

LU-TP 22-40
June 2022

**Autoencoder outlier detection
during the Covid-19 vaccination campaign**

Thomas Eriksson

Department of Astronomy and Theoretical Physics, Lund University

Master thesis supervised by Anders Björkelund and Mattias Ohlsson



LUND
UNIVERSITY

Abstract

Outlier detection in high-dimensional data is a complex task, useful in many fields. One major application is in healthcare, where the high-dimensional healthcare registers can be used to detect patterns related to the medical behaviour and state of the population. In this project, neural network based autoencoders were used to study the Covid-19 vaccination campaign, where it was trained on healthcare data from unvaccinated individuals. Outliers were selected by thresholding the reconstruction error from the autoencoder, both for vaccinated and unvaccinated individuals, to compare the data points. The outliers were in principle selected based on the dimensionality of their input vector, and the frequency of representation in the training data set. The previously known myocarditis cases in the young male population, as a side effect of the mRNA Covid-19 vaccines, was used to model the method. The method detected various patterns, mainly vaccination patterns of the population and the vaccine effect. Outlier detection in high-dimensional healthcare registers can possibly detect patterns before statistical models, and search the data for non-linear correlations.

Populärvetenskaplig beskrivning

Data inom hälsovården kan användas för att hitta mönster för att exempelvis ta fram hur många som är sjuka, eller hur många som har fått ta blodprov. Det används ofta statistiska modeller för att göra undersökningar, men i detta projektet används artificiell intelligens, och mer specifikt neuronät. Då tar vi hjälp av artificiell intelligens för att hitta mönster inom datakällor som kan vara svåra att hitta med vanliga statistiska modeller.

Inom detta projekt utforskas en metod för att leta efter mönster inom hälsovårdsdata. Specifikt försöker vi hitta avvikande värden för de som har vaccinerat sig emot Covid-19. Detta gör vi i ett försök att hitta biverkningar av vaccinet, och för att undersöka andra underliggande mönster. Sverige har väl dokumenterad hälsovård, och projektet har tillgång till Skånes hälsoregister. Därför finns det förutsättningar att i den stora datamängden göra nya upptäckter. Den stora datamängden måste hanteras på lämpligt sätt, så att den tillgängliga datorkraften räcker till.

Den artificiella intelligensen som används är ett neuronät, och mer specifikt en autoencoder. För att se vilken data som sticker ut, får en autoencoder ta emot data och återskapa den. Hur den ska återskapa datapunkterna har den lärt sig ifrån ett valt data set. Hur väl den kan återskapa data blir ett mått på vilka mönster autoencodern har lärt sig, och vidare vilka mönster datapunkten passar in i.

Contents

1	Introduction	6
2	Related work	7
2.1	Outlier detection	7
2.2	Autoencoders as outlier detectors	8
2.2.1	Conventional autoencoder description	8
2.2.2	Early deep learning outlier detector	9
2.2.3	Learning feature representation of normality	9
2.2.4	Autoencoder ensemble	10
3	Method	10
3.1	Data management	10
3.1.1	Available and selected data	10
3.1.2	Sampling for <i>control</i> data sets	14
3.1.3	Myocarditis as signal event	15
3.2	Method outline	17
3.3	Autoencoder	17
3.3.1	Autoencoder as outlier detector	17
3.3.2	Hyperparameter selection	19
3.3.3	Outlier threshold	21
3.3.4	Autoencoder ensemble	24
3.4	Outlier analysis	24
3.4.1	Feature based analysis	24
3.4.2	Group based analysis	25
4	Results	25
4.1	Autoencoder as outlier detector	25
4.2	Outlier analysis	25
4.2.1	Feature-based analysis	25
4.2.2	Group based analysis	30

4.2.3	Summary of myocarditis results	32
5	Discussion and conclusions	33
5.1	Limitations	33
5.2	Future work	34

List of acronyms

AE Autoencoder
DL Deep learning
OD Outlier detector

List of Figures

1	Sketch of an autoencoder. This sketch provides a conceptual image of an autoencoder. The autoencoder tries to reproduce an input vector, first being encoded down to a lower-dimensional representation in the bottleneck layer, and then decoded to the output layer.	9
2	Each data point was defined by its vaccination date. The medical tests, medical procedures, and diagnoses were added to create the data point. Medical tests, medical procedures and diagnoses were collected for 30 days after the vaccination. The diagnoses history was also collected for 30 days prior to the vaccination date.	11
3	Flow chart of the data selection process. The top layer shows the data sets used. Parts of them were selected and used to create the next layer, a complete data point with all of the selected information. From the complete data point, an input vector was created for the autoencoder, with the wanted format and type for training the autoencoder.	12
4	Lower-dimensional (one-dimensional) data in a higher-dimensional graph (two-dimensional). In the graph lower dimensions are visualized in a higher-dimensional space. The same concept can be applied on higher dimensions and is a fundamental idea behind an autoencoder.	15
5	Illustration of the sampling process. For each <i>case</i> data point, the age, gender and sickness level is used in order to sample two unvaccinated points with the attributes. These are then placed in $control_{train}$ and $control_{eval}$	16

6	Method outline. This image shows the main parts of the project. The circles are data sets and the squares are algorithms. The autoencoder is trained, the outlier sets created and then analysed.	18
7	The autoencoder. The autoencoder used for the results in Section 4. . .	20
8	Training/validation loss graph The loss on the y-axis is the loss function in equation 4.7. Note the separation between training and validation loss as the training continues.	22
9	The reconstruction errors with (a) the loss function used during training, (b) MSE which was the choice for outlier thresholding and (c) the F1-score for the first output layer. (d) Shows the mathematical difference between log loss and MSE.	23
10	F-1 score for medical tests. Measurement of how well the autoencoder can reconstruct the medical tests. High frequency and lower dimensions should yield better reconstruction. Points of interest noted with circles. . .	28
11	F-1 score for medical procedures. Measurement of how well the autoencoder can reconstruct the medical procedures. High frequency and lower dimensions should yield better reconstruction. Points of interest are the worst reconstructed medical procedures, and the best reconstructed procedures.	29
12	Visualization of outliers using t-SNE. All of the $case_{outlier}$, blue dots, and $control_{outlier}$, red dots, data points are visualized with t-SNE. The outliers bottleneck layer representation is used along with t-SNE. The myocarditis cases for vaccinated young males can also be seen, with the larger green dots. Various regions were selected for further investigation (i-v).	31
13	Moderna outliers using t-SNE. The data points latent space representation visualized with t-SNE. a) Shows the selected regions of interest (i-v). b) Shows the age. c) Shows the gender. d) Shows the sickness level.	32

List of Tables

1	The source data sets used.	12
2	The data sets created containing the data points and the number of myocarditis cases.	14
3	The five output layers with their activation function, corresponding loss function, and the number of features per layer.	21
4	Outlier data sets.	25
5	Most over-represented and under-represented medical tests in $case_{outlier}$, relative to $control_{outlier}$	26

6	Most over- and under-represented medical procedures in $case_{outlier}$, relative to $control_{outlier}$	27
---	---	----

1 Introduction

As the Covid-19 pandemic swept the world with an unseen force for modern times, the need of a vaccine became critical. Fortunately, the pharmaceutical industry managed to create multiple successful vaccines, which help us manage the pandemic. The effect of the vaccine is undoubtedly positive both for the society and for the individual, weighing risks of infection and side effects. The vaccines had an enormous pressure from society to be delivered in shorter time than the normal procedure. The normal timeline for developing a vaccine is an order of magnitude greater than it took to produce the Covid-19 vaccines. Although rigorous testing was done for the Moderna vaccine [1] [2], some rare but serious side effects could still be noted after the vaccination campaign had started. Perhaps the most severe side effect was the myocarditis and pericarditis, or heart inflammations, within the young male adult population.

The vaccines have still gone through extensive testing, building a statistical foundation of which to base the safety of the vaccines. Still, as the general searches for vaccine side effects will be done in a classical statistical and mathematically linear sense, it begs the question if there might be more side effects for certain subgroups in society. Therefore, it can be useful to develop alternative side effect searches, which can be done as the vaccines are rolled out during the next eventual pandemic. To complement the classical rigorous search, there would be a non-linear search for general healthcare changes. These would have low evidential credit of themselves, but would be used as a searching tool to indicate possible additional side effect-subgroup pairings, which then could be further investigated using classical statistical methods.

This project sets out to test a method, as the first step in creating a non-linear search for vaccine side effects using deep learning. More specifically, using autoencoders to detect individuals with side effects as outliers within the population. The method was developed during the course of the project, and while the method has not yielded any newly discovered vaccine side effects, a lot of knowledge can be found of which to base new methods on. The project is within the *COVERS* initiative [3] and had access to the healthcare data for the population of Scania, Sweden. The population of Scania's contact with the healthcare system for a time after vaccination was studied.

The project detected outliers among the population. Further, the project highlighted various healthcare patterns, and the source of these patterns were speculated about. Part of the patterns could be assumed to be side effect related, while others were more plausibly related to alternative contributing factors, such as the vaccination pattern of the population or the vaccination effect to decrease Covid-19. One of the use cases of the project was to search for myocarditis as a side effect.

2 Related work

2.1 Outlier detection

While this project will focus on outlier detection (OD) using deep learning (DL), it is useful to briefly talk about OD methods in general, in order to understand why DL methods can be useful. Aside from DL techniques there are both statistical methods, as well as non-DL machine learning methods. Hodge et al. (2004) [4] present a survey of various outlier detection techniques. In this section various categories and methods of ODs will be discussed.

Outlier detectors do not have a universal model which is always optimal but it is often a matter of finding the right tool for the task. According to the survey, there are three broader types of outlier detectors. The first is outlier detection without any knowledge of the data sets used. There is no label whether a data point is an outlier or not. The second type of model have labels on both the outliers and non-outliers. The labels are used to train a model to distinguish outliers from non-outliers. The third type is when there are only non-outliers to find a pattern of normality from, and if any new data points do not follow the pattern, they would be considered outliers.

The statistical methods can be applied in various ways, but are often based on calculating distances or proximities. The simplest form of an OD method would be to plot the data and look at what data points would be considered outliers, by calculating the extremities of the data set. This is a straightforward method in low-dimensional data, but becomes harder as the data dimensionality increases. With higher data set dimensionality, the normal data set, and its convex hull will be harder to define. This is known as the curse of dimensionality. There are various methods to reduce the dimensionality of data sets in order to simplify the OD.

In distance based methods, the distance between data points is calculated and used to define outliers. An example would be to look at the k -nearest neighbours, which takes the distance between data points and consider the data point an outlier if not at least k other data points are within a threshold distance. As this does not scale well computationally with increasing number of data points or dimensions, it would not be suitable for this project.

Thudumu et al. (2020) [5] presents *A comprehensive survey of anomaly detection techniques for high dimensional big data*. They mention that there is more work needed for high-dimensional outlier detection, and that there are a lot of challenges. One main challenge is that high-dimensional data increases sparsity of data points. The sparsity in turn results in more equal distance between data points, and lower density. These are the attributes that many of the statistical models are based upon. Visualization methods have problems in high-dimensions and need dimensional reduction techniques. One of them is the principle component analysis, which can be done linearly or non-linearly [6]. Dimensional reduction

can find the intrinsic dimension of the data, but it can also remove information needed for the outlier detection. Another high-dimensional visualization method is the t-distributed stochastic neighbour embedding (t-SNE) [7].

One method that is available in the *Scikit-learn* library [8] is the isolated forest method. The method isolates outliers instead of profiling normal data points. The method uses the fact that outliers are easier to isolate than normal points. It takes the data set and divides it into two parts by selecting a random feature and a random threshold value. This procedure is recursively repeated until data points are isolated, or a pre-defined limit is reached, either way creating a tree. The number of divisions that are needed to isolate a data point gives the outlier score for the same data point. This training procedure is done many times to get an ensemble of trees and receive stable outlier scores. The method works well by sub-sampling when training on large data sets, which reduces the processing time. The curse of dimensionality is still present in the isolated forest method, but by choosing important features to threshold on, the problem can be reduced [9].

2.2 Autoencoders as outlier detectors

2.2.1 Conventional autoencoder description

The autoencoder (AE) was introduced by Rumelhart and McClelland (1987), in *Learning Internal Representation by Error Propagation* [10], which presented a solution to a problem proposed by Ackley, Hintion and Sejnowski (1985) [11] known as the encoding problem. By using an autoencoder structure, a more compact form of data could be learned. Starting off with one-hot encoded data of 8 types, and training the artificial neural network to match the output to the input, a more compact representation of the data could be found. The autoencoder learns to go from one-hot encoded data to binary in the encoding part of the autoencoder. It also learns to go from binary back to one-hot in the decoding part. It is done by updating the weights in the autoencoder with backpropagation. The autoencoder learns representation of the data used for training, in the discussed example a relatively simplistic transformation between data types, but more complex structures can also be stored in the autoencoder.

A straightforward explanation of the autoencoder can be found in most articles discussing autoencoders [12] [13] [14] [15]. The AE consists of two parts, the encoding part, which takes an input and reduces its dimension, and the decoding part which takes the low-dimensional representation to restore it to the high-dimensional state. The restoration can be done in various ways, most commonly by training neural network layers to match the input layer and the output layer. In-between the encoder and decoder, there will be a layer of nodes where the lowest representation will be present. This can be called the bottleneck layer, latent space or the bridge between encoder and decoder. In Figure 1 a sketch of an autoencoder can be seen. Section 3 presents the basis of working with autoencoders as outlier detectors.

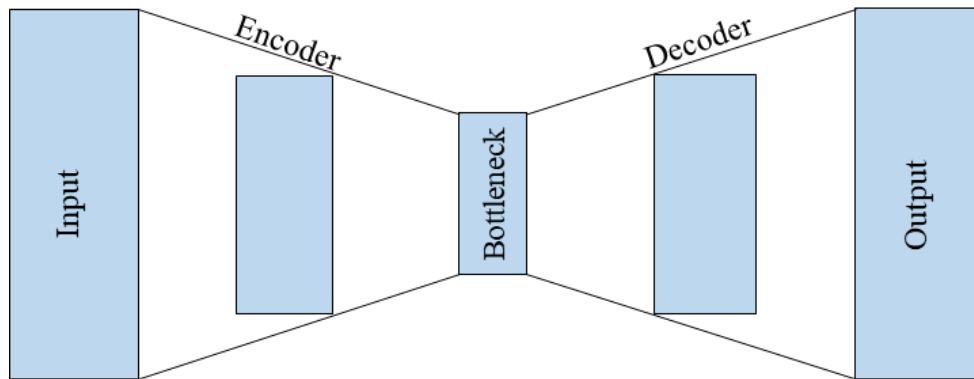


Figure 1: **Sketch of an autoencoder.** This sketch provides a conceptual image of an autoencoder. The autoencoder tries to reproduce an input vector, first being encoded down to a lower-dimensional representation in the bottleneck layer, and then decoded to the output layer.

2.2.2 Early deep learning outlier detector

Hawkins et al. (2002) [16] were the first to present a method similar to the autoencoder outlier detector. Their method relies on a replicator neural network, which is close to the design of the autoencoder. The input layer receives the vector with data, which goes through three hidden layers, and then the network tries to replicate the received input values after the output layer. The bottleneck layer has a specific activation function, creating a step-wise design. Each input has a quantized value in the bottleneck layer, yielding a form of clustering. While the clustering can be used, the main application is to use the reconstruction error to choose the outliers, much as the autoencoder.

2.2.3 Learning feature representation of normality

Pang et al. (2021) [17] presents three categories of deep learning outlier detectors. The first is *Deep learning for feature extraction*. That is, utilizing deep learning to give the data attributing features. Then, separate processes would be required to gain outlier scores. In contrast, the entire process could be done by deep learning with *End-to-end Anomaly Score Learning*. Given input, the output will be an anomaly scoring output function. Finally, which is the method used in this project, deep learning can be used by *Learning Feature Representation of Normality*. These methods will learn normality in a control data set in order to assess the anomaly [17]. For this project, normality would entail the parts of the population without recent Covid-19 vaccinations. As would be expected, the chosen approach of these three will have a fundamental impact on the project's method. The chosen path was as mentioned mainly *Learning feature representation of normality*, but intertwined with *Deep learning for feature extraction*, as broader methods can intersect.

2.2.4 Autoencoder ensemble

Instead of training just one autoencoder and using it to make predictions, Chen et al. (2017) [13], propose to use many autoencoders, all built slightly differently. This is achieved by dropping some of the connections in the autoencoder. The goal is to reduce the effect of overfitting that one autoencoder will be prone to. If the same autoencoder is used instead of different ones, the same overfitting might occur. By using an ensemble of autoencoder ODs and taking the median or average of the trained values, overtraining can be countered. In the article many additional techniques can be found, such as adaptive sampling, adaptive learning rate and pre-training. These make the training faster and increase the probability of finding the global optimum.

3 Method

3.1 Data management

In order to understand the method, an understanding of the data is first needed. At the start of the project, several data point definitions were contemplated. The definition of the data point is central as the methodology must be changed based on it. The data point that was used was defined as a vaccination occurrence. For each vaccination occurrence, a vector of data was gathered. The vector contained information about the individuals healthcare for a time after the vaccination, with the assumption that information about side effects was also stored in the data points. In Figure 2 an illustration of an individual data point is shown.

3.1.1 Available and selected data

When selecting data, there were a couple of considerations. The selected data should plausibly have information related to side effects, otherwise it should not be selected. As the computational power is limited, the data needs to be limited as well. If it is unlikely that a feature has valuable information about side effects, it should be omitted. Across the source data sets some of the information is repeated, which is unnecessary. While myocarditis is used as a model signal event, the search should be general, and as such, data should be selected for any possible side effect and not just for myocarditis.

The first step was going from different data sets to one large matrix with each row consisting of a data point. Some of the matrix columns were then used as input in the autoencoder, while other columns were only used to correctly identify the data point. To have a matrix ready for the autoencoder, the columns used as inputs were selected into a separate matrix as well. This concept is shown in Figure 3. Table 1 shows the number of data points, and how many individuals there are in each source data set.

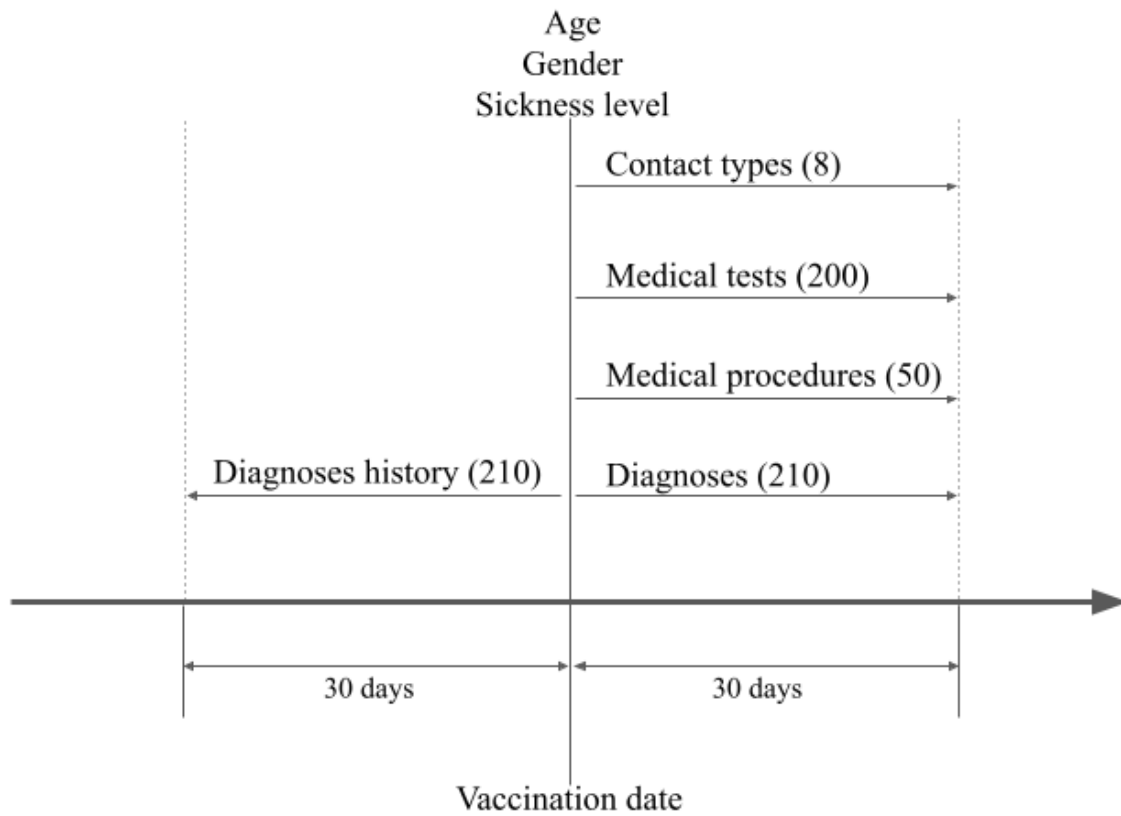


Figure 2: Each data point was defined by its vaccination date. The medical tests, medical procedures, and diagnoses were added to create the data point. Medical tests, medical procedures and diagnoses were collected for 30 days after the vaccination. The diagnoses history was also collected for 30 days prior to the vaccination date.

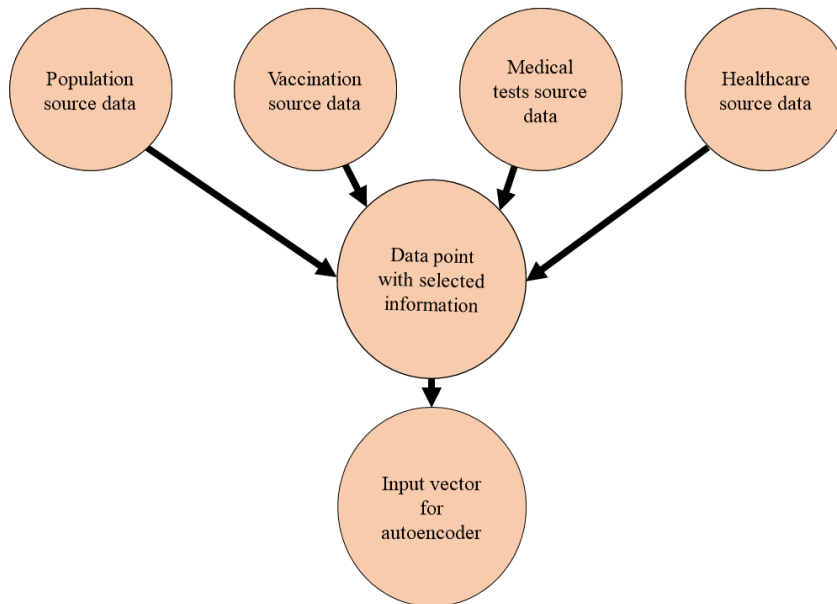


Figure 3: **Flow chart of the data selection process.** The top layer shows the data sets used. Parts of them were selected and used to create the next layer, a complete data point with all of the selected information. From the complete data point, an input vector was created for the autoencoder, with the wanted format and type for training the autoencoder.

Table 1: The source data sets used.

Type of source data set	# Data points	# Individuals
Population	1,385,479	1,385,479
Vaccination	2,417,287	998,156
Medical tests	56,092,732	561,995
Healthcare	13,687,565	1,159,587

The first data set contains general population data, such as gender, age, country of birth and so forth. From this data the gender and age were selected. A constant age for each patient was defined as by how old they were when the vaccinations against Covid-19 started on December 27th 2020. Information of socio-economical types were omitted, to limit the scope of the project. Ages were normalized between zero and one, and gender was one-hot encoded.

The second data set contains Covid-19 vaccinations, including time and type. Each vaccination occurrence was used as the foundation for a data point, defining the temporal limits of what should be selected from the other data sets. The type of vaccine is highly relevant in the search of side effect but since a *control_{train}* data set will be created to train the autoencoder, it cannot be included in the input vector. The vaccination data set contained

information about the batch number of the vaccine, but since a search for general vaccine side effects was the goal, and not a search for defect batches, the batch number was not selected for the data point array. No information from the vaccination data set is used in the actual input vector to the AE, but are used in the data management to create data points, and in the outlier analysis.

The third data set included medical tests, mostly blood tests. From these, the most commonly used were selected. Covid-19 tests were removed, most individuals in the healthcare system would have been tested for Covid-19 and including it would create unnecessary noise. Two types of troponin test were merged into one feature, in an attempt to eliminate noise for myocarditis related features. Troponin is a biomarker for cardiovascular diseases. The medical tests were counted for 30 days after vaccination, and then binarized. In total 196 blood tests were selected.

The fourth and final data set contains data for when patients received healthcare. There were many various variables, but in order to keep the number of features down, the selected attributes were limited to the type of contact that the patient made with the healthcare, diagnoses set by doctors, and medical procedures. There were eight types of contact with the healthcare, each type of contact was counted and normalized between zero and one. The number of times the patient received in-patient care during the last six months defined the sickness level. The three levels selected of zero times, one time or more than one time were one-hot encoded.

The diagnoses were labelled by the ICD-10 system, and to keep the number of features down, each ICD-10 code was not an individual feature, but was considered to be within a range of ICD-10 codes. Each ICD-10 code gave a count increase in its corresponding range. In the input vector the counters were binarized. The ICD-10 ranges that had no values in the entire data set were removed as features. The diagnoses were collected 30 days after vaccination, but also 30 days prior to vaccination, in order to give the AE features that reflect the status of the patient prior to vaccination. The final result for the diagnoses were 210 binarized counters of the ICD-10 ranges, and the same for the diagnoses history. Two specific codes were also added, for Covid-19 and myocarditis.

The medical procedures, encoded by KVÅ-codes ¹, were sorted by frequency. The 50 most common and relevant KVÅ-codes were selected. These were not ranges as ICD-10 codes, but the direct codes for the procedures. The number of occurrences were counted for 30 days after vaccination, before being binarized. The result was a vector of 50 features. Note that the only information that is added is if a specific procedure has been done, and not the results of the procedure. Nevertheless, the fact that an examination has been done is in itself an indicator of a suspicion by the medical expertise.

Within the population, only individuals between 16 and 65 years old, at the start of vaccination in Scania (2020-12-27), were accepted as data points. The lower limit of 16,

¹In contrast to ICD-10, which is an international coding, KVÅ is a national coding system defined by *The Swedish National Board of Health and Welfare*.

is selected based on that the vaccination in Sweden was divided into age groups, and 16 was the lowest of which complete records were obtained. The higher limit of 65 was selected in order to decrease the computational time, and in an attempt to simplify the project. Individuals over 65 years would be considered more complex data points and less predictable for side effect of vaccines. Moreover, ages above 65 were presented by the Swedish healthcare agency as significant for severe Covid-19 [19].

Three data sets were constructed, the first contained the *cases*, where an individual had just received a vaccine against Covid-19. At the same time, two *control* data sets were created, one used for training, *control_{train}* and one used for evaluation of the outliers, *control_{eval}*. The project was modelled for a search of myocarditis as a side effect of the Covid-19 vaccine, and as a result, the number of myocarditis cases in each of the data sets was highly relevant, see Table 2.

Table 2: The data sets created containing the data points and the number of myocarditis cases.

Type of data set	# Data points	# Myocarditis, all ages	# Myocarditis, age 16-24
<i>Case</i>	1,399,321	41	12
<i>Control_{train}</i>	1,399,206	38	8
<i>Control_{eval}</i>	1,399,321	45	5

Both the general data set and each data point can be considered to have a dimensionality. The data set might follow patterns correlating certain dimensions. A data point has an intrinsic dimensionality of its own, still relative to the data set, proportional to the number of uncorrelated features. An example of lower-dimensional data points in a higher-dimensional space can be found in Figure 4.

3.1.2 Sampling for *control* data sets

For each *case* data point vector collected, two corresponding *control* data point vectors were needed, as shown in Figure 6. One was needed for training, and one to select outliers to use as normalization in the outlier analysis. The *case* data set consist of vaccinated data points. To get a valid *control* data set, matching a number of attributes was required. Therefore, for each data point, two random persons who did not get the Covid-19 vaccine, neither one month before nor after the specific date, were selected. The sampled persons had the same gender, age and sickness level. As a large data set was obtained, only one *control_{train}* data point was selected per *case* data point for training. In Figure 5, the process of sampling points is illustrated.

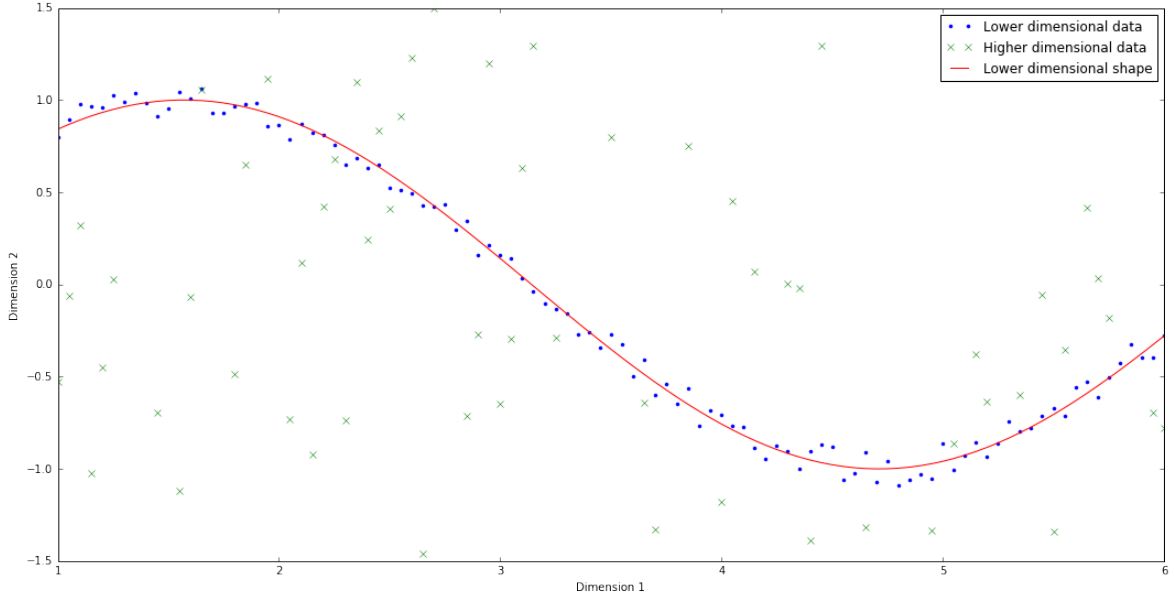


Figure 4: **Lower-dimensional (one-dimensional) data in a higher-dimensional graph (two-dimensional).** In the graph lower dimensions are visualized in a higher-dimensional space. The same concept can be applied on higher dimensions and is a fundamental idea behind an autoencoder.

3.1.3 Myocarditis as signal event

One of the most significant side effect of the Covid-19 mRNA vaccines is considered to be the myocarditis cases occurring at a higher frequency within the young male population [18]. Myocarditis is an inflammation of the heart muscle. This side effect was first discovered after parts of the population were vaccinated, and as a result the data contains vaccine-induced cases of myocarditis. Karlstad et al. (2022) [18] clearly show that there are myocarditis cases in the nordic countries and that it occurs more frequently after a second vaccination dose of mRNA vaccine. Creating a method that can find the myocarditis cases, and using the same method for a general search was one of the corner stones of this project. However, there are in total very few cases to constitute a signal on which to base the method on. In the *case* data set, there are a total of 12 myocarditis in young males, while there are 8 in the *control_{train}* and 5 in *control_{eval}*.

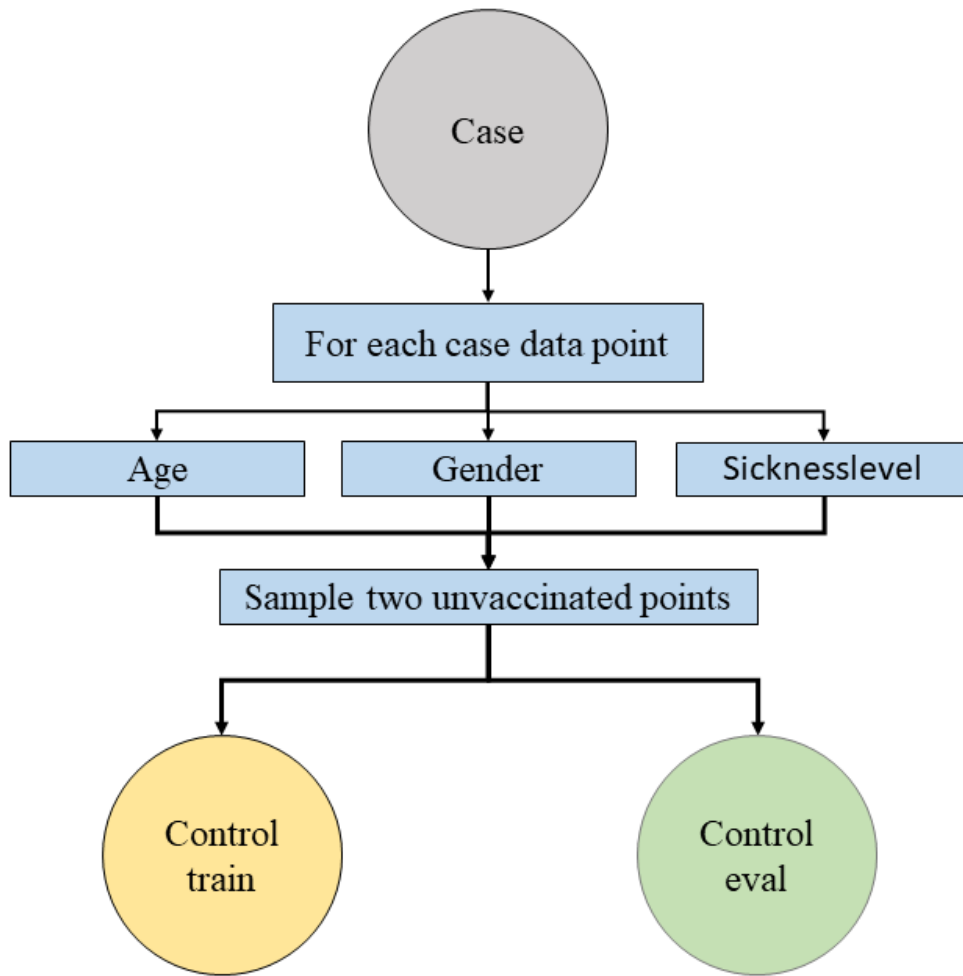


Figure 5: **Illustration of the sampling process.** For each *case* data point, the age, gender and sickness level is used in order to sample two unvaccinated points with the attributes. These are then placed in $control_{train}$ and $control_{eval}$.

3.2 Method outline

There are three distinct parts to this project. The first entails management of the data from the data sources, and creating usable vectors with a subset of the total data. This part have been presented in Section 3.1. The second part is the outlier detection which selects a part of the data as outliers. In this project an autoencoder was used, but similar work can be done with various outlier detection methods, some of them mentioned in Section 2. The third part is the analysis of the outliers, going from a selected outlier data set to information about the Covid-19 vaccine patients. A central part of the project is the definition of a data point, which will influence the rest of the project.

Figure 6 shows the general structure of the project, containing the mentioned three parts. In order to train and use the autoencoder, the input data must be selected. Similarly, to analyse the outliers, they must be selected as such. As a result, these three parts are intrinsic to the project and each part needs to be completed before moving on to the next. Each of these parts are noted with a box in Figure 6. Various combinations of data sets and outlier sets were tested in preliminary experiments, before a decision was made to move on with one of them. As many decisions had to be made throughout the project, it was not possible to continuously go back and re-define the entire project.

The autoencoder uses the $control_{train}$ data set to train, resulting in a trained AE as can be seen in Figure 6. The $control_{train}$ is also used both as an input and reference when training the AE. Now the trained autoencoder is completed, and the data sets $case$ and $control_{eval}$ can both be used to get reconstruction by running them through the autoencoder. To assess the reconstruction, mean squared error (MSE) was used, and values over a certain threshold were selected as outliers. These outliers were then used directly in ratio of inputs in order to analyse the outliers. The outliers were also analysed by taking their encoded counterpart found in the bottleneck layer in the AE, and then visualized by using t-SNE.

3.3 Autoencoder

3.3.1 Autoencoder as outlier detector

In Section 2, a few methods of outlier detection were discussed, with a focus on autoencoder as outlier detectors. This project was fundamentally different from most projects discussed in the theory section, as side effects are considered the outliers, but strictly their input vector may not be outliers, as the same sickness may occur regardless in the $case$ data set. Therefore, it is the number of patients with a sickness in certain parts of the population, relative to a control or normal value, that would be the outlier. As such, the search is not as trivial as to train an autoencoder on the $control$ data, and use the signal data on the trained autoencoder to detect outliers. Instead, the method of finding high-dimensional outlier data points of both the $case$ and $control_{eval}$ was constructed. The $control_{eval}$ data is the sampled unvaccinated data points not used for training, while the $case$ data set is

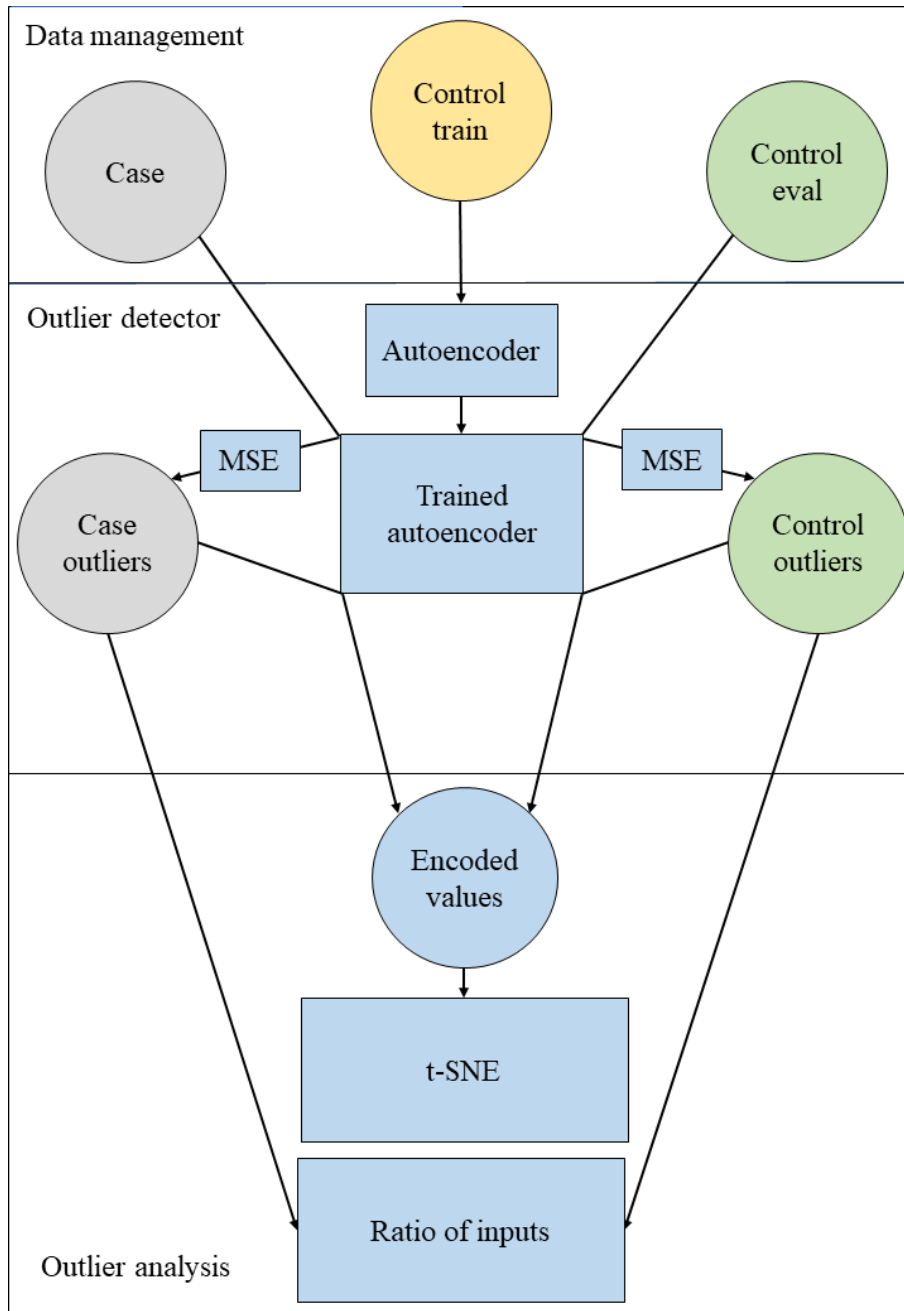


Figure 6: **Method outline.** This image shows the main parts of the project. The circles are data sets and the squares are algorithms. The autoencoder is trained, the outlier sets created and then analysed.

the vaccinated data points.

The high-dimensional outlier data points will not be reconstructed well with a low bottleneck layer dimensionality, and as such, have a higher reconstruction error. These can then be collected for further analysis. The AE reconstruction loss will mainly depend on two parameters, the first being how high-dimensional the data points are. The second being the frequency of those data points in the *control_{train}* data set, that will also determine how much training the AE will receive for such a data point. By looking at the high-dimensional, low frequency data points, the more irregular side effects should be detectable. It follows that low-dimension, high frequency data should be reconstructed well. However, the balance between frequency and dimensionality is not known and, as a result both should be considered when creating a methodology.

Changing the dimensionality of the AE model is straightforward, as the number of nodes at the bottleneck layer creates an upper limit. The frequency of data points can be affected by using the *control_{train}* data set for training, instead of *case* data set. As the project idea was a blind search, the *case* data set cannot be altered with, as the wanted signal is unknown. The AE would reconstruct the myocarditis cases worse if the non-vaccination related occurrence cases were removed from the *control_{train}* data set, but since that cannot be done in a blind search, there is no point in doing so for the myocarditis cases. After selecting appropriate hyperparameters, and training the AE, the next step was to select the outliers with a reconstruction error.

3.3.2 Hyperparameter selection

This section describes how the AE was modelled. Pang et al. (2021) [17] give an outline of how many layers are suitable, with most of the AE OD built for tabular data having less than five layers in total. For simplicity, only symmetric AE were tested, that is, only AE with the same number of layers and size of layers on each side of the bottleneck layer. The AE can have zero, one, two or three layers between input and bottleneck layer and still remain relatively consistent with other AE OD in literature.

The next step was to determine the number of nodes per layer, relative to the bottleneck layer size. Each layer pair of the encoder and decoder can be observed to be an AE by themselves. Then the work of encoding and decoding should be equally spread between the layers. However, there are two ways to define an even change of the number of nodes: Either the number of nodes goes down with a fixed number of nodes in between each layer, or the number of nodes goes down with a fixed factor. For example, if the feature vector has a dimensionality of $Feat_{dim}$, and the bottleneck has a dimensionality of B_{dim} , the fixed change in number of nodes, ΔN , can be calculated as follows:

$$\Delta N = \frac{Feat_{dim} - B_{dim}}{n} \quad (3.1)$$

where n is the number of layers. Thus, as mentioned, the fixed factor change of the nodes

can be calculated as follows:

$$N_i = \left(\frac{Feat_{dim}}{B_{dim}}\right)^{1/n} \cdot N_{i-1} \quad (3.2)$$

Where N_i is the layer one layer away from N_{i-1} , where $N_0 = B_{dim}$. The factor method of calculating the number of nodes spreads the amount of the work more evenly throughout the AE. However, both methods were tested, and as expected, the factor method gave lower reconstruction error. Other than the two mentioned methods of calculating the number of nodes, a larger number of nodes per layer, and a lower number of nodes were tested.

The autoencoder was constructed with the open-source NN library Keras [20]. The activation function used was ReLU, and dense layers were used. The bottleneck layer dimensionality is a crucial part of the method, and was tested multiple times as the methodology evolved. It became apparent that most of the data had an intrinsic dimensionality of around 20, as an increase in the bottleneck layer size did not yield any rapid improvement of the validation loss. Therefore, the number of nodes in the bottleneck layer was set to 20, and from that and Equation 4.3, the number of nodes could be calculated. In Figure 7 the structure of the autoencoder can be seen, with the selected number of layers and number of nodes for each layer. The optimizer was ADAM, and the learning rate was left as default. The output layer was at first just one layer with MSE as its output function, but was later revised to manage the various data types adequately. Hence, the output layer was split into 5 layers, with suitable activation functions in the output layers. The various output layers were weighted such that each node had an equal weight in the joint loss function.

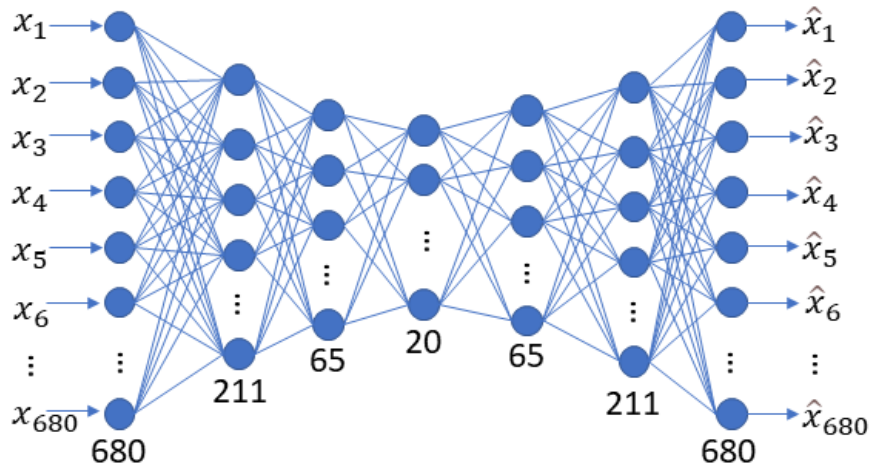


Figure 7: **The autoencoder.** The autoencoder used for the results in Section 4.

The activation functions in the output layers should be selected such that they can easily reconstruct their corresponding input data. The first output layer will try and recreate

binarized values, and as such, the sigmoid output function was chosen. The second and third output layer was trying to recreate a float value between zero and one, and linear output function was chosen for this. For the two one-hot encoded values of gender and sickness level, a softmax function was chosen. For the sigmoid output activation function, binary cross entropy was used as the loss function. MSE was used for both linear layers, and categorical crossentropy for the softmax output activation function. These choices are summarized in Table 3.

Table 3: The five output layers with their activation function, corresponding loss function, and the number of features per layer.

Description	Output function	Loss function	# Features
Medical tests, ICD-10, KVÄ	Sigmoid	Binary crossentropy	666
Contact	Linear	MSE	8
Age	Linear	MSE	1
Gender	Softmax	Categorical crossentropy	2
Sickness level	Softmax	Categorical crossentropy	3

The total loss function can then be defined as follows:

$$Loss = w^T \cdot L$$

where $L = [L_1, L_2, L_3, L_4, L_5]$ and $w^T = [w_1, w_2, w_3, w_4, w_5] = \frac{1}{666}[666, 8, 1, 2, 3]$, which yields the following expression:

$$\begin{aligned}
Loss = & -\frac{666}{666} \frac{1}{666} \sum_i^{666} (y_i \log \hat{y}_i + (1 - y_i)(\log(1 - \hat{y}_i))) + \frac{8}{666} \frac{1}{8} \sum_i^8 (y_i^{contact} - \hat{y}_i^{contact})^2 \\
& + \frac{1}{666} (y^{age} - \hat{y}^{age})^2 - \frac{2}{666} \sum_i^2 y_i^{gender} \log \hat{y}_i^{gender} - \frac{3}{666} \sum_i^3 y_i^{sickness} \log \hat{y}_i^{sickness} \quad (3.3)
\end{aligned}$$

When training Equation 4.4 was used. The $control_{train}$ data set was split into two parts where 80% of the data points were used for training and 20% were used for validation. Batch size of 512 was used, and after about 20-30 epochs the validation loss seem to plateau. The training loss continued downwards, which could plausibly be attributed to overtraining. In Figure 8 this trend can be seen, with an increasing separation between the training and validation loss as the training continues. As the data set contains 1.4 million data points, each epoch will have many points to learn from, which explains the low total number of epochs.

3.3.3 Outlier threshold

Once the AE has been trained, it can be used to make reconstructions of the *case* data set. If the original input is denoted x , the AE reconstruction is denoted \hat{x} , and the AE is

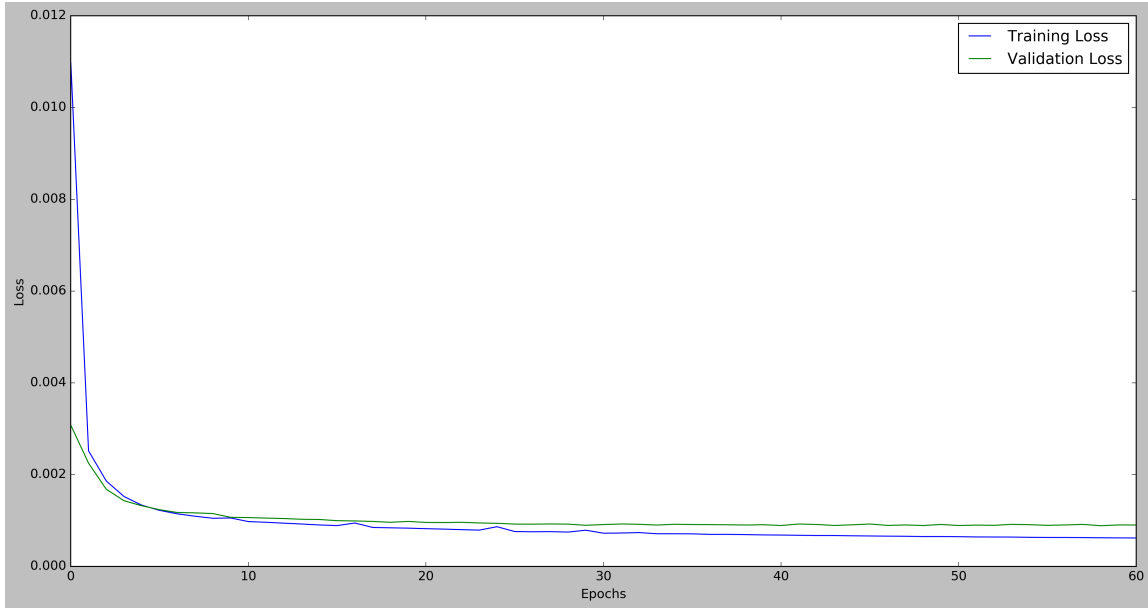


Figure 8: **Training/validation loss graph** The loss on the y-axis is the loss function in equation 4.7. Note the separation between training and validation loss as the training continues.

denoted as the function $f(x)$, the reconstruction process can be written as:

$$\hat{x} = f(x) \quad (3.4)$$

The reconstructions can be used as a measurement of how much of an outlier that data point is. In order to go from the reconstruction vector to a measurable reconstruction error, an error function is needed. There were a number of possible candidates that were tested.

We considered three ways to calculate the reconstruction error: The loss function, MSE and F1-score. Using MSE and the loss function gives similar results, but there were still differences. Calculations of the *case* data set can be found in Figure 9a for loss function and Figure 9b for MSE. Note that the range of values on the x axis start slightly above zero in order to see any detail. As expected they both follow roughly the same structure. The main difference is within the balance between various features. The loss function has a few features that used MSE while training, but most of them used cross entropy. The cross entropy terms has a loss that gives a $-\log(\hat{x})$ error, while MSE gives $(1 - \hat{x})^2$. These are not equal mathematically, with the log terms giving a higher reconstruction loss for the same reconstruction difference. This is visualized in Figure 9d, which has the cross-entropy/log loss math besides the MSE. The log loss always gives a higher error, and its derivative does not approach zero at one, as is the case for MSE. If each feature should be weighted equally in the reconstruction loss, the more reasonable choice is MSE.

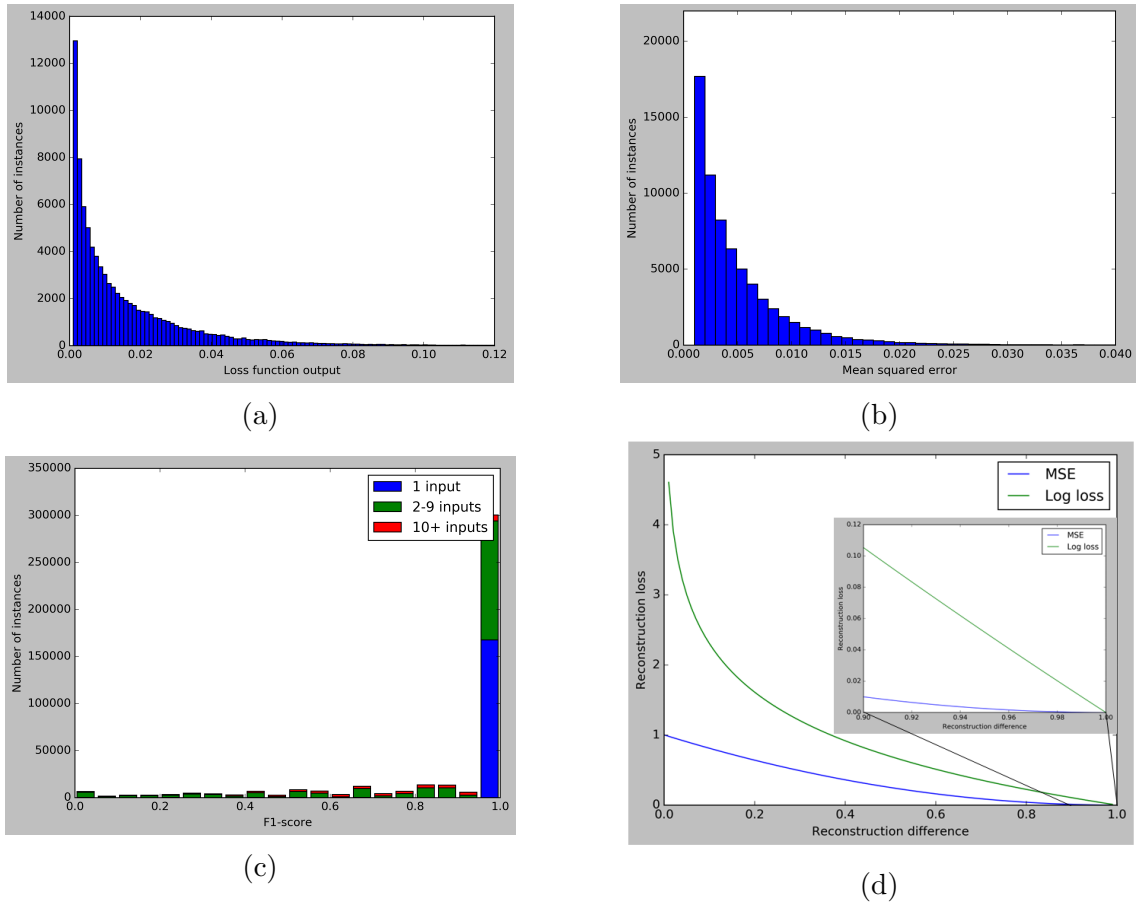


Figure 9: **The reconstruction errors** with (a) the loss function used during training, (b) MSE which was the choice for outlier thresholding and (c) the F1-score for the first output layer. (d) Shows the mathematical difference between log loss and MSE.

Furthermore, the data point based F1-score was also presented. As binarized values are easily used to get a F1-score, only the first output layer was used. The F1-score calculation for all the non-zero vectors can be found in 9c. Note that 1.0 is a perfect score, which means that almost all input vectors with one non-zero input is perfectly reconstructed. This reconstruction error calculation yielded significantly different outliers than the loss function or MSE. Specifically, the F1-score thresholding will manage data points with few non-zero features differently.

For each of these a threshold value was decided to determine outliers. This was modelled after the myocarditis cases of young males. As MSE was selected as the reconstruction error to use, the threshold for $MSE > 0.006075$ was used as this was the best reconstruction error that any of the young male myocarditis cases had. Certain sicknesses should crudely map to certain regions in the graph, but one region can contain several different sicknesses. Therefore, a subset of sicknesses should now be selected.

3.3.4 Autoencoder ensemble

Each time the AE was trained, the $case_{outlier}$ and $control_{outlier}$ turned out differently. As the results had large variations, the method needed an extra step. Taking inspiration from Chen et al. (2017) [13], an ensemble was used. The method presented by Chen, as discussed in Section 2.2.4, was to alter the autoencoder many times, and collecting the reconstruction error from each to find a better global optimum. This method is not needed in this project, as each time the autoencoder is trained, it reaches roughly the same loss as it stabilizes. However, each time it gives a different set of data points as outliers. Therefore, in this project, it would not make sense to reconstruct the autoencoder each time, but rather stabilize the results by running the same autoencoder many times. The autoencoder was trained 10 times, and the median value was selected for each data point. These values were then used for thresholding the data set to find the outliers.

3.4 Outlier analysis

Two different methods were used to analyse the selected outliers. The first is an analysis based on the features, with the goal to discover features that are more frequent in the $case$ outlier data set than in the $control_{eval}$ outlier data set. The second tries to look at the outliers as a group, by using t-SNE.

3.4.1 Feature based analysis

Each individual feature in the input vector was reconstructed with varying success. In order to see how well the various inputs were reconstructed, the F1-score can be calculated for each feature, rather than data point. Secondly, an indicator of the outliers properties is the ratio of inputs. The ratio of inputs would be how frequent various inputs occurred, in the $case_{outlier}$ set compared to $control_{outlier}$ set. By taking the ratio, R_{feat} , of the sum of the $case_{outlier}$ feature column, $Feat_{case}$, and the sum of the $control_{outlier}$ feature column, $Feat_{eval}$, the following can be calculated:

$$R_{feat} = \frac{Feat_{case}}{Feat_{eval}}. \quad (3.5)$$

Each ratio would ideally indicate the effect of the vaccination on each input feature. Note that each ratio is taken individually for each feature. The ratios with high values will indicate possible side effects, as for example, a myocarditis patient might take Troponin tests to find the nature of their chest pain. At the same time, ratios with low values, could indicate the vaccine effect, or other side effects with a positive effect on the patients health. Only features with at least 50 tests in both outlier data sets were used.

3.4.2 Group based analysis

The group analysis was based on t-SNE [7]. The first t-SNE plot of interest shows how $case_{outlier}$ and $control_{outlier}$ data points spread, and if there were groups with a significantly larger amount of $case_{outlier}$ points. It is also of interest to see how the age, gender and sicknesslevel is spread in t-SNE, where only the Moderna vaccinated patients were selected for further analysis. The encoded representation was used found in the bottleneck layer was used in t-SNE. Various perplexity and learning steps were tested, in an attempt to maximize structure. Specific regions of interest were selected by visual inspection and the outlier data points in the regions were selected and analysed for common traits.

4 Results

4.1 Autoencoder as outlier detector

The number of outliers selected can be seen in Table 4. The number of data points are roughly the same, but there is 3 % more in $case_{outlier}$. It was expected that $case_{outlier}$ would be larger than $control_{outlier}$, since the AE should have identified more of the features in $control_{outlier}$ because it was trained on a similar data set.

Table 4: Outlier data sets.

Type of data set	# Datapoints	# Myocarditis, all ages	# Myocarditis, age 16-24
$Case_{outlier}$	17,807	29	10
$Control_{outlier}$	17,276	33	3

4.2 Outlier analysis

4.2.1 Feature-based analysis

There were two main methods in the feature-based analysis. The first was ratio of inputs, as discussed in the method section. In Table 5, the most over- and under-represented features in the medical tests can be found. The placement is relative to their ratio. Note that only tests where the total number of tested individuals were above 50 in both $cases_{outlier}$ and $control_{outlier}$ were accepted, since the ratio does not work well on a low number of data points.

The method was selected with side effects in mind, and the final method can detect side effect patterns. It can also detect some of the actual vaccine effects, vaccination patterns within the population and patterns in the healthcare system along with noise.

Table 5: Most over-represented and under-represented medical tests in $case_{outlier}$, relative to $control_{outlier}$.

Placement	Type	Ratio	# $Cases_{outlier}$
1	B-HbA1c	1.5	145
2	B-Takrolimus	1.4	449
3	P-Prolactin	1.4	206
4	P-Carbon dioxide	1.3	608
5	P-hCG and beta-chain	1.3	123
.			
.			
135	P-Troponin	0.9	1778
.			
.			
186	aB-Oxygenconcentration	0.7	186
187	P-Myoglobin	0.7	164
188	B-Metamyelocyte	0.7	140
189	Avd U-OXY Oxycodon	0.7	153
190	P-Procalcitonin	0.7	358
191	P-Fibrinogen	0.6	231

In Table 5, the highest ratio is the HbA1c-test, which is a medical test that measures blood sugar levels within the last three months and is frequently taken to identify diabetics. It is documented that there is an increased levels of HbA1c for a time after vaccination [21]. The next test is tacrolimus, a drug mainly used during organ transplants, and the test checks the levels of the drug in the blood. A connection between tacrolimus tests and Covid-19 vaccines is that kidney transplant patients have been recorded gaining much less effective vaccine protection [22]. However, this gives no obvious explanation of the high ratio. A possible explanation may be that operations on vaccinated patients could be more frequent than unvaccinated, resulting in the high ratio. The next is prolactin, a hormone normally at high level during and some time after pregnancy. It has been observed having higher levels after vaccination in a few cases [23]. There is no obvious explanation for an increase of carbon dioxide measurements after vaccination, as no significant connection between carbon dioxide and Covid-19 vaccinations seem to be reported. The next test is hCG and beta-chain, which is often used as a pregnancy test. Again, there are no obvious connections to Covid-19, or Covid-19 vaccines. Perhaps it can be related to vaccination patterns, or a behavioural change after vaccination.

Now looking at the under-represented medical tests, which would contain more tests within $control_{outlier}$ than in $cases_{outlier}$. Starting at the bottom, fibrinogen is present as a test during care of Covid-19 patients [24]. As fewer fibrinogen tests are taken on $case_{outliers}$, that could be an effect from the vaccines working as intended. The same reasoning can

be applied on procalcitonin [25], a blood test often used to diagnose sepsis. Oxycodon, an opioid pain killer, could be from vaccination patterns related to addicts. Metamyelocyte and myoglobin have been recorded as useful biomarkers for Covid-19 [26][27]. Troponin is used within Covid-19 in-patient care [28], and heart related diseases.

Table 6: Most over- and under-represented medical procedures in $case_{outlier}$, relative to $control_{outlier}$

Placement	Type	Ratio	# $Cases_{outlier}$
1	Hemodialysis, chronic	1.7	252
2	Cervical screening test	1.6	118
3	Hemodiafiltration	1.3	111
4	PET scan	1.2	116
5	FeNO measurement	1.2	76
.			
.			
.			
21	Doppler echocardiography, transthoracic, extensive	1.0	689
22	Doppler echocardiography, transthoracic, simple	0.9	241
23	Vaginal ultrasound examination	0.9	70
24	Measures related to Covid-19	0.8	1651
25	Blood transfusion, erythrocytes, allogenic	0.8	161
26	Cardiotocography	0.6	158

Just as the medical tests can have more frequently used tests, there can be more frequent medical procedures. The most over- and under-represented medical procedures can be found in Table 6. The most over-represented medical procedure between vaccinated and unvaccinated is chronic hemodialysis. A possible explanation could be that these patients may have a high vaccination rate, while only a fraction of them may have been sampled in *control* sets. Cervical screening tests decreased greatly during the pandemic, and much as the hemodialysis might be from vaccination patterns, so might the cervical screening tests. Hemodiafiltration should follow a similar pattern as hemodialysis. An increase in PET scan might be side effect related, as enlarged lymph nodes has been reported after Covid-19 vaccination [29]. There is no obvious explanation for fractional exhaled nitric oxide (FeNO) measurements to increasing with vaccination.

Regarding the under-represented medical procedures, patterns could be seen, but they are generally harder to identify than medical tests. Cardiotocography, which is fetal heart rate measurement, could be related to the vaccination patterns of pregnant patients, who vaccinate to a lesser degree [30]. Some of the lowest ratios in medical procedures cannot be explained without further investigation, such as the vaginal ultrasound examinations and blood transfusion. Measures against Covid-19 should detect the same pattern as many of the Covid-19 related healthcare medical tests. As extensive Doppler echocardiography is

part of the characteristics of myocarditis, it is unfortunate that it is dominated by other types of patients, mainly other heart problems. As simpler Doppler echocardiography has roughly the same ratio, it should entail that Doppler echocardiography does not follow a pattern set by the myocarditis cases.

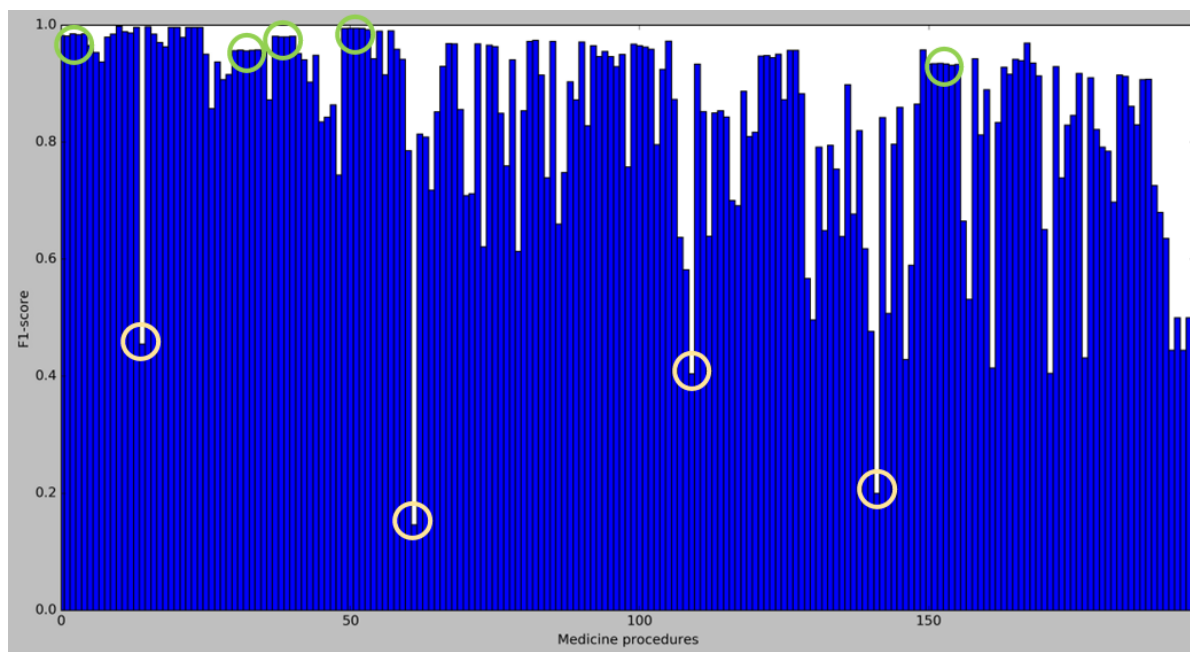


Figure 10: **F-1 score for medical tests.** Measurement of how well the autoencoder can reconstruct the medical tests. High frequency and lower dimensions should yield better reconstruction. Points of interest noted with circles.

In addition to the ratio of inputs analysis, investigation of how well the various features can be reconstructed was added. This was done for both the medical tests and the medical procedures for the *cases_{outlier}* data set. The medical tests F1-score can be found in Figure 10. The tests are sorted by the frequency of use within the healthcare system. There are a few tests that are significantly worse than the others, in particular there are four tests that dip down lower than their neighbouring tests, each of them noted with a yellow circle. The first, is P-Glucose (PNA), and second is Avd P-Glucose. This is the same type of test taken at two different points within the healthcare system. As mentioned in Section 3, frequency of the data points, and how high-dimensional the data points are, should determine the reconstruction error. The glucose tests are frequent, which means that the pattern of the glucose tests are irregular or high-dimensional. This might be because of the combined use of glucose tests, both for diabetics but also for severity of Covid-19 infection [31]. The third dip at test number 114 is P-Procalcitonin, which also can be related to Covid-19 infection [25]. The final dip noted with a circle is the fetal hemoglobin test, which has no obvious explanation.

Another observation that can be done with the graph in Figure 10, is that there is a general trend of worse reconstruction as the medical tests number increases. As the medical tests are sorted by frequency, this observation is consistent with the idea that frequency is central in the reconstruction error. A final observation is that there are areas in the graph that form plateaus of roughly the same reconstruction error. These are noted with small green circles in Figure 10. The tests are highly correlated, and also has a lower dimensionality. Each plateau will have a lower intrinsic dimension, and high frequency as there are more tests. As a result, they are generally reconstructed more accurately than single tests. Troponin can be seen at position 46 to have a F1-score of 0.84.

A similar analysis can be done on medical procedures F1-score. As these are also sorted by frequency, the downward trend can be observed here as well. There were not as many highly correlated procedures as there were tests. The dips at medical procedure number 13, 15, 26, 30 and 48 are as follow: hyposensitization, other specified investigation, photo documentation UNS, coronary angiography and radiation therapy. It is intuitive that it will be hard to find patterns for the medical procedure *other specified investigations*. However, the other medical procedures are hard to give a possible explanation to.

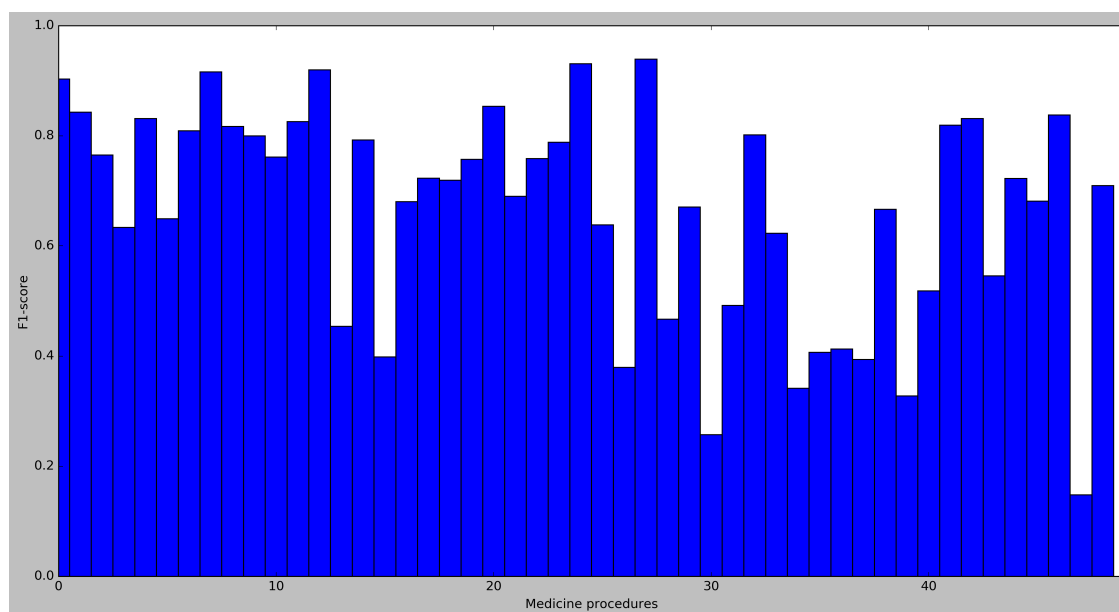


Figure 11: **F-1 score for medical procedures.** Measurement of how well the autoencoder can reconstruct the medical procedures. High frequency and lower dimensions should yield better reconstruction. Points of interest are the worst reconstructed medical procedures, and the best reconstructed procedures.

Some of the peaks are equally challenging to explain, while others are reoccurring from the ratio of inputs table. The peaks are as follows: Measures related to Covid-19, Eye biomicroscopy, cervical screening test, blood transfusion, and dermatoscopy. Measurements

against Covid-19 and blood transfusion both appear with low ratios of inputs. At the time cervical screening test also have a high reconstruction error, so there does not seem direct correlation between reconstruction error and ratio of inputs.

4.2.2 Group based analysis

In Figure 12 a t-SNE visualization of $cases_{outlier}$ and $control_{outlier}$ can be found. The regions in the plot will have some over-representation of certain sicknesses. A few regions with more $cases_{outlier}$ than $control_{outlier}$ were further investigated. For the first region noted by (i), most of the data point have kidney failure and many are under chronical hemodialysis, and CPR tests are frequent. Region (ii) nearly all data points have cancer, and kidney failure is still common. Region (iii) has more mental health related issues, while region (iv) has an over-representation of heart related issues. High blood pressure and myocardial infarction are common. In this region two myocarditis cases are also present. Region (v) has an over-representation of mental health related issues, heart problems and diabetes. It is difficult to see clusters in the t-SNE visualization, possibly as a result of that many data points were used within the visualization, in total 35,083 data points.

Next we focus on the Moderna vaccine, as the myocarditis cases are the most prevalent when using the Moderna vaccine. The t-SNE plot for the Moderna vaccine outliers can be found in Figure 13a. Note that only data points from the $case_{outlier}$ data set were visualized here. There are a few clearly defined clusters. The most clearly defined ones have been noted with a box, and the data points within the boxes were investigated. In region (i), all are the chronical hemodialysis patients noted in Figure 12 as well. Here they are more closely grouped. These patients had kidney failure and the associated medical tests. The region noted as (ii) in Figure 13a also had many chronical hemodialysis patients, but more cancer and diabetes than region (i). There were also more hemodiafiltration and blood pressure measurements. These two regions are still quite close both visually within the figure, and in regards of tests. The close-by region (iv) also have many kidney failure and chronical hemodialysis patients. Region (iii) is dominated by cancer, and not as close as the previous three regions, probably as a result of being primarily cancer patients. A region further away is (v), which indeed has no kidney failures. Instead, it is dominated by mental illnesses.

Further information can be gathered by showing the various attributes used during sampling. The first is age, and in order to keep the visualization clean, only three age ranges were used, instead of 10 in the sampling process. The age related separation can be found in Figure 13b. Here it can be seen that almost all clusters investigated are dominated by the 41-65 years range. The one exception is the mental health related cluster, number(v) in Figure 12. All clusters can be seen to have both genders within them in Figure 13c. However, regions outside heavily clustered regions can be seen to cluster with gender rather heavily. Finally, the t-SNE visualization with the sickness level can be found in Figure 13d. The kidney failure and cancer clusters have a higher sickness level, while the mental illness

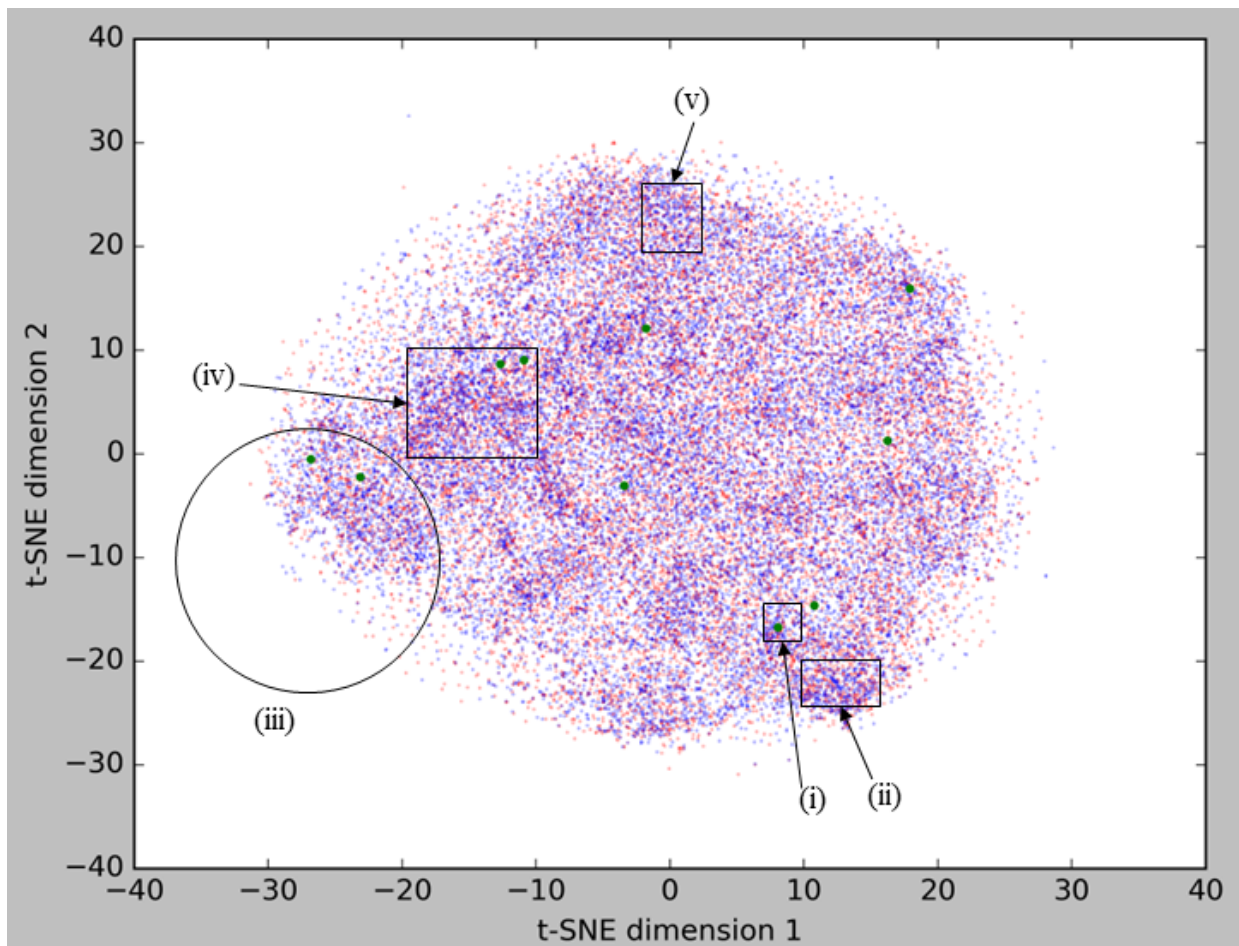


Figure 12: **Visualization of outliers using t-SNE.** All of the $case_{outlier}$, blue dots, and $control_{outlier}$, red dots, data points are visualized with t-SNE. The outliers bottleneck layer representation is used along with t-SNE. The myocarditis cases for vaccinated young males can also be seen, with the larger green dots. Various regions were selected for further investigation (i-v).

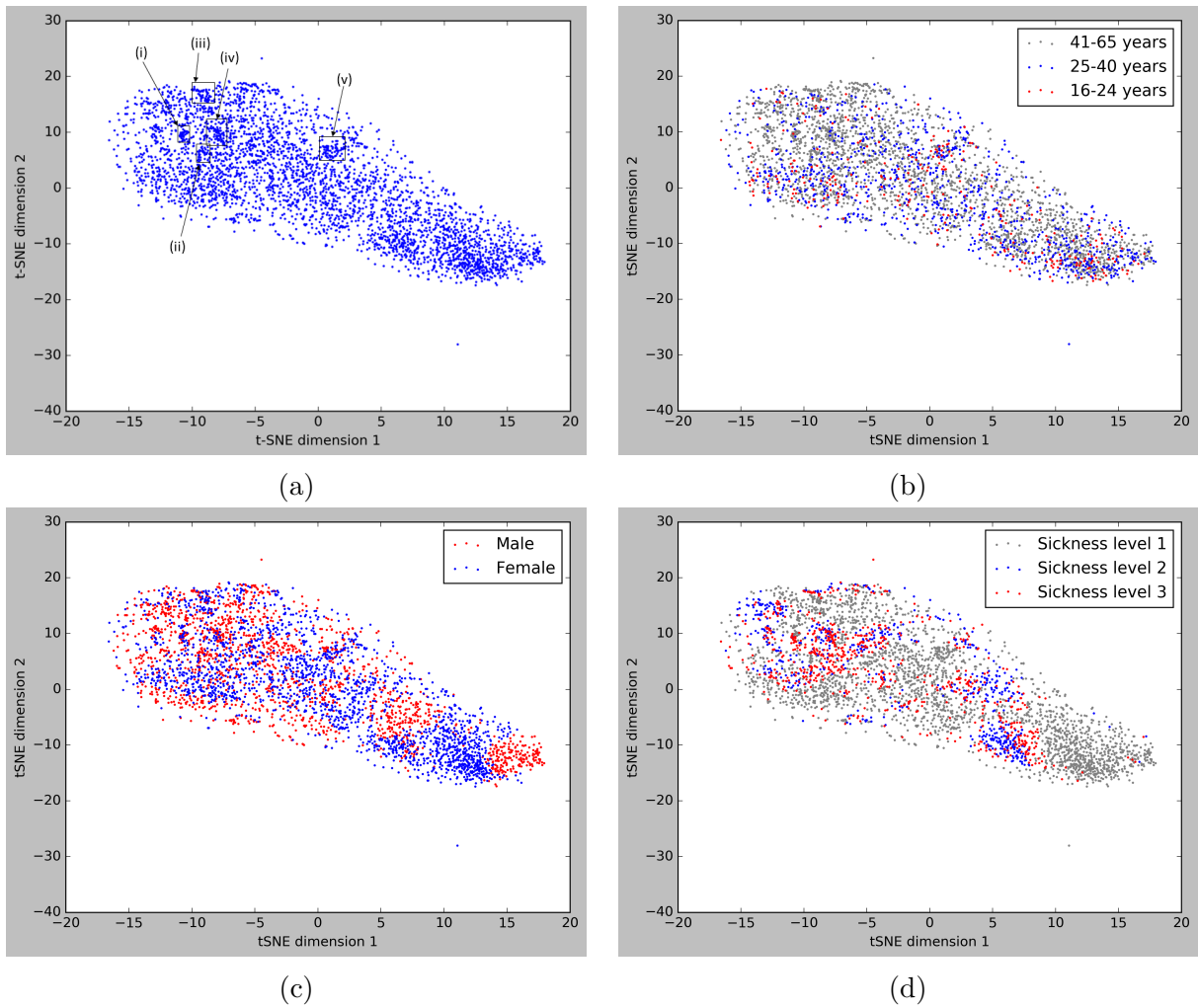


Figure 13: **Moderna outliers using t-SNE.** The data points latent space representation visualized with t-SNE. a) Shows the selected regions of interest (i-v). b) Shows the age. c) Shows the gender. d) Shows the sickness level.

cluster have a lower sickness level. Generally, regions will group heavily with sickness level, possibly because of the diagnoses history being included in the input vectors.

4.2.3 Summary of myocarditis results

The tests that are most frequent for young males with myocarditis after vaccination, have been overshadowed by other diseases. Troponin seems to be correlated with both Covid-19 care and other heart problems than myocarditis. The other strong characteristic of myocarditis, extensive Doppler echocardiography, was also shared with other diseases. In t-SNE they do not cluster significantly.

5 Discussion and conclusions

The information gathered, from feature-based outlier analysis, was a combination of various patterns. The vaccination effect seems to show clear results in the data. There are some results that indicate that vaccine side effects were observed, such as an increase of PET scans after vaccination. Then there are results that indicate that the vaccination patterns within the population can be detected, such as for cardiocography being less frequent as a result of pregnant patients being vaccinated to a lesser degree. In the group-based outlier analysis some clusters were observed. The most significant ones were related to kidney problems, but cancer, diabetes and mental illness also formed notable clusters. A large cluster of mental illness could be seen, which plausibly can be related to Covid-19 and increased isolation.

Features without a proper explanation for their ratio can be from any of the discussed patterns. Ratios that are not explained can be undiscovered side effects, but could also be from any of the other patterns and would need to be further investigated for each case. Furthermore, there may be results related to patterns within the administrative side of healthcare. These patterns will be hard to reveal without further information about what type of patterns can be detected within the administrative systems.

One of the most crucial choices throughout the project was the choice of how to define a data point. There are many possible definitions, and the type of data point definition will form the rest of the project. This project set out to capture myocarditis as a side effect of the Covid-19 mRNA vaccines, but was unable to do that. With the used definition the individual myocarditis cases within young males after vaccination will not be outliers.

The difference between looking at the entire population, and selecting certain outliers, would be that the tests become more representative to a specific disease the fewer diseases that share the same test. Even though myocarditis was still occluded by other diseases, the myocarditis cases will have a larger impact on the ratio of inputs when analysing outliers than analysing the entire population. This becomes one of the use cases of the method, to try and make certain tests more correlated to a disease of interest. One use case in this project is how the correlation between troponin and myocarditis increases with the method.

5.1 Limitations

This method comes with limitations. If a side effect of the vaccine is a frequent sickness within society, or if the sickness has a low-dimensionality, the reconstruction error will be low, and the side effect will not be detectable. Therefore, this project may not be viewed as a complete search for side effects, but rather for cases of sickness where the frequency is lower in society, and the dimensionality is high. As a side note, the dimensionality of the input vector should indicate how difficult the sickness is to diagnose, since more tests

would increase the dimensionality.

There are additional complicating factors with having myocarditis as a model event. One is that myocarditis increase from Covid-19 infection [32], and another that there is no separation between the sickness as a side effect from vaccination and sickness from other causes.

There are a few complications within the data sets as well. The first would be that ICD-10 codes are not as reliable from some parts of the healthcare system as others. When a medical doctor meets a patient, they might note down possible ICD-10 codes that are mere suspicions. Also, doctors working with inpatient care will have a much higher certainty. In this method, these values were equally regarded as a legitimate ICD-10.

A possible limitation is that within the project, the data points that are considered outliers are collected and used in t-SNE. These data points were per definition reconstructed poorly, which means that their bottleneck layer vector might not be representative to the input vector. As the bottleneck vector is used in t-SNE, the t-SNE visualization might not be representative to the outlier data point. Perhaps another dimensionality reduction should be used, that can reduce the data points with less information loss.

5.2 Future work

There are a number of possible alternatives to this projects data point definition. One of them might be that each vaccination date is its own data point, and all the vaccinated individuals during that day constitute the data point. This data point will have the temporal structure of the pandemic, which was not used for this projects method. The new data point definition could be further developed into defining data points as a vaccination date and personal characteristics combination. The personal characteristics include age, gender and sickness level. Adding personal characteristics as part of the data point definition would amplify the possibility of finding the myocarditis cases, as they are only detectable within certain personal characteristics combinations. The new type of data point would be a time series for each personal characteristics combination, which might invite the use of LSTM architectures. This would take the project in an alternative direction, but such a project could potentially have a larger chance of detecting side effects. The advantage of the new data point definition is that there can be actual outliers that can be tested with more traditional methods.

If the same definition of a data point would be used again, there could be value in investigating the use of variational autoencoders, as they have become increasingly popular for outlier detection. Instead of learning a representation in latent space, it learns a distribution, and then draws values from the distribution as it reconstructs the values again. They become more dynamic, as each type of data point might have its own distribution.

If the wanted signal is the side effect of the vaccine, all other patterns within the results should be minimized. A number of steps could improve the side effect signal. There is a

difference between the vaccine effect and the side effect, and in order to enhance the side effect signal, a possible step would be to decrease the effect signal. A simple way to remove the vaccine effect pattern would be to remove as many Covid-19 patients as possible from the source data. As Covid-19 was wide-spread throughout society, it would be difficult to remove it altogether, but the severe Covid-19 cases and an attempt of removing moderate Covid-19 cases would probably increase the side effect signals. By matching the vaccination pattern with the *control* sampling, the vaccination pattern effects in the results may be minimized.

A possible way to separate various types of sicknesses, could be the quantity of tests. This follows from that a patient with myocarditis might take more or less tests, for example troponin, than individuals with other diseases. The information about the counter values were erased when the counters were binarized. A possible design could be a counter with a cut-off, to ease the training and not get too low average values when normalizing. Another possible step would be to look at the result of the tests, to give more information about the patient, in order to give the AE information to separate the diseases.

The medical tests could possibly be more compactly represented, some tests are taken almost exclusively together, while others are the same tests taken in various parts of the healthcare system, and thus resulting in multiple entries for the same medical test. These could for many purposes be merged into one feature, as they have a dimensionality close to one. This would serve as a high quality dimensional reduction, as a part of the pre-processing of the data.

Even though a single trained AE gives irreproducible results, looking at the outliers might reveal patterns that will not be present in the ensemble method. Therefore, it might be more useful to look at single trained AE in projects aiming at hypothesis generation, such as this.

The idea of using a continuous outlier detector on a stream of data, to detect irregular behaviour within the populations healthcare, is an idea that is worth further investigation. If a successful method is constructed, the potential gain within society is large. This project can be seen as the first steps creating such systems.

Acknowledgments

I would like to thank *COVERS* deployed by the Region of Scania for giving me the opportunity to do this project. Thanks to my supervisors, Anders Björkelund and Mattias Ohlsson for guiding me through the project. Finally, thanks to Jonas Björk, Fredrik Kahn, Axel Nyström and Patrik Edén for discussion and feedback.

References

- [1] Jackson, L. A. et al. 2020. An mRNA Vaccine against SARS-CoV-2 — Preliminary Report. *New England Journal of Medicine*. 383. 20. 1920-1931.
Available from: <https://doi.org/10.1056/NEJMoa2022483>
- [2] Baden, L. R. et al. 2021. Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *New England Journal of Medicine*. 384. 5. 403-416.
Available from: <https://doi.org/10.1056/NEJMoa2035389>
- [3] COVERS, accessed 2022-05-13.
Available from:
<https://www.skane.se/organisation-politik/forskning/pagaende-forskning/covid-19-vaccinationsuppfoljningsstudie-covers/>
- [4] Hodge, V.J., Austin, J. A. 2004. Survey of Outlier Detection Methodologies. *Artif Intell Rev* 22. 85–126.
Available from: <https://doi.org/10.1007/s10462-004-4304-y>
- [5] Thudumu, S., Branch, P., Jin, J. et al. 2020. A comprehensive survey of anomaly detection techniques for high dimensional big data. *J Big Data* 7, 42.
Available from: <https://doi.org/10.1186/s40537-020-00320-x>
- [6] Kramer, M.A. 1991. Nonlinear principal component analysis using autoassociative neural networks. *AIChE J.*, 37: 233-243.
Available from: <https://doi.org/10.1002/aic.690370209>
- [7] Maaten, L., Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9. 2579-2605.
- [8] Pedregosa, F. et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 12. 2825-2830.
- [9] Liu, F. T. et al. 2008. Isolation forest. 2008 eighth IEEE international conference on data mining. IEEE.
- [10] Rumelhart, D. E., McClelland, J. L. 1987. Learning Internal Representations by Error Propagation,. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. MIT Press. 318-362.
- [11] Ackley, D. H., Hinton, G. E., Sejnowski, T. J. 1985. A Learning Algorithm for Boltzmann Machines. *Cognitive Science*. 9. 147-169.
Available from: https://doi.org/10.1207/s15516709cog0901_7
- [12] Betechuoh, B. L. et al. 2006. Autoencoder networks for HIV classification. *Current Science*. 1467-1473.

- [13] Chen, J. et al. 2017. Outlier Detection with Autoencoder Ensembles. Proceedings of the 2017 SIAM International Conference on Data Mining (SDM). 90-98
- [14] Zhou, C., Paffenroth R. C. 2017. Anomaly Detection with Robust Deep Autoencoders. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17). Association for Computing Machinery. New York, NY, USA. 665–674.
Available from: <https://doi.org/10.1145/3097983.3098052>
- [15] Aygun, R. C., Yavuz, A. G. 2017. Network Anomaly Detection with Stochastically Improved Autoencoder Based Models. 2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud). 193-198.
Available from: <https://doi.org/10.1109/CSCloud.2017.39>
- [16] Hawkins, S., et al. 2002. Outlier Detection Using Replicator Neural Networks. In: Kambayashi, Y., Winiwarter, W., Arikawa, M. (eds) Data Warehousing and Knowledge Discovery. DaWaK 2002. Lecture Notes in Computer Science, vol 2454. Springer, Berlin, Heidelberg.
Available from: https://doi.org/10.1007/3-540-46145-0_17
- [17] Pang, G. et al. 2021. Deep Learning for Anomaly Detection: A Review. ACM Comput. Surv. 54. 2. Article 38 (March 2022).
Available from: <https://doi.org/10.1145/3439950>
- [18] Karlstad Ø., Hovi P., Husby A., et al. 2022. SARS-CoV-2 Vaccination and Myocarditis in a Nordic Cohort Study of 23 Million Residents. JAMA Cardiol. Published online April 20, 2022.
Available from: <https://doi.org/10.1001/jamacardio.2022.0583>
- [19] Folkhälsomyndigheten, accessed 2022-04-29.
Available from:
<https://www.folkhalsomyndigheten.se/smittskydd-beredskap/utbrott/aktuella-utbrott/covid-19/skydda-dig-och-andra/rad-och-information-till-riskgrupper/>
- [20] Chollet, F. et al. 2015. Keras. Available at: <https://github.com/fchollet/keras>.
- [21] Liu, J., Wang, J., Xu, J. et al. 2021, Comprehensive investigations revealed consistent pathophysiological alterations after vaccination with COVID-19 vaccines. Cell Discov 7, 99.
Available from: <https://doi.org/10.1038/s41421-021-00329-3>
- [22] Reischig, T., Kacer, M., Vlas, T., et al. 2022. Insufficient response to mRNA SARS-CoV-2 vaccine and high incidence of severe COVID-19 in kidney transplant recipients during pandemic. Am J Transplant, 22. 801–812.
Available from: <https://doi.org/10.1111/ajt.16902>

- [23] Murvelashvili N., Tessnow A. 2021. A Case of Hypophysitis Following Immunization With the mRNA-1273 SARS-CoV-2 Vaccine. *Journal of Investigative Medicine High Impact Case Reports*.
Available from: <https://doi.org/10.1177/232470962111043386>
- [24] Sui J., Noubouossie D., Gandotra S., Cao L. 2021. Elevated Plasma Fibrinogen Is Associated With Excessive Inflammation and Disease Severity in COVID-19 Patients. *Front Cell Infect Microbiol.* 3;11:734005.
Available from: <https://doi.org/10.3389/fcimb.2021.734005>
- [25] Hu. R et al. 2020. Procalcitonin levels in COVID-19 patients. *International Journal of Antimicrobial Agents.* 56. 2.
Available from: <https://doi.org/10.1016/j.ijantimicag.2020.106051>.
- [26] Zini G., Bellesi S., Ramundo F., d’Onofrio G. 2020. Morphological anomalies of circulating blood cells in COVID-19. *Am J Hematol.* 95(7). 870-872.
Available from: <https://doi.org/10.1002/ajh.25824>
- [27] Jia-Sheng, Y., et al. 2021. Myoglobin Offers Higher Accuracy Than Other Cardiac-Specific Biomarkers for the Prognosis of COVID-19. *Frontiers in Cardiovascular Medicine*, 8.
Available from:
<https://www.frontiersin.org/article/10.3389/fcvm2021.686328>
- [28] Al Abbasi B. et al. 2020. Cardiac Troponin-I and COVID-19: A Prognostic Tool for In-Hospital Mortality. *Cardiol Res.* 11(6). 398-404.
Available from: <https://doi.org/10.14740/cr1159>
- [29] Minamimoto R, Kiyomatsu T. 2021. Effects of COVID-19 vaccination on FDG-PET/CT imaging: A literature review. *Glob Health Med.* 3(3). 129-133.
Available from: <https://doi.org/10.35772/ghm.2021.01076>
- [30] Razzaghi H, Meghani M, Pingali C, et al. 2021. COVID-19 Vaccination Coverage Among Pregnant Women During Pregnancy — Eight Integrated Health Care Organizations, United States. *MMWR Morb Mortal Wkly Rep.* 70. 895–899. Available from: <https://doi.org/10.15585/mmwr.mm7024e2external>
- [31] Wang, W., Shen, M., Tao, Y. et al. 2021. Elevated glucose level leads to rapid COVID-19 progression and high fatality. *BMC Pulm Med* 21, 64.
Available from: <https://doi.org/10.1186/s12890-021-01413-w>
- [32] Block J.P., Boehmer T.K., Forrest C.B. et al. 2022. Cardiac Complications After SARS-CoV-2 Infection and mRNA COVID-19 Vaccination — PCORnet. United States. *MMWR Morb Mortal Wkly Rep* 71. 517-523.
Available from: <http://dx.doi.org/10.15585/mmwr.mm7114e1>