Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records, an analysis

Dennis Scheper

2020-09-28

Table of content

1 Results

1.1 Data acquisition

All information used in this article was obtained from a database, consisting of 41 tables and totals 117 features, such as demographics (like gender, race, and age), inpatient or outpatient, and (in-hospital) mortality. [1] Data came from 130 hospitals in the USA for over ten years (1998-2008) and contained around 74 million unique visits by 18 million unique patients. This research used information that needed to accede to the following specifications:

1. The record is a hospital admission;
2. The encounter is diagnosed with "diabetes," any kind will satisfy;
3. The length of admission was at least one day up to eighteen days;
4. Laboratory test results are available; and
5. Medications were administered.

Of all encounters, 101.000 records fulfilled all specifications and were taking into account for further analysis. We took all 55 existing attributes, filtered and adjusted them until the dataset was better structured, had less missing values, and contained only the satisfying information. The final dataset is constructed by a preliminary analysis and pre-processing of the data. All steps involved are described in the following subsections.

1.2 Tackling missing values

First, we take a look at how missing values are distributed in our dataset. This is depicted in figure 1, where the color black means a value is missing. The figure shows this per attribute, and we observe that some attributes have many missing values, and others have contained less. To get a better picture of these attributes and their missing values, we zoom in on these attributes in table 1.

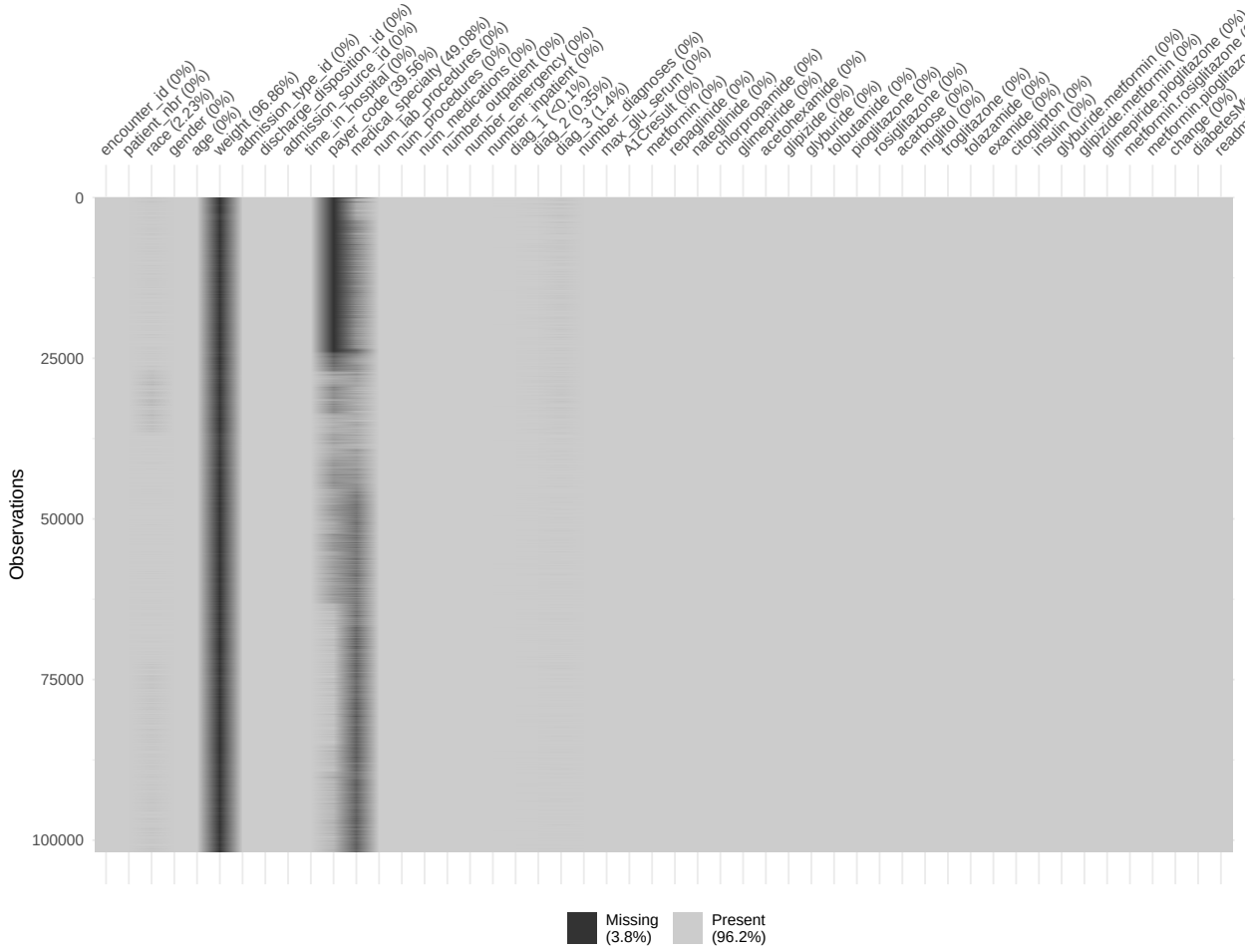


Figure 1: Missing values per attribute (depicted in black).

Attribute name	Missing values	Percentage missing (%)
race	2273	2.23%
weight	98569	96.86%
payer_code	40256	39.56%
medical_specialty	49949	49.08%
gender	3	0.00%
diag_3	1423	1.40%

Table 1: All attributes containing missing values, presented in the amount of missing values and its presence expressed in percentage

We observe, while looking at table 1, a multitude of different amounts of missing values. For example, the weight attribute is missing a staggering 97% of its valuation and gender, only missing three values. Having such a range of missing values, we decided to look at each attribute's contribution to our research goals and how much data is missing. If that contribution is minimal, we may decide to remove the attribute entirely. An example of such an attribute is payer_code, which is made up of payment; therefore, it is not really interesting to fill up the missing values here, and we removed it from the dataset. Besides removing payer_code, we proposed the same treatment for attribute weight; as mentioned earlier, it almost entirely contains missing values. The original authors stated that it is a result of government protocol. Knowing this, we can remove the attribute safely. We also remove medical_specialty for being too sparse.

For attributes with lesser amounts of missing values, we had the choice of doing two things: remove the instance with the missing values or revalue them on, for example, the most common valuation or label them as 'missing' instead of NA or as a question mark. For attribute race, we relabel the missing values to 'Missing/unknown' as removing has a potentially bigger impact on the outcome of the machine learning process. We handled diag_3 in the same manner. As for attribute gender, it holds three valuations: male, female, and missing or unknown, with only three instances falling into the latter category. For this reason, we decided to remove these records entirely.

1.3 Duplicates

The dataset contains multiple observations of inpatients visits. This may cause long-term issues as encounters, which is the unit of our analysis, are not statistically independent. Table 2 presents how many duplicates we have in our dataset, and its the percentage to the total records.

Total records	Total duplicates	Percentage of total
101.766	16.773	16.48%

Table 2: Number of duplicates and its percentage in the total records

The total number of duplicates is 16.773 or 16.48% of total records. To ensure a statistically independent dataset, we remove all of these instances. As a result, we are left with 84.993 unique records, each record corresponding to one patient.

1.4 Recoding attributes and introducing new ones

As our dataset contains many categorical attributes, it presents some potential for attributes to be incorporated into one another, thereby creating a more robust and more efficient data structure.

First, we need to introduce a variable that is essential to our research goals - and the authors also adapt that. The variable 'HbA1c test' represents what the original authors describe as a unique opportunity to assess hospital protocol's current efficacy surrounding diabetes testing. So it is vital for our research goals to implement this variable, and has the following four valuations: 1) no HbA1c test performed, 2) HbA1c performed and the result is in the normal range, 3) HbA1c performed and the result is higher than 8%, with no changes in diabetic medications, and 4) HbA1c test performed, with a result greater than 8%, and diabetic medication was changed.

Moving on to the three attributes that describe admission type (admission_type_id), disposition type (disposition_type_id), and admission source (admission_source_id), present us with the potential to move these attributes into a single one as we are interested in only the difference between ICU and non-ICU protocols. These attributes are made up of a digit reference to, for example, what the source of admission was. These are described in a separate ID codebook, shown in tables 4, 5, and 6. Concerning the number of references to an ICU or non-ICU related instance, the admission type has three values representing ICU related instances, the disposition type five, and the admission source also has five valuations regarding ICU; all of these are labeled as 'yes,' the rest of the values will be labeled as 'no.' All of this will be contained by a new attribute called 'icu_related' After this implementation, the three original attributes are up for removal. Additionally, we removed all instances of patients dying, which is depicted as yellow in table 5.

ID	Description
1	Emergency
2	Urgent
3	Elective
4	Newborn
5	Not Available
6	NULL
7	Trauma Center
8	Not Mapped

Table 3: Admission type ID and their description. Depicted in red are ICU related; no color means non-ICU related.

ID	Description
1	Discharged to home
2	Discharged/transferred to another short term hospital
3	Discharged/transferred to SNF
4	Discharged/transferred to ICF
5	Discharged/transferred to another type of inpatient care institution
6	Discharged/transferred to home with home health service
7	Left AMA
8	Discharged/transferred to home under care of Home IV provider
9	Admitted as an inpatient to this hospital
10	Neonate discharged to another hospital for neonatal aftercare
11	Expired
12	Still patient or expected to return for outpatient services
13	Hospice / home
14	Hospice / medical facility
15	Discharged/transferred within this institution to Medicare approved swing bed
16	Discharged/transferred/referred another institution for outpatient services
17	Discharged/transferred/referred to this institution for outpatient services
18	NULL
19	"Expired at home. Medicaid only, hospice."
20	"Expired in a medical facility. Medicaid only, hospice."
21	"Expired, place unknown. Medicaid only, hospice."
22	Discharged/transferred to another rehab fac including rehab units of a hospital.
23	Discharged/transferred to a long term care hospital.
24	Discharged/transferred to a nursing facility certified under Medicaid but not certified under Medicare.
25	Not Mapped
26	Unknown/Invalid
27	Discharged/transferred to a federal health care facility.
28	Discharged/transferred/referred to a psychiatric hospital of psychiatric distinct part unit of a hospital
29	Discharged/transferred to a Critical Access Hospital (CAH).
30	Discharged/transferred to another Type of Health Care Institution not Defined Elsewhere

Table 4: Disposition type ID and their description. Depicted in red are ICU related; no color means non-ICU related.

ID	Description
1	Physician Referral
2	Clinic Referral
3	HMO Referral
4	Transfer from a hospital
5	Transfer from a Skilled Nursing Facility (SNF)
6	Transfer from another health care facility
7	Emergency Room
8	Court/Law Enforcement
9	Not Available
10	Transfer from critical access hospital
11	Normal Delivery
12	Premature Delivery
13	Sick Baby
14	Extramural Birth
15	Not Available
17	NULL
18	Transfer From Another Home Health Agency
19	Readmission to Same Home Health Agency
20	Not Mapped
21	Unknown/Invalid
22	Transfer from hospital inpt/same fac reslt in a sep claim
23	Born inside this hospital
24	Born outside this hospital
25	Transfer from Ambulatory Surgery Center
26	Transfer from Hospice

Table 5: Admission source ID and their description. Depicted in red are ICU related; no color means non-ICU related.

Attributes depicting the primary (diag_1), secondary (diag_2), and final diagnosis (diag_3) consists of a three-digit referring to an ICD code. All the different codes can be categorized into a smaller range of valuations. Table 4 shows what categories are used. For example, ICD codes 390 to 459 are in circulatory diseases, according to [2]. This attribute now contains eight valuations instead of a much wider range. This gives the opportunity to analyze more -potential- correlations between, for example, time spent in hospital and disease categories.

ICD-9 codes	Categorical valuation
Code 240-279: endocrine, nutritional and metabolic diseases	Diabetes
Code 390-459: disease of the circulatory system	Circulatory
Code 460-519: disease of the respiratory system	Respiratory
Code 520-579: disease of the digestive system	Digestive
Code 580-629: disease of the genitourinary system	Genitourinary
Code 710-739: disease of the musculoskeletal system and connective tissue	Musculoskeletal
Code 800-999: Injuries	Injury
Other codes	Others

Table 6: ICD-9 codes and their new corresponding categorical valuation.

1.5 Normalization

Since our dataset does not contain many numeric attributes, it makes normalization up for debate. In figure 2, we observe all numeric data without normalization represented in histograms. The distribution is very diverse across all plots; for example, plot A ranges between 0 and 15, whereas B's distribution is much wider (between 0 and >100 for some outliers). We notice across almost all histograms that counts of a valuation increases in the beginning and decrease slowly over higher valuations. This creates, in some plots, many outliers. Examples of this are plots E, F, and G, where there seem to be many instances regarding lower values and almost no higher valuations (in all of the examples, this is after the value of ten). For the reasons mentioned, we introduce a log-2 scale to reduce the range of valuation between numeric attributes; the results are shown in figure 2.

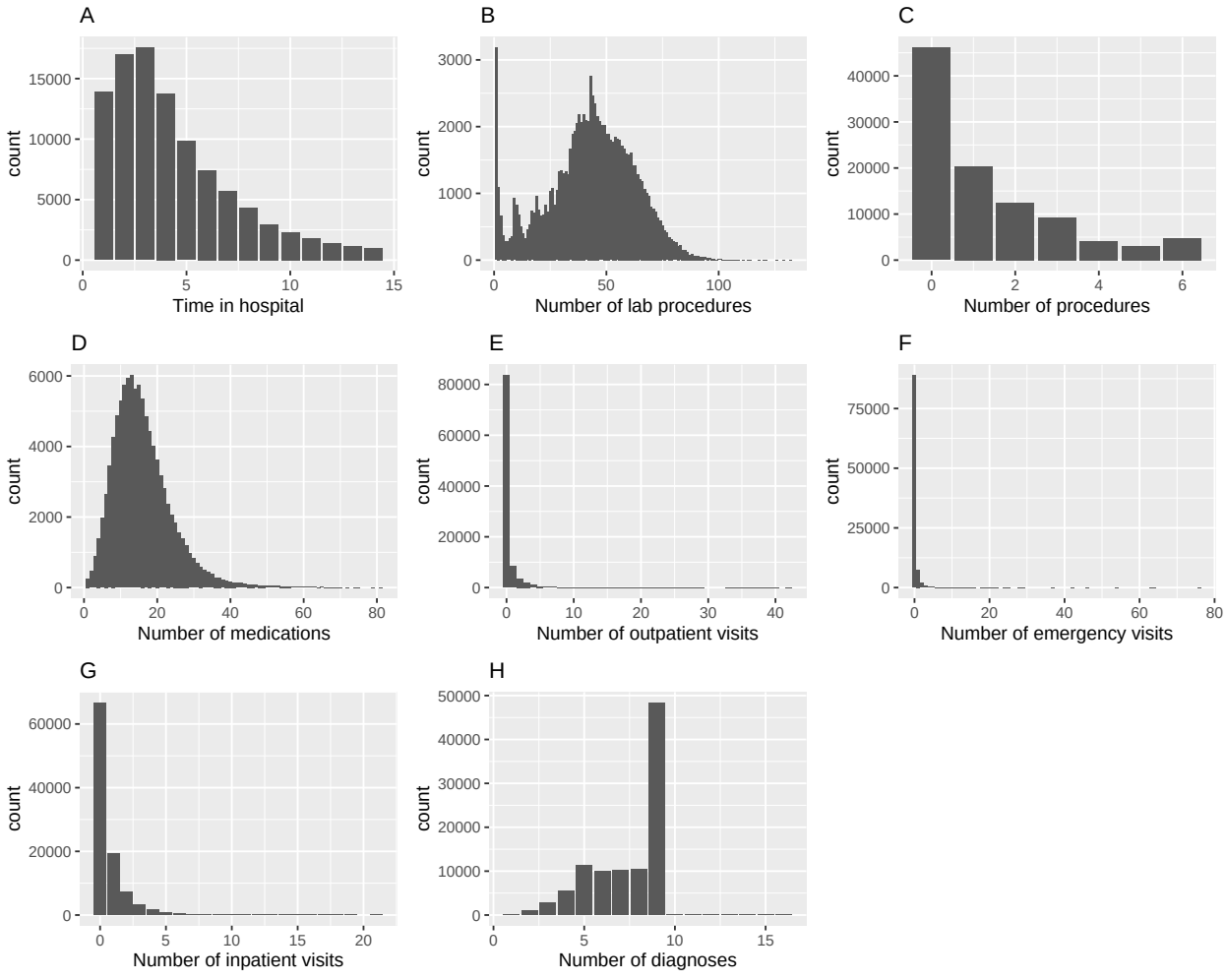


Figure 2: Numeric data presented in histograms without any normalization.

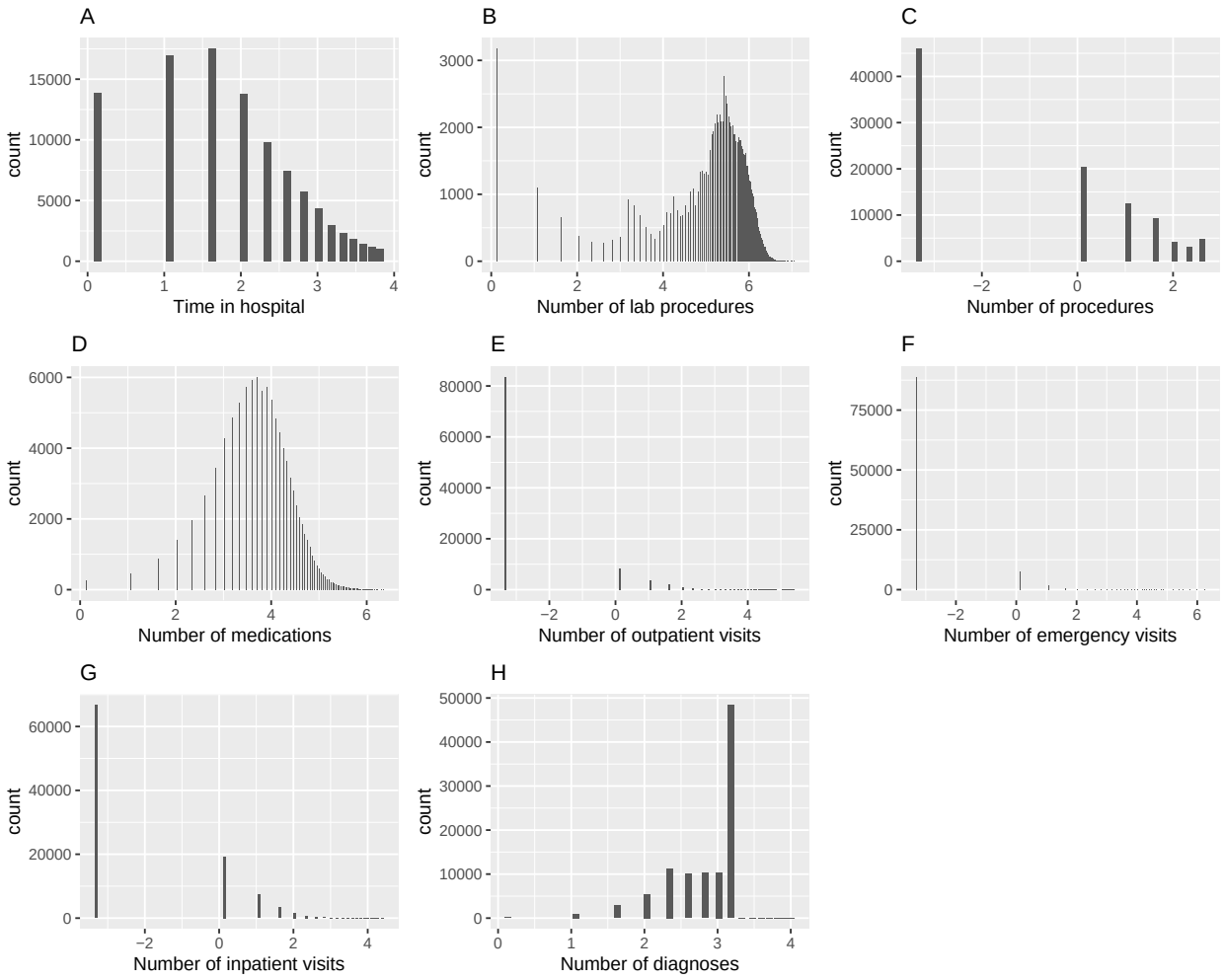


Figure 3: Numeric data presented in histograms on a log-2 scale.

Many instances have a valuation of zero, and with the introduction of a log-2 scale, we are about to transform these values into '-Inf' if nothing is done. To combat this, we added 0.1 to every value so that the valuation is still representable to its original valuation, and we do not have to remove all of these values. Looking at figure 3, we observe that the range of valuation between attributes is much smaller.

1.6 Correlation

Correlation between attributes is important to determine. Highly correlated attributes may cause problems with machine learning speed, and it can reduce the precision of estimated coefficients. It is, therefore, necessary to look at the correlation between attributes. We do this for every numeric attribute.

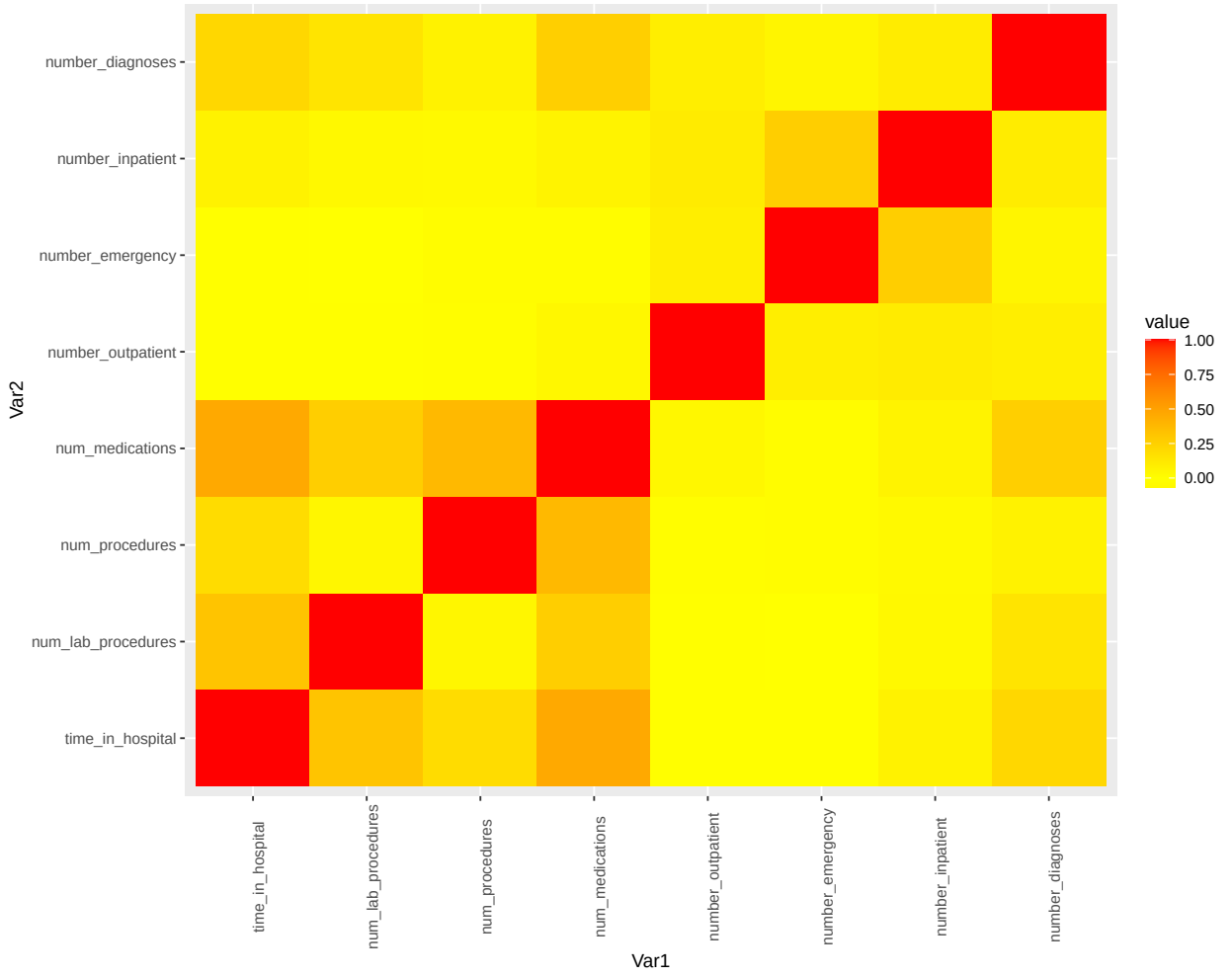
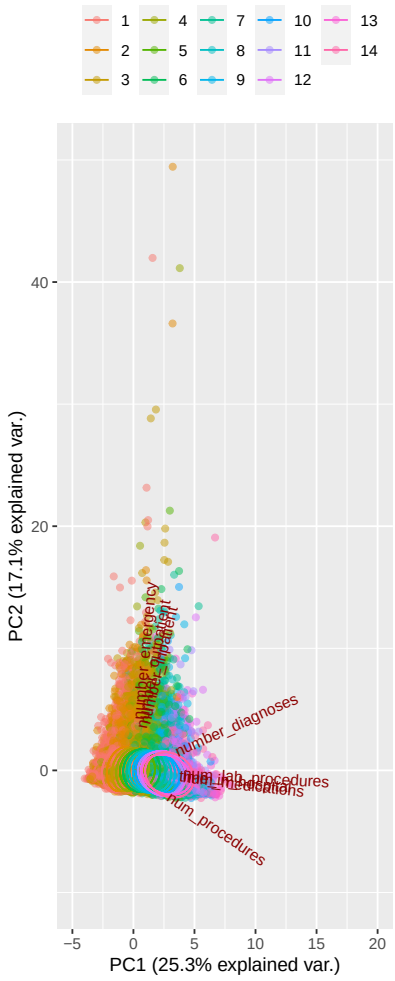


Figure 4: Heatmap of numeric attributes, with red depicted as a strong correlation, orange as a (small) correlation, and yellow as no correlation.

Our heatmap (shown in figure 4) shows a strong correlation as red, a small correlation as orange, and no correlation as yellow. The combination of attributes that stand out is, for example, num_medications-time_in_hospital and num_procedures-num_medications. As we can see, it is mostly the num_medication attribute that has some correlation. Others do not stand out that much. We see that most attributes do not correlate much with each other, at least for the numeric attributes.



\begin{figure}
\caption{ PCA plot of every numeric attribute, grouped around the attribute time_in_hospital.}
\end{figure}

While observing figure 5, which is a PCA plot with every numeric attribute and grouped around attribute time_in_hospital (the most important attribute surrounding our research goals), there is not much we can conclude about the correlation between attributes. The data seems all over the place, and clustering does not give a clear picture of correlation. We can see that attributes are mostly independent.

1.7 Final dataset

After removing whole attributes and some instances, our final dataset and recoding and revaluing others contain 70.443 records and 50 attributes. We parted ways with a total of 5 attributes and 31.323, which is a net loss of 30.8%.

2 Discussion

There has been a couple of recoding of attributes, for example, the introduction of the icu_related attribute and, as a consequence, the removal of three redundant attributes. Some other attributes still can be recoded, thereby creating a more efficient data structure. An example of such an occurrence is the attributes for primary, secondary, and final diagnosis; these attributes are now categorized according to their ICD code. As we are mostly interested in the diabetes category, constructing a new structure with the categories 'Diabetes-related' and 'Others' would create a binary valuation, which is very efficient for machine learning algorithms. The danger of removing so many instances is a great reduction in available information, which could be useful in further analysis down the road. Another example could be a new attribute regarding the number of total visits, which would retire the attributes for the number of inpatients, outpatient, and emergency visits. Removal of these attributes would create a more vague interpretation of visits of patients.

The attribute medical_specialty is one that was retired due to its large percentage of missing values. However, it has the potential to give more insight into where exactly diabetes protocols have been lackluster. Some may state that the attribute for admission source could be used for the same intentions. However, medical_specialty gives a lot more valuations and thereby was a great candidate for a more in-depth analysis to improve hospital protocol regarding diabetes testing. Its large part of missing values could have been relabeled to 'Missing' and its information been used. Since the original authors did not continue with the attribute, it was a strong belief to part ways with it.

The normalization process conducted had some minor implications; we used a log-2 scaling normalization technique were many '-Inf' values were introduced due to zeroes' presence. A temporary solution was to add 0.1 to every instance in numeric attributes. This choice may cause issues with that; not every log-transformed value is representable to its original valuation. Although the added value is relatively low compared to some instance values, it can cause a difference in machine learning outcomes.

After normalization, we still notice some outliers in numeric attributes. After the introduction of the log-2 scale, all valuations have an improved range in comparison with one another, but the amount outliers could still be of harm when coming to the machine learning process. Outliers can decrease the speed at which an algorithm is working and give poorer outcome results in a worst-case scenario. In contrast, having some outliers can have no serious implications. The question is if the normalization process we defined was effective enough to get rid of most outliers.

Looking at the dataset in a broader picture, we cleaned up many redundant instances and removed complete attributes or merged some to improve the dataset's structure. While the dataset still contains some outliers, most of them were removed in some way or another. Keeping this in mind, we can safely consider this -final- dataset as being ready for the machine learning process.

3 Conclusion and futher work

This article's goal was to clean up and better the structure of the dataset, which was retrieved from an outside source. We removed many instances and sometimes whole attributes, whereas others got merged and, therefore, helped create a better dataset structure. We can conclude that the cleanup and bettering of the structure is accomplished. The dataset is now ready for the machine learning process.

For new research or analysis accompanying this dataset, we like to address some additional changes to better the dataset structure. As discussed, some attributes can be merged or filled according to a different set of rules. For example, the attribute medical_specialty could be used in a different research setting - it contains much information but was ultimately removed due to its enormous amount of missing values. Other research initiatives could be more interested in this particular attribute.

4 References

References

[1] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014. DOI: https://doi.org/10.1155/2014/781670

[2] Centers for Disease Control and Prevention, National Center for Health Statistics, ICD-9, https://www.cdc.gov/nchs/icd/icd9.htm, November 6, 2015.