

# Temporal Illness Prediction using a Bayesian Model

Esben Pilgaard Møller, Thomas Kobber Panum,  
and Bjarke Hesthaven Søndergaard

*June 4th, 2014*



**AALBORG UNIVERSITY**  
STUDENT REPORT





**AALBORG UNIVERSITY**  
STUDENT REPORT

Aalborg University  
Department of Computer Science  
Selma Lagerlöfs Vej 300  
9220 Aalborg East  
<http://www.cs.aau.dk>

**Title:**

*Temporal Illness Prediction using a  
Bayesian Model*

**Theme:**

Master Thesis

**Project Term:**

P10, Spring 2014

**Project Group:**

SW104F14

**Students:**

Thomas Kobbler Panum  
Esben Pilgaard Møller  
Bjarke Hesthaven Søndergaard

**Supervisor:**

Nicolaj Søndberg-Jeppesen

**Copies:** 5

**Finished:** June 4th, 2014.

*This report and its content is freely available, but publication (with source) may only be made by agreement with the authors.*



## **Abstract**

The medical sector gathers and digitizes a lot of quantitative information on patients. This quantitative information is used to describe health properties of patients, such as analysis samples and diagnoses. Doctors are known to utilize information from analysis samples to diagnose illnesses of patients, but this relationship is not preserved when the information is digitized. This project aims to test the existence of a relationship between analysis samples and illnesses. To test this existence, a non-parametric bayesian model is constructed, which aims to predict the illness of a diagnosis based on analysis samples. This model uses Kernel Density Estimation to estimate normality spaces for medical properties of analysis samples given illnesses of diagnoses. These normality spaces contain densities based on temporal data of analysis samples and are used for estimating likelihoods of illnesses through analysis samples. The model is evaluated on different sets of illnesses and compared to naive prediction approaches. For each set of illnesses the model outperformed the naive approaches. Based on this, it is assumed that there exists a relationship between analysis samples and diagnoses.



## 1. Introduction

The Danish healthcare sector uses digital tools to store information about its procedures. This has led to large amounts of data being available, which creates possibilities for data mining. Data mining in this context could increase the quality of healthcare services through deeper understanding of the medical procedures conducted in healthcare facilities. This deeper understanding can improve decision making in medical facilities and thereby help increase the quality of service in the medical sector. This is due to the fact that when a doctor diagnoses a patient, it is done based on his interpretation of a set of observations that has been made with regards to the patient. The interpretation of these observations can be biased by the doctor's medical experience, or affected by the extent of the doctor's knowledge. Therefore by analyzing the data stored about medical procedures, objectifying the information in it, and making it available to doctors, diagnostic procedures conducted at medical facilities could be improved. This project examines a medical data warehouse (Boyesen and Nielsen, 2013) that contains information about medical procedures for diagnostic purposes (Panum and Møller, 2013), with the goal of creating decision support for doctors based on existing data.

Medical procedures can be viewed as actions or events, e.g. measuring a patient's blood pressure or diagnosing a patient (Stedman, 2000). There exist various types of medical procedures, e.g. analytical procedures and diagnostic procedures. Some medical procedures return a result, e.g. the measurement of a patient's blood pressure returns a numeric measurement indicating the blood pressure. Different types of medical procedures might yield different types of results, e.g. measuring blood pressure yields a numeric measurement, while diagnosing a patient identifies an illness. The medical data warehouse available contains information on two types of medical procedures, analytic- and diagnostic procedures, and it is these medical procedures which are examined throughout this project. A diagnostic procedure involves the act of determining whether a patient is healthy or ill based on observations and information about the patient and, if the patient is ill, determining which illness the patient is affected by. An analytical procedure is the act of gathering health related information about a patient.

Medical procedures can be interconnected, e.g. a procedure that involves measuring the blood pressure of a patient may lead to the diagnosis of high blood pressure for that patient. This interconnection stems from the result of the diagnostic procedure often being based on the results of a set of analytical procedures, i.e. a diagnosis is given based on a set observations. In short, a set of analytical procedures can be used to determine a health state for a patient through a diagnostic procedure.

The set of analytical procedures which may prove useful in a diagnostic procedure is dependent on the illness a given patient has. As an example, a patient may go to the emergency room with high fever and, through a set of analyses, a doctor may determine that the patient has the flu. In this case the analytical procedures that are relevant for the diagnostic procedure will most often lie within a short time frame of the patient going to the emergency room, as the flu in most cases is cured within a relatively short time frame (MUSC, 2011). Another example is a patient with a chronic disease. For such an illness, it might be necessary for the doctor to use the result of analytical procedures within a large time frame in order to give the correct diagnosis, since evidence which supports that the patient is affected by a chronic illness might be available in the result of older analytical procedures. The set of analytical procedures that are relevant for a given diagnostic procedure in the data warehouse is therefore dependent on the illness the patient was diagnosed with, as there is no explicit link between diagnostic and analytical procedures. It is important to note that, while the representation in the data warehouse suggests diagnosing based on performed analytical procedures, this is only possible due to the fact that illnesses can impact the measurements through analytical procedures (Strimbu and Tavel, 2010).

Determining the interconnection between diagnoses and analyses requires discovering patterns or association rules (Diamond and Forrester, 1979). The event of diagnosing a patient can be seen as the problem of finding patterns or association rules related to some illness, which makes it distinguishable from other illnesses. The relevance of an analysis sample in regards to some illness depends on how well the analysis sample conforms to an observed pattern of the medical property being measured in

regards to the illness. An observed pattern can be a rise in blood pressure for patients, shortly before being diagnosed with high blood pressure.

Constructing models for support in diagnostic procedures has been done before (Shwe et al., 1991) (Middleton et al., 1991). These describe the work of creating decision-theoretic version of Quick Medical Reference (*QMR*) called Quick Medical Reference Decision Theoretic (*QMR-DT*). *QMR* is a support tool for use in general medicine and can be used by medical practitioners, either as a knowledge base in which they can look up information or to assist them in diagnostic procedures. The intention with *QMR-DT* was to focus on the diagnostic aspects of *QMR* using the knowledge base *INTERNIST-1*. *INTERNIST-1* is a knowledge base covering over 600 diagnoses and 4000 findings where findings, i.e. analysis samples, and diagnoses are already linked. This means it is known beforehand which analysis samples are relevant for each diagnosis. In order to reduce the complexity of the model for *QMR-DT* a set of assumptions were made and covered in (Shwe et al., 1991). These involve e.g. diagnoses being marginally independent and findings being conditionally independent with regards to diagnoses.

The results covered in (Middleton et al., 1991) show that *QMR-DT* performed equally to *QMR* even with many simplifying assumptions used. *QMR-DT* is however a reformulation of the *INTERNIST-1* knowledge base, which took 20 person-years to create (Middleton et al., 1991). This means that, with a sufficient knowledge base, diagnostic support can be created using a probabilistic model.

The data warehouse available differs from the *INTERNIST-1* knowledge base in that there exists no explicit link between analytical- and diagnostic procedures. Based on this, it is assumed the data warehouse contains hidden information, which, if found and made available, can strengthen the decision making for doctors. This project aims to examine the relationship between results of analytical procedures and diagnostic procedures. If this relationship exists then it is assumed that it is possible to predict the result of a diagnostic procedure based on the results of a set of analytical procedures, which can be used in decision support for diagnosing patients.

This relationship will be examined based on 22 057 845 given diagnoses and 97 497 758 analysis samples distributed across 1 098 988 patients in the data warehouse. The data has been recorded over a time period of 48 years.

## **2. Abstract Problem Statement**

The assumptions in Section 1 state that analysis samples are used to determine an illness and that, as a result of this, there exists a relationship between analysis samples and illnesses. This lead to the following abstract problem statement:

*Given a medical data warehouse, containing patient information regarding analysis samples and diagnoses, is it possible to construct a model for predicting illnesses based on measured medical properties with higher accuracy than naive approaches?*

In this context, naive approaches refer to either predicting the illness using arbitrary guessing or predicting using a static guess which is the illness with the maximum likelihood in regards to the distribution of diagnosis frequencies for illnesses. This relationship is assumed to exist if it is possible to predict illnesses better than these naive approaches.

In order to concretize the problem further, it is necessary to gain a deeper understanding of the available data. This is covered in Section 3 and 4 through an analysis of how analysis samples and diagnoses are represented in the data warehouse, along with a statistical analysis of their distributions and their interconnection.

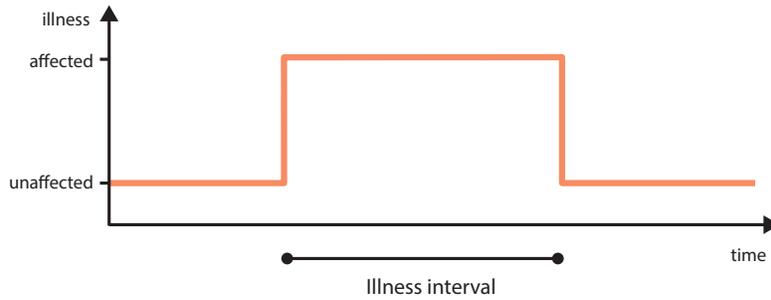
## **3. Data Representation**

The data in the data warehouse originates from the digitization of medical procedures. In order to utilize the data to prove that there exists a relationship between analytical and diagnostic procedures, it is required to understand how these medical procedures are digitized and what the available data represents in relation to real world patients. The result of each medical procedure is stored in the data warehouse. For diagnostic procedures this is the given diagnosis, and for analytical procedure an analysis sample.

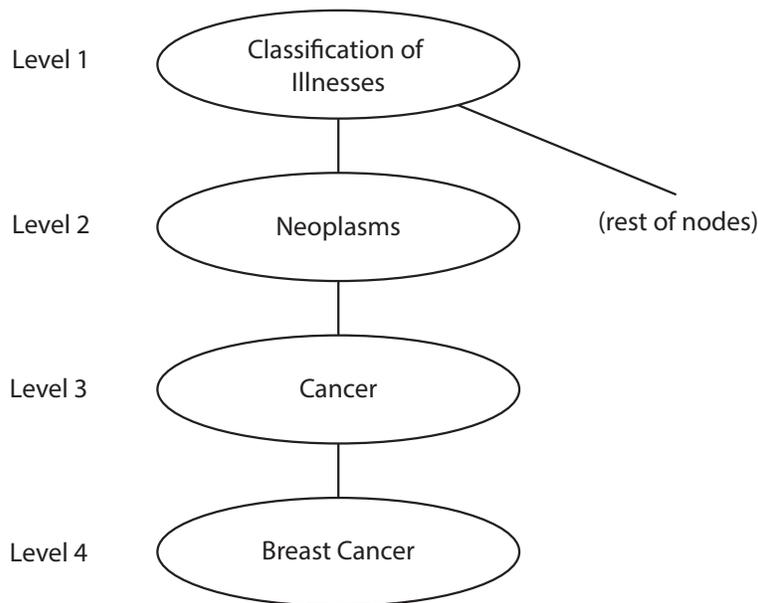
When a patient is given a diagnosis at a hospital it is a confirmation by a doctor that the patient is affected by an illness. In the medical sector, a diagnostic procedure is used to confirm or disprove that the patient is affected by the illness at the time of conduction. This is digitized by an entry for a diagnosis in the data warehouse, stating that a patient was affected by an illness at a certain time. Therefore the illnesses are represented as a binary state, meaning the patients either have the illnesses or they do not have the illnesses. In a medical sense, an illness for a patient is timeboxed, meaning patients are ill for a certain period of time, which could in some cases mean the patient's entire life. This representation, for a given patient, is illustrated in Figure 1. The time period in which a patient is ill will be referred to as the illness interval. This representation means that an entity of a diagnosis, in the data warehouse, does not represent the start nor the end of a given illness, but only the point in time the patient was given a diagnosis. This means that how long the patient has been ill when a diagnosis is given is unknown, and the severity of the illness is unknown. As an example, there is a difference between the severity of stage 1 and stage 4 lung cancer (Cancer Research UK, 2014), but this distinction is not modeled in the data warehouse.

Illnesses are organized in a tree structure, named SKS (Statens Serum Institut, n.d.), where each level represents a certain level of concreteness, with leaf nodes being concrete illnesses. The level of concreteness used is referred to as the granularity. An example is that in level 3, there is a category named *cancer*, while on level 4 there exist various sub types of cancer, e.g. *breast cancer*, *lung cancer* etc. An example of a branch in the diagnosis tree structure for breast cancer is illustrated in Figure 2. For this project, the leaf nodes in the SKS structure will be used to define the granularity of illnesses.

Analytical procedures measure a medical property of a patient, which can have been influenced by the health of the given patient. In this context, a medical property is an attribute which can be measured on a patient, such as blood sugar. The results of the measurements of these medical properties are usually represented as numeric values, e.g. blood sugar level. The results will be referred to as measurements. The measurements might be different at various points in time, indicating that the health state of



**Figure 1:** Example of a patient's health state over time. The time interval in which a patient is ill is referred to as the illness interval.

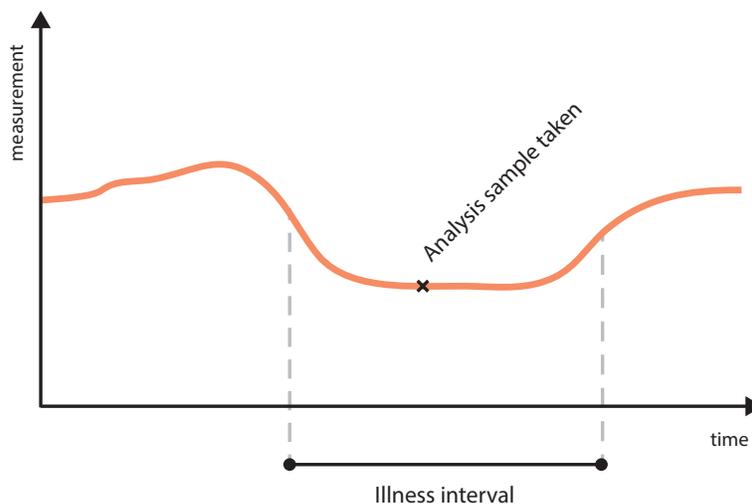


**Figure 2:** Branch in the SKS structure for breast cancer.

the patient has changed. A single analysis sample thereby only represents a medical property at a certain point in time, and not the given medical property's change over time during an illness. An example of this is illustrated in Figure 3.

The data warehouse only contains analysis samples for medical properties which are measured through quantitative analysis. Qualitative information gathered through

conversations with the patient, such as the patient complaining of a stomach ache or the doctor observing an inflamed area, is not recorded in the available data. This is discussed in Section 13.



**Figure 3:** Example of an analysis sample taken during a patient’s illness interval for a medical property.

### 3.1. Data Warehouse

The data for analysis samples and diagnoses are stored in the data warehouse in a tabular representation. They each have a related fact table, in which events of the given type are stored as entities.

Analysis samples and diagnoses are related to a patient. They each have an associated timestamp which represents the time the analysis sample was taken or the diagnosis given. There exists no explicit relationship in the data which defines the analysis samples that were used to give a specific diagnosis.

The attributes assumed to influence the health state of a patient in the entities; patients, analysis samples, and diagnoses are presented and discussed in the following paragraphs.

*Patient – age and gender.* The age of a patient might influence the health state of a patient, and thereby the medical properties measured through analysis samples. An example of this is the level of growth hormone, which is high for young patients, but decreases as the patient grows older (Bowen, 2006)(Cain et al., 2009). Furthermore the values measured in analysis samples can differ based on the gender of the patient, due to the biological structure of men and women being different (Holdcroft, 2007).

*Analysis sample – medical property, timestamp, measurement, and patient.* Analysis samples have a related medical property. This is referred to as the class attribute for analysis samples, which is the attribute used to group analysis samples. The timestamp states when the analysis sample was taken in time, and is recorded with a granularity in days. The measurement is a numeric value for the medical property. Each analysis sample is related to the patient, from which the analysis sample was taken.

*Diagnosis – illness, timestamp, and patient.* The illness attribute describes the condition of the patient being diagnosed, e.g. breast cancer. The illness of a diagnosis is referred to as the class attribute for diagnoses. The timestamp states when a diagnosis was given and it has a granularity level of days. Each diagnosis has a related patient to which the diagnosis was given.

A complete overview of attributes that the data warehouse contains for both analysis samples and diagnoses can be seen in Appendix B.

Some patients have had diagnoses given to them without having had analysis samples taken, and some patients have had analysis samples taken without being given any diagnoses. These patients and their related diagnoses and analysis samples are filtered out, since no relationships between diagnoses and analyses are present for these patients.

Diagnosis entities in the data warehouse do not only represent when a patient has had an illness, but also various kinds of contact a patient has had with the medical

sector. For example, a diagnosis given to a patient might have the illness *went to the emergency room*, since diagnosis entities cover all contact with the medical sector. Of the total amount of diagnosis entities, 78.6 % have been excluded due to not returning a concrete illness such as the flu. This leaves 4 726 003 given diagnoses and a total of 70 684 854 analysis samples. These diagnoses and analysis samples are gathered over a time period starting on the 8th of January 1965 and ending on the 1st of April 2013, thereby spanning a time period of 48 years.

With regards to the SKS structure, the amount of illnesses patients have been diagnosed with for each level of the SKS structure is illustrated in Figure 4. Each bar in Figure 4 represents the amount of distinct illnesses patients have been diagnosed with in the data warehouse, for that given level. As described, a part of the given diagnoses have been filtered out due to the diagnosed illness not referring to a concrete illness. This filtering has been done on granularity level 2 of the SKS structure, meaning diagnosis entities for illnesses in entire sub tree of the SKS structure have been removed. This leaves us with 31 105 distinct illnesses.

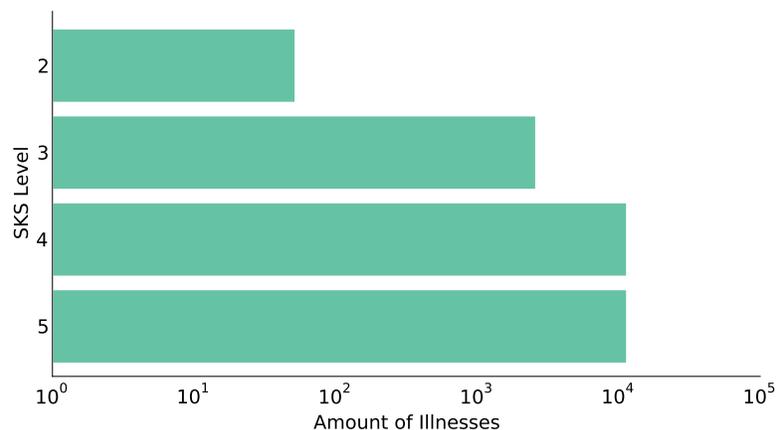


Figure 4: Amount of illnesses at each level in the SKS structure.

## 4. Data Analysis

In order to get a deeper understanding of the data, the data is examined through qualitative statistical analysis. The qualitative statistical analysis includes demographics of the data, analyzing class distributions, and gaining an understanding of how time affects the relationship between diagnoses and analyses. The demographic analysis consists of investigating the amount of analysis samples and diagnoses, and their attributes. Analysis samples and diagnoses have their respective class attributes, medical property and illness, examined in Section 4.1. The time attribute relationship between analysis samples and diagnoses is covered in Section 4.2.

As mentioned in Section 1, diagnoses and analysis samples are connected through a patient. There exists 405 133 patients, shared across diagnoses and analysis samples. The average amount of diagnoses per patient is 11.7 diagnoses, with a standard deviation of 13.8 diagnoses. On average a patient has had 174.5 analysis samples taken, with a standard deviation of 293.7 analysis samples.

### 4.1. Classes

This section examines the individual classes of analysis samples and diagnoses, through examining demographic information and qualitative analysis. The class of a diagnosis refers to the illness a patient has been diagnosed with, and the class of an analysis sample refers to the medical property measured for the analysis sample.

As covered in Section 3, there are 31 105 unique illnesses patients have been diagnosed with following the filtering of illnesses from the SKS model, and 1357 unique medical properties.

In order to examine the distribution of diagnosed illnesses across diagnoses, a subdistribution will be examined. Figure 5 is a histogram of this subdistribution of diagnoses grouped by diagnosed illnesses, containing only the 100 most frequently diagnosed illnesses in descending order. As seen in the figure, the distribution is not uniform and Pearson's chi-squared test for goodness of fit (McDonald, 2009) test also confirms this theorem, since it returns a p-value of  $0.1 \times 10^{-6}$  when the distribution is

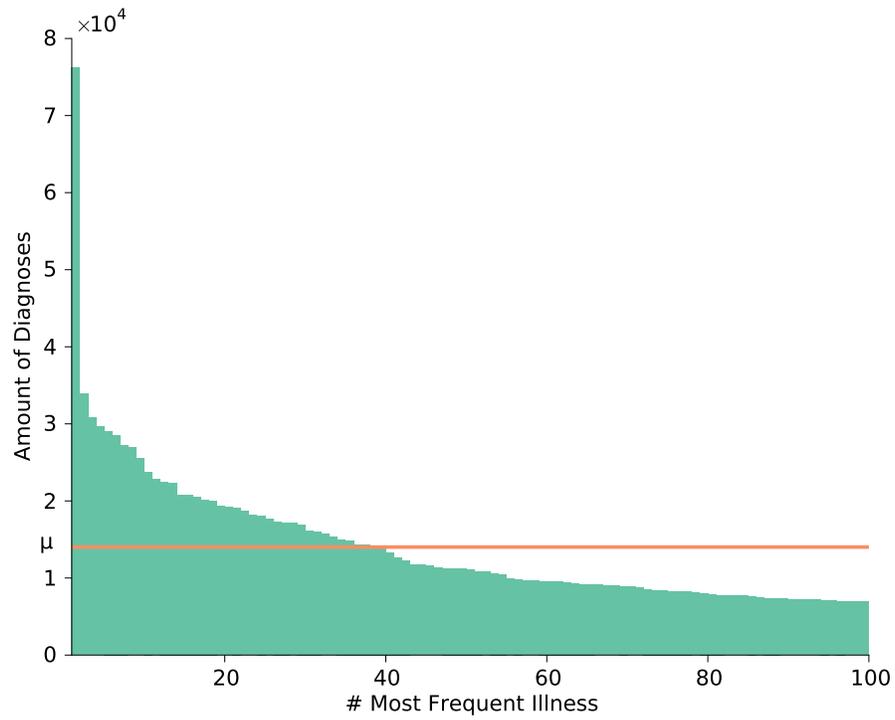


Figure 5: The amount of times the 100 most frequently diagnosed illnesses have been diagnosed.

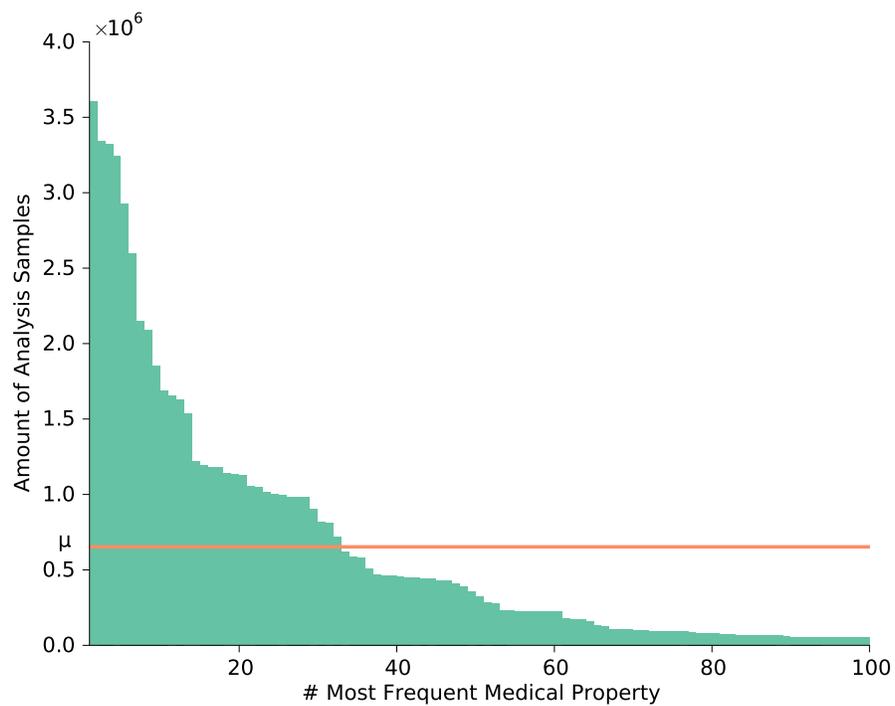
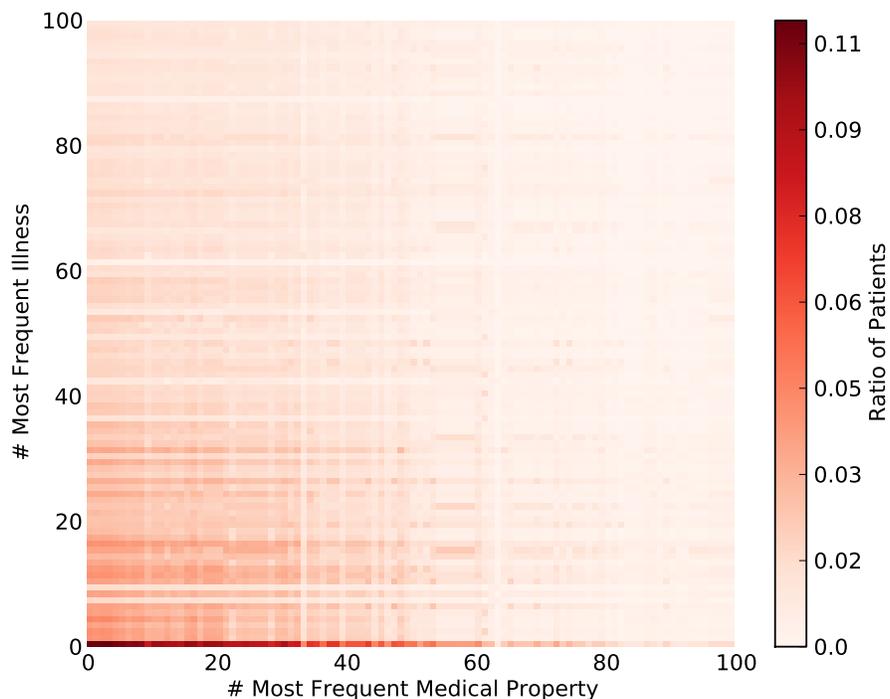


Figure 6: The amount of times analysis samples have been taken for the top 100 most frequently measured medical properties.

run against a uniform distribution. The p-value in this case presents a *goodness of fit*, with the value of 1 being perfectly uniform and being less than 0.05 is considered significantly non-uniform (Fisher, 1925). This means that the distribution of diagnoses across illnesses is not uniform, and thereby it is assumed the probability distribution across illnesses for a patient when he receives a diagnosis is not uniform.



**Figure 7:** Heat map illustrating the ratio of total patients being diagnosed with an illness of the 100 most frequently diagnosed illnesses at some point in time, while having had an analysis sample taken for one of the 100 most frequently measured medical properties.

The same analysis is conducted for how medical properties are distributed across analysis samples. Figure 6 is a histogram of a subdistribution containing the 100 largest analysis samples grouped by medical property in descending order. Of the given class distribution, Pearson's chi-squared test returns a p-value of  $0.1 \times 10^{-6}$ . Based on this p-value and Figure 6, it can be concluded that the distribution of medical properties measured across analysis samples is not uniform and thereby it is assumed that the

probability distribution across medical properties when a patient has an analysis sample taken is not uniform.

Figure 7 illustrates how these frequent illnesses and medical properties are related. It is a heat map of the illnesses and medical properties from Figure 5 and 6. The heatmap shows the ratio of all patients in the data warehouse that has been diagnosed with a given illness while having had a medical property measured, for each combination of illnesses and medical properties for the 100 most frequently diagnosed illnesses and the 100 most frequently measured medical properties.

## 4.2. Time

This section covers the analysis of how the two classes, illnesses and medical properties, are related based on their common attribute time through a patient. As described in Section 1 it is assumed that diagnoses are given based on a set of analyses. Through the representation of the data, there exists no explicit link showing which analysis samples were used as indicators for diagnoses. This relationship can thereby only be established based on their shared attribute time through patients.

Visualizing the relationship, for a given patient, can be done by constructing a timeline. This timeline contains the analysis samples and diagnoses of a given patient, mapped out in relative time to each other. An example of such a timeline is illustrated in Figure 8, with  $d$  representing diagnoses and  $a$  representing analysis samples. This timeline will be the foundation for the visualizations and the qualitative analysis.



Figure 8: Example of a patient timeline with analysis samples  $a$  and diagnoses  $d$ .

Prior to examining the relationship between diagnoses and analysis samples, their individual distribution of time values will be examined. Averaging the time interval be-

tween diagnoses of patients with multiple diagnoses yields a result of 255.4 days, with a standard deviation of 691.6 days. This is related to the mean frequency with which a patient is ill. The standard deviation being relatively large, suggests that the frequency with which a patient is ill depends largely on the given patient.

Based on the assumption stated in Section 1, that analysis samples are used to determine an illness, which results in a diagnosis, we construct the two following hypotheses: analysis samples are primarily taken prior to the time of diagnosis ( $H_1$ ), and analysis samples are primarily taken close to the time of diagnosis ( $H_2$ ).

In order to test the validity of  $H_1$ , we will examine the density of analysis samples of discrete time intervals prior and posterior to a diagnosis. Figure 9 illustrates an example for a discrete time interval of length  $r$  prior to diagnoses, ranging from time of diagnosis minus  $r$  to time of diagnosis, for a patient. These time intervals are examined by counting the amount of analysis samples within them. The mean amount of analysis samples within a time interval prior or posterior to a diagnosis is illustrated in Figure 10. The mean amount of analysis samples posterior to a diagnosis is consistently larger than prior to a diagnosis. This rejects  $H_1$ .

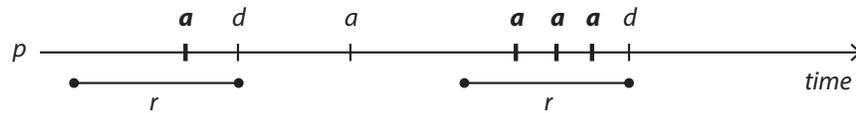
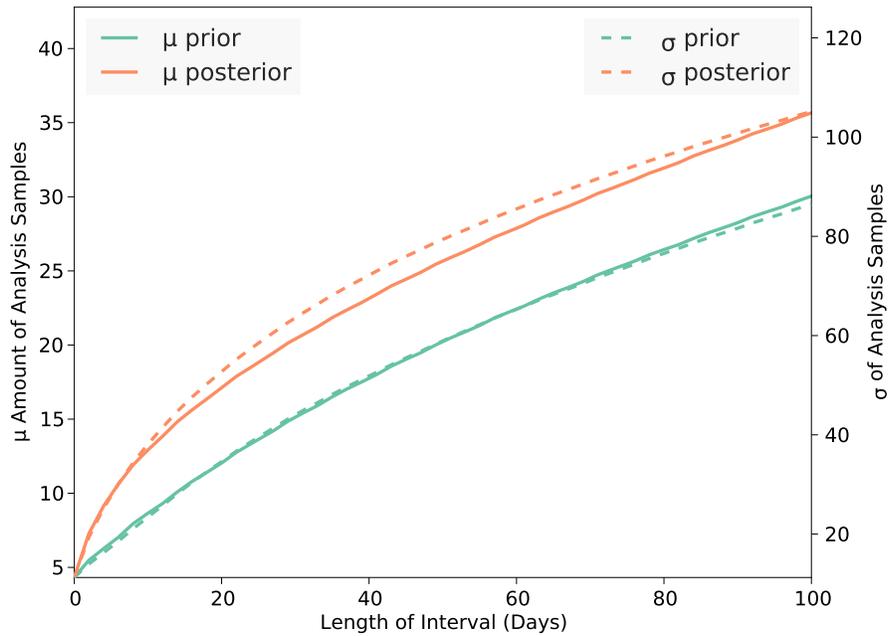


Figure 9: Example of a patient timeline, with a relative discrete time interval  $r$  prior to diagnoses  $d$ .

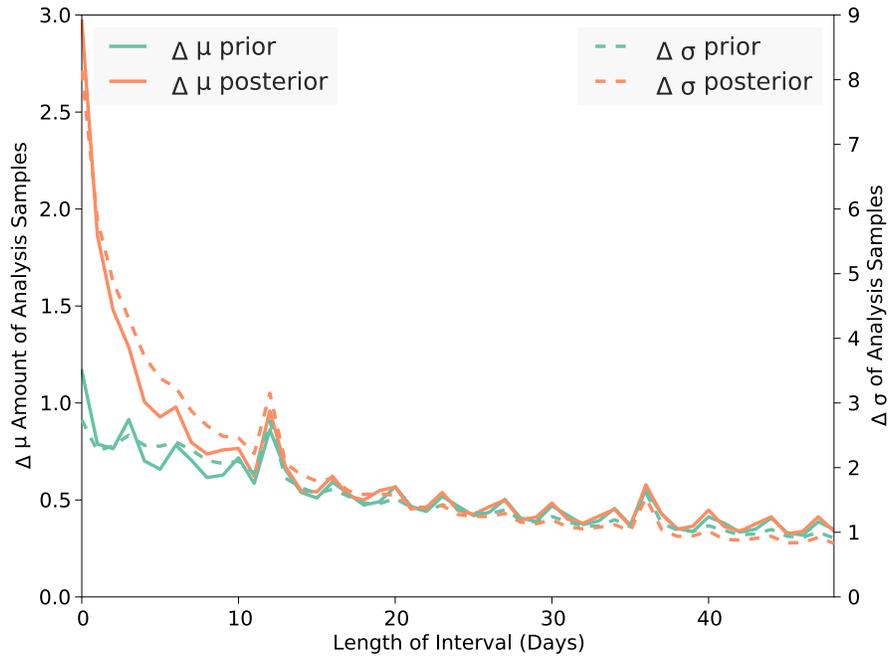
$H_2$  can be tested by examining the slope of Figure 10, due to the slope being correlated to the mean increase in size of one time interval to next time interval. If the analysis samples are uniformly distributed across time the slope should be constant across every time interval. The slope is illustrated in Figure 11, and it can be seen that it is not constant. The slope is larger at smaller time intervals, meaning that in general more analysis samples are taken close to the time of the diagnosis. This confirms  $H_2$ .



**Figure 10:**  $\mu$  and  $\sigma$  of the amount of analysis samples taken prior and posterior to diagnoses for a range of relative discrete time intervals.

Knowing that  $H_1$  is rejected, leads to the question of why the mean amount of analysis samples taken is larger after a diagnosis. One hypothesis is that posterior analysis samples are used for monitoring the health state of a patient. Diagnosing a patient involves determining the illness of a patient, meaning analysis samples are investigated as indicators for a given illness (Strimbu and Tavel, 2010). In the case of monitoring, the illness has been classified, and thereby it is known which medical properties are affected during the state of illness. Monitoring therefore involves periodically taking analysis samples of the relevant medical properties, in order to determine when it returns to values that are considered healthy, and thereby the patient is no longer affected by the illness.

This hypothesis can be formalized as the following: the mean amount of different medical properties in relation to the mean amount of analysis samples of a time interval should be smaller for analysis samples taken posterior to a diagnosis than prior



**Figure 11:** Slope for  $\mu$  and  $\sigma$  of the amount of analysis samples taken prior and posterior to diagnoses for a range of relative discrete time intervals.

( $H_3$ ). Figure 12 illustrates the mean amount of different medical properties of analysis samples in a time interval. It can be seen that the mean amount of different medical properties posterior and prior are very similar. When considering this fact, and with regards to the mean amount of analysis samples posterior to a diagnosis being higher than prior,  $H_3$  can be confirmed.

The confirmation of  $H_2$  means analysis samples are not uniformly distributed across time, and are thereby frequently taken within a time interval prior to a diagnosis being given. Based on this confirmation it can be assumed that diagnoses are often given based on a set of recently taken analysis samples. Knowing this relationship exists, it can be examined if there exists a relationship between analysis samples for specific medical properties and illnesses based only on the time attribute. An example of such a relationship is that knowing when an analysis sample was taken for blood sugar gives evidence about when the patient is diagnosed with diabetes. In order for this relationship to exist there must be a correlation between when analysis samples are taken and

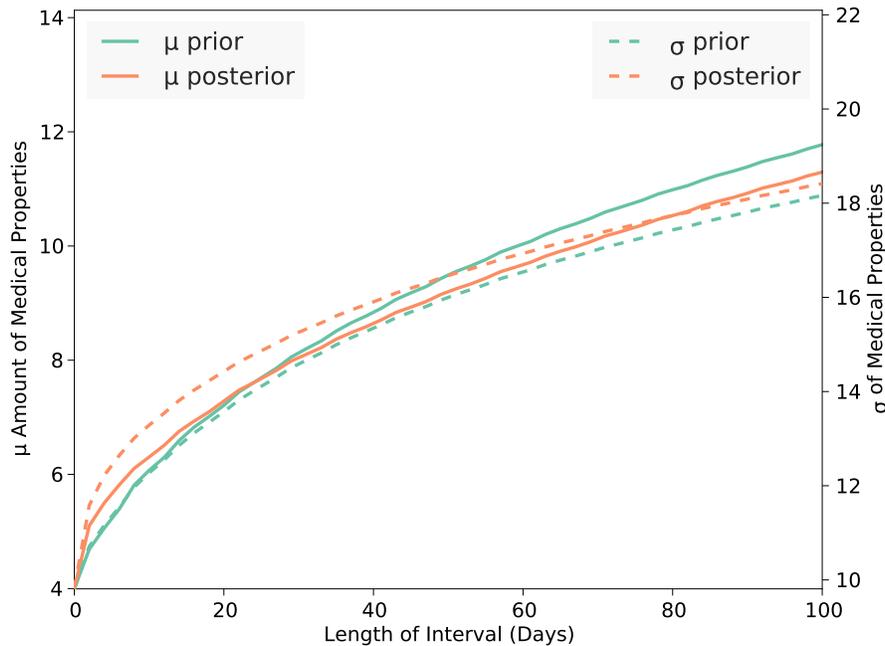


Figure 12:  $\mu$  and  $\sigma$  of the amount of medical properties a patient has had taken prior and posterior to a diagnosis for a range of relative discrete time intervals.

diagnoses are given to patients. If the relationship does not exist, the analysis samples will be relatively uniformly distributed on the time dimension prior to a diagnosis being given. Due to the uncertainty of when, during an illness, a patient is diagnosed, this relationship is assumed not to exist. The hypothesis  $H_4$  is therefore defined as follows: given an analysis sample and a diagnosis, there exists no strong correlation across only the time attribute.

$H_4$  is investigated by examining if there exist trends for when a specific medical property was measured given that the patient is affected by a specific illness. The correlation will be examined for the top 100 most frequently diagnosed illnesses and measured medical property. The top 100 is selected as a subset, since the most data will be available for the most frequent illnesses and measured medical properties. For each combination of illnesses and medical properties within the top 100 for each class, an entropy measure is used to analyze if there is a trend based on time. Entropy is the measure of uncertainty in a random variable (Russell and Norvig, 2009). The entropy

measure which will be used is Shannon Entropy, which will be referred to as entropy. A simple example is a coin flip which has a 50 % chance to take either the value heads or tails. The formal definition for calculating the entropy is as follows:

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log_2(p(x)),$$

where  $X$  is a discrete random variable which has the finite set of states  $\mathcal{X}$  (Lesne, 2011). Entropy can be normalized to a value between 0 and 1 using the following method:

$$H_n(X) = \frac{- \sum_{x \in \mathcal{X}} p(x) \log_2(p(x))}{\log_2(|\mathcal{X}|)}$$

described in (Masisi et al., 2008). The chance that a random variable has to take on a value can be expressed via a probability distribution.

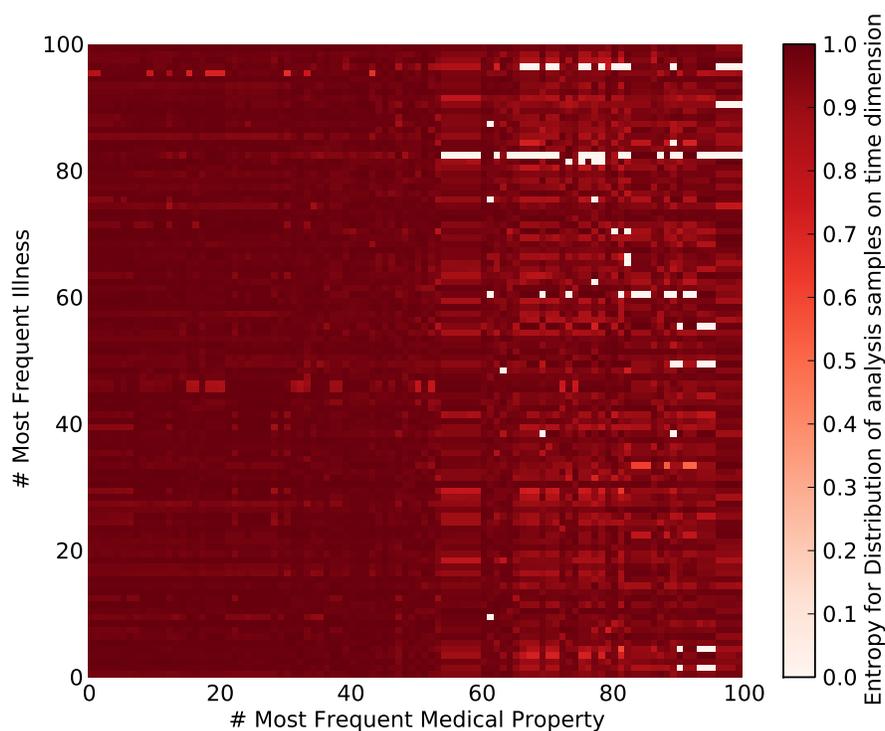
The entropy measure is calculated based on data extracted in the following manner. For a given combination of an illness and a medical property, all analysis samples of the given medical property taken within 30 days prior to a patient being diagnosed with the given illness is extracted. The choice of 30 days is based on  $H_2$ , which confirmed that analysis samples are primarily taken close to a diagnosis being given. Based on this, it can be assumed that prior analysis samples are primarily taken as part of the process of giving a diagnosis to a patient. The time interval prior to a diagnosis, from which analysis samples were used to give the diagnosis, is unknown. It can however be assumed that the shorter the time difference is between when an analysis sample was taken and a diagnosis given, the more likely it is that the analysis sample was used to diagnose the patient. Thereby the time difference between a diagnosis and a prior analysis sample can be seen as a relevance measure for that given analysis sample given the diagnosis. The time interval, from which to include analysis samples when diagnosing a patient, might vary based on the illness the patient is affected by. As an example, it might be possible to detect cancer for a patient in an analysis sample taken six months prior to the patient being diagnosed with cancer, but it is not possible to detect if a patient has a cold, based on a six months old analysis sample. However, in order to simplify the analysis, a general time interval of 30 days prior to a diagnosis, is used across all illnesses.

The extracted analysis samples are divided into bins based on which time interval prior to their related diagnosis they lie in. This means all analysis samples taken within 1 – 5 days prior to a diagnosis being given are placed in one bin, while all analysis samples taken within 6 – 10 days prior to a diagnosis are placed in another bin and so on up until the 30 day range being examined. The result is a set of 6 bins, one for each range of length 5 in the 30 day range. Based on the number of analysis samples in each bin, a probability distribution is created for which the entropy is calculated. A large entropy means there is high uncertainty, and would indicate that there are no clear trends in the distribution of analysis samples based on the time dimension and confirms  $H_4$ . The length of the time interval, which the bins are created based on, has an impact on the resulting entropy. For small time intervals a trend has to lie within a short time interval to be detectable, whereas larger bins make it easier to detect general trends.

Based on this examination of  $H_4$ , it can be seen as investigating if analysis samples are distributed non-uniformly in the relative time interval prior to a diagnosis, as opposed to  $H_2$  which investigated if analysis samples were distributed uniformly across the entire time dimension for patients.

A heat map for the entropy measurements is illustrated in Figure 13. This shows that for the majority of the combination of analysis samples and illnesses the entropy is relatively high, indicating no specific trends for when analysis samples are taken based on the time dimension. Therefore  $H_4$  is confirmed.

A scenario that the previous method of examination does not account for, is the fact that two diagnoses might be given within a small period of time of each other, thereby share analysis samples. Two illnesses can affect the same medical property (Shwe et al., 1991), and thereby lead to conflicting scenarios when determining one specific illness based on analysis samples. Examining the frequency of this scenario is not possible without knowing which medical properties are affected by which illness. However, it is possible to examine a more abstract scenario, which is the frequency of two diagnoses within a time interval. This frequency can be seen as the probability of receiving an additional diagnosis posterior to receiving a diagnosis. Figure 14 illustrates the proba-



**Figure 13:** Heat map over the Distribution of Analysis Samples on the Time Dimension

bility of a patient being given an additional diagnosis posterior to a diagnosis.

It can be seen that this abstract scenario occurs frequently, with the probability of receiving an additional diagnosis on the same day being 78%. The trend is also clear when plotting the mean amount of diagnoses for a time interval posterior to a diagnosis, as seen in Figure 15.

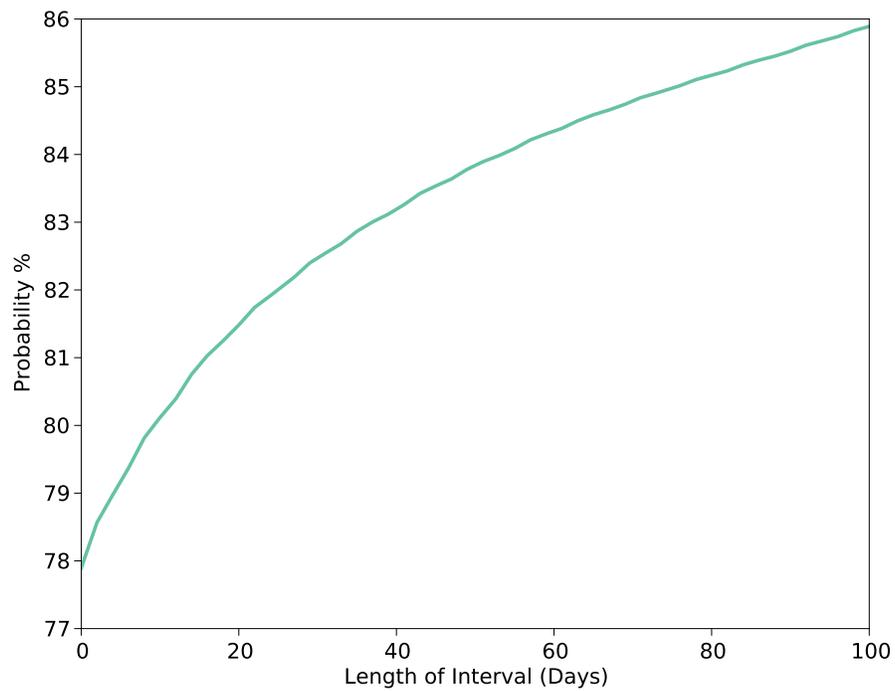


Figure 14: Probability of an additional diagnosis being given to a patient within a range of time intervals into the future for diagnoses.

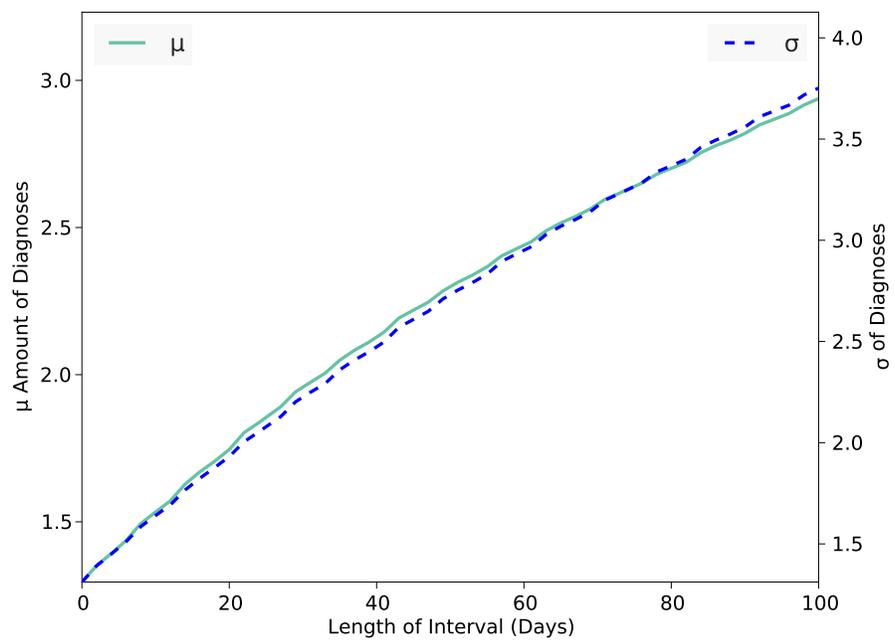


Figure 15:  $\mu$  and  $\sigma$  amount of additional diagnoses given prior to a diagnosis being given to a patient within a range of time intervals.

## 5. Problem Definition

The abstract problem statement being investigated is the following:

*Given a medical data warehouse, containing patient information regarding analysis samples and diagnoses, is it possible to construct a model for predicting illnesses based on measured medical properties with higher accuracy than naive approaches?*

The abstract problem statement lead to the need for a deeper understanding of the available data in order to concretize it. Sections 3 and 4 covered the data structure of analysis samples and diagnoses in the data warehouse, and an analysis of the respective classes' distribution and their interconnection. This has lead to knowledge that allows us to concretize the abstract problem statement.

Section 3 covers issues which arise from the digitization of diagnoses and analysis samples. An example of this is that the point in time where a patient becomes ill does not necessarily match the time a patient was diagnosed. Thereby in order to use diagnoses as reference points for an illness, one has to be aware of the existence of uncertainty for which point in time the diagnosis was given, within the illness interval. Furthermore, in regards to the time relationship between analysis samples and diagnoses, there is no information about when an analysis sample was taken, in regards to the illness interval. However, the degree of uncertainty is limited by the length of the illness interval for the patient. When constructing a predictive model, this uncertainty is a factor that should be accounted for.

In Section 3 it is mentioned that the presentation of illnesses is discretized through a binary state. The effect of having illnesses as represented as binary states is discussed in (Shwe et al., 1991), and it concludes that while it might weaken the model, it increases simplicity. This means that the data will be used in the form it is stored in the data warehouse, and will not require transformation into a more dynamic model for illnesses.

The assumption of relationship between analysis samples and illnesses, stated in Section 1, can be transformed into the assumption of: there exist certain global patterns in some medical properties prior to, or when affected by, a given illness. These patterns are global in the sense that they are to some degree common across all patients, and they are defined by how the measurements for a medical property change prior to or when a patient is affected by a given illness. Analysis samples are a sparse representation of a medical property, which these patterns must be based upon. As covered in Section 4.2, the patterns cannot be defined based on only the time attribute for analysis samples. Therefore it is important to account for both measurement and time, with regards to temporal aspects, when constructing a model for prediction. As an example, if a set of analysis samples indicates that a patient's blood pressure has consistently risen over the course of a week, then it is important to preserve the temporal aspect of this increase in the model, since it might be a symptom for a specific illness. It is the existence of these patterns which must be proved, in order to confirm that there exists a relationship between analysis samples and illnesses.

Section 4.2 investigated the frequency of two diagnoses occurring within a time interval for a patient. The latter diagnosis in such an interval, can be a complication, in the medical sense. This means that the second diagnosis was caused by the patient having the firstly diagnosed illness. When constructing a model for predicting illnesses, it would mean that a prior diagnosis might be the cause of a future diagnosis. This can mean that prior knowledge of a patient's health state affects the probability distribution of illnesses for a posterior diagnosis. For simplicity, as suggested by (Shwe et al., 1991), we assume that diagnoses are marginally independent to reduce complexity of the model while potentially increasing error of the model.

Analysis samples are primarily taken close to the time of a diagnosis as covered in Section 4.2. This strengthens the assumption that time is an important attribute for analysis samples which must be accounted for. The assumption that medical properties are affected by a patient's illness, leads to measurements of medical properties in analysis samples being affected by the patient's illness. This allows for prediction of illnesses based on analysis samples. Medical properties might be correlated, but to

reduce complexity and suggested by (Shwe et al., 1991) we use the assumption of conditional independence of medical properties given an illness.

There are two major factors to account for when predicting the illness of a diagnosis: the time or time interval of the prediction diagnosis, and the time distance between the analysis samples and the prediction diagnosis.

This leads to two assumptions: firstly, that diagnoses are given based upon a scenario where the patients observed changes in their health state, meaning that the patients are ill prior to being given a diagnosis. Secondly, that when the medical sector has diagnosed the illness, doctors start a treatment that shortens the length of the illness interval. The treatment might involve procedures, such as the use of medicine, which could affect medical properties and lead to noise in the analysis samples.

In Section 4.2 it was described how the time difference between an analysis sample and a diagnosis can be seen as a relevance measure for the analysis sample given the diagnosis. The horizon, the time interval in which the future diagnosis and prior analysis samples lie, must therefore be accounted for when constructing a model.

Based on the representation of the data, where there exists no explicit link between analysis samples and diagnoses, there are two statistical approaches to consider when constructing a model. These two options are either to perform extensive analysis of the data in order to gather knowledge of the parameters and thus use a parametric method, or to use a non-parametric method that relies on the data. Non-parametric approaches are generally better for large sets of data, since these data sets are difficult to generalize across without introducing errors (Doshi-velez et al., 2009). Thereby non-parametric approaches are deemed more suitable for the given data set.

Based on these simplifying assumptions, and the knowledge gained through analysis of the data, it is possible to create a model based on existing data that, by objectifying the data in the data warehouse, can assist in showing that there exist a relationship between analysis samples and diagnoses, which can be used to predict the illnesses of

diagnoses. With this in mind we now propose the following concretized problem statement:

*Given a set of patients, which have had analysis samples taken and been given diagnoses, is it possible to construct a non-parametric temporal model for classifying the illness of a future diagnosis, based on a set of analysis samples, within a horizon with higher accuracy than naive methods?*

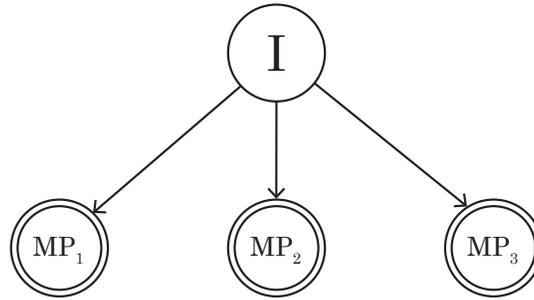
In this context, temporal model refers to a model that accounts for temporal aspects during inference.

## 6. Abstract Model

This section covers the intuition behind the constructed model. The act of diagnosing a patient, involves acting under uncertainty, since it is not possible to directly observe the illness a patient is affected by. However, there are observable indicators that affect a doctor's belief about which illness the patient might be affected by. As an example, observing an inflamed ear canal in a patient complaining of earache, increases the belief that the patient is affected by an outer ear infection. To handle this uncertainty, a probabilistic approach is chosen for the model.

When modeling illnesses and medical properties it can either be seen as illnesses causing changes in medical properties or it can be seen as a state of medical properties causing an illness. This can be clarified with these two examples: Firstly, an ear infection causes an inflamed ear canal, thus the illness causes changes in medical properties. Secondly, the state where the patient has an inflamed ear canal is diagnosed as an ear infection, therefore the state of the medical properties causes the conclusion that the patient suffers from the illness. We choose to model it as illnesses causing changes in medical properties.

How medical properties are affected by the illness can thereby be mapped as a belief network, also known as a Bayesian network. Figure 16 illustrates a Bayesian network representing the relationship between illnesses and medical properties.



**Figure 16:** Abstract model represented as a Bayesian Network, with  $I$  being a discrete node with a state for each illness,  $MP_1$ ,  $MP_2$ , and  $MP_3$  being continuous nodes for medical properties.

The network in Figure 16 uses three medical properties as an example, and thereby contains four nodes. The first node is the illness node  $I$ , which represents a random variable that has a state for each illness  $i_1, \dots, i_m$ . Illnesses are modeled in a single node, due to the assumption of marginal independence of illnesses introduced in Section 5. The node can be used for representing a probability distribution  $P(I)$  across its discrete amount of states, which holds a probability for each given illness. This probability distribution is used to represent the probabilities a patient has of being affected by certain illnesses.

The nodes  $MP_1, MP_2, MP_3$ , each represent a medical property. Having edges going from  $I$  to  $MP_1, MP_2, MP_3$  can be read as: *Given information about illnesses  $I$ , this information influences our belief of the random variables  $MP_1, MP_2, MP_3$ .* An example of this is, given knowledge of a patient being affected by the illness high blood pressure, it influences our belief of the medical property blood pressure, in terms of that it should be high for the patient. Based on the assumption of conditional independence of medical properties given an illness, covered in Section 5, there are no edges connecting medical property nodes. These medical property nodes are continuous random variables meaning they can be in a continuous amount of states, and this is illustrated through a double circle. Having a continuous state space, in this scenario, refers to that the attributes which describe a medical property are continuous values. An example of this

it that the measurement of a medical property, such as blood sugar, can be a numeric value and therefore does not have a discrete amount of states. Thereby  $P(MP_1)$  represents a space of normality measure for  $MP_1$  with a dimension for each attribute used to describe  $MP_1$ . Consider the example of measurements for patients' blood sugar being normally distributed with  $\mu$  being 150 and a low variance. Receiving a measurement of 3000 is highly unlikely based on the distribution of the data, and would thereby have a low normality value for that point in the normality space of the medical property blood sugar.

While  $P(I)$  represents a probability distribution over illnesses  $I$ ,  $P(MP_1 | I)$  represents a conditional probability distribution. This refers to that given information about the random variable  $I$ , we can construct a conditional probability distribution for  $MP_1$ , in which there exists a probability distribution of  $MP_1$  for each state in  $I$ . This means that given information about an illness, it affects our belief of a medical property. An example of this is, given the illness high blood pressure, our belief is influenced in terms of what would be considered normal measurements for blood pressure. This leads to the intuition that: *Given a set of observations for medical properties  $MP_1, MP_2, MP_3$  that defines points in the normality space, it is possible to deduce the probability distribution for illnesses based on this  $P(I | MP_1, MP_2, MP_3)$  given knowledge of normality for medical properties given illnesses  $P(MP_1, MP_2, MP_3 | I)$ .*

The abstract model is a naive Bayes model, based on the information variables,  $MP_1, MP_2, MP_3$ , are independent given the hypothesis variable  $I$  (Jensen and Nielsen, 2007). Time is not accounted for in the presented representation, and thereby the model only illustrates the intuition behind prediction of illness for an immediate state. Thereby, in a concrete model, time affects every node of the system, which has to be accounted for. The explained example, shown in Figure 16, is applicable to a model with any number of medical properties  $MP_1, \dots, MP_n$ .

The requirement of  $P(MP_1, \dots, MP_n | I)$  for determining  $P(I | MP_1, \dots, MP_n)$  can be described through Bayes' rule:

$$P(I | MP_1, \dots, MP_n) = \frac{P(MP_1, \dots, MP_n | I)P(I)}{P(MP_1, \dots, MP_n)},$$

and with the assumption of medical properties being conditionally independent given illnesses, it leads to the following:

$$P(I | MP_1, \dots, MP_n) = \frac{P(MP_1|I)P(MP_2|I)\dots P(MP_n|I)P(I)}{P(MP_1, \dots, MP_n)}.$$

$P(MP_1, \dots, MP_n)$  can be omitted, since the evidence for  $P(MP_1, \dots, MP_n)$  is identical across all observations, and thereby only functions as a factor for normalization.  $P(I)$  can be seen as set of relative weights for the illnesses. These weights can be seen as a tuning parameter, and can either be estimated through the distribution in the data set or tuned based on performance. Since this is an abstract representation and in a concrete scenario  $P(I)$  is influenced by time, the weights are initially defined through a uniform distribution, for simplicity, and thereby it is possible to omit  $P(I)$ . The remaining probability distribution is  $P(MP_1, \dots, MP_n | I)$ , and the method used for estimating it is covered in Section 7.

## 7. Medical Property-Illness Distributions

This section describes how to obtain the probability distribution  $P(MP_1, \dots, MP_n | I)$ . The probability distribution must be estimated based on the given data set, since the normality spaces for  $MP_1, \dots, MP_n$  given a state in  $I$  are unknown. Thereby a model will be constructed, which utilizes the available data to return an estimate of  $P(MP_1, \dots, MP_n | I)$ . Constructing such a model requires an understanding of how medical properties and illnesses are related.

The intuition for predicting an illness for a patient, based on a set of analysis samples, is that medical properties change in detectable patterns as the patient's health deteriorates during or prior to having an illness. Consider the synthetic example in Figure 17: given continuous measurements of a medical property for a set of patients  $p_1, p_2, p_3$ , becoming affected by illness  $i$  aligned to the same point in time. The patients'

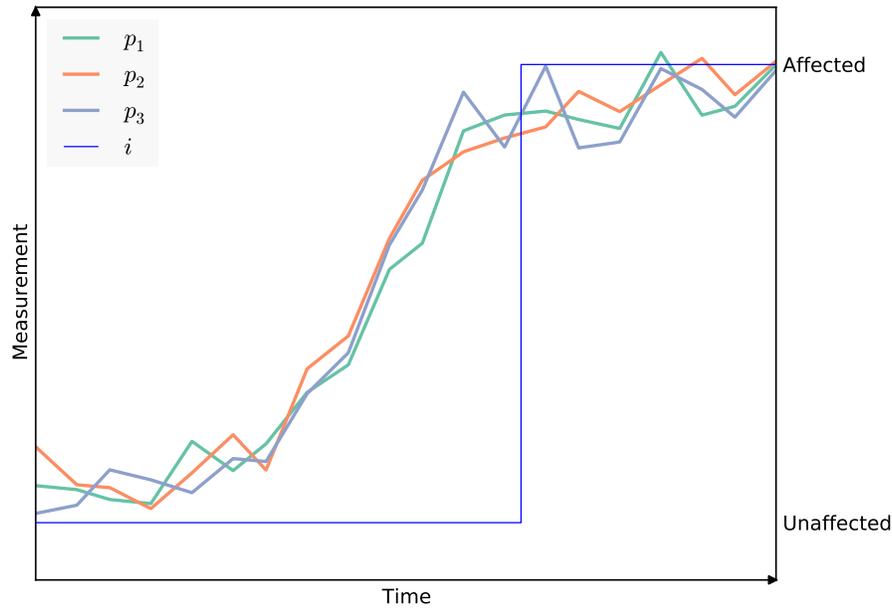


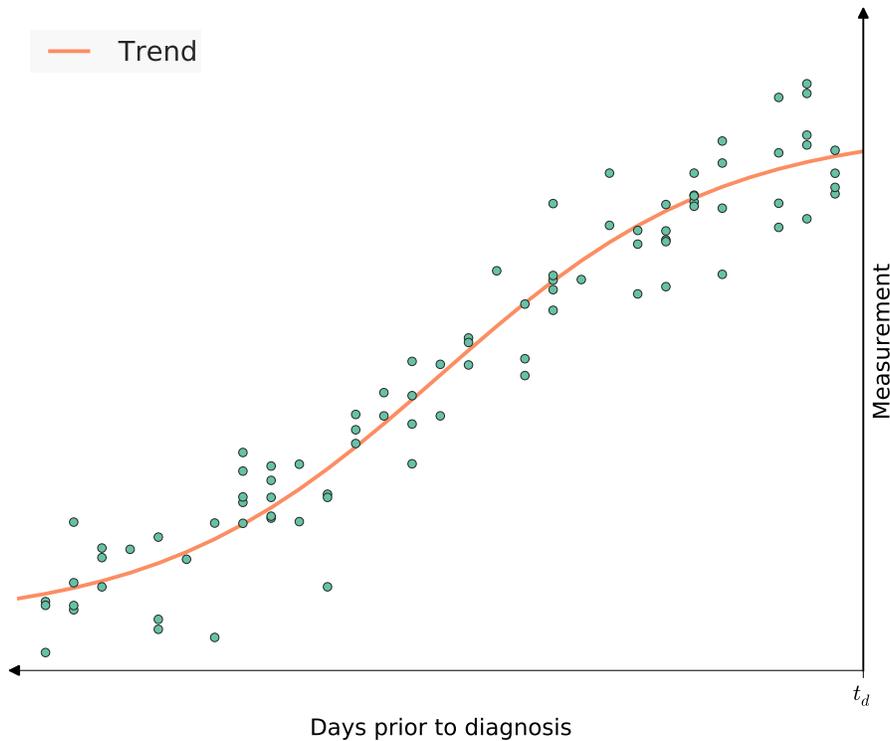
Figure 17: Change in value over time for a medical property for a set of patients,  $p_1, p_2, p_3$ , prior to and after becoming affected by an illness  $i$ .

medical property values rise prior to the patients becoming affected by  $i$ . This illustrates that there exists a correlation between the given medical property and the given illness, which reflects a belief of how the given medical property should behave during or prior to having the given illness.

The intuition is that the more similar a set of analysis samples is to the previous medical property observations, for a given illness  $i$ , the stronger the belief is that the patient is affected by  $i$ . Observations in this context refer to a set of analysis samples. The strength of the belief represents  $P(i | MP_1, \dots, MP_n)$ . The probability  $P(MP | i)$  can be estimated by measuring how normal a given analysis sample of  $MP$  is compared to the normality space for  $MP$  given  $i$  based on previous observations.

The example illustrated in Figure 17 assumes that continuous measurements of a medical property are available, however in the given data set it is represented as snapshots. These snapshots are analysis samples, which represent a value for a medical property in a given time for a given patient.

Using these snapshot values taken from all patients, we can plot an example for how a medical property behaves prior to patients being diagnosed with a given illness. An example of this, based on synthetic analysis samples, can be seen in Figure 18.

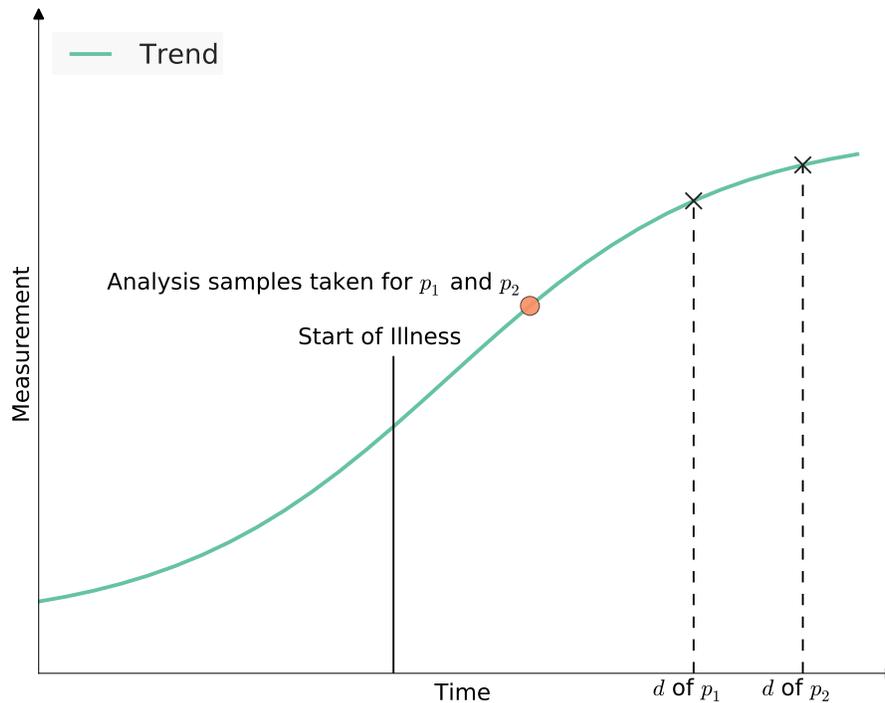


**Figure 18:** Example of an MPID for a medical property  $MP$  and an illness  $i$ , illustrating the trend for how  $MP$  changes prior to patients being diagnosed with  $i$ .

The visualized medical property changes according to the illustrated trend up to the point in time at which the patient is diagnosed with the given illness. Trends for a medical property given an illness are defined by how the analysis samples are distributed in the two dimensional space of time and value. The time dimension is used in order to preserve temporal aspects of observations. An example of this could be observations containing information about whether a medical property has increased or decreased over the given horizon, e.g. an increase in blood pressure. In this context the time for an analysis sample refers to the relative time prior to the diagnosis, when

the analysis sample was taken, with time discretized in days. Thereby the time values in the distributions are defined using diagnoses as reference points. The distribution of analysis samples of a given medical property  $MP$  and for illness  $i$  is referred to as a *Medical Property-Illness Distribution* (MPID) and corresponds to the normality space of  $MP$  given  $i$ .

The independence between two MPIDs, for two illnesses  $i_1$  and  $i_2$  and a medical property  $MP$ , correlates to how well  $i_1$  and  $i_2$  can be distinguished based on  $MP$ . This is because the measure of independence of the distributions describes the degree with which analysis samples for  $MP$  given  $i_1$  generally differ from analysis samples for  $MP$  given  $i_2$ .



**Figure 19:** An example of two patients that have had an analysis sample taken at the same point in time and becomes ill at the same point in time, while being diagnosed at different points in time.

Section 3 described that it is uncertain which relative point in time the diagnosis is given compared to the illness of the patient. This uncertainty is reflected in the diagno-

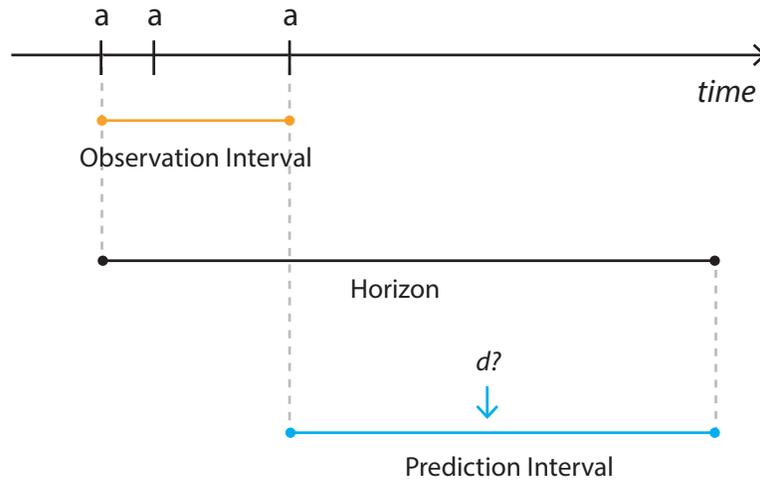
sis reference points used to create the MPIDs, which causes uncertainty on the relative time dimension across the analysis samples given an illness. An example of this is illustrated in Figure 19. Two patients  $p_1$  and  $p_2$  have had two similar analysis samples taken in terms of measured value and relative time to their illness interval. However, due to the difference in their time of diagnosis, when considering the illness interval, the analysis samples will be different in the respective MPID. This means that there exists uncertainty on the time dimension of MPIDs, which has to be accounted for.

Thereby  $P(MP \mid i)$  is estimated by the normality measure for an analysis sample given previous observations of the medical property  $MP$  given the MPID for  $MP$  and  $i$ . The methods used for handling the uncertainty and finding the normality measure for observations in regards to an illness are covered in Section 9.

## 8. Horizon

In order to reason sensibly about future illnesses, it is necessary to limit the time interval for which a future illness can be predicted. This is because increasing the time interval increases the range of possible time intervals between the observations and the future diagnosis. This increase lowers the relevance of the observations, since the time between an observation and a diagnosis influences how relevant an observation is for a diagnosis, as described in Section 4.2. An example of this is that measuring a patient's current temperature does not influence the belief in whether that patient is going to get a fever four years into the future. We thereby propose the concept of a horizon which is a defined time interval, for which it holds that a set of analysis samples and future diagnosis are contained in the interval. An example of a horizon is illustrated in Figure 20.

This example illustrates observations, in terms of analysis samples, on a patient timeline. The minimum time interval in which all of the analysis samples lie is referred to as the *Observation Interval*. When subtracting the observation interval from the time interval the horizon covers, the *Prediction Interval* is retrieved. The prediction interval is the time interval in which the future diagnosis, which we wish to predict the illness of, lies. Thereby, the distance into the future in which you can predict an illness of a



**Figure 20:** Example of a horizon imposed on a patient timeline. The time interval in which the analysis samples are taken is referred to as the observation interval, and the time interval in which the diagnosis being predicted was given is referred to as the prediction interval.

future diagnosis is dependent on the prior information obtained through the time interval, in terms of observations.

The concept of a horizon is related to MPIDs, since it defines the time dimension of the MPIDs. Therefore the horizon limits the information available in the MPID, by excluding analysis samples taken outside a relative time interval prior to diagnoses. The length chosen for the horizon affects the information contained within the MPIDs, since it delimits the analysis samples included in each MPID.

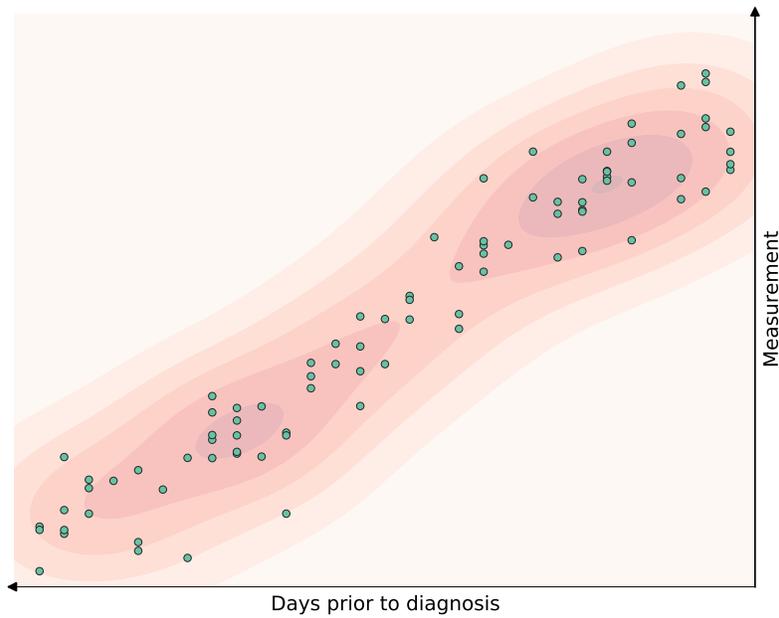
Based on this, limiting the information in MPIDs might affect the ability to predict certain illnesses. An example is the contrast between cancer and the flu; a six month old analysis sample might contain information that is valuable when diagnosing cancer, while the opposite holds for the flu, due to the differences in length of the illness intervals. Thereby, the horizon primarily affects the two following aspects: the information contained in the model and the desired prediction interval.

The chosen horizon will be discussed in regards to these aspects in Section 10.

## 9. Kernel Density Estimation

This section covers how  $P(MP | i)$  is estimated for a given medical property  $MP$  and illness  $i$  through a normality measure. The medical property  $MP$  is a random continuous variable, and therefore does not have a countable amount of states (Jensen and Nielsen, 2007). As mentioned in Section 6, every node is affected by time, which causes  $P(MP | I)$  to be dependent on the relative time to diagnosis and can thereby be seen as  $P(MP | I, \tau)$  where  $\tau$  is the relative time to diagnosis. Since we need to estimate  $P(MP | I)$  for prediction,  $\tau$  must be eliminated. The methods considered for eliminating  $\tau$  are covered in Section 10.1. Based on this we want to construct the *Probability Density Function* (PDF) for  $P(MP | i)$  given values for time and measurement. The PDF is unknown, however there are two approaches of estimating it. The first one involves the use of a parametric approach, which is based on assumptions of distributions in the data. The second uses a non-parametric approach, which involves using a set of data points to estimate it. Based on the problem definition in Section 5, the non-parametric approach *Kernel Density Estimation* (KDE) (Parzen, 1962)(Hansen, 2009) is used.

Using KDE it is possible to estimate the density, also referred to as the relative likelihood, of  $MP$  for patients having  $i$  given the analysis sample was taken at a certain time prior to the diagnosis and having a certain measurement. An example of this can be seen in Figure 21, which is based on the data of Figure 18. The example shows how the densities of the observations are centered around the trend line seen in Figure 18. The densities are depicted using contours, where the most red contour is the highest density. The density at a given value and time represents how frequently that observation has been made compared to other observations.



**Figure 21:** A contour plot of a KDE, estimated based on samples from an example MPID.

The KDE for an MPID is made in the following manner. Given an input vector  $X = x_1, \dots, x_n$  the multivariate kernel estimator is defined as follows:

$$\hat{f}(x) = \frac{1}{n*|H|} \sum_{i=1}^n K(H^{-1}(X_i - x)),$$

where  $K(u)$  is a multivariate kernel function and  $H = (h_1, \dots, h_n)'$  is a bandwidth vector with  $|H| = h_1 h_2 \dots h_n$ , with  $h_i$  being the bandwidth for the  $i$ 'th dimension. The multivariate kernel density estimator  $\hat{f}(x)$  integrates to one, and given the input vector  $X = x_1, \dots, x_n$  it returns the estimated density at point  $X$ .

A Kernel function is a function  $K(u) : \mathbb{R} \rightarrow \mathbb{R}$  which satisfies the following:

$$\int_{-\infty}^{\infty} K(u) du = 1,$$

where a multivariate Kernel function has the form  $K(u) = k(u_1)k(u_2)\dots k(u_n)$ , and is the product of  $n$  one dimensional kernel functions. Thereby the kernel functions fulfill the same criterion as probability distributions, in regards to their sum being 1.

The bandwidth variable is used to control the certainty with which a diagnosis can be predicted. How the KDE is affected by the bandwidth is illustrated in Figure 22 for a one dimensional synthetic data set. For  $h = 1$ , the KDE clearly peaks at the most dense concentration of data points.  $h = 4$  results in an oversmoothing which causes the KDE to not fully reflect the distribution of the underlying data.

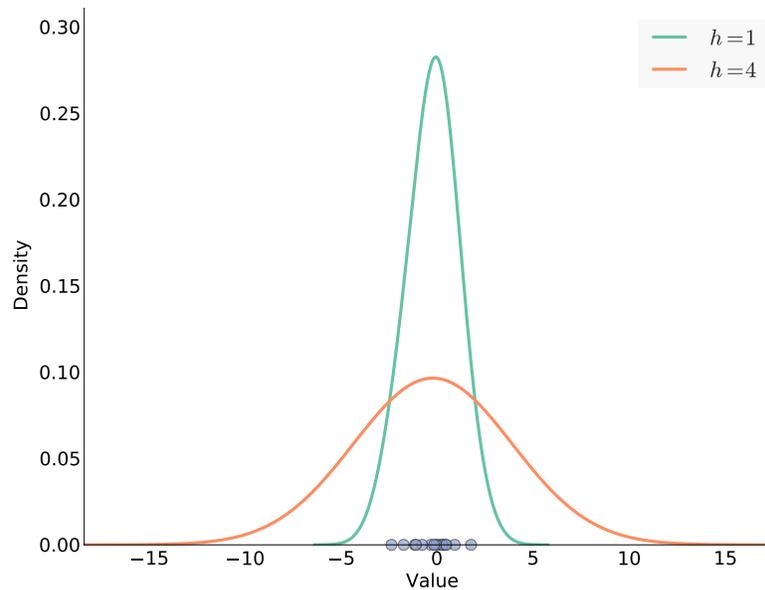


Figure 22: Two kernel density estimations for a synthetic data set, with the two bandwidths  $h = 1$  and  $h = 4$ .

Oversmoothing of the KDEs makes it more difficult to predict diagnoses, since  $P(MP | I)$  is lowered across all illnesses  $I$  when the KDE does not reflect the distribution of the underlying data. This is because oversmoothing of the KDE causes the probability distribution  $P(I | MP)$  to be more uniform, since with generally lowered probabilities for  $P(MP | I)$ , the chance that an illness has a significantly higher probability than other illnesses is lower. A more uniform probability distribution for  $P(I | MP)$  increases the entropy, which means that it is more difficult to predict an illness.

There exist cases where  $MP$  given  $i$  have never been observed before. When this occurs the density cannot be calculated and is therefore returned as 0, which means

the probability  $P(MP | i)$  is 0. Since there is always a chance of observing  $MP$  given  $i$ , this chance is represented using additive smoothing (Manning et al., 2008). Additive smoothing is used by adding a pseudocount to all densities.

The normality measure, introduced in Section 6, is thereby calculated through the use of KDE with additive smoothing applied to all densities. Section 10 covers how this is implemented in the model.

## 10. Model Overview

This section will cover how the methods and concepts presented in Section 7, 8, and 9 can be used to construct a model based on the intuition of the abstract model covered in Section 6. The abstract model showed that  $P(MP_1 | I), P(MP_2 | I), \dots, P(MP_n | I)$  are required in order to calculate  $P(I | MP_1, \dots, MP_n)$ . The probability  $P(MP | i)$ , for each medical property  $MP$  and each illness  $i \in I$ , can be estimated using Kernel Density Estimation based on prior time to diagnosis and measurements for analysis samples. The prior time is equivalent to the distance into the future the given diagnosis, which illness is being predicted, lies relative to the analysis sample. However, as described in Section 8, one wishes to predict the illness of a diagnosis within a time interval. A time interval is equivalent to a range of distances, and is determined by the prediction interval.

### 10.1. Probability Distribution

This Section covers how the probability distribution over a set of illnesses  $I$  is calculated given a set of observations in the form of analysis samples.

Consider the following example (E1): a doctor has one analysis sample of  $MP$  and based on this, he wants to know the probabilities of illnesses  $I$  that a future diagnosis for a patient might have. The analysis sample has a measurement, and thereby the only information missing in order to estimate  $P(MP | i)$  for each  $i \in I$ , is the time in which the future diagnosis lies. This point in time is unknown. However, we know there is a limited amount of possible points in time (days), based on the prediction interval de-

terminated by the horizon. Estimating over a time interval can be seen as the problem of combining multiple probability distributions for each discrete point in time into one probability distribution.

In order to solve this, the two following approaches have been considered. The first approach is to use principles of Bayesian Model Averaging (Opitz and Maclin, 1999), which involves the use of output from multiple models for the same problem and meaning across outputs for each class. In this case, the model could return an output for each point in time and then mean across the outputs for each illness. The second approach involves estimating a time of the future diagnosis for each  $i \in I$  based on principles from maximum likelihood estimation (Cam, 1990). This can be seen as using the time for which  $P(MP \mid i)$  maximizes, given a measurement such that  $P(MP \mid i) \approx \max_{\tau} P(MP \mid i, \tau)$ . Figure 23 illustrates an example of this, through a contour plot of kernel density estimation for  $P(MP \mid i)$ , based on prior time to diagnosis and measurement.

In Figure 23, the measured value for the analysis sample is marked by a line, and thereby only leaving the time into the future at which the diagnosis lies unknown. A horizontal slice across the time dimension in the KDE at the measured value for the analysis sample can be seen in Figure 24. The density can be seen as the likelihood, and with the time dimension containing a discrete amount of possible values, the maximum likelihood can be estimated by determining for which point in time the density is largest. Thereby  $P(MP \mid i)$  is the density at the point in time where the density is largest given a measurement for  $MP$ .

The first approach corresponds to averaging the densities, and thereby the probabilities, for illnesses across the time dimension. This can potentially hide valuable information in scenarios where an illness has a large density for a certain point in time but small densities for other times, due to the nature of arithmetic mean.

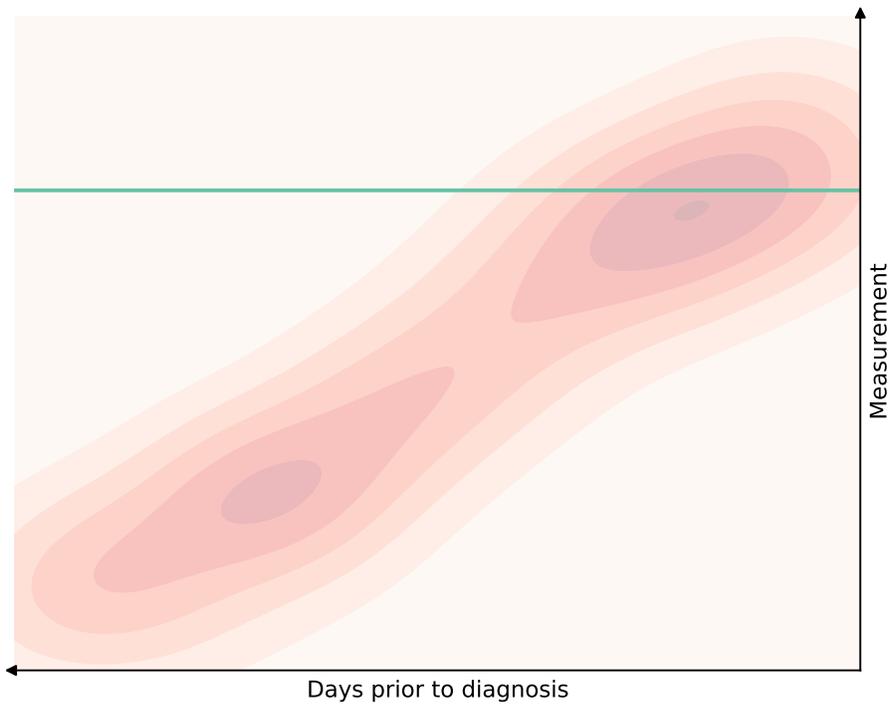


Figure 23: A contour plot of a KDE, estimated based on samples from an example MPID, with a line for a measurement of an analysis sample.

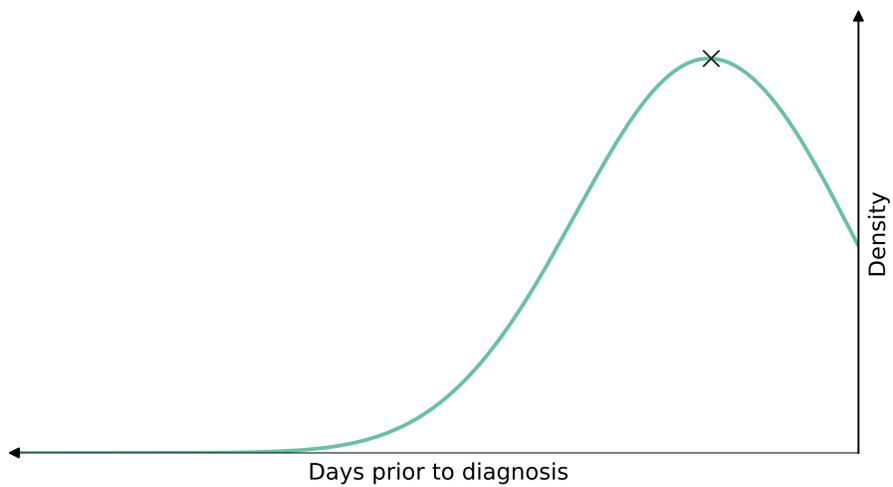
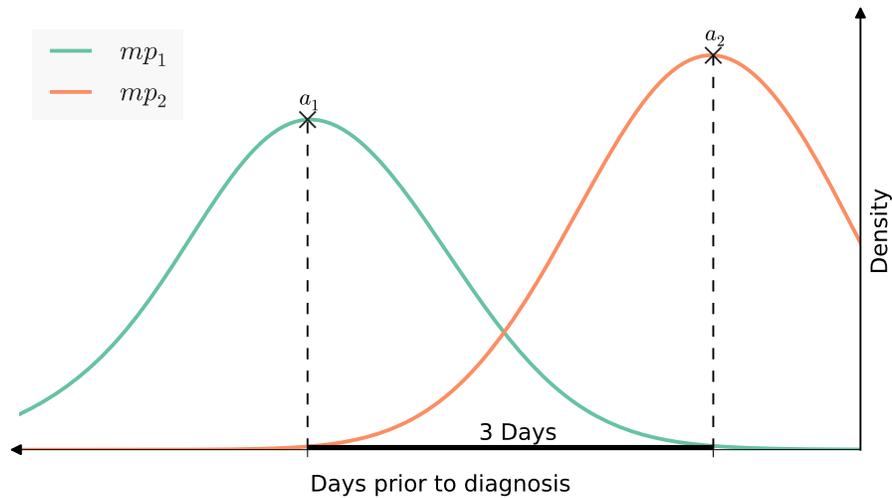


Figure 24: A density slice over time for a medical property based, on the line illustrated in Figure 23.

The second approach's intuition is based on answering the following question: *If a given patient were to get affected by illness  $i$  within a defined horizon, what is the most likely scenario for that to happen?* Thereby with the use of this approach,  $P(I | MP_1, \dots, MP_n)$  can be seen as weights for maximum likelihood scenarios over illnesses, if the patient were to get a diagnosis. The second approach is chosen based on this intuition, and the risk that using the first approach might potentially hide information.

E1 covers the case of a single analysis sample being used for prediction. However, there exist cases in which multiple analysis samples are available. Consider the following example (E2): a doctor has two analysis samples,  $a_1$  and  $a_2$ , measuring respectively the medical properties  $MP_1$  and  $MP_2$ , with  $a_1$  being measured three days prior to  $a_2$ . The doctor wants to know the probabilities of illnesses  $I$  that a future diagnosis for the patient might have. Given this case the, probability distribution being calculated is  $P(I | MP_1, MP_2)$ , which is equal to  $P(MP_1 | I)P(MP_2 | I)$  as per Bayes' rule, assuming a uniform probability distribution over  $I$ , and assuming conditional independence of medical properties given  $I$ , which is described in Section 6.  $P(MP_1 | I)$  and  $P(MP_2 | I)$  are calculated as described in E1, but with an additional temporal constraint. This temporal constraint is based on  $a_1$  and  $a_2$  being taken three days apart, and thereby they create an observation interval within the horizon. The relative time aspects of observations should be withheld, and therefore the goal is to find a set of points in time for which it holds that the relative time between the analysis samples is preserved, while maximizing the product of the densities. An example of this is shown in Figure 25.

The example shows the horizontal density slices across the time dimension of the KDEs for the medical properties  $MP_1$  and  $MP_2$  at the measured values for  $a_1$  and  $a_2$  respectively, for a given illness. The crosses mark the points in time for  $a_1$  and  $a_2$  which maximizes the product of both densities, while preserving the relative time aspect between them. Since the relative times between analysis samples are known, because they can be determined by knowing a single point in time, it is possible to determine the entire set of points in time, which maximizes the joint density for a set of analysis samples  $a_1$  and  $a_2$ , and thereby represents the maximum likelihood.



**Figure 25:** Two density slices over time for medical properties  $MP_1$  and  $MP_2$  given two measurements from two analysis samples respectively, in which the two analysis samples maximize the density based on their relative time being to each other is 3 days apart.

For this given example the newest analysis sample is used as a reference point. This reference point will be used to align the MPIDs on the time dimension, based on the relative times between each analysis sample. An example of the densities aligned over time based on the MPIDs is shown in Figure 26.

Based on the densities aligned on time, it is possible to construct a distribution which represents the joint density for  $MP_1$  and  $MP_2$  aligned based on  $a_1$  and  $a_2$  for a given illness  $i$ . The joint density is, as per  $P(i | MP_1, MP_2) = P(MP_1 | i)P(MP_2 | i)$ , the product of the densities for the individual analysis sample. Thereby, in this case, the point in time of the maximum density in the joint density represents the point in time for the newest analysis sample that maximizes the likelihood across the given analysis samples.

The approach covered in E2 is capable of handling any amount of analysis samples. Using the approaches explained through E1 and E2 it is possible to calculate the probability distribution for a set of illnesses  $I$  given a set of analysis samples. The prob-

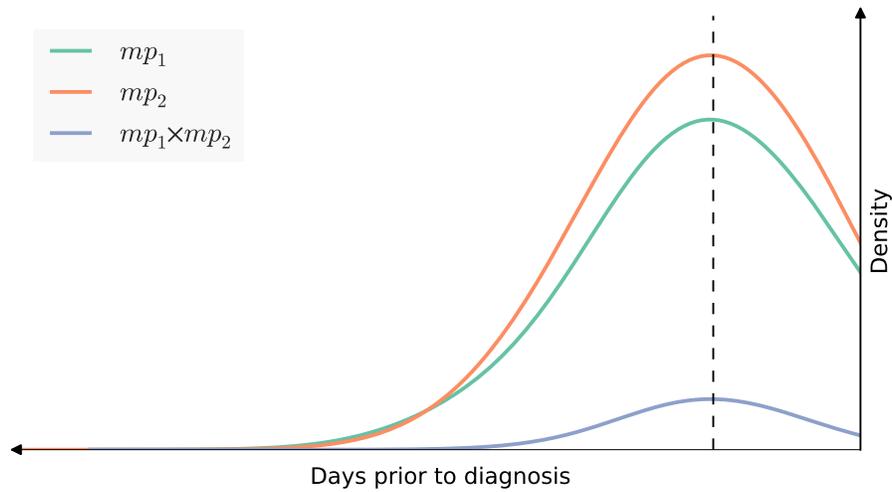


Figure 26: Two density slices based on Figure 25, in which they are aligned based on the analysis sample's relative time to each other and the product of the two density slices.

ability distribution can be seen as a hypothesis returned by the model for which illness or illnesses a patient is affected by.

## 10.2. Configuration

This subsection covers how the model is configured through defining the bandwidth used in KDE and the length of the used horizon.

The bandwidth selected for the KDEs has an impact on the resulting densities, and thereby on the calculated probabilities, as discussed in Section 9. Therefore the bandwidth can be seen as a tuning parameter, which can be adjusted to improve performance. In order to remain non-parametric, we wish to use a method which can estimate an optimal bandwidth. For this purpose, the bandwidth will be defined based on Silverman's rule described by (Silverman, 1982) and with the considerations in (Jones and Lotwick, 1984). Silverman's rule has been shown to be the optimal choice of bandwidth given that the underlying data is Gaussian distributed and is in other cases considered a good starting point for tuning the bandwidth (Silverman, 1986).

It is assumed that the chosen kernel function has little influence on the resulting densities. This hypothesis is investigated in Section 11. The kernel function used for KDE is the Gaussian kernel function, which is defined as:

$$k(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

The length of the horizon determines the time interval in which the illness of a future diagnosis can be predicted. As discussed in Section 8, the chosen length for the horizon can impact the ability to predict some illnesses, since a short horizon might remove evidence that a patient is affected by a given illness. Furthermore if the horizon is too large, then the risk of including analysis samples taken outside an illness interval for a given diagnosis is increased, which is a potential source of noise. Due to the fact that analysis samples are primarily taken close to a diagnosis being given, as shown in Figure 11, it can be assumed that diagnoses are often given based on a set of recent analysis samples, and thereby a relatively short horizon is preferred. Therefore the chosen length of the horizon is 30 days.

## 11. Model Evaluation

Based on the model presented in Section 10, this section covers the reasoning behind the evaluation of the model and the results showing its performance. The problem statement covered in Section 5 defines a binary success criterion, which is evaluated based on a performance measure. The potential use context for the model as a supportive tool for diagnosing patients, is evaluated based on qualitative analysis of performance.

To evaluate if the constructed model fulfills the problem statement, a set of illnesses is selected in order to define the class context for which illnesses will be predicted. The class context is the set of illnesses  $I$  for which the model creates a probability distribution. In order to limit the computational resources required to run the model, the evaluation is performed on a subset of the total 31 105 different illnesses. The illnesses in this subset are selected based on the amount of data concerning them, in terms of how frequently they are diagnosed, since it is assumed that more frequently diagnosed illnesses have a larger amount of medical property information available. The medi-

cal property information consists of the amount of observations within the horizon of a given diagnosed illness. This means that the illnesses are being arbitrarily selected, in a medical sense, since they are not selected based on any medical knowledge. The classes used for the evaluation consists of a set of 20 illnesses, which are among the top 100 most frequently diagnosed illnesses. The patients which have had some analysis sample taken, and have been diagnosed with at least one of these illnesses, form the data set used for evaluating the model.

The intuition for evaluating the model, based on a set of illnesses that are arbitrary in a medical sense, is that if the method performs better than naive approaches on such a set of illnesses, then the model should be generalizable across any set of illnesses. Furthermore if the model performs better than naive approaches on these illnesses, then it should be possible to select a set of illnesses, using medical knowledge, for which the model performs significantly better.

In order to test the model, the intuition is to exclude a subset of diagnoses of the selected illnesses prior to training and use the subset exclusively for testing. This test subset is referred to as the testing set, while the remaining data is referred to as the training set. One thing to consider is that patients might have some individual characteristics in their medical properties which could potentially skew the evaluation of the model. Consider the case where information about a patient is shared across both the training set and the test set. This violates the independence of the two sets, which could lead to overfitting of the model. Thereby the data set is divided across patients, while the testing and training is still performed across diagnoses. The test set consists of 20% of the patients diagnosed with one of the selected illnesses, while the remaining 80% are included in the training set.

Dividing the data on patients leaves us with two sets of patients, and thereby two sets of patient timelines. The timelines for the patients in the test set need to be divided into diagnosis samples with corresponding analysis samples. This represents a test sample. Test samples are created using the principles described in Section 7, such that: for each diagnosis a patient has, the illness of that diagnosis defines the desired

class output for that sample. The diagnoses have a set of analysis samples related to them, which is the analysis samples within a horizon prior to the diagnosis. The horizon used for gathering these analysis samples is of equal length to the one used in the model, which is 30 days. Therefore a test sample consists of an illness and a set of analysis samples, which is referred to as the observation set. However, there exist cases where the observation set is empty, since no analysis samples lie within the horizon of the diagnosis. These cases will be excluded for testing, since without observations it is not possible to reason about the illness of a diagnosis. This results in the test set consisting of 3504 test samples, and the training set consisting of 64 369 diagnoses across 64 369 patients. The test samples are extracted from 29 150 diagnoses across 16 092 patients.

The performance of the model is evaluated through the use of the performance measure *accuracy*. This performance measure is defined as follows:

$$Accuracy = \frac{CP}{TS},$$

where  $CP$  is the amount of correct predictions, equivalent to the amount of times the model was able to predict the illness of the test sample and  $TS$  being total amount of test samples. An illness is considered correctly predicted by the model, if it is returned as the most likely illness for the test sample.

Since the model is evaluated in relation to naive approaches, it is relevant to consider the best possible performance for a naive approach, given the selected illnesses. Consider the frequencies of illnesses across diagnoses of the test set illustrated in Figure 27. Assuming that the class distribution across the training set and the test set are similar, a naive approach would at maximum be able to achieve an accuracy of 15.2%, as covered in Section 5. This is because the most frequent class, represents 15.2% of the test samples. Therefore in order to answer the problem statement, by confirming the posed question, the model must have a higher accuracy than the naive approach. The model correctly predicts 778 out the possible 3504 test samples in the test set, and thereby has an accuracy of:

$$Accuracy = \frac{778}{3504} = 22.2\%.$$

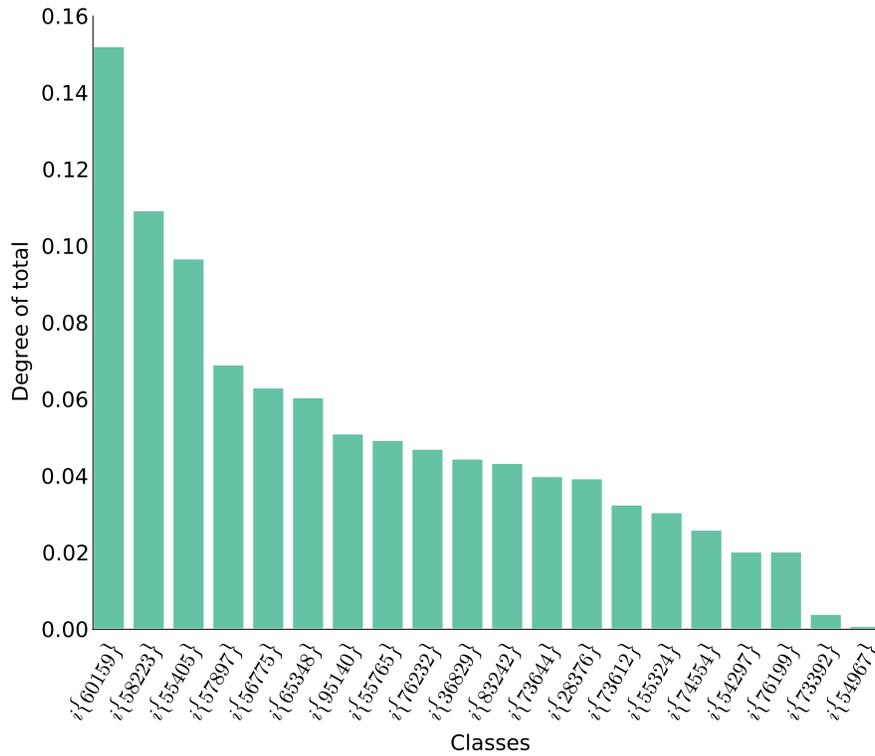


Figure 27: The distribution of the classes within the test set.

Since the naive approach would achieve an accuracy of 15.2%, the model thereby performs better than the described naive approaches on a set of illnesses selected arbitrarily, in a medical sense.

Evaluating the model solely based on accuracy, and thereby only considering the most probable illness returned by the model, might not be reasonable in the context of using the model as a decision support tool for diagnosing patients. Consider the example of a doctor wanting to diagnose a patient. In this example the doctor might not only be interested in the illness with the highest probability, since the hypotheses returned by the model are estimates and thereby may not be exact. The amount of information the model is given as input, in terms of analysis samples, also influences its ability to distinguish certain illnesses. This is because distinguishing between certain illnesses might require observations for specific medical properties, and there-

fore the more observations that are available, the better hypotheses the model should return. The doctor may use the model in a continuous process of gathering information and consulting hypotheses returned by the model to decide which information to gather next. The goal of this process is to increase distinguishability in the hypotheses returned by the model by gathering additional information, and thereby in the end isolate a single illness which the patient will be diagnosed with. In this scenario, it is desirable to examine the model's ability to isolate a group of illnesses which contains the illness to predict.

This is examined by using principles from accuracy. However, a class is in this context considered correctly classified if it is amongst the  $k$  classes with the largest probability. This new accuracy measure is referred to as  $k$ -Accuracy and can be calculated using the following formula:

$$k\text{-Accuracy} = \frac{CP^k}{TS},$$

where  $CP^k$  is the amount of test samples, in which the illness of the test sample is among the  $k$  illnesses with the largest probability returned by the model.

Based on this, the accuracy at the different values of  $k$  can be determined, and is illustrated in Figure 28.

The orange line illustrates a benchmark for accuracy, which is the expected accuracy of arbitrary prediction. It can be seen that the model predicts significantly better than the arbitrary prediction benchmark and with  $k = 5$  it reaches an accuracy above 50%.

Throughout these results the kernel used for Kernel Density Estimation has been a gaussian kernel, as described in Section 10.2. This is due to simplicity and based on the hypothesis that the type of kernel would have a minor influence on the result. Figure 29 illustrates the  $k$ -Accuracy of the model using three types of kernels for Kernel Density Estimation, with the types being: uniform, gaussian, and triangular.

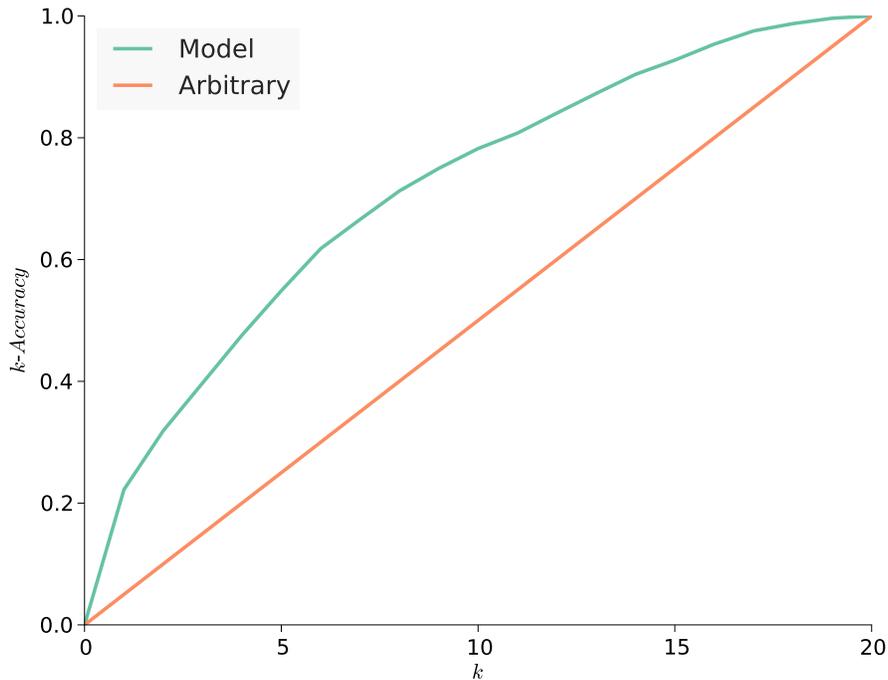


Figure 28:  $k$ -Accuracy for the model at a range of values of  $k$

By examining the standard deviation ( $\sigma$ ) of the  $k$ -Accuracy of each kernel at various values of  $k$ , it can be seen that the performance does not vary across kernels to a large degree. This thereby confirms the hypothesis that the type of kernel used has a minor influence on the performance of the model.

The approach of arbitrarily selecting illnesses as classes, for the model, might not be suitable in the scenario where reaching high measures of accuracy is the objective. Since accuracy is related to classes being distinguishable based on a set of features, this leads to the problem that some illnesses might not be distinguishable based on analysis samples. In order to examine the occurrence of this in the current set of classes, the individual class accuracy is measured and illustrated in Figure 30.

It can be seen that the individual accuracy for the classes varies to a large degree, and thereby lowers the accuracy of the model. This variance is assumed to be based on two problems: as mentioned, the classes are selected arbitrarily and therefore there

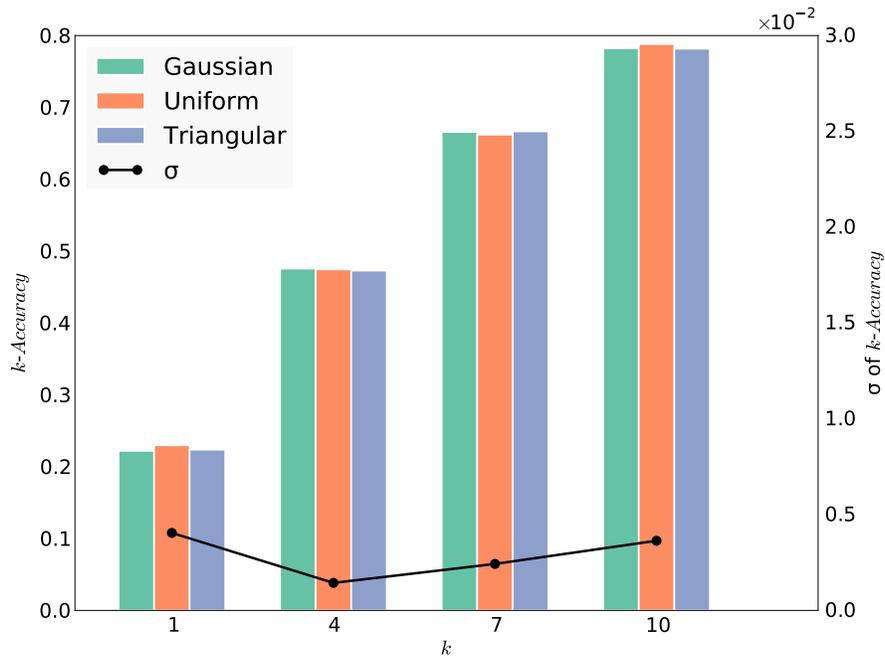


Figure 29:  $k$ -Accuracy for three different kernels used for KDE for four different values of  $k$ , with standard deviation of the  $k$ -Accuracy at the different  $k$  values.

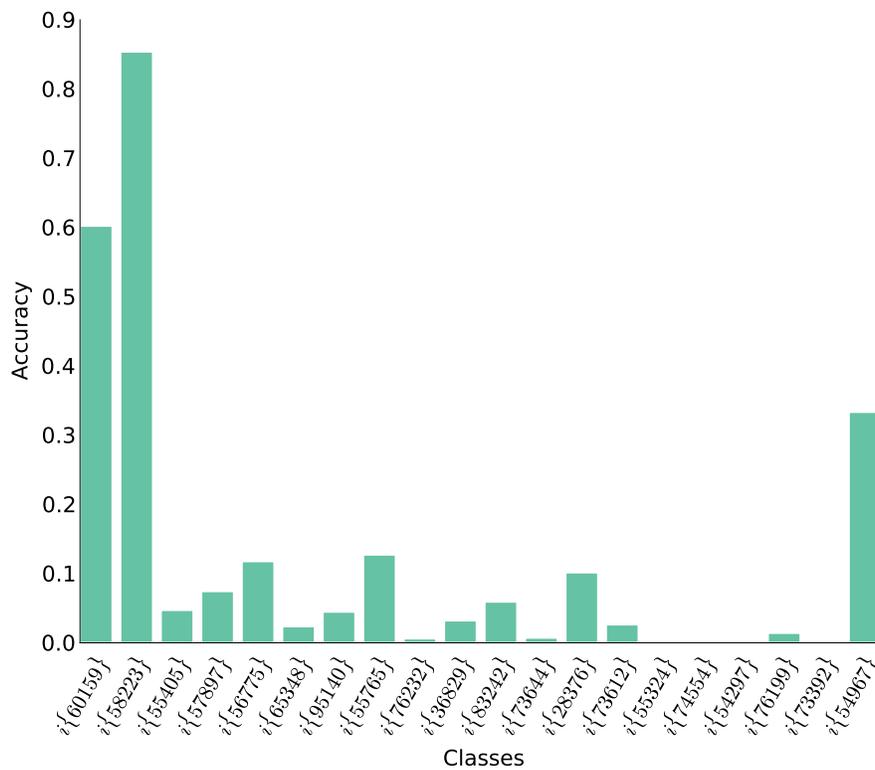


Figure 30: Accuracy for each individual class

is no guarantee that it is possible to achieve a high accuracy for the set of classes. The distinguishability for a class in a set of classes depends largely on the other classes included in the set. This is because the class must be separable from all the other classes in the set, thereby the more classes in the set, the harder it becomes to remain distinguishable. The names of the illnesses in the set used in this evaluation can be seen in Appendix A. This concept of determining the set of classes for achieving higher accuracy is covered in Section 12.

## 12. Class Context

This section covers methods for defining the classes used by the model in order to increase the performance, without using medical knowledge. The set of classes, and its size, influences the performance of the model, as described in Section 11. This set of classes is referred to as the class context for the model. In a use context as a decision support tool, the performance might be used as a measure of validity. This means that if the performance is low, it might not be adequate for use as a decision support tool for doctors diagnosing patients. Thereby, as a measure of the model's potential for use as decision support, its performance is evaluated on a set of classes selected with the purpose of maximizing performance in terms of accuracy. This will be done by estimating a performance measure for each individual class, which is used to define a set of classes that yields a high accuracy for the model. The performance measure is based on the accuracy of the individual class.

Measuring the general performance of a class is complex, since it is affected by each other class included in the model, as described in Section 11. This means that the prediction problem that the model attempts to solve becomes more complex as the size of the set of classes increases. This causes the amount of hypotheses returned by the model to increase and thereby the entropy of the prediction increases. This means that having more classes, and thereby a potentially increased entropy, is likely to lower accuracy and vice versa. Therefore to estimate a set of classes which would increase the accuracy of the model, the individual class accuracy of a large set of classes will be examined. This examination consists of running the model on a large set of classes and

then using the individual performance for each class to determine a new class context. This is done by examining the individual class accuracy and selecting an amount of illnesses with the largest class accuracy, in order to show correlation between accuracy and class context.

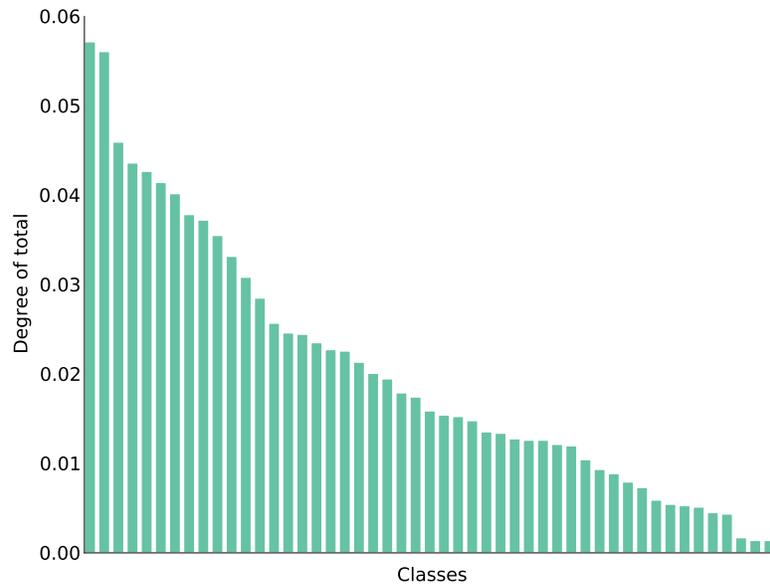
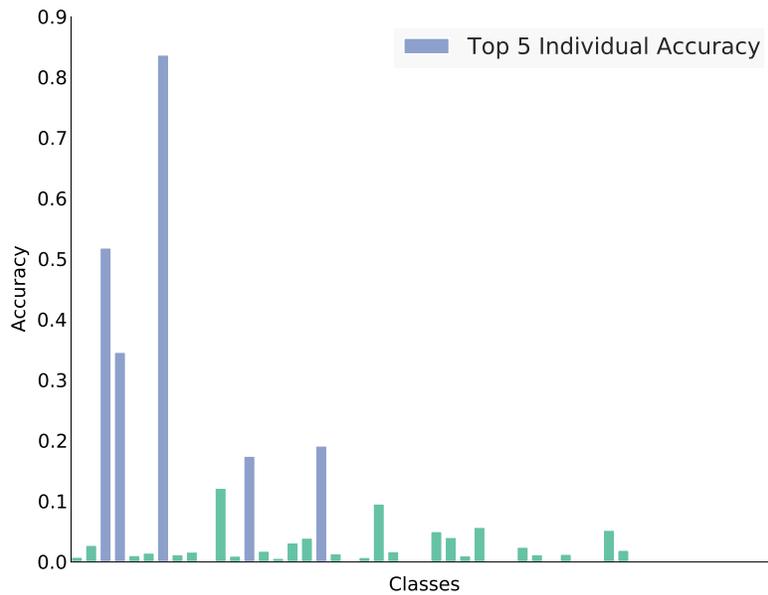


Figure 31: Distribution of classes in the test set for the class context of the 50 classes.

The initial set for this examination consists of 50 illnesses. Figure 31 is a histogram of the 6419 test samples in the test set across the illnesses. In order for the model to perform better than naive methods, its accuracy should be larger than 5.7%, since this is the maximum accuracy that can be achieved by predicting based on the most occurring illness. The accuracy is 9.9%, and thereby better than the naive approach.

In Figure 32 the class accuracy for each class can be seen. Based on these, the five classes with largest class accuracies are chosen as the new class context. The model is then examined with this new class context.

This examination is based on a test set with the size of 2265 test samples. The accuracy of the model is 48%, with naive being 24.6%. The class accuracy for this class



**Figure 32:** Accuracy for each individual class of the 50 classes with the five classes with highest accuracy highlighted in blue.

context can be seen in Figure 33. Comparing these class accuracies with those for the initial class context shows that the accuracy for each class has changed in this new class context. This confirms that the accuracy for each class is influenced by the remaining classes in the class context.

Using this knowledge to examine the performance in regards to a use context leads us to examine the  $k$ -Accuracy for the mentioned class contexts. The performance for the class context used in Section 11 combined with the two class contexts examined in this section is illustrated in Figure 34. This shows the  $k$ -Accuracy for each class context at a relative value of  $k$ . It can be seen that the  $k$ -Accuracy is similar across different class contexts.

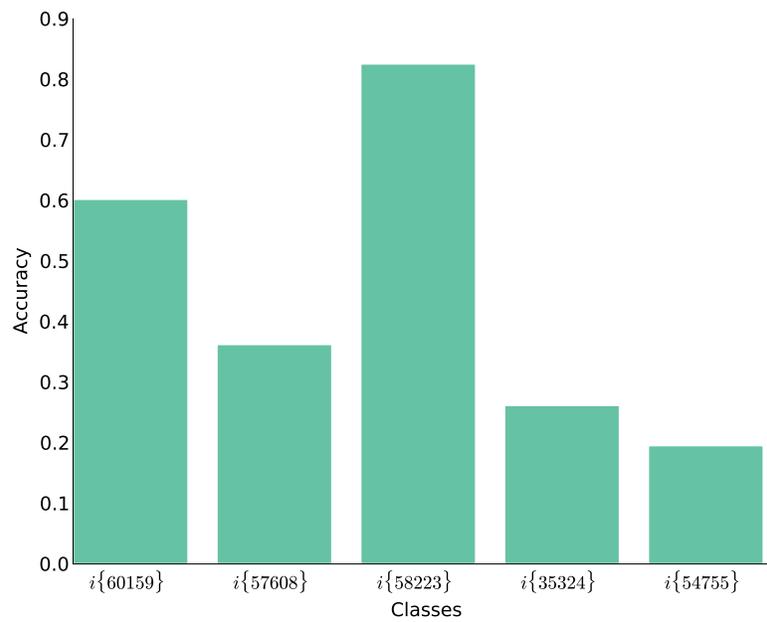


Figure 33: Accuracy for each individual class of the five classes.

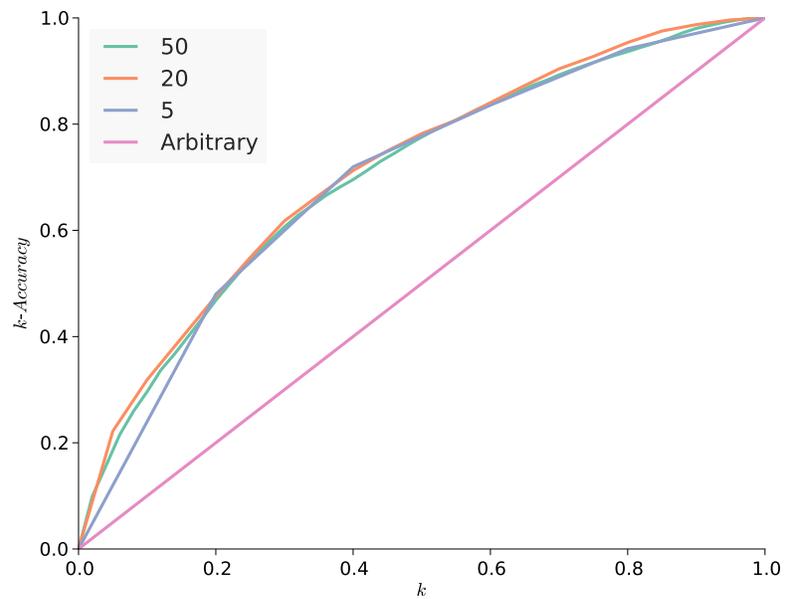


Figure 34:  $k$ -Accuracy of the model given a range of values of  $k$ , for the test set for each class context.

### 13. Discussion

The results presented in Section 11 and Section 12 show how the model performs across sets of classes that are arbitrarily selected and chosen based on performance. The scenario of using a set of classes chosen based on performance for the model achieves an accuracy of 48%. This indicates that for certain illnesses it is possible to define a relationship between analysis samples and illnesses, which makes it possible to predict or get an indication of which illness a given patient is affected by. However, the quality of the predictions or indications depends on the given illnesses. This is because some illnesses cannot be predicted with a high accuracy using the available data, as shown in Section 11. This is possibly caused by lack of information available to the model, since some illnesses might not be predictable solely based on the available medical properties. In summary, the model is not capable of achieving a high accuracy on the set of every illness, so in order for it to achieve high accuracy it is required to use a subset of illnesses.

This thereby limits the medical contexts in which the model is applicable for predicting a single illness. However, as shown in Section 11, when evaluating the model based on *k-Accuracy* for an arbitrarily selected set of illnesses, the model is capable of achieving an accuracy of almost 80% for  $k = 10$  with the model running on 20 illnesses. In this scenario the model rejects half of the possible illnesses, while having high accuracy and thereby low uncertainty for the remaining illnesses. Given this, a possible usage for the model in a diagnostic decision support scenario is as a method for eliminating possible illnesses from a set of illnesses that a doctor suspects a patient is affected by. For this usage, a doctor would configure the model with a set of illnesses he suspects the patient might be affected by. Based on the high *k-Accuracy* for a set of  $k$  most likely illnesses, the doctor eliminates the illnesses not in this set from the original set of illnesses that he suspected the patient might be affected by.

Despite the performance of the model, it may be useful in a scenario where the *k*-Accuracy performance measure is used to determine its usefulness. However, the model predicts incorrectly more than half of the times in a limited class context, when only the illness with the single highest probability is considered.

The current performance is most likely limited by three factors: the quality of the data, the amount of features contained in the model, and the assumptions covered in Section 5.

The quality of the data in this context refers to potential errors in the data. There exist various types of errors, however the ones that are assumed to affect the model are the three following: incorrect diagnoses, uncertainty of when in the illness interval the diagnosis was given, and erroneous analysis samples. Incorrect diagnoses are diagnoses which contain false information in regards to the patient's health state, such as: being diagnosed with the incorrect illness, or receiving a diagnosis while not being affected by an illness and vice versa. The uncertainty between diagnoses and illness intervals across patients was introduced in Section 7 as uncertainty on the time dimension from using diagnoses as reference points. As an example, an analysis sample taken three days prior to a diagnosis might contain the same information as an analysis sample taken five days prior to a diagnosis, given they were taken at the same point in the illness interval. This issue is related to the fact that patients might be diagnosed with a given illness at different points in time in their illness interval, meaning using time as a feature across patients might yield some false information, since similar analysis samples can be displaced on the time dimension. Erroneous analysis samples refer to analysis samples with an incorrect measurement, which influence the model by introducing false information. This can be caused by the analysis sample being mistreated when it was processed.

Currently the model utilizes only features related to analysis samples, i.e. the time the analysis samples was taken and its measurement, for determining the illness of a patient. However, there exist other features in the data warehouse which might contain information that can increase accuracy of the model. These features could be fea-

tures containing information about the patient, such as: age, gender, chronic illnesses, and family relations. When considering features that are not available through the data warehouse, patient records could contain valuable information. Patient records can contain information about symptoms that are not detectable through the available medical properties.

The model is currently based on a set of simplifying assumptions, such as marginal independence of illnesses and that patients are only affected by a single illness at a time. The assumption of marginal independence of illnesses affects the model, since other illnesses that a patient might be affected by are not considered when predicting a new illness. As indicated in Figure 14, the case that a patient is affected by more than one illness simultaneously can occur for patients. Through medical knowledge, it is known that there exist correlations between illnesses through complications (National Eye Institute, n.d.). Knowing this, the information that a patient has recently been diagnosed with an illness can be utilized by the model for prediction to potentially increase its accuracy.

Knowing that the presence of one illness might lead to another could construct a scenario of a patient affected by two illnesses at the same point in time. However, in the current model, the illnesses are modeled as a single random variable, meaning that it is assumed that patients are only affected by a single illness at a time. Therefore the model does not account for the scenarios where the patient might be affected by multiple illnesses at the same point in time. These multiple illnesses, at the same point in time, can be seen as a joint illness that consists of a set of illnesses in which each illness is referred to as a sub illness. This joint illness might be observed as a health state, which differs from the individual health state of each of the sub illnesses. Thereby, the medical properties which influence the belief for the individual sub illnesses might not be equivalent to medical properties which affect the joint illness. Therefore a joint illness can potentially be seen as a new illness, which is influenced by its own set of medical properties. Including this in the model increases the complexity, however it may be useful in a use context as a supportive tool for diagnosing patients.

The foundation for the research conducted in this report is based on the assumption that there exists a relationship between analysis samples and illnesses that makes it possible to predict illnesses based on analysis samples. This assumption is a simplification of the process a doctor goes through when diagnosing a patient. The doctor's belief of the health state for patient is influenced by the analysis samples. A doctor does however use additional information when determining the health state of a patient such as medical history.

The results from the tests of the model hint towards it being possible to use statistical analysis and machine learning to determine the illnesses of a patient based on medical data. This can support the current medical system, since the model functions as a tool for objectifying the diagnostic procedure and knowledge sharing across doctors.

## **14. Conclusion**

The motivation behind this project was based on the belief that it was possible to extract additional information from a medical data warehouse containing data on analysis samples and diagnoses through data mining. This was based on the belief that there exists a relationship between the analysis samples a patient has had taken and the diagnoses he has been given.

This lead to the investigation of an abstract problem statement, which is used throughout analysis of the data in the data warehouse. This analysis showed that analysis samples are distributed close to diagnoses on a patient timeline, indicating that there exists a relationship between analysis samples and diagnoses based on time. Based on the use context in which analysis samples are taken, it was assumed that the measure of an analysis sample is significant for this relationship. The information gained through the qualitative analysis lead to the following problem statement:

***Given a set of patients, which have had analysis samples taken and been given diagnoses, is it possible to construct a non-parametric temporal model for classifying the illness of a future diagnosis, based on a set of analysis samples, within a horizon with higher accuracy than naive methods?***

This problem statement was researched through construction of a probabilistic model, due to the uncertainty of diagnosing patients with illnesses. The intuition of the probabilistic model was based on a Bayesian network, describing the relationship between medical properties and illnesses. This network defines the probability that a patient is affected by a given illness within a defined time interval, based on a set of analysis samples, by a normality measure for each of the medical properties for the given illness. The normality measure is defined using *Kernel Density Estimation*, for each medical property based on the time and measurement of the analysis samples of that medical property. Throughout the model, the length of the time interval, referred to as the horizon, is used to determine the time interval in which the illness is being predicted. The horizon is used to determine the amount of information that is used for the *Kernel Density Estimations* and defines a time interval in which the diagnosis, which we wish to predict the illness of, lies.

The probabilistic model is evaluated through a set of performance measures. The first one being accuracy, which is used to measure the model's ability to predict illnesses of future diagnoses that lie within the horizon for a given set of analysis samples. Evaluating the model based on accuracy can answer the problem statement, since it makes it possible to compare the model to naive approaches. The model was evaluated on three different sets of illnesses of various sizes, which were proper subsets of the set containing all illnesses. For all three sets the model performed better than the best naive approach.

The second performance measure is called *k-Accuracy* and is a measure of the model's ability to isolate a set of illnesses in which the correct illness is. This measure was used to evaluate the model for a use context of doctors using it as decision support, and showed that the model is capable of isolating a relatively small set of illnesses for prediction while achieving high accuracy.

It was shown that the set of illnesses for prediction used for the model influences the performance of the model. Choosing the set of illnesses for prediction, in regards to performance, was shown to be complex since each chosen illness influences the performance of other illnesses in the set.

Through this model driven approach, it is concluded that there exists a relationship between analysis samples of medical properties and illnesses of diagnoses, which can be used to predict the illness a patient will be diagnosed with.

## 15. Future Work

This report covers the intuition and reasoning behind the construction of a model for use in decision support for diagnosing patients. The evaluation of the model covered in Section 11 and Section 12 shows that, when used on a small set of manually selected illnesses, the model's performance increases in comparison to larger sets of illnesses that are arbitrarily selected. However, in order to increase performance, some aspects of the model should be investigated more thoroughly. This section covers some of these aspects and ideas, which are relevant to investigate in order to increase performance.

In the current model, the usefulness a given medical property has for diagnosing a given illness is equal across all medical properties. This causes every medical property to be considered a biomarker (National Cancer Institute, n.d.), meaning no medical property influences the belief of a given illness more than another. This simplification could potentially lead to false evidence. Consider the example of a patient affected by the illness high blood pressure that has had analysis samples taken for the medical properties blood pressure and hemoglobin. Given medical properties are weighted

equally, the analysis sample for hemoglobin influences our belief of the patient having high blood pressure with the same weight as the actual blood pressure. Thereby the influence a certain medical property has for a given illness is a topic that should be researched further.

Using Kernel Density Estimation requires defining the parameter bandwidth. In the model, this parameter is approximated through the use of Silverman's rule, as described in Section 10.2. The parameter bandwidth can influence the performance of the model in terms of accuracy, as discussed in Section 9. Therefore the model can be tuned based on bandwidth in order to increase performance. Tuning the bandwidth parameter could also account for the uncertainty of the time dimension in MPIDs, based on the uncertainty in the diagnosis reference points described in Section 8. Therefore the aspect of tuning the bandwidth parameter for performance is a topic that requires further research.

The model has only been evaluated with a horizon configured to 30 days. As discussed in Section 8, the chosen length of the horizon can influence the performance of the model, since it might cause analysis samples containing relevant information to be discarded. Since the horizon determines the amount of information available to use for prediction of illnesses, the relationship between the length of the horizon and performance should be investigated further.

Determining illnesses solely based on analysis samples for medical properties might not be possible for every illness. This is because some illnesses are determined based on symptoms which are not detectable in medical properties. Therefore, by making the model capable of utilizing qualitative patient information, this might increase the performance of certain illnesses and thereby increase the overall performance of the model. This qualitative patient information could be gathered from sources such as patient records, medical journals, etc. Making this data available to the model and utilizing it for prediction is a topic for further research.

## 16. Acknowledgements

We, as authors of this paper, would like to thank the following people:

*Nicolaj Søndberg-Jeppesen*, our supervisor, for supervising and giving feedback throughout the entire project.

*Bo Thiesson* for assisting with selection of theoretical methods and technical insight.

*Jacob Høy Berthelsen*, for motivating the project and making it possible through access to the medical data warehouse.

A special thanks to the company Enversion for allowing use of their office space, and for supervising with use context related problems.

## A. Illness Names

28376	DC209 - KRÆFT I ENDETARMEN
28842	DC349 - KRÆFT I LUNGE UNS
35324	DD649 - ANÆMI UNS
36829	DE109 - TYPE 1-DIABETES UDEN KOMPLIKATIONER
43122	DF171 - SKADELIG BRUG AF TOBAK
43434	DF200 - PARANOID SKIZOFRENI
47960	DG473 - SØVNAFNØ
54160	DH833 - STØJSKADE PÅ INDRE ØRE
54297	DH911 - ALDERSBETINGET HØRETAB
54343	DH919 - HØRETAB UNS
54396	DH931 - TINNITUS
54755	DI109 - ESSENTIEL HYPERTENSION
54967	DI208 - ANDEN FORM FOR ANGINA PECTORIS
55288	DI251 - STABIL ANGINA PECTORIS
55324	DI252 - GAMMELT MYOKARDIEINFARKT
55405	DI259 - KRONISK ISKÆMISK HJERTESYGDOM UNS
55765	DI350 - AORTASTENOSE
56751	DI489 - ATRIEFLAGREN EL ATRIEFLIMREN UNS
56775	DI489B - ATRIEFLIMREN
57608	DI652 - OKKLUSION EL STENOSE AF A. CAROTIS U HJERNEINFARKT
57897	DI694 - SENFØLGE EFTER TIDLIGERE APOPLEXIA CEREBRI
58223	DI739A - CLAUDICATIO INTERMITTENS
59080	DI849 - HÆMORROIDER UNS UDEN KOMPLIKATION
59086	DI849 - HÆMORROIDER UNS UDEN KOMPLIKATION
60159	DJ189 - PNEUMONI UNS
61023	DJ441 - KRONISK OBSTRUKTIV LUNGESYGDOM M AKUT EKSACERBATION UNS
61079	DJ449 - KRONISK OBSTRUKTIV LUNGESYGDOM UNS

61139	DJ459 - ASTMA UNS
64820	DK309 - FUNKTIONEL DYSPEPSI UNS
65348	DK449 - DIAFRAGMAHERNIE UDEN ILEUS EL GANGRÆN
66247	DK573 - DIVERTIKULOSE ELLER DIVERTIKULIT I TYKTARM U PERF EL ABSCES
72616	DM170 - PRIMÆR DOBBELTSIDIG KNÆLEDSARTROSE
73311	DM232 - GAMMEL MENISKLÆSION
73392	DM238 - ANDEN LIDELSE I KNÆLED
73612	DM255 - LEDSMERTER
73644	DM258 - ANDEN LEDLIDELSE
74554	DM472 - ANDEN SPONDYLOSE M RADIKULOPATI
74560	DM472 - ANDEN SPONDYLOSE M RADIKULOPATI
75246	DM545 - LÆNDESMERTER UNS
76199	DM754 - AFKLEMNINGSSYNDROM I SKULDER
76232	DM759 - SKULDERLIDELSE UNS
81792	DN811 - CYSTOCELE HOS KVINDE
82890	DN92 - KRAFTIG, HYPPIG OG UREGELMÆSSIG MENSTRUATION
83242	DN950 - POSTMENOPAUSAL METRORAGI
95109	DR100 - AKUTTE MAVESMERTER
95140	DR102 - MAVESMERTER LOKALISERET TIL BÆKKEN OG BÆKKENBUND
95189	DR108 - ABDOMINALIA, ANDEN OG IKKE SPECIFICERET
96027	DR319 - HÆMATURI UNS
96203	DR391 - VANDLADNINGSBESVÆR UNS
96944	DR559 - BESVIMELSE EL KOLLAPS

## B. Table Dump

<b>Patient</b>				
id_borger	aendringsdato_sup	DataCentAjourDato	Kommune	Tlf
CPRNummer	BefolkningSted	DiskretionsFlag	LoadTM	UdrejseDato
CPRNummerCrypt	ByNavn	FlygtningeNr	Lokalitet	UdRejseLand
FoedDato	CivilDato	Foedested	Paaroerende1	YderAdresse
navn	CivilStatus	FoedselsAarhundrede	Paaroerende2	YderNavn
Adresse	CoNavn	FolkeKirke	SocialDistrikt	YderNummer
PostBy	CPRAegtefaelle	ForskBeskytDato	StatusDato	YderPostNr
PostNr	CPRgaeldende	FraFlytDato	StatusKode	
AdrBeskytSlut	CPRNummerMor	FraKommune	Stilling	
AdrBeskytStart	CPRSup	HjemStedKommune	SuppDato	
aendringsdato_opl	CPRtype	Koen	TilFlytDato	

<b>Diagnosis</b>				
id_diagnose	level1Kodetekst	level5Kodetekst	level9tekst	GRP23
DiagnoseKode	level1tekst	level5tekst	AFLPOS	GRP99
DiagnoseKodeTekst	level2kode	level6kode	DelKode	AendringsDato
DiagnoseSKSType	level2Kodetekst	level6Kodetekst	AmtAnd	ORIG_FRADAT
DiagnoseTekst	level2tekst	level6tekst	BiOperPris	ORIG_TILDAT
DiagnoseType	level3kode	level7kode	IndGrepPris	LoadDT
Gyldig	level3Kodetekst	level7Kodetekst	KodeType	DiagnoseKodeTekstHist
GyldigFra	level3lkode	level7tekst	Koen	DiagnoseTekstHist
GyldigTil	level3tekst	level8kode	LnkNbr	level1Kodetekst_hist
Lukket	level4kode	level8Kodetekst	DiagGruppe	level1tekst_hist
MaxAlder	level4Kodetekst	level8tekst	OperGruppe	level2Kodetekst_hist
MinAlder	level4tekst	level9kode	SKSUdg	level2tekst_hist
level1kode	level5kode	level9Kodetekst	USBPRI	level3Kodetekst_hist

<b>Diagnosis</b>				
level3tekst_hist	level5tekst_hist	level7tekst_hist	level9tekst_hist	IndskrivningsDato
level4Kodetekst_hist	level6Kodetekst_hist	level8Kodetekst_hist	id_Borger	UdskrivningsDato
level4tekst_hist	level6tekst_hist	level8tekst_hist	HenvDato	
level5Kodetekst_hist	level7Kodetekst_hist	level9Kodetekst_hist	HenvAfsDato	

<b>Analysis</b>		
id_AnalysetypeLeverandoer	OevreAcceptgraense	ArbejdsstationNavn
InvestigationsID	NedreOBSgraense	Fra
LeverandoerID	OevreOBSgraense	Til
LeverandoerAnalyseID	Version	GyldigFra
NedreAldersgraense	StandardRelationstegn	GyldigTil
OevreAldersgraense	FastInterval	Gyldig
ContainerInfoID	ErReflekstest	LeverandoerNavn
KliniskInfoID	LeverandoerType	BestiltInvestigationsID
LaboratorieID	LeverandoerBeskrivelse	RekvisitionsID
Analysenummer	LeverandoerAktiv	id_RekvisitionHandler
Analysenavn	LeverandoerMasterInstrument	id_AnalyseHandler
Forkortelse	LeverandoerHarModuler	id_ResultatLab
NumeriskForkortelse	Koen	id_OrgRekvirerendeOrganisation
NPUKode	LeverandoerGraenseSlutAlder	id_ResultatModtagerOrganisation
System	LeverandoerGraenseStartAlder	id_BetalendeOrganisation
SystemSpec	LeverandoerGraenseEnhedaAlder	id_Borger
Prefix	LeverandoerAlarmSymbol	id_AnalyseTypeLeverandoer
Komponent	LeverandoerAlarmKommentar	id_BestiltSamplingsDato
KomponentSpec	LeverandoerAlarmNote	id_BestiltSamplingsTid
Egenart	LeverandoerNedreGraense	id_SampleModtagelseDato
Fremgangsmaade	LeverandoerOevreGraense	id_SampleModtagelseTid
Enhed	LeverandoerDecimalerNedreGraense	id_RigtigSamplingsDato
NedreAcceptgraense	LeverandoerDecimalerOevreGraense	id_RigtigSamplingsTid

<b>Analysis</b>		
id.RapportSamplingdato	id.ResultatReporteresDato	Resultat
id.RapportSamplingTid	id.ResultatReporteresTid	NumeriskResultat
id.FoersteResultatGodkendelsesDato	id.SidsteOpdateringsDato	ResultatEnhed
id.FoersteResultatGodkendelsesTid	id.SidsteOpdateringsTid	InterntSvar
id.SidsteResultatGodkendelsesDato	id.RekvisitionOprettetDato	ReferenceVaerdi
id.SidsteResultatGodkendelsesTid	id.RekvisitionOprettetTid	OevreReference
id.SamplingSlutDato	id.PrioritetStatusArketype	Nedrereference
id.SamplingSlutTid	ReferenceIntervalTypeID	NormaliseretNumeriskResultat

## C. Summary

The danish medical sector digitizes information about conducted medical procedures. These medical procedures include information about patients' analysis samples and diagnoses. Doctors use observations gathered from medical properties as indicators for beliefs about health states of patients. These indications are based on the knowledge of the individual doctors, and are thereby not necessarily shared across all doctors. Identifying the indicators used across all doctors to diagnose a given illness could provide a means of knowledge sharing and improve the understanding of medical procedures. Thereby, this project aims to examine a medical data warehouse in order to test the hypothesis of the existence of a relationships between medical properties and illnesses.

The medical data warehouse available for this project contains information about analysis samples and diagnoses. This data is used to train a non-parametric temporal model with the purpose of predicting the illness of a diagnosis based on a set of analysis samples of a patient within a horizon with better performance than naive prediction methods. This performance is measured through the accuracy of predictions, and if the performance is better than naive approaches this is seen as a confirmation of the hypothesis about the existence of a relationship between medical properties and illnesses.

The model is based on the intuition of a naive Bayes model in which methods for handling temporal aspects are applied. The prediction is based on the use of Kernel Density Estimation for measuring normality of analysis samples in regards to an illness. These normality measures represent a likelihood that a patient is affected by an illness. The model returns a likelihood for each illness. These likelihoods are normalized into a probability distribution over each illness for a patient.

The evaluation of the model was performed on three sets of illnesses of different sizes, and for each set of illnesses the model outperformed the naive approaches. This indicates that there exists a relationship between analysis samples and diagnoses, and

thereby confirms the hypothesis. This lead to an investigation of the model's ability to be used in diagnostic procedures, by evaluating its performance as a diagnostic decision support tool. It was shown that the model was capable of filtering out a subset of illnesses, which had a large likelihood for containing the correct illness.

## References

- Bowen, Richard A., "Growth Hormone and Aging," <http://arbl.cvmb.colostate.edu/hbooks/pathphys/endocrine/hypopit/ghaging.html> 2006. Accessed: 2014-03-10.
- Boyesen, Mathias Abitz and Julian Birkemose Nielsen, *Dimensionel modellering af laboratoriedata og implementering af modellen i et datavarehus*, Aalborg University, 2013.
- Cain, Peter, Ragnhild Ahl, Erik Hedstrom, Martin Ugander, Ase Allansdotter-Johnsson, Peter Friberg, and Hakan Arheden, "Age and gender specific normal values of left ventricular mass, volume and function for gradient echo magnetic resonance imaging: a cross sectional study," *BMC Medical Imaging*, 2009, 9 (1), 2.
- Cam, L. Le, "Maximum likelihood: An introduction," *Intl. Stat. Rev.*, 1990, 58, 153–171.
- Cancer Research UK, "More About Staging for Lung Cancer," <http://www.cancerresearchuk.org/cancer-help/type/lung-cancer/treatment/more-about-lung-cancer-staging> 2014. Accessed: 2014-05-18.
- Diamond, George A. and James S. Forrester, "Analysis of Probability as an Aid in the Clinical Diagnosis of Coronary-Artery Disease," *New England Journal of Medicine*, 1979, 300 (24), 1350–1358. PMID: 440357.
- Doshi-velez, Finale, Shakir Mohamed, Zoubin Ghahramani, and David A. Knowles, "Large Scale Nonparametric Bayesian Inference: Data Parallelisation in the Indian Buffet Process," in Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams, and A. Culotta, eds., *Advances in Neural Information Processing Systems 22*, Curran Associates, Inc., 2009, pp. 1294–1302.
- Fisher, R.A., *Statistical methods for research workers*, Edinburgh Oliver & Boyd, 1925.
- Hansen, Bruce E., "Lecture Notes on Nonparametrics," <http://www.ssc.wisc.edu/~bhansen/718/NonParametrics1.pdf> 2009. Accessed: 2014-02-25.
- Holdcroft, Anita, "Gender bias in research: how does it affect evidence based medicine?," *Journal of the Royal Society of Medicine*, 2007, 100.
- Jensen, Finn V. and Thomas D. Nielsen, *Bayesian Networks and Decision Graphs*, 2nd ed., Springer Publishing Company, Incorporated, 2007.
- Jones, M. C. and H. W. Lotwick, "Remark AS R50: A Remark on Algorithm AS 176. Kernel Density Estimation Using the Fast Fourier Transform," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 1984, 33 (1), pp. 120–122.

- Lesne, Annick, "Shannon entropy: a rigorous mathematical notion at the crossroads between probability, information theory, dynamical systems and statistical physics," <http://www.lptmc.jussieu.fr/user/lesne/MSCS-entropy.pdf> 2011. Accessed: 2014-03-19.
- Manning, Christopher D, Prabhakar Raghavan, and Hinrich Schütze, *Introduction to information retrieval*, Vol. 1, Cambridge university press Cambridge, 2008.
- Masisi, L, V Nelwamondo, and Tshilidzi Marwala, "The use of entropy to measure structural diversity," in "Computational Cybernetics, 2008. ICC 2008. IEEE International Conference on" IEEE 2008, pp. 41–45.
- McDonald, John H, *Handbook of biological statistics*, Vol. 2, Sparky House Publishing Baltimore, MD, 2009.
- Middleton, Blackford, Michael Shwe, David Heckerman, Max Henrion, Eric Horvitz, Harold Lehmann, and Gregory Cooper, "Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base PART II," *Medicine*, 1991, 30, 241–255.
- MUSC, "Student Health Influenza Information," <http://academicdepartments.musc.edu/studenthealth/flu/duration.htm> 2011. Accessed: 2014-02-25.
- National Cancer Institute, "NCI Dictionary of Cancer Terms," <http://www.cancer.gov/dictionary?cdrid=45618>. Accessed: 2014-05-28.
- National Eye Institute, "Facts About Diabetic Retinopathy," <http://www.nei.nih.gov/health/diabetic/retinopathy.asp>. Accessed: 2014-03-18.
- Opitz, David and Richard Maclin, "Popular Ensemble Methods: An Empirical Study," *Journal of Artificial Intelligence Research*, 1999, 11, 169–198.
- Panum, Thomas Kobber and Esben Pilgaard Møller, *FALCON: Feature Analysis for Diverse High-Cardinality Multiclass Classification*, Aalborg University, 2013.
- Parzen, Emanuel, "On Estimation of a Probability Density Function and Mode," *The Annals of Mathematical Statistics*, 1962, 33 (3), pp. 1065–1076.
- Russell, Stuart and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed., Upper Saddle River, NJ, USA: Prentice Hall Press, 2009.
- Shwe, Michael A, B Middleton, DE Heckerman, M Henrion, EJ Horvitz, HP Lehmann, and GF Cooper, "Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base PART I," *Meth. Inform. Med*, 1991, 30, 241–255.

Silverman, B. W., "Algorithm AS 176: Kernel Density Estimation Using the Fast Fourier Transform," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 1982, 31 (1), pp. 93–99.

Silverman, Bernard W, *Density estimation for statistics and data analysis*, Vol. 26, CRC press, 1986.

Statens Serum Institut, "Hvad er SKS?," <http://www.ssi.dk/sks>. Accessed: 2014-05-09.

Stedman, Thomas Lathrop, *Stedman's Medical Dictionary*, Lippincott Williams & Wilkins, 2000.

Strimbu, Kyle and Jorge A Tavel, "What are biomarkers?," *Curr Opin HIVAIDS*, 2010, 5 (6), 463–6.



## Declaration

*Aalborg University, June 4th 2014*



---

Bjarke Hesthaven Søndergaard



---

Esben Pilgaard Møller



---

Thomas Kobber Panum

