Patterns of Bias in the American Medical System

Research Question:

Can clustering techniques effectively identify patterns of bias within historical datasets containing features that are not specifically related to bias?

Introduction

The information load contained in large datasets contains more than the features themselves in many cases. Most modern data science techniques take advantage of this, particularly with unsupervised models. These datasets contain patterns that are informative of attitudes, behaviours, and context; the **ABC**s. It stands to reason that implicit or explicit bias may be detectable within these datasets. The difficulty in tackling these biases is that the people themselves may not be aware that they have a bias and/or may feel attacked if it were suggested. In the case where one accepts a bias, any questionnaire or self-reporting may be skewed because of that same bias or to diminish the perception of wrong doing (Adams et al. 1999; Gorber & Tremblay 2016). This project aims to provide evidence of this bias in pre-collected data that is untarnished by attempts to disguise or diminish the ABCs so that a collaborative and cooperative discussion based on evidence and logic, and not one of emotion, can begin.

To demonstrate and test the technique, the demonstration question '*Can patterns in American medical data reveal evidence of individual and systemic biases in the American medical system using clustering models?*' will be used. Medical professionals are expected to abide by the values of the Hippocratic oath. However, these professionals are still just people with their own sense of morality and worldview. To obfuscate the situation further, many of these professionals work in a society familiar with discrimination and ignorance and are part of a medical system that has been influenced by the same features for centuries (Feagin & Bennefield 2014). Some of these societal gradients include ABCs associated with race, intersectionality, and socioeconomic status (though the latter is commonly tied to racial bias). While tolerance is more common today, the systems themselves are no less difficult to change (Feagin & Bennefield 2014; Galvan & Payne 2024). If medical professionals, each with their own internal or external biases, are working in a medical system constructed in a setting that lacked tolerance and understanding, is it possible that not all patients are receiving the same level of care as a result of these persisting biases?

Clustering models are unsupervised models that group similar data points/ instances together, revealing clusters from patterns in the data (Ezugwu et al. 2021). Each cluster in this case may represent a different group of people, treatment facility, treatment type, etc. However, with analysis of these clusters, it may demonstrate that there are patterns indicative of bias in treatment towards certain patient groups. This research demonstration proposes that medical data may hold traces of bias in the American medical system and, while it cannot prove a bias, it may be able to provide compelling evidence of one in ways that self-reported audits cannot. In completing this research, it is hoped that there will be an effective and novel technique for evidencing bias and discrimination. While similar techniques have been used to assess fairness of machine learning or artificial intelligence outputs, it is yet to be used outside of a machine learning setting (Schäfer & Wiese 2022). The benefit of this technique is that it is low cost, using data that is already available. It confronts a sensitive topic without placing blame, perhaps making it simpler for policies to be updated and agreed upon. And it may uncover inequalities that, when corrected, provide an economic advantage (eg, by reducing the cost for patients and hospitals, and easing the burden on practitioners in the American medical system as patients are readmitted less frequently or require shorter stays).

Proposal Experiment Design

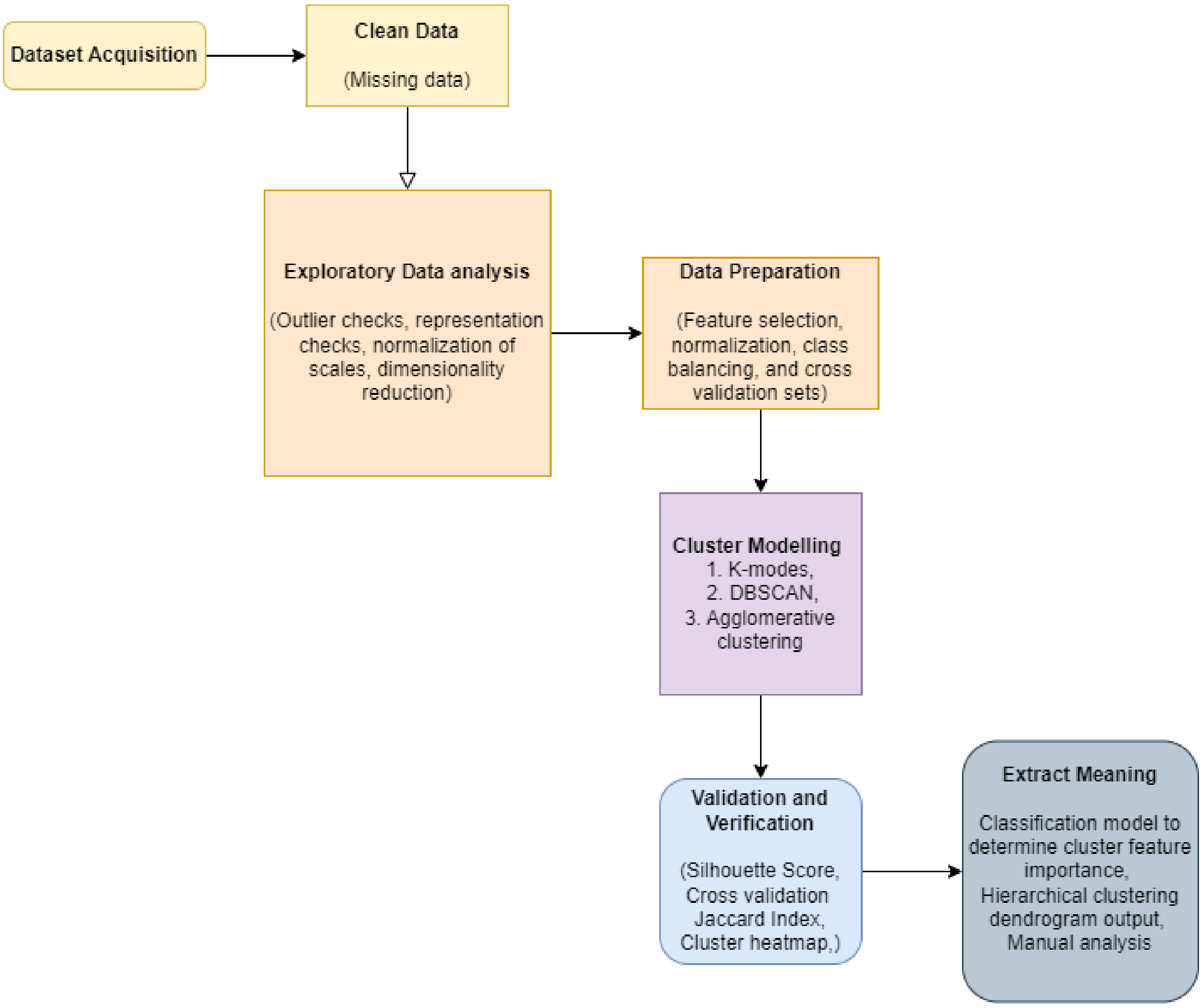
Dataset Acquisition
Clean Data
(Missing data)
Exploratory Data analysis
(Outlier checks, representation checks, normalization of scales, dimensionality reduction)
Data Preparation
(Feature selection, normalization, class balancing, and cross validation sets)
Cluster Modelling
1. K-modes,
2. DBSCAN,
3. Agglomerative clustering
Validation and Verification
(Silhouette Score, Cross validation Jaccard Index, Cluster heatmap,)
Extract Meaning
Classification model to determine cluster feature importance, Hierarchical clustering dendrogram output, Manual analysis

Figure 1: A possible experimental design in the form of a flow chart.

- Data source

The dataset is available from the *UC Irvine Machine Learning Repository* and was sourced from a paper titled '*Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records'* published in *BioMed Research International (Strack et al. 2014)*. It has 101,766 instances with 47 features, but the number of instances is misleading because it is inclusive of readmittance. The data has been deidentified, providing only a patient number.

Table 1: Some of the features that may hold patterns of bias. See Appendix for full data dictionary.

Feature	**Reason for inclusion**
Admission type	Patterns of admission are contextually important and may indicate attitudes.
Discharge disposition	The way in which a patient was discharged is indicative of how they were treated.
Admission source	More informative than admission type, providing exact context of admission.
Time in hospital	The number of days between admission and discharge. Indicative of patient care.
Medical speciality	Context of the doctor the patient was seen by. (May group these into *appropriate specialist* and *other specialist*)
Num lab procedures	The number of procedures indicates how difficult a patient may have been to diagnose but also how much attention they received.
Num procedures	The number of non-lab specific procedures performed. See explanation for num lab procedures. May also contain information of socioeconomic status.
Num_medications	The number of medications a patient received is also indicative of what they could afford. This is an inherent bias in the medical system.
Number inpatient	The number of times a patient visits a hospital in the year proceeding this visit may indicate a lack of appropriate treatment in previous visits.
Number_diagnoses	The number of diagnoses is relevant because it may indicate difficult cases or lack of attention.
Change	Indicates whether there was a change in the diabetic medication
DiabetesMed	Indicates if a person was treated with medication for their diabetes. This may provide socioeconomic information or indicate a lack of care.

Other informative features will be used in the analysis phase including race, gender, age, and weight. The model will not be trained with these because this would influence how it perceives the underlying patterns. The patterns in treatment are expected to generate clusters that are representative of specific groups, such as 'ethnic males aged X and weighing Y kg', etc. It would then be up to a clinician to determine if this difference in treatment is reasonable or the result of bias in the system.

Data source link: https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008

A data dictionary is available in the Appendix.

- Subject treatment

The dataset has been licenced with a Creative Commons Attribution 4.0 international. The researchers initially obtained the data from the *Health Facts database* which is a national data warehouse that collects clinical records In America. Presumably, patients were provided all the protections of the Nuremburg code and gave consent to their information being recorded and used by these researchers and then by the public more generally under this creative commons licence. Strack et al do make mention that they deidentified the data in accordance with the *'Health Insurance Portability and Accountability Act of 1996'* (2014).

- Subject selection

This dataset only includes instances of diabetic diagnoses during hospital admissions with lengths of stay no longer than 14 days that also involved laboratory tests and medication administration.

- Data cleaning and Preprocessing

Missing Data will be imputed were possible with mean values, otherwise these instances will be excluded. Numerical features will be normalized to ensure a standard and comparable scale. As clustering models generally use numerical data, one-hot encoding will be used where appropriate for categorical variables. Cross validation techniques will also be used to assess cluster stability, and so validation sets will be prepared. It may also be necessary to have each ethnic group represented equally and prevent a class imbalance from skewing the model.

- Analysis

Different clustering techniques will be used to identify clusters in the data. Three have been selected and each of these will observe the data in different ways (*Table 2*). If these three perspectives can all highlight similar grouping structures, then this would provide greater weight to the argument that there is bias present in the American medical system. Once clear clusters have been found, the common features of each will be determined. In this way, the research is protected from outside bias as the clusters up until this point have not been associated with any particular subgroup; clusters were protected through their anonymity. Then, using the clusters as a target variable, a classification model (eg, random forest) could be used to tease feature importances from each cluster. Dendrograms and manual analysis may also be used in order to clearly identify these subgroups and their treatment. With this information a careful analysis will be conducted to determine how the clusters differ and potential causes for that difference.

Table 2: The Clustering models selected and a justification for each.

Clustering Model	Justification
K-modes	- Can handle mixed data - Simple implementation - Useful model parameters - Easily interpreted
DBSCAN	- Handles noise and outliers well - Can handle mixed data - Not sensitive to class imbalance - Useful model parameters - No need to set the number of clusters - Easily interpreted
Hierarchical clustering	- Simple implementation - Captures small clusters - Provides information about clusters with a dendrogram - No need to pre-set the number of clusters

Distance metrics will need to be carefully considered as we have mixed data types. Some distance metrics that meet this criterion include Gower's distance, Daisy distance, or extended Jaccard Index. In terms of verification, more research is needed to determine what value of silhouette score or Jaccard index would indicate clear separation to a point that may indicate differing treatment. Various visualizations may also be used to demonstrate any findings.

Ethical considerations

Reidentification

Personal data including age, ethnicity, and gender in conjunction with hospital information could reveal subject identities. However, the information required to reidentify subjects are protected by American medical institutes and American data laws.

The Quality of Representation

This only includes people that have been granted admission and provided with medication. If the goal is to identify an inherent bias, some of the evidence for this bias may not be present in this dataset as it ignores all those that were turned away. However, patterns of bias in the data may still be visible even without representation from this extreme situation. The only hinderance this may pose is if a subgroup(s) has lost all faith in the medical system and no longer presents to hospital with the expectation that they will not be treated.

Another situation is that the subgroup is a low socioeconomic community, meaning they do not present to hospital because they cannot afford the care. Health care affordability is outside the scope of this research, but this systemic inequality is acknowledged as another way in which the American medical system can be discriminatory (Jindal et al. 2023).

Correlation and Not Causation

There is the possibility that particular ethnic subgroups have more difficult cases or are harder to diagnose because of genetic or cultural factors. The clustering model may create the impression that patterns in the data must equate to a bias, but it only provides evidence of bias. Further research would be needed to determine the cause and provide a solution to ensure all patients receive a similar quality of care.

Limitations

- There are not many clustering models that can accurately cluster mixed data.
- There are no bias specific features (eg, the time a patient waited, the duration between medication changes, or other features that indicate more behavioural or systemic failures). However, expecting a system to measure its own failure accurately, or using pre-existing measurements may be less informative (Adkins-Jackson et al. 2021). This technique may also be the ignition needed for further investigation.
- The data fails to provide category keys, providing only numerical category labels; presumably to protect anonymity. They key may exist and efforts are being made to identify these categories.
- There are many confounding factors (eg, language barriers, financial concerns, difficult patient behaviours, etc) that are not available and need to be addressed in future research. Further research will be needed to identify where the failures in the system arise.
- The assumption has been made that a person looks the way they have been categorised in the data, but this is not necessarily true. This occurrence may present something akin to a false positive or false negative, presenting noise to the data. One solution to counter this would be to collect more instances in future to dilute the problem.

References

Adams, A, Soumerai, S, Lomas, J & Ross-Degnan, D 1999, 'Evidence of self-report bias in assessing adherence to guidelines', *International Journal for Quality in Health Care*, vol. 11, no. 3, pp. 187-192.

Adkins-Jackson, PB, Chantarat, T, Bailey, ZD & Ponce, NA 2021, 'Measuring Structural Racism: A Guide for Epidemiologists and Other Health Researchers', *American Journal of Epidemiology*, vol. 191, no. 4, pp. 539-547.

Ezugwu, AE, Shukla, AK, Agbaje, MB, Oyelade, ON, José-García, A & Agushaka, JO 2021, 'Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature', *Neural Computing and Applications*, vol. 33, no. 11, 2021/06/01, pp. 6247-6306.

Feagin, J & Bennefield, Z 2014, 'Systemic racism and U.S. health care', *Social Science & Medicine*, vol. 103, 2014/02/01/, pp. 7-14.

Galvan, MJ & Payne, BK 2024, 'Implicit Bias as a Cognitive Manifestation of Systemic Racism', *Daedalus*, vol. 153, no. 1, pp. 106-122.

Gorber, SC & Tremblay, MS 2016, 'Self-Report and Direct Measures of Health: Bias and Implications', in RJ Shephard & C Tudor-Locke (eds), *The Objective Monitoring of Physical Activity: Contributions of Accelerometry to Epidemiology, Exercise Science and Rehabilitation*, Springer International Publishing, Cham, pp. 369-376.

Jindal, M, Chaiyachati, KH, Fung, V, Manson, SM & Mortensen, K 2023, 'Eliminating health care inequities through strengthening access to care', *Health Services Research*, vol. 58, no. S3, pp. 300-310.

Schäfer, J & Wiese, L 2022, 'Clustering-Based Subgroup Detection for Automated Fairness Analysis', in Springer International Publishing, Cham, pp. 45-55.

Strack, B, Deshazo, JP, Gennings, C, Olmo, JL, Ventura, S, Cios, KJ & Clore, JN 2014, 'Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records', *BioMed Research International*, vol. 2014, pp. 1-11.

Appendix

Table 3: A data dictionary with information about each feature available in the dataset.

Name	Type	Description	Feature values
encounter_id		Unique identifier of an encounter	
patient_nbr		Unique identifier of a patient	
race	Categorical	Values: Caucasian, Asian, African American, Hispanic, and other	'Caucasian', 'AfricanAmerican', nan, 'Other', 'Asian', 'Hispanic'
gender	Categorical	Values: male, female, and unknown/invalid	'Female', 'Male', 'Unknown/Invalid'
age	Categorical	Grouped in 10-year intervals	'[0-10)' '[10-20)' '[20-30)' '[30-40)' '[40-50)' '[50-60)' '[60-70)' '[70-80)' '[80-90)' '[90-100)'
weight	Categorical	Weight in pounds.	Nan '[75-100)' '[50-75)' '[0-25)' '[100-125)' '[25-50)' '[125-150)' '[175-200)' '[150-175)' '>200']
admission_type_id	Categorical	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available	[6 1 2 3 4 5 8 7]
discharge_disposition_id	Categorical	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available	[25 1 3 6 2 5 11 7 10 4 14 18 8 13 12 16 17 22 23 9 20 15 24 28 19 27]
admission_source_id	Categorical	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital	[1 7 2 4 5 6 20 3 17 8 9 14 10 22 11 25 13]

time_in_hospital	Integer	Integer number of days between admission and discharge	[1 3 2 4 5 13 12 9 7 10 6 11 8 14]
payer_code	Categorical	Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay	[nan 'MC' 'MD' 'HM' 'UN' 'BC' 'SP' 'CP' 'SI' 'DM' 'CM' 'CH' 'PO' 'WC' 'OT' 'OG' 'MP' 'FR']
medical_specialty	Categorical	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon	['Pediatrics-Endocrinology' nan 'InternalMedicine' 'Family/GeneralPractice' 'Cardiology' 'Surgery-General' 'Orthopedics' 'Gastroenterology' 'Surgery-Cardiovascular/Thoracic' 'Nephrology' 'Orthopedics-Reconstructive' 'Psychiatry' 'Emergency/Trauma' 'Pulmonology' 'Surgery-Neuro' 'Obsterics&Gynecology-GynecologicOnco' 'ObstetricsandGynecology' 'Pediatrics' 'Hematology/Oncology' 'Otolaryngology' 'Surgery-Colon&Rectal' 'Pediatrics-CriticalCare' 'Endocrinology' 'Urology' 'Psychiatry-Child/Adolescent' 'Pediatrics-Pulmonology' 'Neurology' 'Anesthesiology-Pediatric' 'Radiology' 'Pediatrics-Hematology-Oncology' 'Psychology' 'Podiatry' 'Gynecology' 'Oncology' 'Pediatrics-Neurology' 'Surgery-Plastic' 'Surgery-Thoracic' 'Surgery-PlasticwithinHeadandNeck' 'Ophthalmology' 'Surgery-Pediatric' 'Pediatrics-EmergencyMedicine' 'PhysicalMedicineandRehabilitation' 'InfectiousDiseases' 'Anesthesiology' 'Rheumatology' 'AllergyandImmunology' 'Surgery-Maxillofacial' 'Pediatrics-InfectiousDiseases' 'Pediatrics-AllergyandImmunology' 'Dentistry' 'Surgeon' 'Surgery-Vascular' 'Osteopath' 'Psychiatry-Addictive' 'Surgery-Cardiovascular'

			'PhysicianNotFound' 'Hematology' 'Proctology' 'Obstetrics' 'SurgicalSpecialty' 'Radiologist' 'Pathology' 'Dermatology' 'SportsMedicine' 'Speech' 'Hospitalist' 'OutreachServices' 'Cardiology-Pediatric' 'Perinatology' 'Neurophysiology' 'Endocrinology-Metabolism' 'DCPTEAM' 'Resident']
num_lab_procedures	Integer	Number of lab tests performed during the encounter	[41 59 11 44 51 31 70 73 68 33 47 62 60 55 49 75 45 29 35 42 66 36 19 64 25 53 52 87 27 37 46 28 48 72 10 2 65 67 40 54 58 57 43 32 83 34 39 69 38 56 22 96 78 61 88 50 1 18 82 9 63 24 71 77 81 76 90 93 3 103 13 80 85 16 15 12 30 23 17 21 79 26 5 95 97 84 14 74 105 86 98 20 6 94 8 102 100 7 89 91 92 4 101 99 114 113 111 129 107 108 106 104 109 120 132 121 126 118]
num_procedures	Integer	Number of procedures (other than lab tests) performed during the encounter	[0 5 1 6 2 3 4]
num_medications	Integer	Number of distinct generic names administered during the encounter	[1 18 13 16 8 21 12 28 17 11 15 31 2 23 19 7 20 14 10 22 9 27 25 4 32 6 30 26 24 33 5 39 3 29 61 40 46 41 36 34 35 50 43 42 37 51 38 45 54 52 49 62 55 47 44 53 48 57 59 56 60 63 58 70 67 64 69 65 68 66 81 79 75 72 74]
number_outpatient	Integer	Number of outpatient visits of the patient in the year preceding the encounter	[0 2 1 5 7 9 3 8 4 12 11 6 20 15 10 13 14 16 21 35 17 29 36 18 19 27 22 24 42 39 34 26 33 25 23 28 37 38 40]

number_emergency	Integer	Number of emergency visits of the patient in the year preceding the encounter	[0 1 2 4 3 9 5 7 6 8 22 25 10 13 42 16 11 28 15 14 18 12 21 20 19 46 76 37 64 63 54 24 29]
number_inpatient	Integer	Number of inpatient visits of the patient in the year preceding the encounter	[0 1 2 3 6 5 4 7 8 9 15 10 11 14 12 13 17 16 21 18 19]
diag_1	Categorical	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values	['250.83' '276' '648' '8' '197' '414' '428' '398' '434' '250.7' '157' '518' '999' '410' '682' '402' '737' '572' 'V57' '189' '786' '427' '996' '277' '584' '462' '473' '411' '174' '486' '998' '511' '432' '626' '295' '196' '250.6' '618' '182' '845' '423' '808' '250.4' '722' '403' '250.11' '784' '707' '440' '151' '715' '997' '198' '564' '812' '38' '590' '556' '578' '250.32' '433' 'V58' '569' '185' '536' '255' '250.13' '599' '558' '574' '491' '560' '244' '250.03' '577' '730' '188' '824' '250.8' '332' '562' '291' '296' '510' '401' '263' '438' '70' '250.02' '493' '642' '625' '571' '738' '593' '250.42' '807' '456' '446' '575' '250.41' '820' '515' '780' '250.22' '995' '235' '250.82' '721' '787' '162' '724' '282' '514' 'V55' '281' '250.33' '530' '466' '435' '250.12' 'V53' '789' '566' '822' '191' '557' '733' '455' '711' '482' '202' '280' '553' '225' '154' '441' '250.81' '349' nan '962' '592' '507' '386' '156' '200' '728' '348' '459' '426' '388' '607' '337' '82' '531' '596' '288' '656' '573' '492' '220' '516' '210' '922' '286' '885' '958' '661' '969' '250.93' '227' '112' '404' '823' '532' '416' '346' '535' '453' '250' '595' '211' '303' '250.01' '852' '218' '782' '540' '457' '285' '431' '340' '550' '54' '351' '601' '723' '555' '153' '443' '380' '204' '424' '241' '358' '694' '331' '345' '681' '447' '290' '158' '579' '436' '335' '309' '654' '805' '799' '292' '183' '78' '851' '458' '586' '311' '892' '305' '293' '415' '591' '794' '803' '79' '655' '429' '278' '658' '598' '729' '585' '444' '604' '727' '214' '552' '284' '680' '708' '41' '644' '481' '821' '413' '437' '968' '756' '632' '359' '275' '512' '781' '420' '368' '522' '294' '825' '135' '304' '320' '250.31' '669' '868'

	'496' '250.43' '826' '567' '3' '203' '53' '251' '565' '161' '495' '49' '250.1' '297' '663' '576' '355' '850' '287' '250.2' '611' '840' '350' '726' '537' '620' '180' '366' '783' '11' '751' '716' '250.3' '199' '464' '580' '836' '664' '283' '813' '966' '289' '965' '184' '480' '608' '333' '972' '212' '117' '788' '924' '959' '621' '238' '785' '714' '942' '250.23' '710' '47' '933' '508' '478' '844' '7' '736' '233' '42' '250.5' '397' '395' '201' '421' '253' '250.92' '600' '494' '977' '39' '659' '312' '614' '647' '652' '646' '274' '861' '425' '527' '451' '485' '217' '250.53' '442' '970' '193' '160' '322' '581' '475' '623' '374' '582' '568' '465' '801' '237' '376' '150' '461' '913' '226' '617' '987' '641' '298' '790' '336' '362' '228' '513' '383' '746' '353' '911' '506' '873' '155' '860' '534' '802' '141' 'V45' '396' '310' '341' '242' '719' '239' '533' '616' '519' '301' 'V66' '5' '989' '230' '385' '300' '853' '871' '570' '848' '463' '9' '934' '250.21' '236' '361' '594' '501' '810' '643' '430' '528' '205' '791' '983' '992' '490' '172' '171' '622' '306' '863' '864' '474' '660' '759' '356' '634' '967' '551' '695' '187' '732' '747' '323' '308' '370' '252' '152' '846' '164' '365' '718' '48' '266' '720' '94' '344' '797' '170' '878' '904' 'V56' '882' '843' '709' '973' '454' '686' '939' '487' '229' '991' '483' '357' '692' '796' '693' '935' '936' '800' '920' 'V26' '261' '307' '262' '250.9' '831' '145' '223' 'V71' '839' '685' 'V54' '35' '34' '179' '964' '136' '324' '389' '815' '334' '143' '526' '588' '192' 'V67' '394' '917' '88' '219' '325' '792' '717' '994' '990' '793' '207' '637' '195' '373' '847' '827' '31' '891' '814' 'V60' '703' '865' '352' '627' '378' '342' '886' '369' '745' '705' '816' '541' '986' '610' '633' '640' '753' '173' '835' '379' '445' '272' '382' '945' '619' '881' '250.52' '866' '405' '916' '215' '893' '75' '671' '928' '906' '897' '725' '867' '115' '890' '734' '521' '674' '470' '834' '146' '696' '524' '980' '691' '384' '142' '879' '250.51' '246' '208' '448' '955' '653' '149' '245' '735' '883' '854' '952' '838' '194' 'V43' '163' '216' '147' '354' '27' '477' '318' '880' '921' '377' '471' '683' '175' '602'

```
'250.91' '982' '706' '375' '417'
 '131' '347' '870' '148' '862' '61' '817'
'914' '360' '684' '314' 'V63'
 '36' '57' '240' '915' '971' '795' '988'
'452' '963' '327' '731' '842'
 'V25' '645' '665' '110' '944' '603'
'923' '412' '363' '957' '976' '698'
 '299' '700' '273' '974' '97' '529' '66'
'98' '605' '941' '52' '806' '84'
 '271' '837' '657' '895' '338' '523'
'542' '114' '543' '372' 'V70' 'E909'
 '583' 'V07' '422' '615' '279' '500'
'903' '919' '875' '381' '804' '704'
 '23' '58' '649' '832' '133' '975' '833'
'391' '690' '10' 'V51']
```

diag_2	Categorical	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values	[nan '250.01' '250' '250.43' '157' '411' '492' '427' '198' '403' '288' '998' '507' '174' '425' '456' '401' '715' '496' '428' '585' '250.02' '410' '999' '996' '135' '244' '41' '571' '276' '997' '599' '424' '491' '553' '707' '286' '440' '493' '242' '70' 'V45' '250.03' '357' '511' '196' '396' '197' '414' '250.52' '577' '535' '413' '285' '53' '780' '518' '150' '566' '250.6' '867' '486' 'V15' '8' '788' '340' '574' '581' '228' '530' '250.82' '786' '294' '567' '785' '512' '305' '729' '250.51' '280' '648' '560' '618' '444' '38' 'V10' '578' '277' '781' '250.42' '278' '426' '584' '462' '402' '153' '272' '733' '34' '881' '203' '250.41' '250.13' '293' '245' '250.12' '558' '787' '342' '573' '626' '303' '250.53' '458' '710' '415' 'V42' '284' '569' '759' '682' '112' '292' '435' '290' '250.93' '642' '536' '398' '319' '711' 'E878' '446' '255' 'V44' '250.7' '784' '300' '562' '162' '287' '447' '789' '790' '591' '200' '154' '304' '117' '847' '852' '250.83' '250.11' '816' '575' '416' '412' '441' '515' '372' '482' '382' 'V65' '572' '283' '78' '250.81' '576' '432' '595' '295' 'V12' '204' '466' '721' '434' '590' '271' '813' '368' '227' '783' '250.5' '258' '253' '309' '250.91' '519' '333' '459' '250.92' '250.4' '179' '420' '345' '433' '661' '537' '205' '722' '405' '437' '714' '211' 'E812' '263' '202' '397' '250.23' 'E932' '201' '301' '723' '614' '568' '861' 'V57' '724' '189' '297' '453' 'E888' '730' '354' '451' '738' 'E939' '805' 'V43' '155' '910' '218' '358' '220' 'E937' '583' '958' '794' '564' '436' '250.22' '620' '621' '331' '617' '596' '314' '378' '250.8' '625' '478' '731' '172' '404' '681' '470' '279' '281' '531' '443' '799' '337' '311' '719' 'E944' '423' 'E870' '465' 'E849' '782' '481' '480' 'V23' '199' '79' '438' '348' '42' 'E950' '473' '627' '726' '54' '490' '317' '332' '508' '369' '600' '349' '485' '208' '922' '431' '296' 'E934' '753' 'E935' '386' '728' '607' 'E915' '344' '716' '289' '191' '873' '850' '611' '377' '352' '616' 'V17' '136' '455' '933' 'E885' '860' '513' '603' '484' '223' 'V72' '291' '151' 'V58' '550' '510' '891' '185' '592' '791' '138' '598' '336' '362' '217' '825' '298' '821' 'E880' '343' '429' 'E879' '579' '225' '250.9' 'V49'

	'696' '233' '658' '969' '275' '250.1' '601' '704' '808' 'E890' 'V18' '920' '380' '570' 'E817' '359' '812' '274' 'V14' '324' '758' 'V66' '911' 'E931' 'E924' '593' '792' '727' 'V46' '394' '532' 'V64' '557' '864' '718' 'E942' '807' '604' '924' '820' '580' '273' '241' '282' '824' 'V61' '646' '701' '736' '565' '383' '250.2' 'E947' '452' '872' '905' 'E930' '921' '131' '448' '389' '421' '214' '705' '494' '752' '623' '9' '299' '959' '365' '967' 'E858' '40' '691' '909' '5' '814' '746' '250.31' '556' '680' '745' '351' '306' '110' '695' '552' '346' '918' '882' '947' '520' '188' '31' '356' '737' 'V08' '322' '182' '517' '974' 'E929' 'V53' '912' '252' '608' '516' 'E933' '94' '702' '923' '594' '647' '111' '934' '430' '487' '709' '796' '156' '977' '915' '756' '840' '341' '259' '693' '725' 'V62' '528' '683' '953' '457' '501' 'E900' 'V09' '522' '919' '461' '506' '193' '483' 'E936' '717' '802' '335' 'V54' '320' '945' '906' '239' '454' '826' '823' 'E941' '226' '795' '684' '844' '250.33' '308' '615' '588' '712' '663' '706' '833' '741' '713' '533' 'E884' '586' '555' '755' 'E928' '742' '869' '962' 'V11' '543' '373' '870' '913' '152' '810' '965' '907' '908' '995' '845' '474' '442' '751' '323' '472' '464' '686' '250.32' '540' '251' '811' '652' '659' '851' '422' '815' '307' '325' '463' '992' '692' '521' '917' 'E965' '524' '916' 'E813' '173' '238' '137' '514' '312' '837' '355' '980' '622' '475' '500' '754' '261' '801' '868' '968' '381' '11' '250.21' '694' '610' '734' 'E814' '310' '130' '246' '892' '846' '634' '75' 'E927' 'E905' '183' '379' 'E917' '163' 'E868' '495' '747' '989' 'E854' '240' '832' '605' '602' '644' 'V16' '35' 'V70' '376' '266' 'E918' '619' '477' '656' '46' '883' '171' 'V13' '698' '842' 'E850' '800' '269' '664' 'E887' '952' '164' 'E881' '527' '685' '366' '836' '27' 'V63' '865' '793' '232' '990' '52' '831' '327' '542' '806' '972' '862' 'E829' 'E919' '944' 'E916' '963' '316' '645' '347' 'V85' '374' 'V02' '748' '256' '186' '866' '975' '96' '395' '262' 'E819' '654' '994' '318' 'E826' '879' '674' '641' '822' '145' '797' '353' 'E938' 'E816' '948' '987' '99' '192' '250.3' 'E906' '534' '115' 'E818' 'E980' '360' '338' '529' '871' '750'

	'212' '302' '955' '141' '88' 'V25' '215' '350' 'V50' 'V03' 'E853' 'E968' 'E882' '140' '703' '991' '893' 'E821' '235' 'V69' '670' '195' 'V55' '388' '268' '894' '114' '260' '853' '7' '880' 'V86' '180' 'E945' '523' '863' '649' '270' '665' '460' '942' '364' '66' 'E883' '123' '884' 'V60' '843' '927']

diag_3	Categorical	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values	[nan '255' 'V27' '403' '250' 'V45' '38' '486' '996' '197' '250.6' '427' '627' '414' '416' '714' '428' '582' 'V43' '250.01' '263' '250.42' '276' '482' '401' '250.41' '585' '781' '278' '998' '568' '682' '618' '250.02' '305' '707' '496' '599' '715' '424' '518' '553' '794' '411' 'V42' '531' '511' '490' '562' '250.8' '250.7' '250.52' '784' '491' '581' '420' '8' '724' '730' '789' '131' '250.82' '999' '41' '493' '250.03' '753' '786' '529' 'E888' '425' '595' '303' '560' '711' '492' '332' '296' '438' '362' '250.4' '654' '244' 'V70' '737' '625' '681' '250.51' '404' 'V10' '810' '280' '440' '785' '588' '569' '272' '997' '250.43' '918' '584' '54' '788' '426' '722' '250.92' '196' '461' '535' '787' '891' '284' '458' '648' '780' '182' '285' '593' '413' '664' '564' '201' '356' 'V15' '292' '782' '473' '455' 'E932' '357' '348' '294' '250.23' '459' 'E878' '437' '733' '507' '525' '250.53' '397' '572' '805' '453' '331' '736' '402' '591' '576' '465' '533' '703' '349' '315' '658' '608' '578' '716' '382' '300' '282' '571' '536' '596' '287' '644' 'V11' '558' 'E885' '162' '198' '218' '412' '396' 'V14' '570' '433' 'E934' '882' '288' '577' '443' '729' '836' '295' '799' '281' '304' '153' '410' '616' '250.83' '601' '291' '75' '512' '660' '250.5' '598' '337' '574' '653' 'V58' '311' '415' '386' '602' '790' '112' '873' '620' '436' '70' '155' '138' '663' '530' '710' '42' '342' '250.91' 'E884' '515' '307' '704' '728' '731' '583' '238' '441' '293' '573' '532' '290' '594' '319' '250.13' '250.12' '519' '346' '380' '135' '642' '698' '924' '905' 'E933' '555' '309' 'E879' '286' '565' '752' '580' '446' '444' '344' '252' '35' '813' '394' '301' '575' '258' 'V17' '802' '435' '746' 'V12' '709' '881' 'E935' '139' '250.81' '718' '365' '202' '334' '185' '398' 'V44' '517' 'E849' '614' '466' '626' '250.9' '368' '605' '883' '289' '478' '617' '429' '442' 'V25' '866' '610' '557' '959' 'E942' '94' '920' '345' '313' '379' '79' '516' '586' '821' '600' '242' '373' '592' 'V64' '487' '253' '706' 'E947' '117' '340' 'E950' '656' 'E949' '590' 'V09' '250.22' '934' '694' '203' '250.93' '995' '726' '923' '958' '275' 'E929' '211' 'V18' 'V66' '199' '665' '53' '279' '522' '791'

	'890' '456' 'E938' 'E816' '122' '721' 'V65' '136' '480' '423' 'E920' '793' '647' '537' '351' '845' '336' '274' '719' '945' '434' '494' '227' '157' '208' '174' 'V57' '812' '734' '150' 'V23' '447' '692' '228' 'V16' '756' '405' 'E928' '823' '552' '528' '389' '240' '454' '792' '366' 'E939' '907' '270' '310' '266' '387' 'E931' '783' '245' '607' '355' 'E930' '705' '372' '369' '611' '283' 'V46' '110' '867' 'E956' '251' '250.2' '820' '712' '695' '567' '343' '723' 'V08' '273' '623' '807' '451' '495' '701' '34' 'V53' '314' '472' 'E945' '11' '189' '534' '354' '333' 'V54' '277' '659' '708' '452' '655' '816' '670' '621' '246' '953' '865' 'E817' '646' '151' '378' '78' '298' '840' '641' '521' '745' '619' '912' '506' 'E904' '259' 'E870' 'E980' '383' '204' '696' '566' '727' '47' 'E943' '358' '191' '965' '921' '432' '27' 'E861' '758' '477' '524' '751' '652' '556' '188' '825' '919' '732' '908' '951' '962' '685' 'E850' 'E944' '527' '341' '693' '250.1' 'V49' '860' '323' 'V55' '579' '508' '969' '205' '462' 'E880' '680' '697' '826' '200' '457' '717' '738' '742' '735' '235' '308' '725' '241' '824' '464' '260' '917' '239' '661' '892' '261' 'E883' '943' '744' 'E936' '796' '318' '967' '350' '854' 'E905' '9' '741' 'E941' '170' '643' '317' '759' '909' 'V22' '831' '713' '180' '801' '360' '359' '501' '335' '250.11' '306' '811' '690' 'V02' '271' '214' '847' '543' 'V63' '906' '842' '686' '445' '808' '861' 'E852' '220' 'E887' 'E858' '915' '970' '256' '747' '395' '243' '815' '481' '5' 'E927' '297' '299' '851' '864' '922' '384' 'E876' '225' '158' 'E937' '871' '88' '966' 'E917' 'E812' 'V62' 'E924' '604' '233' 'E916' '377' '797' 'V72' '172' '7' '421' '852' 'E819' '972' '916' '956' '3' 'E965' '173' '193' '154' '347' '862' '250.3' '987' '470' '262' 'E855' '161' '115' '179' '910' '312' '17' '460' '265' '66' '163' 'V60' '870' 'E906' '514' '944' '844' '417' '152' '183' '991' '216' '385' '164' '935' '510' '814' '485' '850' '250.21' 'E919' '872' '195' '431' '597' '933' '171' '884' '156' '868' '483' 'E815' '542' 'V61' '853' '374' 'E881' 'E882' 'E822' '192' '754' '327' '523' '500' 'V85' '992' '657' '684' '603' 'E826' '550' '913' '376' '755' '361' '186' '720'

			'250.31' '674' '911' 'E813' '226' '365.44' 'E818' '146' '955' 'E894' '475' 'V13' '880' '930' 'E915' '381' '132' '353' '795' '893' 'V01' 'E853' '863' '540' 'E828' '430' '800' 'E865' '148' 'E946' '822' '879' '848' 'V86' 'V03' '338' '989' '388' 'E966' '111' 'E922' '123' '757' 'E901' '141' '268' 'E892' '649' '702' '948' '223' '484' 'E886' '838' '928' '236' '624' '837' 'E987' 'V07' '841' '622' 'E912' 'E955' '463' 'V06' 'E864' '217' '877' '391' 'E825' '952' '669' '875' 'E900' '215' '538' '980' '834' '448' '175' '49' '876' '230' '57' 'E854' '942' '14' '750' '370' '671' '971']
number_diagnoses	Integer	Number of diagnoses entered to the system	[1 9 6 7 5 8 3 4 2 16 12 13 15 10 11 14]
max_glu_serum	Categorical	Indicates the range of the result or if the test was not taken. Values: >200, >300 , normal, and none if not measured	['None' '>300' 'Norm' '>200']

A1Cresult	Categorical	Indicates the range of the result or if the test was not taken. Values: >8 if the result was greater than 8%, >7 if the result was greater than 7% but less than 8%, normal if the result was less than 7%, and none if not measured.	['None' '>7' '>8' 'Norm']
metformin	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	['No' 'Steady' 'Up' 'Down']

repaglinide	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	['No' 'Up' 'Steady' 'Down']
nateglinide	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	['No' 'Steady' 'Down' 'Up']

chlorpropamide	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	['No' 'Steady' 'Down' 'Up']
glimepiride	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	['No' 'Steady' 'Down' 'Up']

acetohexamide	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	['No' 'Steady']
glipizide	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	['No' 'Steady' 'Up' 'Down']

glyburide	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	['No' 'Steady' 'Up' 'Down']
tolbutamide	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	['No' 'Steady']

pioglitazone	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	['No' 'Steady' 'Up' 'Down']
rosiglitazone	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	['No' 'Steady' 'Up' 'Down']

acarbose	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	['No' 'Steady' 'Up' 'Down']
miglitol	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	['No' 'Steady' 'Down' 'Up']

troglitazone	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	['No' 'Steady']
tolazamide	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	['No' 'Steady' 'Up']

examide	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	['No']
citoglipton	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	['No']

insulin	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	['No' 'Up' 'Steady' 'Down']
glyburide-metformin	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	['No' 'Steady' 'Down' 'Up']

glipizide-metformin	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	['No' 'Steady']
glimepiride-pioglitazone	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	['No' 'Steady']

metformin-rosiglitazone	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	['No' 'Steady']
metformin-pioglitazone	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed	['No' 'Steady']

change	Categorical	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: change and no change	['No' 'Ch']
diabetesMed	Categorical	Indicates if there was any diabetic medication prescribed. Values: yes and no	['No' 'Yes']
readmitted	Categorical	Days to inpatient readmission. Values: <30 if the patient was readmitted in less than 30 days, >30 if the patient was readmitted in more than 30 days, and No for no record of readmission.	['NO' '>30' '<30']