faced some challenges and discovered limitations to the development and the use of such CDS.

Thus, the next step in our development would be to proceed to randomized, multi-centric prospective studies in order to prospectively collect data in emergency departments but also outside (in family doctors, ambulance etc.) to reduce all discovered bias.

VIII. References

[1] https://mimic.mit.edu/docs/iv/modules/ed/

[2] https://www.aapc.com/codes/icd9-codes-range/57/

[3] https://icd.who.int/browse10/2019/en#/I21

[4] https://physionet.org/about/citi-course/

[5] Kimblad H, Marklund J, Riva G, Rawshani A, Lauridsen KG, Djärv T. Adult cardiac arrest in the emergency department - A Swedish cohort study. Resuscitation. 2022 Jun;175:105-112. doi: 10.1016/j.resuscitation.2022.03.015. Epub 2022 Mar 18. PMID: 35314209.

[6] Kareemi H, Vaillancourt C, Rosenberg H, Fournier K, Yadav K. Machine Learning Versus Usual Care for Diagnostic and Prognostic Prediction in the Emergency Department: A Systematic Review. Acad Emerg Med. 2021 Feb;28(2):184-196. doi: 10.1111/acem.14190. Epub 2021 Jan 2. PMID: 33277724.

[7] Ahsan MM, Siddique Z. Machine learning-based heart disease diagnosis: A systematic literature review. Artif Intell Med. 2022 Jun;128:102289. doi: 10.1016/j.artmed.2022.102289. Epub 2022 Mar 29. PMID: 35534143.

[8] Jianhua Wu, Chris P Gale, Marlous Hall, Tatendashe B Dondo, Elizabeth Metcalfe, Ged Oliver, Phil D Batin, Harry Hemingway, Adam Timmis, Robert M West, Editor's Choice - Impact of initial hospital diagnosis on mortality for acute myocardial infarction: A national cohort study, *European Heart Journal. Acute Cardiovascular Care*, Volume 7, Issue 2, 1 March 2018, Pages 139–148,

IX. Appendices

Appendix 1: description of MIMIC-IV-ED content

Patient Admissions: The ED data in MIMIC-IV contains information about patients who were admitted to the emergency department. This includes both demographic information (such as age, gender, and ethnicity) and clinical information (such as admission and discharge times, triage notes, and reason for admission).

Triage Information: The dataset may include information related to the triage process, such as the level of urgency assigned to each patient upon arrival, based on their presenting symptoms and clinical condition.

Clinical Assessments: It contains details of the initial assessment and examination conducted by healthcare providers in the emergency department. This may include vital signs (e.g., blood pressure, heart rate, respiratory rate), initial diagnoses, and any immediate interventions or treatments administered.

Diagnostic Tests: The ED data may include information about diagnostic tests ordered during the patient's visit, such as blood tests, imaging studies (e.g., X-rays, CT scans), and electrocardiograms (ECGs).

Procedures and Treatments: Details of any procedures performed or treatments administered in the emergency department, such as medication administration, wound care, or insertion of intravenous lines.

Discharge Information: Information about the disposition of patients after their visit to the emergency department, including whether they were admitted to the hospital, discharged home, or transferred to another healthcare facility.

Clinical Notes: Textual notes written by healthcare providers during the patient's visit, including admission notes, progress notes, and discharge summaries.

Appendix 2: Descriptive statistics

Table 1: Descriptive statistics on the length of stay in the emergency department for men and women in MIMIC-IV-ED dataset. All values are in hours.

Gender	Count	Mean	Standard deviation	Minimum[1]	Median	Maximum	Q1-Q3
Women	229,898	7.16	6.41	-22.73	5.51	341.04	3.61/8.30
Men	195,189	7.16	6.88	-22.43	5.38	493.07	3.43/8.35
Total	425,087	7.16	6.63	-22.73	5.47	493.07	3.53/8.31

1 The negative length of stays are present when the date of admission is bigger than the date of discharge. It corresponds to 4 patients (one left without been seen and 3 that where sent home). They seems to be due to a mistake in the manual entry of the date as the stays admission are close to or around midnight.

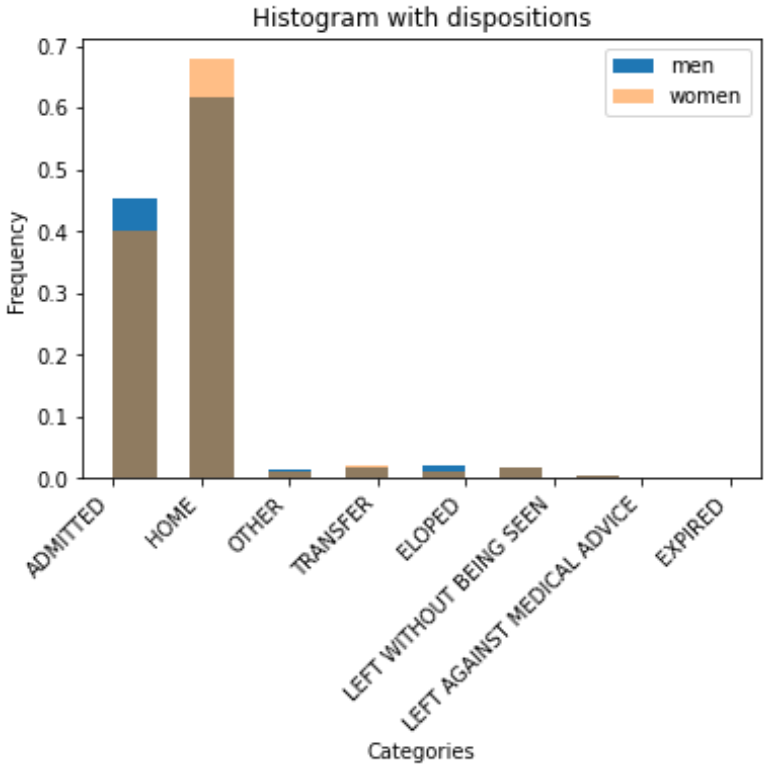


Figure 1: Distribution of the dispositions in MIMIC-IV-ED dataset for men and women.

Table 2: Frequency of the dispositions in MIMIC-IV-ED dataset for men and women.

Disposition	Count (frequency) for men	Count (frequency) for women
TOTAL	195,189 (100%)	229,898 (100%)
HOME	105,189 (53.89%)	136,443 (59.34%)
ADMITTED	77,426 (39.67%)	80,584 (35.05%)
ELOPED	3,454 (1.77%)	4,052 (1.76%)
TRANSFER	2,973 (1.52%)	3,397 (1.15%)
LEFT WITHOUT BEING SEEN	2,758 (1.41%)	2,256 (0.98%)
OTHER	2,183 (1.12%)	2,114 (0.92%)
LEFT AGAINST MEDICAL ADVICE	1,006 (0.51%)	875 (0.38%)
EXPIRED	200 (0.10%)	177 (0.07%)

Appendix 3: dictionaries used for the symptom's extraction

```
dict_side = {'right':['right ','r '],
             'left':['left ','l ']}
```

Figure 2: Dictionary of sides

```
dict_location = {'head':['head'],
                 'eyes':['eye'],
                 'nose':['nose'],
                 'neck':['neck'],
                 'jaw':['jaw'],
                 'ear':['ear'],
                 'back':['back'],
                 'stool':['stool'],
                 'joint':['joint'],
                 'mouth':['tooth','mouth','tongue'],
                 'axillary':['axillary'],
                 'throat':['throat'],
                 'face':['face', 'facial'],
                 'chest':['chest','thorax','c/p'],
                 'abdominal':['abdominal','abd '],
                 'arm':['arm','shoulder'],
                 'leg':['leg','tib','calf'],
                 'knee':['knee'],
                 'foot':['foot','ankle','pedal'],
                 'hand':['hand','wrist','finger','thumb'],
                 'hips':['hip'],
                 'urinary track':['urin', 'pelvis']
                 }
```

Figure 3: Dictionary of locations

```
dict_symptom = {'pain':['pain','c/p','tightness'],
                'jaundice':['jaundice'],
                'hyperglycemia':['hyperglycemia'],
                'dehydration':['dehydration'],
                'Hematemesis':['hematemesis','vomiting blood'],
                'distention':['distention'],
                'nausea':['nausea','vomit','n/v'],
                'swelling':['swelling'],
                'tachycardia':['palpitation','tachycardia'],
                'pain':['pain'],
                'bleed':['bleed','blood','hemorrhage','bld','booldy'],
                'fatigue':['lethargy','weak','fatigue'],
                'fever':['fever'],
                  'cough':['cough'],
                  'itch':['itch'],
                  'paralysis':['paralysis'],
                  'diarrhea':['diarrhea'],
                  'dizzy':['dizzy','syncop','n/v/d','dizziness'],
                  'hemorroids':['hemorroids','brbpr'],
                  'cramps':['cramp'],
                  'neurologic':['delirium','delusional','memory difficulty','confusion', 'hallucination','altered mental'],
                  'lump':['lump','mass'],
                  'numbness':['numbness','heaviness'],
                  'seizure':['seizure','uncontrolled movements'],
                  'migraine':['migraine','headach'],
                  'sore':['sore'],
                  'smelling urine':['smelling urine'],
                  'hearing loss':['hearing loss','decreased hearing'],
                  'rash_redness':['redness','rash'],
                  'hypoglycemia':['hypoglycemia'],
                  'dyspnea':['dyspnea','sob','respiratory distress','shortness of breath'],
                  'anemia':['anemia'],
                  'throat foreign body sensation':['throat foreign body sensation'],
                  'constipation':['constipat'],
                  'dysuria':['dysuria','urinary retention'],
                  'anxiety':['anxiety'],
                  'hematuria':['hematuria']
                 }
```

Figure 4: Dictionay of symptoms

```
dict_diagnosis = {'trauma':['injury','drug',"fall ","fall/","fx","trauma",'wound',
                            "ped struck", "traumatic", "injury",'laceration','strike',
                           "mvc", "od", "s/p arrest", "fracture","lesion",'assault','cat bite',
                              "run over","accident","burn",'torn','dog bite'],
                  'cardiac arrest':['cardiac arrest','infarction'],
                  'infection':['cellulitis','meningitis','infection','septic','septis','bronchitis','pneumonia', 'influenza','ili'],
                  'burn':['burn'],
                  'other':["appendicits","lyme","device","fascitis","allergic","pregnan","witnessed","chemo","post op",
                           "blockage","crohns","s/p",'transfer',
                           'abcess','biopsy','colostomy',"pneumothorax","suture",'disection','detox','tracheostomy',
                            "exposure",'equipment','catheter','medication','drug','overdose','etoh','substance','food']}
```

Figure 5: Dictionary of diagnosis in chief of disclosure

Appendix 4: Descriptive statistics on features

Table 3: Descriptive statistics on the input features

	count	mean	std	min	25%	50%	75%	max
temperature	297695	98.03	3.70	0.10	97.50	98.00	98.60	979.00
heartrate	302232	85.13	18.21	1.00	72.00	84.00	96.00	1228.00
resprate	299825	17.64	5.29	0.00	16.00	18.00	18.00	1797.90
o2sat	299572	98.57	19.27	0.00	98.00	99.00	100.00	9322.00
sbp	301317	135.46	50.82	1.00	120.00	133.00	149.00	19734.00
dbp	300718	81.85	1218.12	0.00	68.00	77.00	87.00	661672.00
jaundice	311654	0.00	0.05	0.00	0.00	0.00	0.00	1.00
hyperglycemia	311654	0.01	0.08	0.00	0.00	0.00	0.00	1.00
dehydration	311654	0.00	0.04	0.00	0.00	0.00	0.00	1.00
Hematemesis	311654	0.00	0.05	0.00	0.00	0.00	0.00	1.00
distention	311654	0.00	0.07	0.00	0.00	0.00	0.00	1.00
nausea	311654	0.07	0.26	0.00	0.00	0.00	0.00	1.00
swelling	311654	0.03	0.16	0.00	0.00	0.00	0.00	1.00
tachycardia	311654	0.02	0.14	0.00	0.00	0.00	0.00	1.00
bleed	311654	0.01	0.11	0.00	0.00	0.00	0.00	1.00
fatigue	311654	0.04	0.20	0.00	0.00	0.00	0.00	1.00
fever	311654	0.04	0.19	0.00	0.00	0.00	0.00	1.00
cough	311654	0.02	0.15	0.00	0.00	0.00	0.00	1.00
itch	311654	0.00	0.03	0.00	0.00	0.00	0.00	1.00
paralysis	311654	0.00	0.01	0.00	0.00	0.00	0.00	1.00
diarrhea	311654	0.01	0.11	0.00	0.00	0.00	0.00	1.00
dizzy	311654	0.06	0.25	0.00	0.00	0.00	0.00	1.00
hemorroids	311654	0.01	0.11	0.00	0.00	0.00	0.00	1.00
neurologic	311654	0.02	0.15	0.00	0.00	0.00	0.00	1.00
lump	311654	0.00	0.04	0.00	0.00	0.00	0.00	1.00
numbness	311654	0.01	0.11	0.00	0.00	0.00	0.00	1.00
seizure	311654	0.01	0.10	0.00	0.00	0.00	0.00	1.00
migraine	311654	0.04	0.19	0.00	0.00	0.00	0.00	1.00
sore	311654	0.01	0.11	0.00	0.00	0.00	0.00	1.00
smelling urine	311654	0.00	0.00	0.00	0.00	0.00	0.00	1.00
hearing loss	311654	0.00	0.01	0.00	0.00	0.00	0.00	1.00
rash_redness	311654	0.01	0.12	0.00	0.00	0.00	0.00	1.00
hypoglycemia	311654	0.00	0.06	0.00	0.00	0.00	0.00	1.00
dyspnea	311654	0.07	0.26	0.00	0.00	0.00	0.00	1.00
anemia	311654	0.00	0.07	0.00	0.00	0.00	0.00	1.00
constipation	311654	0.01	0.07	0.00	0.00	0.00	0.00	1.00
dysuria	311654	0.02	0.13	0.00	0.00	0.00	0.00	1.00
anxiety	311654	0.01	0.10	0.00	0.00	0.00	0.00	1.00
hematuria	311654	0.01	0.09	0.00	0.00	0.00	0.00	1.00
pain_back	311654	0.05	0.21	0.00	0.00	0.00	0.00	1.00
pain_neck	311654	0.01	0.11	0.00	0.00	0.00	0.00	1.00
pain_chest	311654	0.09	0.28	0.00	0.00	0.00	0.00	1.00
pain_joint	311654	0.00	0.02	0.00	0.00	0.00	0.00	1.00
pain_abdomin	311654	0.14	0.35	0.00	0.00	0.00	0.00	1.00
pain_head	311654	0.00	0.02	0.00	0.00	0.00	0.00	1.00
pain_urinary tr	311654	0.00	0.02	0.00	0.00	0.00	0.00	1.00
paralysis_face	311654	0.00	0.01	0.00	0.00	0.00	0.00	1.00
paralysis_arm	311654	0.00	0.00	0.00	0.00	0.00	0.00	1.00
cramps_abdo	311654	0.00	0.02	0.00	0.00	0.00	0.00	1.00
pain_arm_left	311654	0.01	0.09	0.00	0.00	0.00	0.00	1.00

DiagnoSmart Platform
Path to Unbiased Diagnosis

Appendix 5: Explanation of XGBoost

XGBoost, which stands for eXtreme Gradient Boosting, is an optimized and scalable implementation of gradient boosting machines, a powerful ensemble learning technique.

It combines several well-known methods:

Gradient Boosting

- Gradient boosting is a machine learning technique that builds a predictive model in the form of an ensemble of weak learners, typically decision trees.
- It works by sequentially adding weak learners to a model, with each new learner correcting the errors made by the existing ensemble.
- The final prediction is made by aggregating the predictions of all the weak learners.

Decision Trees

- Decision trees are simple models that make predictions based on a series of hierarchical decisions.
- Each decision tree is a weak learner, meaning it performs slightly better than random guessing but is not particularly strong on its own.

XGBoost improves upon traditional gradient boosting by introducing several optimizations and regularization techniques to make the model more accurate and less prone to overfitting. It uses a gradient descent algorithm to minimize a loss function while adding new trees to the ensemble. XGBoost incorporates both a linear model and a tree model, allowing it to capture linear and complex patterns in the data. Regularization techniques such as L1 and L2 regularization are used to control the complexity of the model and prevent overfitting. XGBoost also supports parallel processing and distributed computing, making it highly scalable and efficient for large datasets.

XGBoost has several advantages:

- Fast and efficient: XGBoost is designed for speed and can handle large datasets with millions of samples and features.
- Regularization: Built-in regularization techniques help prevent overfitting and improve generalization performance.
- Flexibility: XGBoost can be used for both regression and classification tasks and supports various loss functions.
- Feature importance: XGBoost provides feature importance scores, allowing users to understand which features are most influential in making predictions.

Overall, XGBoost is a versatile and powerful machine learning algorithm that is widely used in practice due to its effectiveness, efficiency, and scalability.