university of groningen

faculty of science and engineering

Predicting CAD Severity Using Patient Symptom Descriptions and BERT

A.M. Heeres

university of groningen / faculty of science and engineering

University of Groningen

Predicting CAD Severity Using Patient Symptom Descriptions and BERT

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Artificial Intelligence
at University of Groningen under the supervision of

Dr. Fokie Cnossen (Faculty of Science and Engineering)
Prof. Dr. ir. Peter van Ooijen (Faculty of Medical Sciences)

Ama Maria Heeres (s3332403)

September 5, 2024

Contents

Acknowledgments

We thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Hábrók high performance computing cluster.

I would also like to thank Eline Meijer, Moniek Koopman and Tanja de Vries for checking my work, regularly meeting with me to discuss the research and for answering any questions I had about the CONCRETE project.

Finally of course I want to thank Prof. Dr. Fokie Cnossen and Dr. ir. Peter van Ooijen for offering me the opportunity to work on this project and for being my supervisors during what has been the largest task of my studies thus far.

Abstract

Coronary Artery Disease is one of the leading causes of death worldwide. With such a common condition it is incredibly important to be able to quickly diagnose and treat any patient suffering from it. One way to identify the severity of a patient's CAD risk is with a CT scan that looks at the calcium scores of the coronary arteries. These CT scans often take time and resources, so the CONCRETE project seeks to streamline this process by analyzing patient chest pain complaints using different natural language processing models to try and reduce the number of patient that need to be referred for additional testing. In this paper we use the BERTje architecture to try and predict whether a patient has a high or low calcium score based on a patient's complaints. These complaints would be the same as those a general practitioner would hear when interviewing a patient. Our experiments did not produce a model that could reliably predict CAD severity, nor did it indicate whether any specific questions were more indicative of a high calcium score. We can conclude that either the questions analyzed can not be used with the BERTje model to identify high CAD risk cases, or more likely that the current amount of data that the CONCRETE project has collected is not yet sufficient to fine tune a BERTje model.

1 Introduction

ischaemic heart disease, also known as Coronary Artery Disease (CAD), is responsible for about 16% of all deaths, making it the leading cause of death worldwide [1]. In the EU alone the total cost of Cardiovascular disease is €282 billion annually with just CAD being responsible for €77 billion (27%) of that cost [2]. Given both the scale of the suffering that CAD causes and the costs associated with it, it is only natural that we are continually searching for faster and more efficient ways of diagnosing and treating CAD and related conditions.
For this purpose the CONCRETE project was created by the UMCG to use Coronary calcium scoring as a tool to reduce the number of patients that need to be referred for additional testing. Ideally developing methods to exclude patients without CAD at an early stage [3] [4]. This would then cut down on the number of tests that have to be performed and would both reduce costs and free up more resources for the patients who need it the most.
One of the goals of the CONCRETE project is to create a resource that can assist GPs in diagnosing patients. For this reason we will be analyzing the CONCRETE data that has thus far been collected and we will attempt to fine tune a Bidirectional Encoder Representations from Transformers (BERT) model to see if this could be used to differentiate between patients with and without a calcium score that is indicative of CAD.

1.1 Research Questions

This thesis focuses on the following two questions:

Q1. Are there certain symptoms and risk factors that correspond to different levels of calcium scores which can be used for better assessment of CAD risk?
We will be testing this by analyzing the answers to different questions and and seeing how well they can be used to predict calcium score levels. Our expectation is that certain questions will be more important for a predictive model and thus be more valuable in CAD prediction. However, we are not certain if we have enough data to be able to find this information.

Q2. Can a patient's claims and description of their symptoms be used for robust CAD severity prediction and gender stratification?
This question we can answer by making a model that can predict calcium level, and by extension CAD severity, based on the answers that a patient gives on questions about their chest pain complaints. The gender stratification part of the question we can answer by also having gender be one of the factors that the model tries to predict. Our hypothesis is that it should be possible to use a BERT model to achieve both of these aims.

1.2 Thesis Outline

In this thesis we go over the details of CAD, calcium scoring used for CAD diagnosis and the natural language processing method that we use. The model we will be using is the BERT model [5] which is a model based on the Transformer architecture. There has already been some research into using natural language processing to help with CAD diagnosis [6] [7], but most of this has not been on text directly from a patient questionnaire. In this regard our research focuses on providing a tool to evaluate a patient at a very early stage without a lot of processing having to be done by a general practitioner.
The dataset we use consists of answers to questionnaires about chest pain complaints. After pre-processing these we will train a BERT model on them. We use a range of hyper-parameters that have been shown to be effective, the details of which are described in the methods section. For the experiment we use the BERTje model [8] and train it on the HPC cluster of the University of Groningen.
We developed three experimental setups, each with different advantages and disadvantages. The different setups will help us answer different parts of our research questions.
The first setup analyses each answer as a data-point. The model is trained to predict the calcium score that corresponds to the input answers. Though a naive approach, if this works then it could produce a model that can predict CAD severity based on a patient's answers. We test this setup in 4 different configurations based on the balance of the data and whether or not we also try to predict the gender of the patient.
The second setup takes all of the answers of a single patient as input and based on that tries to detect if the patient has a high or low calcium score. This setup would have much less noise from less useful answers, but has the downside of having far fewer data points. For this reason we do not have different configurations for this setup.
Finally the third setup is trained to do the same as the first setup, but instead of one model trained on all of the answers we train a separate model for each question. This might be able to tell us whether certain questions are more or less useful for predicting CAD severity. This again has the downside of greatly reducing the size of the training data.
The results we found were sadly unimpressive. None of the methods produced a fine tuned model that could reliable filter out low calcium scores. From these results we conclude that either the data is insufficient to fine tune a BERTje model on, this method is not capable of finding a relation between the answers and the calcium score, or there is no relation between the answers and the calcium score. Of these we think that the most likely explanation is that the dataset as it is now, does not contain enough data yet.

2 Background Literature

Due to the multidisciplinary nature of this research the background literature required involves both literature on coronary artery disease(CAD) and on the natural language processing method we will be using. We will first look at exactly what CAD entails and how to detect it. Then we will look at the transformer architecture and the BERT model.

2.1 Coronary Artery Disease

CAD is a condition that causes almost 9 million deaths per year [1]. Not only is the cost of these deaths globally enormous, we also see that it is especially poorer regions that are less able to deal with the costs of this condition [9].
CAD is caused by the buildup of fatty deposits on the inner walls of the coronary arteries. This plaque builds up over time and if it becomes stable it will harden and calcify. If this calcified obstruction is large enough it will make it difficult for blood to flow through the coronary artery, which in turn can cause the heart to not receive the resources it needs in periods of high demand. This means that generally, symptoms only show up when a person suffering from CAD is being active. Generally the symptoms disappear when at rest and the heart's demand for blood decreases. In several cases the symptoms can however also show up when at rest [10].
The most common symptom of CAD is angina. This is a type of chest pain specifically caused by insufficient blood flow to the heart. The exact kind of pain differs per patient. It can feel like a squeezing or pressure on the chest, but also like a feeling of fullness of discomfort. This pain can also radiate towards other parts of the body like the arms, neck, jaw or back [11].
In some cases the plaque on the inner wall of the coronary arteries can rupture causing thrombosis. Alternatively this might also be caused by plaque erosion or the calcified node itself causing thrombosis. This then results in Acute coronary syndrome, which can express itself as unstable angina on the most benign end and cell death leading to a heart attack on the more severe end [12].
One way to screen patients for CAD is by using coronary artery calcium screening. This method uses a CT-scan to make images of the coronary arteries and measure the amount of calcium present in those arteries. This method has shown to be an effective indicator of CAD [13] to the point that it can even be used to diagnose asymptomatic cases [14].

2.2 Transformers

Transformers are a machine learning model architecture that is used for sequence-to-sequence tasks. These tasks involve turning an input sequence into an output sequence. They do this using an encoder and decoder which learn to transform one sequence into the other. Before the transformer architecture this was done mainly using complex convolutional or recurrent networks where the encoder and decoder were connected through attention mechanisms. The transformer architecture significantly improves both performance and efficiency by basing the entire architecture on these attention mechanisms. The following section is a general description of the architecture as originally described with more detail in the paper that introduced the transformer architecture [15].
The encoder and decoder design works by encoding an input into a certain state and then decoding it into the desired output. In this process attention is a technique where instead of encoding a whole sentence into a certain state, each word might have a state and these states influence the output at a specific point. Essentially creating a mechanism to have an output where the context matters.

The transformer architecture consists of equally sized encoder and decoder stacks. These stacks consist of a number of layers. A cross-section of a layer can be seen in figure 1 with the encoder layer on the left and the decoder layer on the right. The encoder layer consists of two sub-layers. An attention and a feed forward sub-layer. The decoder has an additional attention sub-layer.
The attention layer takes the input sentence and for every word in that sentence it looks at how important that word is for itself and the other words in the sentence, This allows it to take the context of each word into account. In the figure we see that this is called multi-head attention. What makes it multi-head is that the architecture has multiple of these layers running in parallel that are then summed and normalized afterwards.

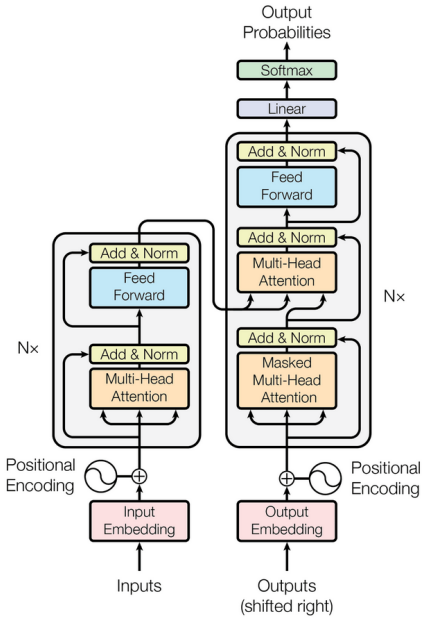


Figure 1: A diagram showing the transformer architecture with the encoder(left) and decoder (right) [15]

Both the encoder and decoder layer have an attention sub-layer that only uses its own input, which we can call a self attention layer. The decoder has an additional attention sub-layer, which uses both the output of the decoder self attention sub-layer and the encoder layer as its input, which is how the encoder is connected to the decoder. After the attention sub-layers the output of is send to a fully connected feed forward network. The output of that network is then normalized and send onward. Either to the next encoder and decoder layer, or from the decoder to the output layer.

2.3 Bidirectional Encoder Representations from Transformers

The BERT model is a model that uses the transformer architecture to pre-train a model that can then be used to accomplish a variety of language processing tasks. It does this by learning language representations from a large body of text and then being fine tuned for a specific task [5].
This model isn't the first to use this method. It had already been shown that pre-training models could be used as a way to better perform natural language processing tasks [16] [17]. Other aspects of the BERT model had also been used before. For example the Generative Pre-trained Transformer [17] also uses the transformer architecture to create a pre-trained model that can subsequently be fine tuned for specific tasks. A limitation of this model is that it only takes the left to right context of the input into consideration. This is what the BERT model attempts to improve upon by looking at the context in both direction instead [5].
The way the BERT model works is that it has both a pre-training and fine tuning architecture which it tries to keep as similar as possible. In figure 2 we can see the training architecture that the BERT model uses to pre-train.
The BERT model itself is a multi layer bidirectional transformer as is also seen in the original paper where transformers were introduced[15] and is also what we described in section 2.2. The base BERT model uses encoder and decoder stacks with 12 layers and 12 attention heads within these layers. The BERT model can accept both single and double sentences as input, making it useful for a variety of tasks[5].

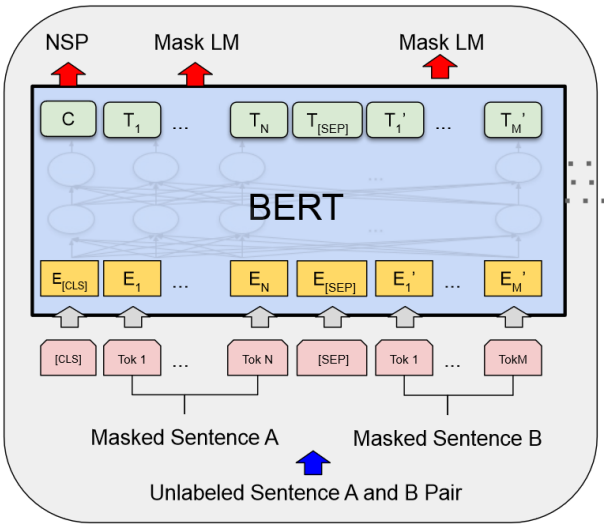


Figure 2: A diagram showing the pre-training architecture of the BERT model [5]

The model is trained using two unsupervised learning tasks called NSP(next sentence prediction) and MaskLM. The NSP task has the model performing a binary prediction of which sentence is supposed to come next. The MaskLM task instead masks random words and has the model predicting which words are supposed to be in those places [5]. Another advantage of using these unsupervised learning methods is that it does not require a large annotated corpus.
After pre-training, the model can be used to perform a variety of tasks. The task we will be focusing on is sentence classification. The architecture for this task can be seen in figure 3 [18]. This architecture uses the pre-trained BERT model with a character sequence as input and a class label as output. As we can see, it is almost identical to the model used for pre-training with only the input and output layers being slightly different.

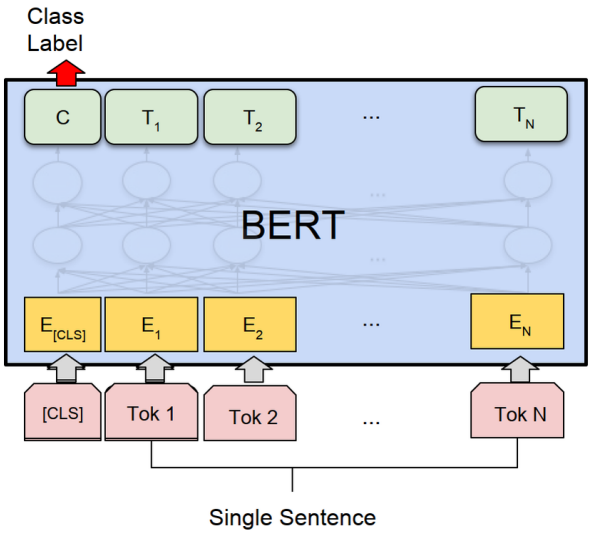


Figure 3: A diagram showing the fine tuning architecture of the BERT model for sentence classification [18]

BERT was trained on an english dataset. However there is also a multilingual BERT model [19] and multiple dutch BERT models [20] [8]. Of these the BERTje model seems to be the most versatile based on the tests they performed and the results that they report [8].

2.4 Machine Learning and CAD

In their paper "Automatic prediction of coronary artery disease from clinical narratives" by Buchan et al [6] researchers try to create an automated system that can predict CAD based on their medical history. The data they trained on consists of 516 patient records of clinical free text. They used apache cTAKES to extract information from this text and for classification they used Naive Bayes, MaxEnt and SVM models. They conclude that it is possible to predict CAD development in patients this way, however they also had multiple limitations in their research. A major limitations being that their dataset is fairly small and only represents a specific part of the patient population.
In their literature review, Alizadehsani et al concluded that there were multiple shortcomings in detecting CAD using ML algorithms [21]. The algorithms that were most successful at that time were KNN, SVM, and ANN. This was around the same time that BERT was first introduced. Around this time we also have Koleck et al [7] highlighting the focus on extracting symptom information for disease classification in ML analysis of health records. They argue that analyzing symptoms themselves in these health records has been neglected and should receive more attention.
On the same dataset of clinical records that Buchan et al performed there experiment, Houssein et all use the BERT architecture and manage to achieve much better performance than previous models [22]. In their research they used BERT and character embeddings stacking to create a very impressive method that outperforms all others used on this specific dataset. It also shows the BERT model can be a very powerful tool when used to analyze such data.
A major difference between this previous research and our current research is the nature of the data we are analyzing. Whereas these previous experiments were done on clinical narratives in medical records, we will be trying to extract features and predict CAD severity based on patient answers only. This is similar to the way a patient would tell them to a general practitioner via a questionnaire. Rather than trying to extract the symptoms from a health record and predicting a CAD diagnosis based on

that, we will be attempting to predict based only on the words the patients themselves use to describe their situation. We believe this is a novel way of utilizing the BERT architecture for analyzing CAD severity.
The CONCRETE data consists of sections with both open and multiple choice questions. The research of which this thesis is a continuation experimented on both the multiple choice and the open questions. The main focus there was on the multiple choice questions. Training models on the multiple choice questions worked, but at that time the dataset was smaller and it was not possible to train a satisfactory model on the open questions [23]. Having more data and focusing solely on the open questions might give us better results now.

3 Methods

In the following section we describe the dataset that we train our models on. We will also be describing some of the challenges that we encountered while cleaning the data and the equipment we used, including which BERT architecture we use and why.

3.1 The CONCRETE data set

The CONCRETE data set is a data set that collects information from patients with chest pain complaints. New patient information is still being collected. Although the collection is still ongoing, we performed our study on what what has thus far already been collected. The data contains the answers to multiple choice questions about quality of life, answers to open questions regarding the specific chest pain complaints and CT-scan Calcium scores with corresponding CAD severity.
The questions themselves were designed according to the ALECOBO method of identifying the main complaint of a patient. ALECOBO is an abbreviation for: Nature of the complaint (Aard van de klacht), localization (Lokalisatie), severity (Ernst), chronology (Chronologie), origin (Ontstaan), influence (Beïnvloeding) and other complaints (Overige klachten). The idea is that each question asked tries to question the patient about a different aspect of the complaints.
Marinov previously analyzed an earlier version of this data set that contained significantly less data and focused on both the multiple choice quality of life questions as the open questions about chest pain complaints.[23]
Our main interest is in the open questions about the chest pain as we are looking for relations and features in the natural language that could be used to help general practitioners improve diagnosis and care. Therefore we will only be using the answers to the open questions for our research.
The data set contains the patients age, gender, calcium scores, CAD severity and 15 answers to open questions about the patient's chest pain complaints. The data was gathered using a questionnaire that was provided to the patient by their GP. This questionnaire can be found in appendix A. Also included in apendix A is a table of just the open questions we used. This table contains a question number for easier referencing, the original (dutch) question text and an English translation of the original text. Note that the English translation is there for the reader of this paper and was not provided to the participants that filled out the questionairre.
After cleaning up the data 183 questionnaires proved to be useful for training, giving us 2745 open answers to work with. Of these patients 144 had a low calcium score of less than 100 and 39 patients had a high calcium score of over 100. In the data there are 65 male and 118 female patients. It is clear that there is limited data available and that the data we have is incredibly skewed in both calcium score and gender, both of these problems are things we will have to take into account for the experimental design.

3.2 Cleaning the data

Certain answers might not have been useful for training our model, therefore We removed any entries that did not have a calcium score or that lacked parts of the information that we would need for the training process. Then the data among the remaining patients also required some cleaning up.
Commonly patients answered questions by referring to their answer on a previous question. For example by answering a question with "see question X". In these cases we replaced the reference with the answer that it is referring to as that is as close as we can get to what the patient would have written as their answer to that question. Another issue was patients interpreting a question different

to how it was intended. For example the question ”Hoe lang duren de door u beschreven pijn- of gevoelsklacht(en)?” which translates roughly to ”How long do the complaints you described last?” can be interpreted in two ways. It can be seen as a question about the duration of the pain during an episode of it occurring, it can also be seen as a question about when the complaints first started occurring. The result is that some answers are in the range of minutes and hours, while others might be in the range of months or years. Unfortunately this issue was not isolated to a few patients, but was a very common misinterpretation of the question. Our options were to either remove this question, remove a significant portion of the answers to this question, or keep these answers in to stay as close to the original data as possible. The decision was made to keep this data with the hope that this would provide additional information for the model to use and train on.

3.3 Equipment

As mentioned before we will be using a transformer-based pre-trained language model from the BERT family.[5] The specific model we will be using is the mono lingual dutch model called BERTje. This model uses the same architecture and parameters as the original mBERT model. The multi lingual BERT model has been trained on just wikipedia, while the BERTJE model has been trained on a much more varied data set of 2.4 billion tokens from several high quality dutch texts. [8]
The reason we are using the mono-lingual dutch BERTje model rather than the multi-lingual BERT model is that although the BERTje model has not been used much on medical data, the BERTje model outperforms the mBERT model in most regards.[8]
To train the model we will be using the Hábrók high Performance computing cluster of the University of Groningen. On this cluster we will be running all of our experiments. For training purposes some form of our data will have to be on the Hábrók system, however all but the most necessary features of the data have been removed from it to ensure that no sensitive information is ever in an insecure environment.

4 Experimental Setup

To test our research questions and take steps to achieving the goals of the CONCRETE project we will be using multiple experimental setups. The first setup is a naive method that simply considers each answer a data point to predict a CAD severity for. The second setup considers each patient a single data point to predict a CAD severity for. The final setup uses the same data points as the first setup, but is limited to only training on the answers to an individual question rather than the answers to all 15 questions.
To make sure that we get the optimal models we will be testing different learning rates and number of epochs that were found to be best for fine tuning a Bert model on a language classification task. These are 5e-5, 4e-5, 3e-5 and 2e-5 for the learning rate and 3, 4, 5 and 10 for the number of epochs.[5][24] We will be running each combination of hyper parameters 10 times and reporting the averages of the results from these runs. For each experiment we will randomly divide the dataset into three smaller sets. One for training, one for validation after each epoch and one for testing the final model. The ratio between these subsets will be 0.8/0.1/0.1 of the complete data.
The threshold at which the coronary calcium score is considered problematic enough for medication is over 100 [25], therefore we will be using 100 as the cutoff point between low and high calcium scores.
For all experiments we will be measuring the model's loss during training and the model's accuracy during and after training. We will also be measuring the precision and recall of the high calcium score predictions during and after training.
The model recall measures what fraction of the inputs that should produce a true positive for high calcium actually produce a positive result in the model. The eventual goal of the project is to offer tools for general practitioners to be better able to send certain people in for additional testing for CAD. Therefore we care much more about catching all of the high calcium score cases rather than just being as accurate as possible which is why recall is the most important metric when looking at model performance.

4.1 Answers Based Model

The first experiment considers each answer to be it's own data point where we train the model to predict whether or not an answer is linked to a high or low calcium score. The goal of this is to answer our first research question on whether or not we can use the answers to open ended questions to predict calcium scores and by extension CAD severity. This approach is however quite naive and it has several limitations.
First of all, this approach does not differentiate between the different questions that were asked, meaning that there are a lot of confounding factors differentiating the data points. For example an answer related to how long the patient suffers from the symptoms is not really comparable to an answer related to the locations of the pain, while both might be related to the complaint, they are very different data points but with this method are used the exact same way to train the model.
The reason to still run this naive experiment is to see if there happens to be some relation that can be found when looking at all of the answers as individual data points. If such a clear relation exists then this method could create a very effective model for predicting calcium score and CAD severity.
Additionally, this dataset is limited by the relatively small number of patients. By using each individual answer we maximize the number of data points we have, which potentially might make for a better model. We will be running 4 variations of this setup.
We will be testing 4 variations of this setup. There are two variables we vary among these.

The first we call Balanced or Unbalanced, which refers to whether or not we randomly reduce the data points related to low Calcium scores to be equal to those related to high calcium scores. This reduces the amount of data we have, but compensates for there being a lot more low calcium patients in the dataset. This could both improve the results by removing the imbalance, or worsen the results by reducing the size of the data.
The second variable we call Gender or noGender. This refers to whether the model is trying to predict whether the input corresponds to a male or female patient. This might tell us something about potential differences between men and women with regard to CAD predictions, but it will also make the task more complex and might worsen the model.

Unbalanced noGender For the first variation we do not balance the training data to compensate for the large difference between high and low calcium scores. We then only have the model try to predict whether the input corresponds to a low or high calcium score.

Unbalanced Gender For the second variation we do not balance the training data to compensate for the large difference between high and low calcium scores. We then have the model try to predict both low or high calcium score and male or female for the input it receives.

Balanced noGender For the third variation we balance the training data so that we have the same number of high and low calcium score data. We then only have the model try to predict whether the input corresponds to a low or high calcium score.

Balanced Gender For the fourth variation we balance the training data so that we have the same number of high and low calcium score data. We then have the model try to predict both low or high calcium score and male or female for the input it receives.

4.2 Patient Based Model

The second setup that we will be running will consider each patient as a single data point, taking the combination of all their answers as a single token list. Instead of having the model trained on single questions that might not actually be related enough to make for good training data, we consider the answers as a single element in these experiments.
The advantage of taking entire patients as the data point means that each data point can be considered more closely related to each other as they are the collection of all the answers of a single patient who has either a high or a low calcium score. Essentially creating a model that predicts whether a patient has a high or low calcium score based on the whole range of answers they have given.
However, the downside is that it reduces an already relatively small data set. Therefore, We might not be able to create an effective model due to not having a lot of data points. The result could be that any relation between the answers and the calcium scores might not be found by the model due to the lack of data.
Unlike with the previous setup we will not be running gendered variations as having only 65 male patients would mean that we would be training a model on 130 data points, which would not be able to produce a useful model. We will also not be running a balanced version as this would leave only 78 data points, which would also not be enough for the fine-tuning of a useful model. Therefore we will only be using the Unbalanced noGender variation of this model and not all 4 variations as we did with the previous setup.

4.3 Question based model

For the third experimental setup we try to find whether certain questions are more reliable predictors of a high calcium score than others. Unlike experiment 1, which considers all of the data points as equal, for this experiment we will be training separate models on the answers of each of the individual questions. This way every data point is actually related to each other in that they are answers to the same question and are either linked to a low or high calcium score. This then also takes a step to achieve the ultimate goal of the CONCRETE project, namely to develop guidelines of what a GP should look out for when talking to a patient and trying to asses the likelihood of CAD.
Unlike the previous setups we will not be using different hyper parameters, but instead we will use the hyper parameters that were most effective for the answer based model as this experiment is simply training models on subsets of the answer based model data. This also reduces the number of redundant models we will be training for this experiment.
This setup does have the same problems as the patient based model in that there is a very small dataset for each individual question. For the same reason as before we will therefore again not be balancing the data or considering gender as we are otherwise even more unlikely to have enough data to fine tune the models.

5 Results

The main results we will be reporting are the final test recall and accuracy of the different models and hyper parameters. We also measured these after each epoch to see how they develop during training. The graphs showing these training curves can be found in the appendix and parts of these will occasionally be referred to and shown for any insights they might give about the training process.

5.1 Answer Based Model

First we will look at the results of the Answer Based model where we take each individual answer as a data point. This model has 4 variations as explained in section 4.1.

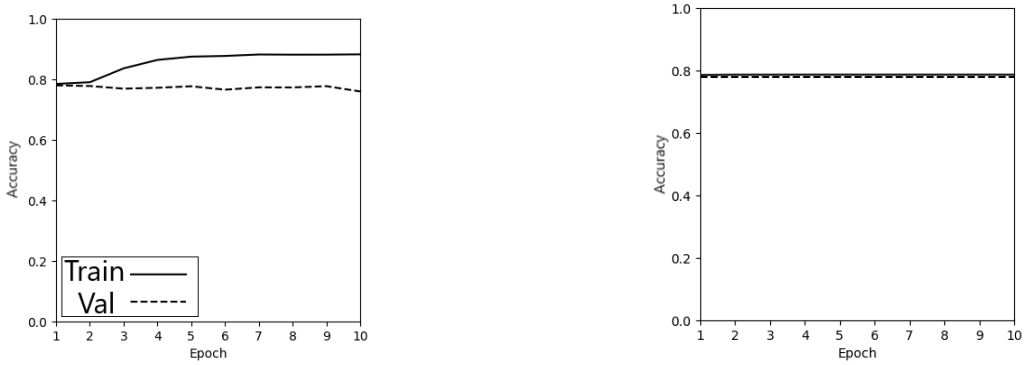
5.1.1 Unbalanced noGender

We will first look at the models where we do not balance the dataset and do not classify based on gender. We will first look at the accuracy of the models with different hyper parameters and then at the recall.

	Learning Rate			
Epochs	2e-5	3e-5	4e-5	5e-5
3	0.766	0.776	0.783	0.791
4	0.766	0.786	0.780	0.802
5	0.765	0.783	0.794	0.768
10	0.780	0.788	0.796	0.791

Table 1: Average Test Accuracy for the Unbalanced NoGender Answer Based Models. Averages taken over 10 runs.

In table 1 we see the accuracy values of the models trained with different hyper parameters when tested against the test data. Here we see that higher learning rates produce slightly higher accuracy. We also observe that having more epochs does not impact the accuracy by much.
The reported model accuracy is between 0.76 and 0.81 which is an accuracy rate that is very close to the rate of low calcium score data points in the training dataset (78.69%). This is a fairly low accuracy and not what we would expect if this data and model were a good predictor for low or high calcium score. We will go deeper into this in the conclusion.
Looking at the learning curves, we see that at the higher learning rates the model is actually doing very little learning. Accuracy pretty much remains consistent at just under 0.8 with no improvement over the epochs. The training curve being just a straight line as we can see in 4b. This means that the learning rate in those cases is too high to have the model change in a way that improves performance on the training data.
At the lower learning rates we see that the model might be learning. In 4a we can see the average learning curve of the models which were trained over 10 epochs with a learning rate of 2e-5. Here we do see that the model is improving on the training data, but that this has a small detrimental effect on the validation data. This could mean that the model is slightly overfitting on the training data and thus not improving on the actual validation data.



(a) low learning rate

(b) High learning rate

Figure 4: The average training and validation accuracy for the unbalanced noGender model. Trained for 10 epochs with a learning rate of 2e-5 (left) and 5e-5 (right).

Next up it is important that we look at the recall for these models. These can be seen in table 2. We specifically see that the higher learning rates where almost no actual learning ended up happening have a recall of zero and thus never correctly guesses a high calcium score.

	Learning Rate			
Epochs	2e-5	3e-5	4e-5	5e-5
3	0.142	0.0	0.0	0.0
4	0.112	0.055	0.0	0.0
5	0.186	0.039	0.0	0.0
10	0.131	0.030	0.0	0.0

Table 2: Average Test Recall for high calcium scores from the Unbalanced NoGender Answer Based Models. Averages taken over 10 runs.

The lower learning rates which do show some learning have higher recall rates. However these rates are still very low and not nearly high enough to be of any use for our purposes.

During training similar observations can be seen as with the accuracy. High learning rates show no learning as would be expected. The low learning rate does actually show some learning but the validation accuracy stagnates below 0.2 while training accuracy goes up much higher.

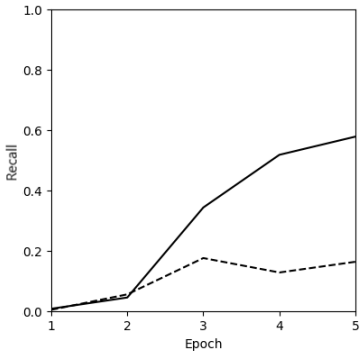


Figure 5: The average training and validation Recall for the unbalanced noGender model. Trained for 5 epochs with a learning rate of 2e-5.

5.1.2 Balanced noGender

Next up we look at the model where the data was balanced before training. Unlike the previous models which were trained on data with about a 78/22 split between low and high calcium scores, these models are trained on data with a 50/50 split between low and high calcium scores.

Looking at the test accuracy we see two distinct differences with the models trained on the unbalanced data. We no longer see the pattern that higher learning rates correlate with a higher accuracy. We also see that the accuracy is now around 0.5 instead of around 0.78, which would be the new frequency of the most common class in the dataset after balancing.

	Learning Rate			
Epochs	2e-5	3e-5	4e-5	5e-5
3	0.518	0.516	0.508	0.509
4	0.530	0.503	0.502	0.498
5	0.491	0.509	0.493	0.493
10	0.497	0.491	0.514	0.477

Table 3: Average Test Accuracy for the Balanced NoGender Answer Based Models. Averages taken over 10 runs.

The accuracy training curves for these models are the same as they were for the Unbalanced-noGender models. A low learning rate corresponding to a slight increase in training accuracy but no increase in valuation accuracy, and higher learning rates showing a flat line where no improvement from learning happens. This could be the reason why in the test results it seem that the lower max epoch values have better accuracies.

Looking at the recall results we see again that the lower learning rate outperforms higher learning rates. Interesting to note is that having a lower max number of epochs also appears to improve the results here. Combined with the same being true for accuracy might indicate some overfitting at higher epoch counts.

	Learning Rate			
Epochs	2e-5	3e-5	4e-5	5e-5
3	0.116	0.0	0.042	0.0
4	0.153	0.0	0.0	0.0
5	0.038	0.0	0.0	0.0
10	0.043	0.0	0.0	0.0

Table 4: Average Test Recall for high calcium scores from the Balanced NoGender Answer Based Models. Averages taken over 10 runs.

Looking at the recall values during training we see that they are quite different from those in the unbalanced data set. As we can see in figure 6, rather than showing the same pattern of a high training recall and a low valuation recall, we see only a small difference between the two with the values being stagnant and the lines not really ascending during training.

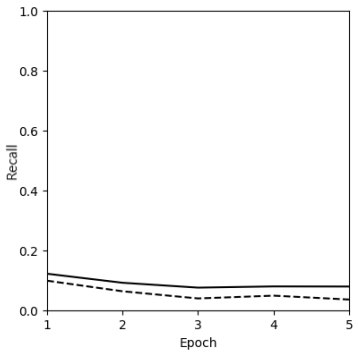


Figure 6: The average training and validation recall for the unbalanced noGender models. Trained for 5 epochs with a learning rate of 2e-5.

5.1.3 Unbalanced Gender

Next is the set of models that take gender into account, but which are trained on the unbalanced dataset.

Looking at the test accuracy in Table 5 it would seem that higher learning rates produce models with slightly higher accuracy scores. The accuracy score is around 0.5 which is also around the ratio of low calcium scores of female participants. This is the same pattern we observed in the other unbalanced dataset, where the accuracy score is about as big as the ratio of the most prominent class in the data.

Looking at the training curves we see the same results as in the first set of models, with the higher learning rates showing a flat line and the lower learning rates showing an increase in training accuracy with no increase in valuation accuracy. The only difference being that the line is lower and the difference between the training and validation lines are slightly larger.

	Learning Rate			
Epochs	2e-5	3e-5	4e-5	5e-5
3	0.499	0.527	0.517	0.529
4	0.507	0.516	0.537	0.524
5	0.520	0.503	0.520	0.520
10	0.498	0.502	0.524	0.521

Table 5: Average Test Accuracy for high calcium scores from the Unbalanced Gender models. Averages taken over 10 runs.

The test recall scores for both women and men is visible in tables 6 and 7 respectively. In these tables we see a similar pattern to previous experiments, with low recall at a learning rate of 2e-5 and 0 recall at higher learning rates. We also see that the male scores are higher than the female scores, this could be due to there being a lot less male participants.
Both the male and the female recall curves mirrors the recall curve of the model that doesn't take gender into account.

	Learning Rate			
Epochs	2e-5	3e-5	4e-5	5e-5
3	0.021	0.0	0.0	0.0
4	0.058	0.0	0.0	0.0
5	0.034	0.025	0.0	0.0
10	0.030	0.021	0.0	0.0

Table 6: Average Test Recall for female high calcium scores from the Unbalanced Gender models. Averages taken over 10 runs.

	Learning Rate			
Epochs	2e-5	3e-5	4e-5	5e-5
3	0.030	0.0	0.0	0.0
4	0.079	0.0	0.0	0.0
5	0.151	0.025	0.0	0.0
10	0.073	0.036	0.0	0.0

Table 7: Average Test Recall for male high calcium scores from the Unbalanced Gender models. Averages taken over 10 runs.

5.1.4 Balanced Gender

Finally we have the models that take gender into account and are trained on the balanced dataset. Looking at the test accuracy in Table 8 we see that the difference between the accuracy scores is relatively small, with the highest accuracy achieved at a learning rate of 2e-5. With the higher accuracies being between 0.350 and 0.396 the model accuracy is again about as high as the largest class.

Looking at the training curves we again see the same behavior as in the previous experiments with most graphs being a flat line and a few graphs at the lower learning rates showing a small increase in accuracy in the training curve, but little to no change in the validation curve.

	Learning Rate			
Epochs	2e-5	3e-5	4e-5	5e-5
3	0.396	0.349	0.350	0.344
4	0.378	0.321	0.330	0.336
5	0.333	0.357	0.321	0.340
10	0.333	0.352	0.344	0.335

Table 8: Average Test Accuracy for high calcium scores from the Balanced Gender models. Averages taken over 10 runs.

The test recall scores for both women and men is visible in tables 9 and 10 respectively. In these tables we see a similar pattern to previous experiments, with low recall at a learning rate of 2e-5 and 0 recall at higher learning rates. For the Female High calcium recall the highest score achieved is only 0.142 which is lower than the proportion of female high calcium score participants. Interestingly, the highest recall for the Male High calcium recall is 0.569 which is actually higher than the accuracy and the proportion of male high calcium participants in the total data.

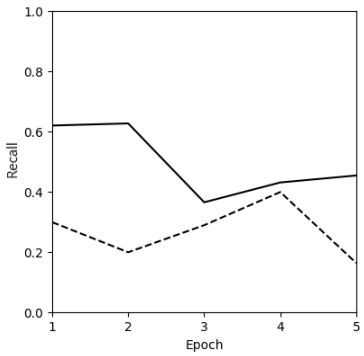


Figure 7: The average male training and validation recall for the balanced gender models. Trained for 5 epochs with a learning rate of 2e-5.

When considering the training curves, the female training curves are very similar to those of the balanced model without gender, which makes sense considering that most of the data in the full dataset is data from female participants. The male recall training curves on the other hand are erratic and to not be following a clear pattern as can be seen in figure 7

	Learning Rate			
Epochs	2e-5	3e-5	4e-5	5e-5
3	0.142	0.0	0.0	0.0
4	0.099	0.0	0.0	0.0
5	0.088	0.003	0.0	0.0
10	0.096	0.0	0.0	0.0

Table 9: Average Test Recall for female high calcium scores from the Balanced Gender models. Averages taken over 10 runs.

	Learning Rate			
Epochs	2e-5	3e-5	4e-5	5e-5
3	0.569	0.0	0.0	0.0
4	0.395	0.0	0.0	0.0
5	0.192	0.100	0.0	0.0
10	0.259	0.0	0.0	0.0

Table 10: Average Test Recall for male high calcium scores from the Balanced Gender models. Averages taken over 10 runs.

5.2 Patient Based Model

The patient based model takes the aggregate of all answers of a patient as a single data point. Though gender and balancing are not the case here due to limited data, the results are still worth discussing. Looking at the accuracy we see that overall they are lower than in the answer based model, but not by much. There also does not appear to be a clear pattern where a lower learning rate or a higher number of epochs results in a better model. The accuracy results are still around the ratio of the low and high calcium score participants.

	Learning Rate			
Epochs	2e-5	3e-5	4e-5	5e-5
3	0.762	0.756	0.745	0.792
4	0.770	0.731	0.748	0.784
5	0.742	0.765	0.734	0.740
10	0.770	0.764	0.784	0.756

Table 11: Average Test Accuracy for high calcium scores from the Patient Based models. Averages taken over 10 runs.

Looking at the training curves we see again that they depict the same behavior as the other models. The training accuracy is going up while the validation accuracy stays about the same.
Recall for this set of models does show a pattern in the test results. It would seem that a lower learning rate corresponds to better recall and that 5 epochs is the optimal number as training for less or more

epochs does not seem to improve the recall. We also see that the maximum recall achieved is very close to the ratio of high calcium participants.

	Learning Rate			
Epochs	2e-5	3e-5	4e-5	5e-5
3	0.153	0.136	0.028	0.066
4	0.130	0.084	0.124	0.144
5	0.250	0.129	0.131	0.188
10	0.252	0.120	0.118	0.011

Table 12: Average Test Recall for high calcium scores from the Patient Based models. Averages taken over 10 runs.

Looking at the training curves, we see a more extreme version of the recall training curve that we found in section 5.1.1. We see a sharply rising training training result, but validation remains very low. The only different this time is that training goes up much higher and validation slightly increases at the lowest learning rate when only training for 5 epochs.

5.3 Question Based Model

Looking mainly at recall, we got the best results in the unbalanced Answer Based Model for a learning rate of 2e-5 with 5 epochs of training. Because of this we decided to use those hyper parameters for the final experiment as we will be analyzing the same data. In figure 8 we can see the test recall and accuracy per question of the models that were trained for the Question Based Models.

The results we can observe are not much better than the ones we observed in the model with the same hyper parameters but with all answers used for training. This is mainly true for the accuracy which appears to hover around 0.75. In the recall we do see more variance with some questions like question 6 and question 11 having a recall of almost 0, while other questions like 4 and 12 are closer to 0.25. However, these recall rates are still much lower than what would be ideal.

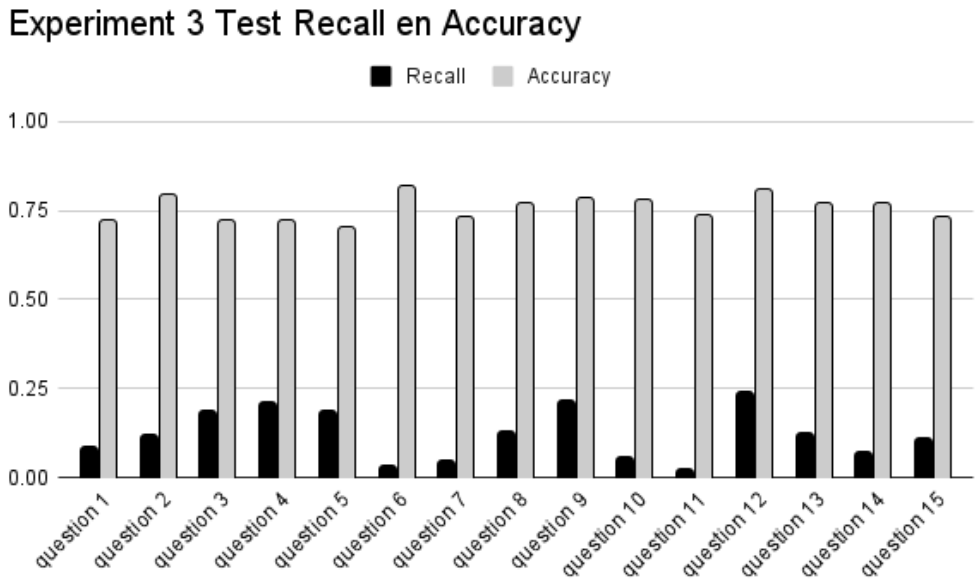


Figure 8: The final test accuracy and high calcium recall for each question. Average taken over 10 training runs.

If we then look at the training process we can see the same behavior as with the previous models, where training scores accuracy increase without much improvement in the valuation scores. We also observe the same for recall, with the more aggressive increase in training score but a very low validation score during training.

6 Discussion

The results show that we were unable to develop a satisfactory model for predicting CAD severity based on the current data. Both the accuracy and the recall were far below any value that could be usable in practice. This happens regardless of whether the model was balanced based on calcium scores or had classes that were differentiated by gender.

We have to draw an unfortunate conclusion from the results as they are now. In each situation the best accuracy value we can find is around the frequency of the largest class in the dataset. Additionally the highest high-calcium recall values we can find are not much higher than the frequency of high calcium data in the training dataset. This is an indication that instead of the models training to recognize patterns in the contents of the data, the models are simply training to predict the classes based on their frequency in the training data.

Looking at both the patient and question based datasets we see the exact same issue with accuracy around the frequency of low calcium scores in the training data (78.69%) and high calcium recall in even the most ideal circumstances reaching only a level similar to the frequency of high calcium entries in the data (21.31%). With certain hyper parameters we can achieve slightly better results than this, but at best this could be an indication of an avenue for future research as these results are still very low and are not even close to being useful for any kind of clinical applications.

Potentially it might be possible to draw some very careful conclusions from the question based model. Some questions appear to result in better recall, which might be an indication that the answers to those question contain more information that might indicate a high calcium score. However, even the best performing models here are still terrible and at best this only highlights potential future directions to research without providing a working model.

6.1 Conclusion

From our results we see that the models are very likely overfitting on the data, as some models have training accuracy increasing with each epoch, but valuation accuracy not following suite. It would appear that the models are closer to guessing with frequencies than actually training to predict CAD severity. This leaves four explanations for what is causing these results.

One explanation is that not every data point in the dataset might be useful. As we saw in the result on the question based model experiment, some questions appear to have less value than others. This could mean that not all of the answers contain useful information and these answers might be contaminating the data adding too much noise for proper training.

Another potential factor could be the model architecture itself. Potentially this might simply not be the right method to try and extract any useful information from the CONCRETE dataset.

Another possible explanation is the size of the dataset being too limited. The total dataset contains 144 questionnaires containing 2745 answers. This is already fairly limited for training a BERT model. Additionally only 585 (or 21.3%) of the answers were related to a high calcium score. This means that there is a lot of pressure for the model to prefer one class over the other which we clearly see in the unbalanced models. Balancing for this even further reduces the already limited training data to only 1170 answers and does not produce good models either.

Finally there is the possibility that these questions and answers are simply not enough to accurately predict CAD severity. The lack of results means either the method failed to find a relation between the answers and CAD severity or that these answers are not a good indication of CAD severity.

6.2 Summary

In conclusion, there is very little we can conclude from our experiments. The main reason for this that the results point to, is that the amount of data is not large enough to properly fine tune a BERTje model on. Neither balancing the training data, differentiating between the genders, merging answers per patient or analyzing only specific question succeeded in training a BERTje model that could perform much better than guessing based on frequency. This leaves us with the following answers to our research questions.

Are there certain symptoms and risk factors that correspond to different levels of calcium scores which can be used for better assessment of CAD risk? Very carefully we can suggest that some questions in this data set might be better suited than others for future research into this. However, no actual good models were produced and it might be the case that these risk factors can not be deduced based only on the answers to these questions.

Can a patient's claims and description of their symptoms be used for robust CAD severity prediction and gender stratification? Had we found a model that could predict high calcium scores with good accuracy then we could have concluded that patient claims and descriptions provide sufficient information to predict CAD severity. Sadly this did not happen and with the current dataset and the BERTje model it appears that robust CAD severity prediction and gender stratification is not possible. This can be either due to insufficient data, limitations of the method or it could be that patient claims and description are not sufficient on their own to determine CAD severity.

6.3 Reflection and future work

The CONCRETE project is far from finished and the dataset is still being expanded. This could potentially help deal with the biggest issue that we found in this research, namely the lack of sufficient data. So for future research it would be good to expand upon the experiments performed in this paper with a larger dataset.

In hindsight it should not have been a surprise that we did not have enough data to perform the task we set out to perform. Having access to only 183 questionnaires for training means that performing any kind of task using a large language model would be incredibly difficult. Even if there is relation between the questionnaires and the calcium scores, this dataset might not be enough to find that connection.

Considering the data limitation, is it worth the resources to continue this line of research or is that effort better used elsewhere? Continuing with the method performed by us might not be the most productive course of action. The most successful NLP experiments with similar goals have been on much larger data sets that were also of a higher quality. The main datasets used are often health records which contain a lot of information about the patient. The answers to our questionnaires only concern a few symptoms and include a lot of noise as the information from the patient is not filtered through a healthcare provider who might function as a filter that removes some unnecessary information.

Additionally Koleck et al [7] make a valid point about the burden related to symptoms and the lack of NLP research into symptoms themselves. Most NLP research focuses on extracting symptoms for the purpose of classifying diseases. Though it might make for more interesting results to be able to automate a diagnosis, studying symptoms themselves and the way they are documented in health records may very well be just as valuable. Looking at the concrete data it could then perhaps be better

to focus more on what the answers say about the symptoms patients are experiencing rather than just trying to use the answers to get to a diagnosis.
This also highlights the danger of enthusiasm about AI having a detrimental effect on the allocation of resources. A project that might have insufficient data and maybe not add as much as it could, but which sounds very impressive, might get prioritized over more fruitful projects that might appear more boring in what they set out to achieve. For this reason it would help if AI researcher are not just alert about which projects they work on, but actively go out of their way to inform other researchers about the realistic limitations and possibilities of current AI methods.
When looking at our research, another problem is that the population from which data can be gathered is somewhat limited. We want to research dutch patients but because of this we are limited by the dutch CAD patient population. When we compare this to being able to gather data in more common languages it makes sense that it would be much harder to create a sizable dataset. The CONCRETE project itself was also not just for training large language models but is a broader project. This means that the data gathering setup was not optimized for gathering data that would work well for training an NLP model.
Considering this it is vital that future research into this topic makes sure that the dataset used is of sufficient quality to perform these experiments on. It could also be worthwhile to reevaluate the value of this line of research and the value of perhaps switching the focus towards something more achievable.
In the case that there is still a desire to continue this method and research, one oversight of our research that was not taken into consideration was balancing the genders as well as the calcium scores. Though this would further reduce the data, it could maybe provide better results as it would take away the imbalance of male and female participants. Further analyzing the individual questions and seeing whether some might be adding more noise than they are adding information could perhaps also assist in future research.
Finally the kinds of models and methods that could be released upon the CONCRETE data is far from exhausted and the state of the art is ever changing. Other ways of using BERT models or other NLP models could potentially proof to be more useful for the CONCRETE project's goals.
In conclusion, reevaluating the value of the line of research we have explored here might be worthwhile. If it is concluded that continuing with this research is still the best option, then there are still multiple possibilities that have not yet been explored in which this data can be analyzed and processed.

Bibliography

[1] WHO, "The top 10 causes of death." https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death, December 2020. Accessed June 2024.

[2] R. Luengo-Fernandez, M. Walli-Attaei, A. Gray, A. Torbica, A. P. Maggioni, R. Huculeci, F. Bairami, V. Aboyans, A. D. Timmis, P. Vardas, and J. Leal, "Economic burden of cardiovascular diseases in the European Union: a population-based cost study," *European Heart Journal*, vol. 44, pp. 4752–4767, August 2023.

[3] "Concrete." https://doi.org/10.34760/65265f7516662, 2019 - 2024. Accessed June 2024.

[4] M. Y. Koopman, J. J. W. Reijnders, R. T. A. Willemsen, R. van Bruggen, C. J. M. Doggen, B. Kietselaer, M. J. O. Wolcherink, P. M. A. van Ooijen, J. W. C. Gratama, M. O. R. Braam, P. van der Harst, G. J. Dinant, and R. Vliegenthart, "Coronary calcium scoring as first-line test to detect and exclude coronary artery disease in patients presenting to the general practitioner with stable chest pain: protocol of the cluster-randomised concrete trial," *BMJ open*, vol. 12, April 2022.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[6] K. Buchan, M. Filannino, and Özlem Uzuner, "Automatic prediction of coronary artery disease from clinical narratives," *Journal of Biomedical Informatics*, vol. 72, pp. 23–32, 2017.

[7] T. A. Koleck, C. Dreisbach, P. E. Bourne, and S. Bakken, "Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review," *Journal of the American Medical Informatics Association*, vol. 26, no. 4, pp. 362–379, 2019.

[8] W. de Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, and M. Nissim, "Bertje: A dutch bert model," December 2019.

[9] U. Ralapanawa and R. Sivakanesan, "Epidemiology and the magnitude of coronary artery disease and acute coronary syndrome: A narrative review," *Journal of epidemiology and global health*, vol. 11, no. 2, pp. 169–177, 2021.

[10] R. D. Shahjehan and B. S. Bhutta, "Coronary artery disease." https://www.ncbi.nlm.nih.gov/books/NBK564304/, August 2023. StatPearls [Internet], Accessed June 2024.

[11] C. Hermiz and Y. R. Sedhai, "Angina." https://www.ncbi.nlm.nih.gov/books/NBK557672/, June 2024. StatPearls [Internet], Accessed June 2024.

[12] B. A. Bergmark, N. Mathenge, P. A. Merlini, M. B. Lawrence-Wright, and R. P. Giugliano, "Acute coronary syndromes," *Lancet*, vol. 399, pp. 1347–1358, April 2022.

[13] M. Y. Koopman, R. T. A. Willemsen, P. van der Harst, R. van Bruggen, J. W. C. Gratama, R. Braam, P. M. A. van Ooijen, C. J. M. Doggen, G. J. Dinant, B. Kietselaer, and R. Vliegenthart, "The diagnostic and prognostic value of coronary calcium scoring in stable chest pain patients: A narrative review.," *oFo : Fortschritte auf dem Gebiete der Rontgenstrahlen und der Nuklearmedizin*, vol. 194, p. 257–265, March 2022.

[14] D. Shreya, D. I. Zamora, G. S. Patel, I. Grossmann, K. Rodriguez, M. Soni, P. K. Joshi, S. C. Patel, and I. Sange, "Coronary artery calcium score - a reliable indicator of coronary artery disease?," *Cureus*, vol. 13, 12 2021.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[16] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (M. Walker, H. Ji, and A. Stent, eds.), (New Orleans, Louisiana), pp. 2227–2237, Association for Computational Linguistics, June 2018.

[17] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding with unsupervised learning," *Technical report, OpenAI*, 2018.

[18] L. Korel, U. Yorsh, A. Behr, N. Kockmann, and M. Holena, "Text-to-ontology mapping via natural language processing with application to search for relevant ontologies in catalysis," *Computers*, vol. 12, p. 14, January 2023.

[19] S. Wu and M. Dredze, "Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), (Hong Kong, China), pp. 833–844, Association for Computational Linguistics, Nov. 2019.

[20] B. van der Burgh and S. Verberne, "The merits of universal language model fine-tuning for small datasets – a case with dutch book reviews," 2019.

[21] R. Alizadehsani, M. Abdar, M. Roshanzamir, A. Khosravi, P. M. Kebria, F. Khozeimeh, S. Nahavandi, N. Sarrafzadegan, and U. R. Acharya, "Machine learning-based coronary artery disease diagnosis: A comprehensive review," *Computers in biology and medicine*, vol. 111, p. 103346, August 2019.

[22] E. H. Houssein, R. E. Mohamed, and A. A. Ali, "Heart disease risk factors detection from electronic health records using advanced nlp and deep learning techniques," *Scientific Reports*, vol. 13, May 2023.

[23] B. Marinov, "Machine learning classification of the coronary artery disease and clustering of free-form medical complaints," Master's thesis, University of Groningen, 2021.

[24] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?," February 2020.

[25] I. Golub, S. Lakshmanan, S. Dahal, and M. J. Budoff, "Utilizing coronary artery calcium to guide statin use," *Atherosclerosis*, vol. 326, pp. 17–24, June 2021.

Appendices

Appendix A: Health Questionnaire

code

Gezondheidsvragenlijst

Nederlandse versie voor Nederland

Zet bij iedere groep in de lijst hieronder een kruisje in het hokje dat het best past bij uw gezondheid VANDAAG.

1. MOBILITEIT

Ik heb geen problemen met lopen ❑
Ik heb een beetje problemen met lopen ❑
Ik heb matige problemen met lopen ❑
Ik heb ernstige problemen met lopen ❑
Ik ben niet in staat om te lopen ❑

2. ZELFZORG

Ik heb geen problemen met mijzelf wassen of aankleden ❑
Ik heb een beetje problemen met mijzelf wassen of aankleden ❑
Ik heb matige problemen met mijzelf wassen of aankleden ❑
Ik heb ernstige problemen met mijzelf wassen of aankleden ❑
Ik ben niet in staat mijzelf te wassen of aan te kleden ❑

3. DAGELIJKSE ACTIVITEITEN *(bijv. werk, studie, huishouden, gezins- en vrijetijdsactiviteiten)*

Ik heb geen problemen met mijn dagelijkse activiteiten ❑
Ik heb een beetje problemen met mijn dagelijkse activiteiten ❑
Ik heb matige problemen met mijn dagelijkse activiteiten ❑
Ik heb ernstige problemen met mijn dagelijkse activiteiten ❑
Ik ben niet in staat mijn dagelijkse activiteiten uit te voeren ❑

4. PIJN / ONGEMAK

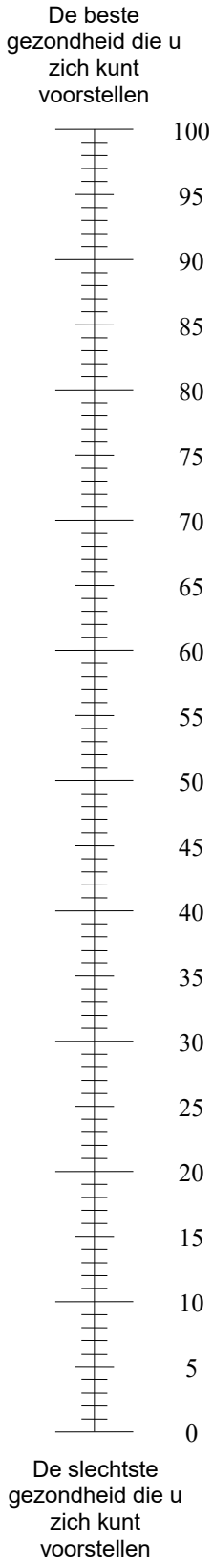
Ik heb geen pijn of ongemak ❑
Ik heb een beetje pijn of ongemak ❑
Ik heb matige pijn of ongemak ❑
Ik heb ernstige pijn of ongemak ❑
Ik heb extreme pijn of ongemak ❑

5. ANGST / SOMBERHEID

Ik ben niet angstig of somber ❑
Ik ben een beetje angstig of somber ❑
Ik ben matig angstig of somber ❑
Ik ben erg angstig of somber ❑
Ik ben extreem angstig of somber ❑

- We willen weten hoe goed of slecht uw gezondheid VANDAAG is.
- Deze meetschaal loopt van 0 tot 100.
- 100 staat voor de beste gezondheid die u zich kunt voorstellen.
  0 staat voor de slechtste gezondheid die u zich kunt voorstellen.
- Markeer een X op de meetschaal om aan te geven hoe uw gezondheid VANDAAG is.
- Noteer het getal waarbij u de X heeft geplaatst in onderstaand vakje.

UW GEZONDHEID VANDAAG = ☐



De volgende vragen zijn meerkeuze vragen, waarbij u per vraag 1 antwoord kunt geven. Kruis het antwoord aan die het beste bij u ervaring past.

7. Heeft u door uw hartprobleem in de voorbije 4 weken hinder ondervonden met het doen van binnenshuis wandelen (gelijkvloers)?

- ☐ Heel veel hinder ondervonden
- ☐ Veel hinder ondervonden
- ☐ Een beetje hinder ondervonden
- ☐ Geen hinder ondervonden

8. Heeft u door uw hartprobleem in de voorbije 4 weken hinder ondervonden met het doen van tuinieren, stofzuigen of boodschappen dragen?

- ☐ Heel veel hinder ondervonden
- ☐ Veel hinder ondervonden
- ☐ Een beetje hinder ondervonden
- ☐ Geen hinder ondervonden

9. Heeft u door uw hartprobleem in de voorbije 4 weken hinder ondervonden met zonder te stoppen één heuvel of één trap op te lopen?

- ☐ Heel veel hinder ondervonden
- ☐ Veel hinder ondervonden
- ☐ Een beetje hinder ondervonden
- ☐ Geen hinder ondervonden

10. Heeft u door uw hartprobleem in de voorbije 4 weken hinder ondervonden met het in een stevig tempo meer dan 100 meter lopen?

- ☐ Heel veel hinder ondervonden
- ☐ Veel hinder ondervonden
- ☐ Een beetje hinder ondervonden
- ☐ Geen hinder ondervonden

11. Heeft u door uw hartprobleem in de voorbije 4 weken hinder ondervonden met zware dingen optillen of verplaatsen?

- ☐ Heel veel hinder ondervonden
- ☐ Veel hinder ondervonden
- ☐ Een beetje hinder ondervonden
- ☐ Geen hinder ondervonden

12. Heeft u door uw hartprobleem in de voorbije 4 weken hinder ondervonden van gevoelens van kortademigheid?

- ☐ Heel veel hinder ondervonden
- ☐ Veel hinder ondervonden

☐ Een beetje hinder ondervonden
☐ Geen hinder ondervonden

13. Heeft u door uw hartprobleem in de voorbije 4 weken hinder ondervonden van zich in het algemeen lichamelijk belemmerd voelen?

☐ Heel veel hinder ondervonden
☐ Veel hinder ondervonden
☐ Een beetje hinder ondervonden
☐ Geen hinder ondervonden

14. Heeft u door uw hartprobleem in de voorbije 4 weken hinder ondervonden van zich moe of uitgeput voelen of weinig energie hebben?

☐ Heel veel hinder ondervonden
☐ Veel hinder ondervonden
☐ Een beetje hinder ondervonden
☐ Geen hinder ondervonden

15. Heeft u door uw hartprobleem in de voorbije 4 weken hinder ondervonden van zich niet ontspannen of rustig voelen?

☐ Heel veel hinder ondervonden
☐ Veel hinder ondervonden
☐ Een beetje hinder ondervonden
☐ Geen hinder ondervonden

16. Heeft u door uw hartprobleem in de voorbije 4 weken hinder ondervonden van zich depressief voelen?

☐ Heel veel hinder ondervonden
☐ Veel hinder ondervonden
☐ Een beetje hinder ondervonden
☐ Geen hinder ondervonden

17. Heeft u door uw hartprobleem in de voorbije 4 weken hinder ondervonden van zich gefrustreerd voelen?

☐ Heel veel hinder ondervonden
☐ Veel hinder ondervonden
☐ Een beetje hinder ondervonden
☐ Geen hinder ondervonden

18. Heeft u door uw hartprobleem in de voorbije 4 weken hinder ondervonden van zich ongerust maken?

- ☐ Heel veel hinder ondervonden
- ☐ Veel hinder ondervonden
- ☐ Een beetje hinder ondervonden
- ☐ Geen hinder ondervonden

19. Heeft u door uw hartprobleem in de voorbije 4 weken hinder ondervonden van beperkt zijn in sporten of lichaamsbeweging?

- ☐ Heel veel hinder ondervonden
- ☐ Veel hinder ondervonden
- ☐ Een beetje hinder ondervonden
- ☐ Geen hinder ondervonden

20. Heeft u door uw hartprobleem in de voorbije 4 weken hinder ondervonden van werken in huis of in de tuin?

- ☐ Heel veel hinder ondervonden
- ☐ Veel hinder ondervonden
- ☐ Een beetje hinder ondervonden
- ☐ Geen hinder ondervonden

De laatste vragen van deze vragenlijst zijn open vragen en gaan over de klachten waarmee u naar de huisarts bent gegaan.

21. Kunt u aangeven met wat voor soort pijn of gevoelsklacht(en) u naar de huisarts bent gegaan?

……
……
……
……

22. Heeft u naast de pijn- of gevoelsklacht(en) nog andere klachten ervaren?

……
……
……
……

23. Hoe lang duren de door u beschreven pijn- of gevoelsklacht(en)?

……
……
……
……

24. Waar zit de door u beschreven pijn- of gevoelsklacht(en) precies?

……
……
……
……

25. Stralen de door beschreven pijn- of gevoelsklacht(en) uit? (zo ja, waar naar toe)

……
……
……
……

26. Hoe ernstig zijn de door u beschreven pijn- of gevoelsklacht(en) op een schaal van 1 (mild) tot 10 (zeer ernstig)?

……

……

27. Moet u door de door u beschreven pijn- of gevoelsklacht(en) bepaalde activiteiten laten welke u gewoonlijk wel deed? En welke activiteiten zijn dit?

……

……

……

……

28. Hoe vaak heeft u de door u beschreven pijn- of gevoelsklacht(en)?

……

……

……

29. Hoe verlopen de door u beschreven pijn- of gevoelsklacht(en) per keer?

……

……

……

……

30. Kunt u de situatie beschrijven waarbij u voor het eerst de door u beschreven pijn- of gevoelsklacht(en) voelde?

……

……

……

……

31. Was er naar uw mening een duidelijke aanleiding voor de door u beschreven pijn- of gevoelsklacht(en)?

……
……
……
……

32. Waar denkt u dat de door u beschreven pijn- of gevoelsklacht(en) mee te maken hebben?

……
……
……
……

33. Waardoor worden de door u beschreven pijn- of gevoelsklachten minder?

……
……
……

34. Waardoor worden de door u beschreven pijn- of gevoelsklachten erger?

……
……
……
……

35. Wat betekenen de door u beschreven pijn- of gevoelsklacht(en) voor uw dagelijks leven en/of maakt u zich hier zorgen over?

……
……
……
……

36. Datum waarop de vragenlijst is ingevuld: *……...*

Hartelijk dank voor het invullen van deze vragenlijst!

Met vriendelijke groet,
M.Y. (Moniek) Koopman

Appendix B: Open questions with Translation

Question Number	Original question (dutch)	Translated question (English)
Question 1	Kunt u aangeven met wat voor soort pijn of gevoelsklacht(en) u naar de huisarts bent gegaan?	Can you indicate the type of pain or sensation complaint(s) you went to the doctor with?
Question 2	Heeft u naast de pijn- of gevoelsklacht(en) nog andere klachten ervaren?	Have you experienced any other symptoms besides the pain or sensation complaint(s)?
Question 3	Hoe lang duren de door u beschreven pijn- of gevoelsklacht(en)?	How long do the pain or sensation complaint(s) you described last?
Question 4	Waar zit de door u beschreven pijn- of gevoelsklacht(en) precies?	Where exactly is the pain or sensation complaint(s) you described located?
Question 5	Stralen de door u beschreven pijn- of gevoelsklacht(en) uit? (zo ja, waar naar toe)	Do the pain or sensation complaint(s) described by you radiate (if so, where to)?
Question 6	Hoe ernstig zijn de door u beschreven pijn- of gevoelsklacht(en) op een schaal van 1 (mild) tot 10 (zeer ernstig)?	On a scale of 1 (mild) to 10 (very severe), how severe are the pain or sensation complaint(s) you described?
Question 7	Moet u door de door u beschreven pijn- of gevoelsklacht(en) bepaalde activiteiten laten welke u gewoonlijk wel deed? En welke activiteiten zijn dit?	Do the pain or sensation complaint(s) you described require you to no longer do certain activities that you used to do? And what activities are these?
Question 8	Hoe vaak heeft u de door u beschreven pijn- of gevoelsklacht(en)?	How often do you have the pain or sensation complaint(s) you described?
Question 9	Hoe verlopen de door u beschreven pijn- of gevoelsklacht(en) per keer?	In what manner do the pain or sensation complaint(s) you described occur at each time?
Question 10	Kunt u de situatie beschrijven waarbij u voor het eerst de door u beschreven pijn- of gevoelsklacht(en) voelde?	Can you describe the situation in which you first felt the pain or sensation complaint(s) you described?
Question 11	Was er naar uw mening een duidelijke aanleiding voor de door u beschreven pijn- of gevoelsklachten?	In your opinion, was there a clear cause for the pain or sensation complaint(s) you described?
Question 12	Waar denkt u dat de door u beschreven pijn- of gevoelsklacht(en) mee te maken hebben?	What do you think the pain or sensation complaint(s) you described are related to?
Question 13	Waardoor worden de door u beschreven pijn- of gevoelsklachten minder?	What reduces the pain or sensation complaint(s) you described?
Question 14	Waardoor worden de door u beschreven pijn- of gevoelsklachten erger?	What increase the pain or sensation complaint(s) you described?
Question 15	Wat betekenen de door u beschreven pijn- of gevoelsklacht(en) voor uw dagelijks leven en/of maakt u zich hier zorgen over?	How do the pain or sensation symptom(s) you described affect your daily life and/or are you worried about this?

Appendix C: Results Answer Based Model

Training curves: Unbalanced noGender

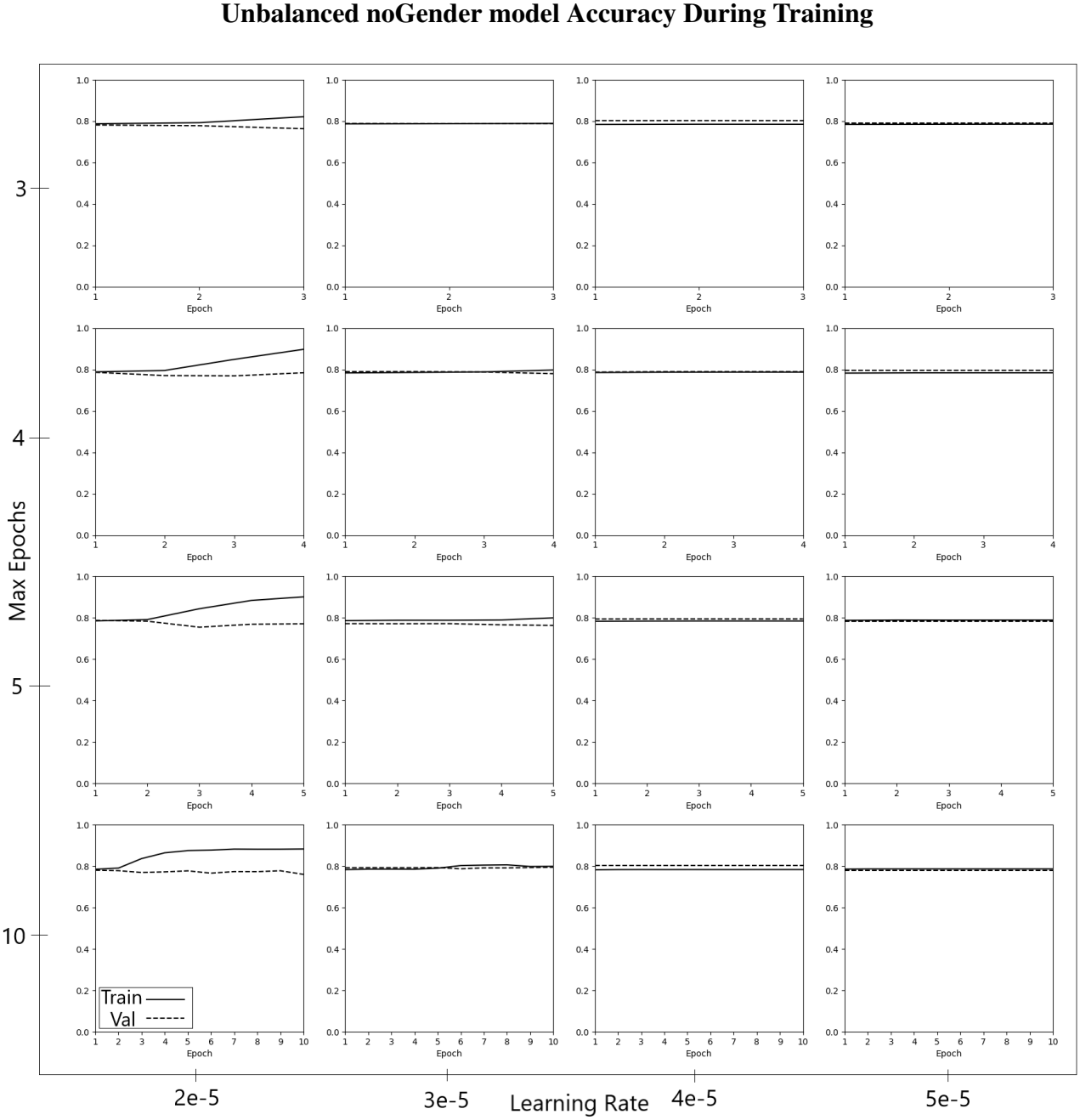


Figure 9: The average training a validation accuracy per learning rate and max epochs combination. Average taken over 10 training runs.

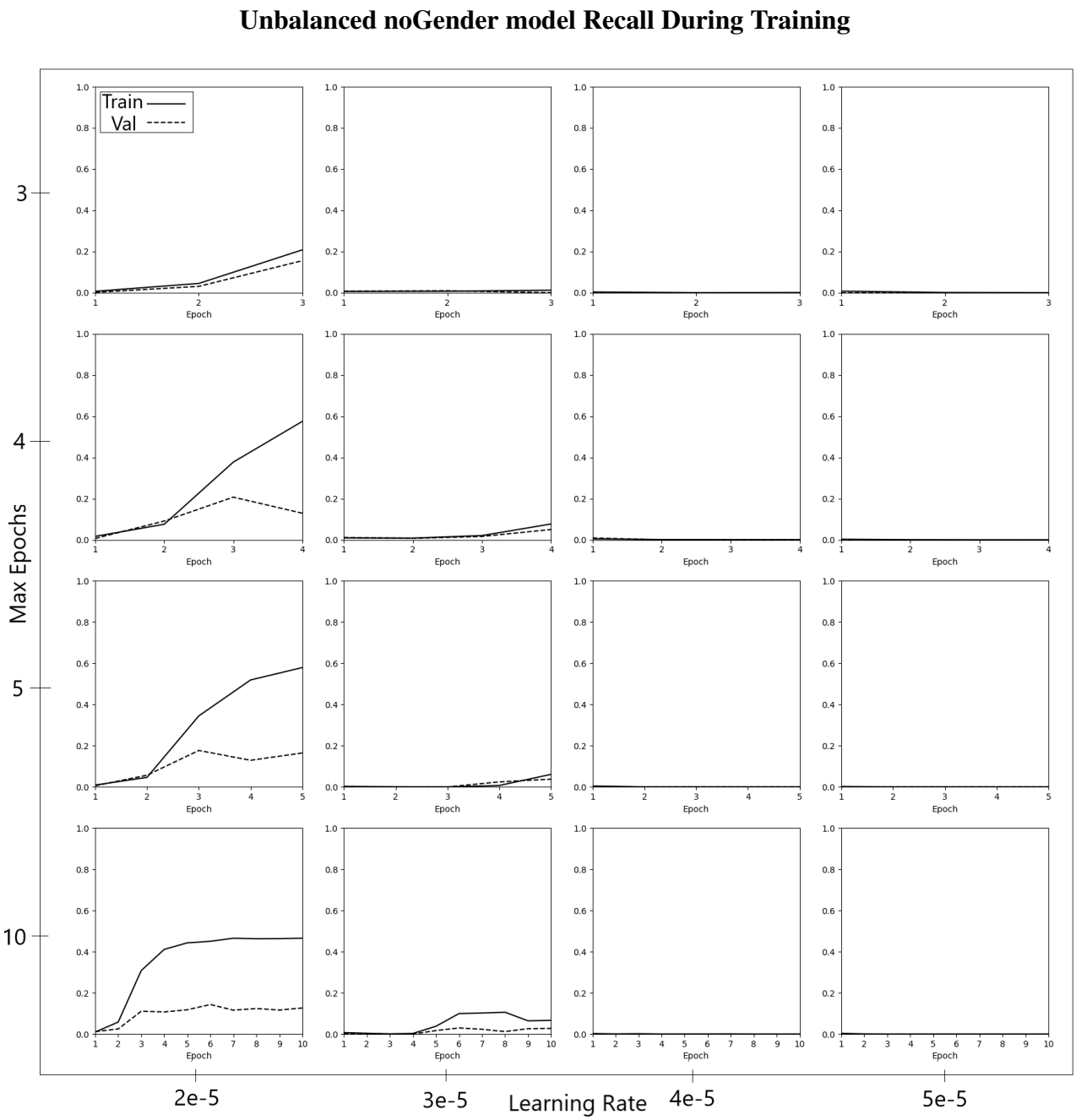


Figure 10: The average training and validation recall per learning rate and max epochs combination. Average taken over 10 training runs.

Training curves: balanced noGender

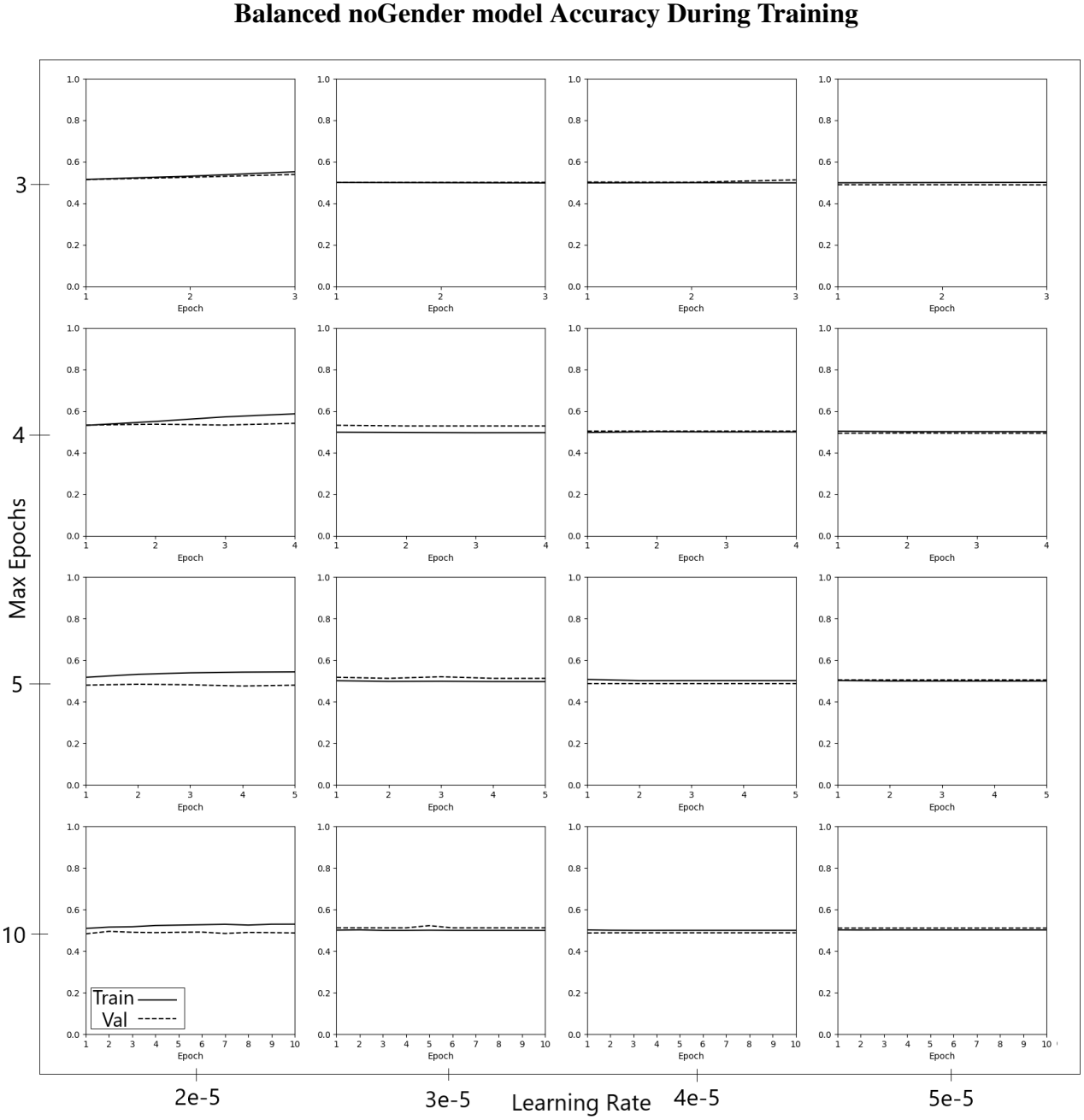


Figure 11: The average training and validation accuracy per learning rate and max epochs combination. Average taken over 10 training runs.

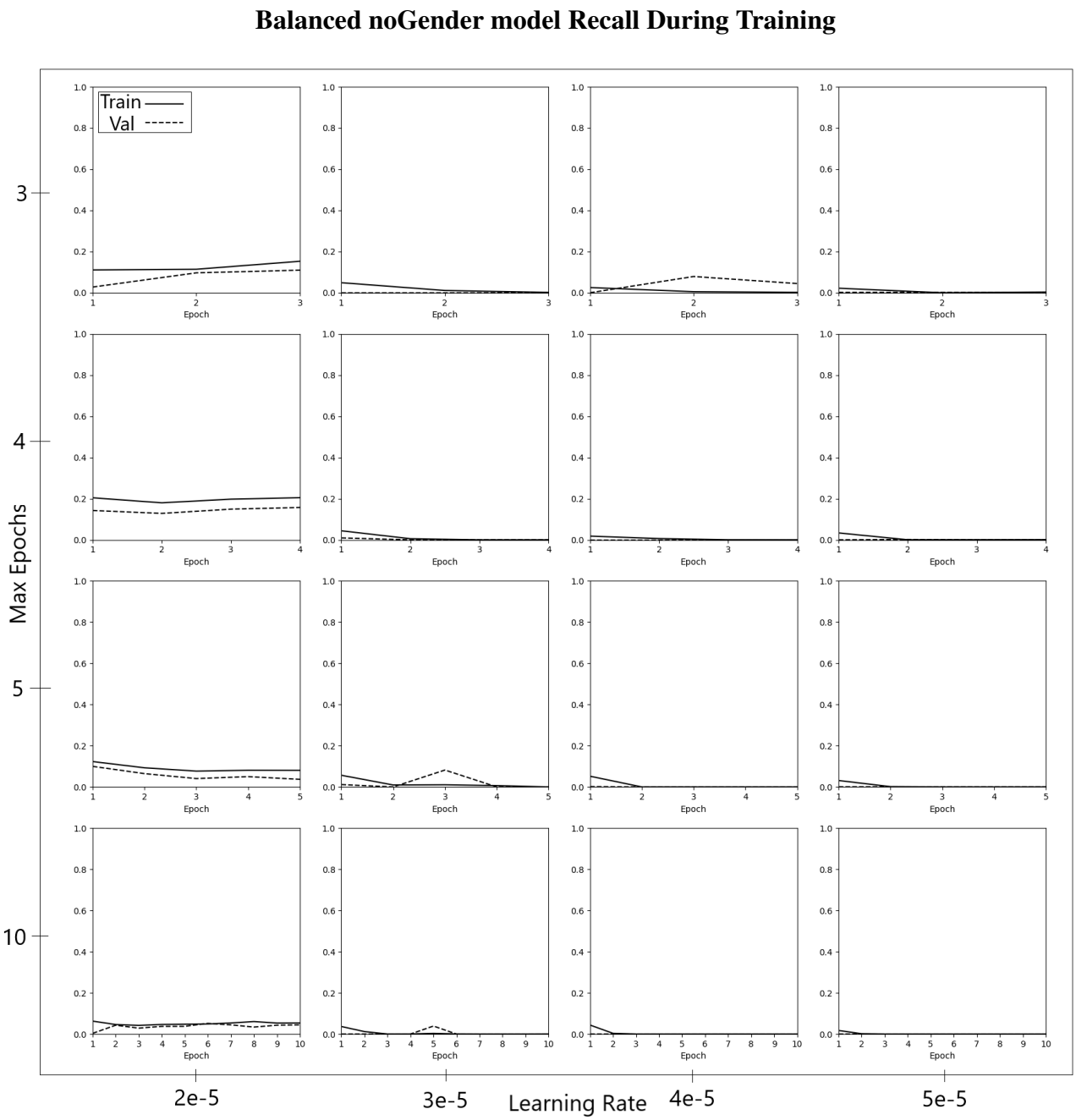


Figure 12: The average training and validation Recall per learning rate and max epochs combination. Average taken over 10 training runs.

Training curves: unbalanced gender

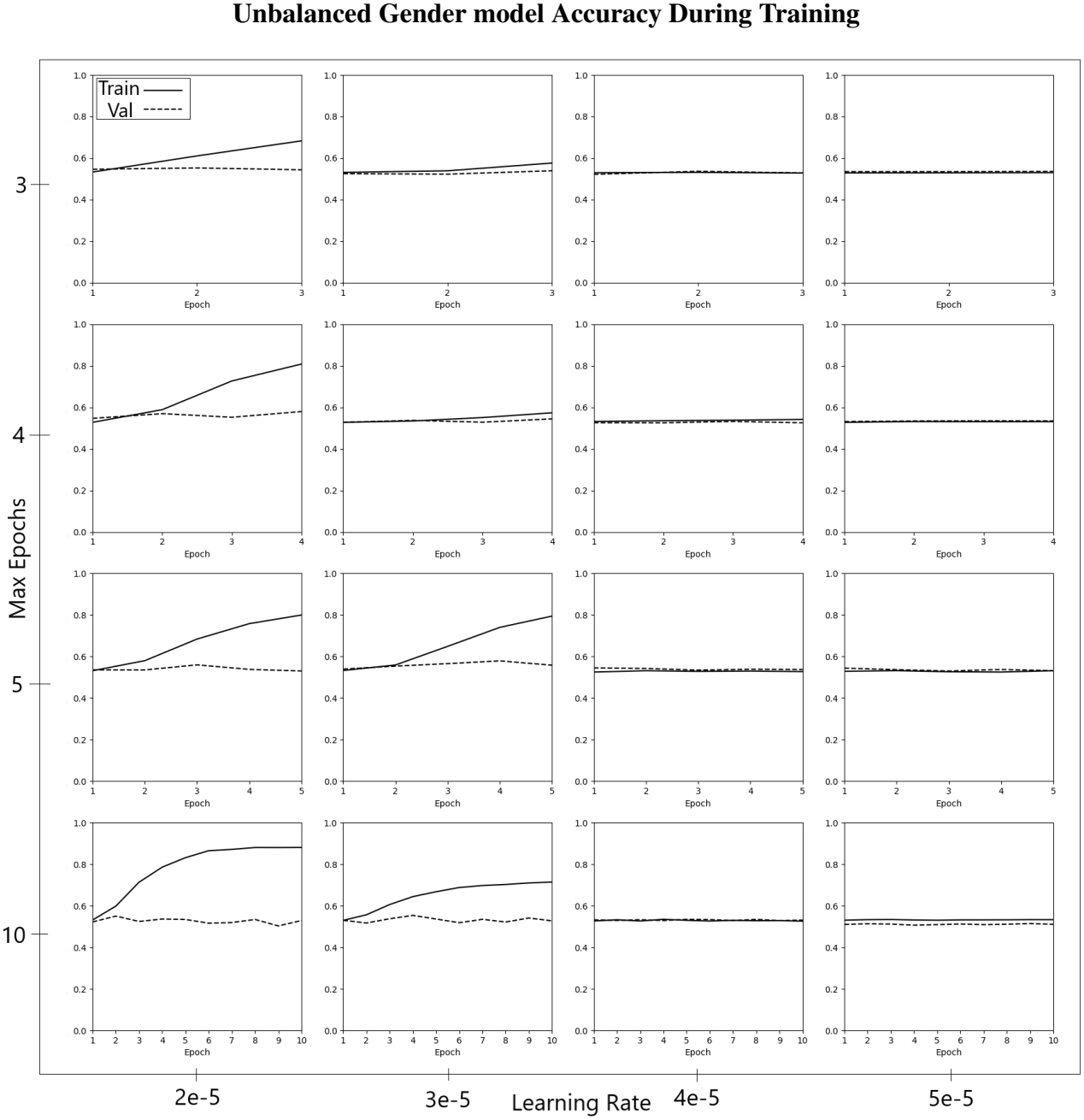


Figure 13: The average training and validation accuracy per learning rate and max epochs combination. Average taken over 10 training runs.

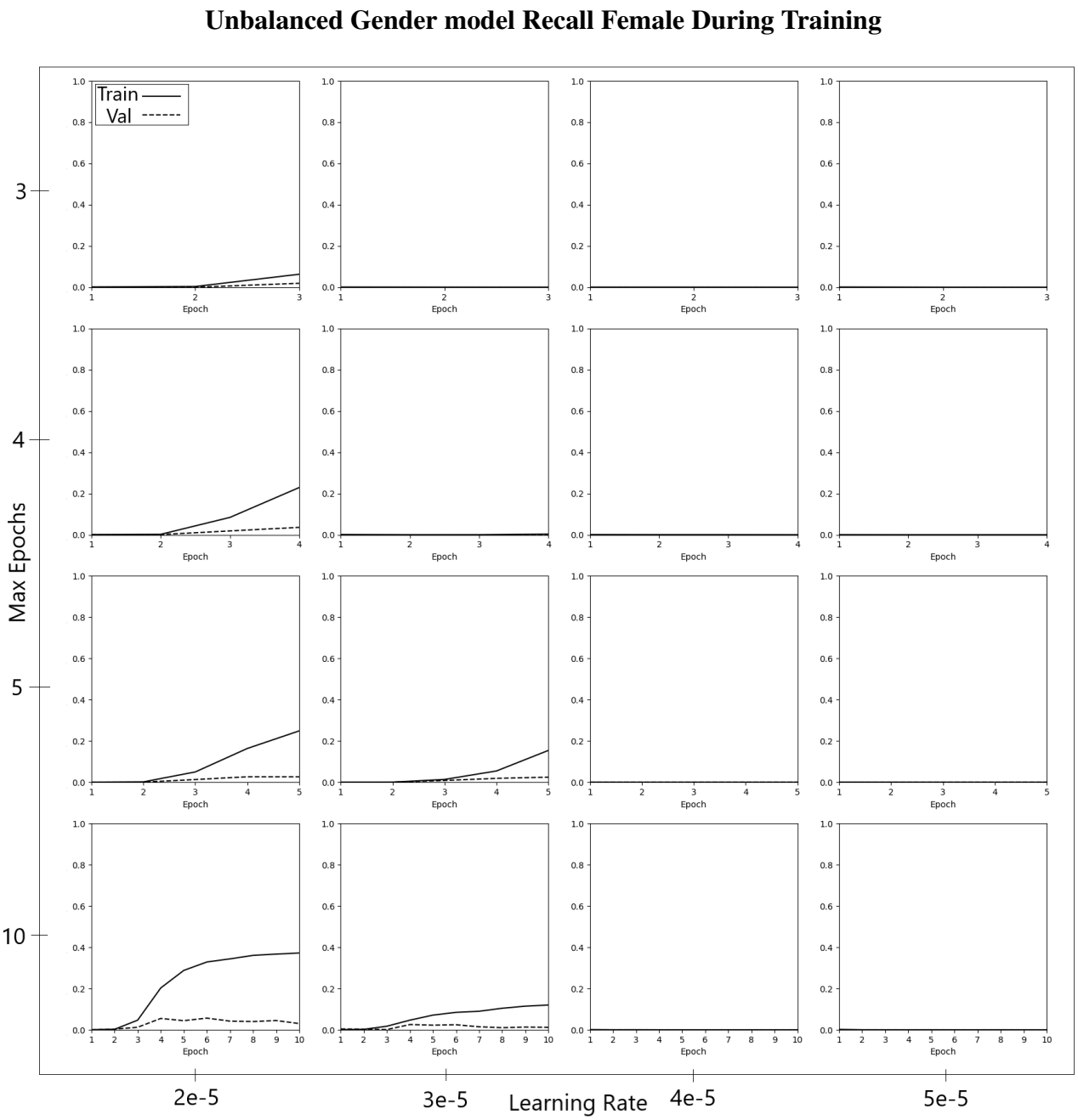


Figure 14: The average training and validation Recall for female high calcium scores per learning rate and max epochs combination. Average taken over 10 training runs.

Unbalanced Gender model Recall Male During Training

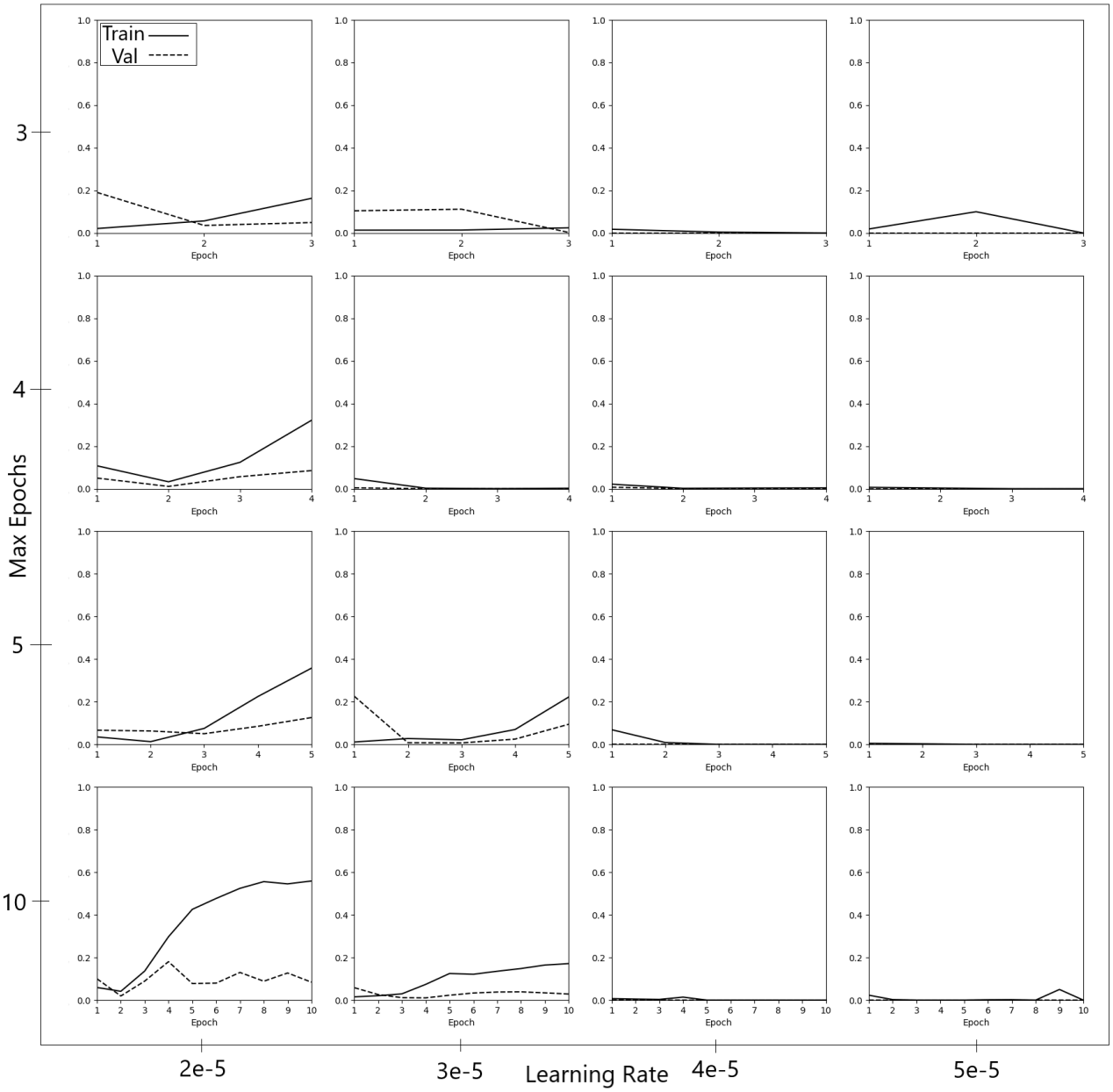


Figure 15: The average training and validation for male high calcium scores per learning rate and max epochs combination. Average taken over 10 training runs.

Training curves: balanced gender

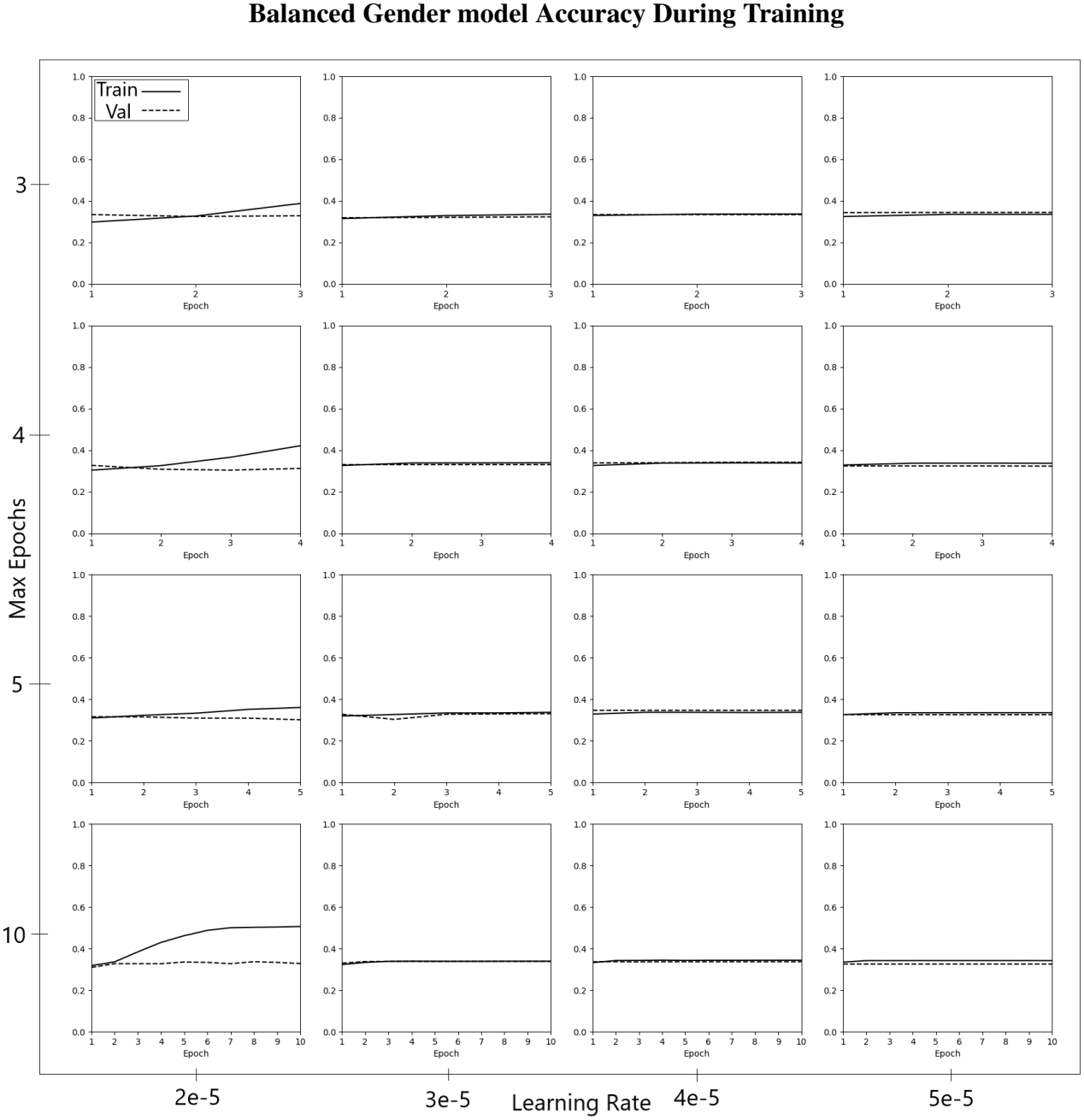


Figure 16: The average training and validation accuracy per learning rate and max epochs combination. Average taken over 10 training runs.

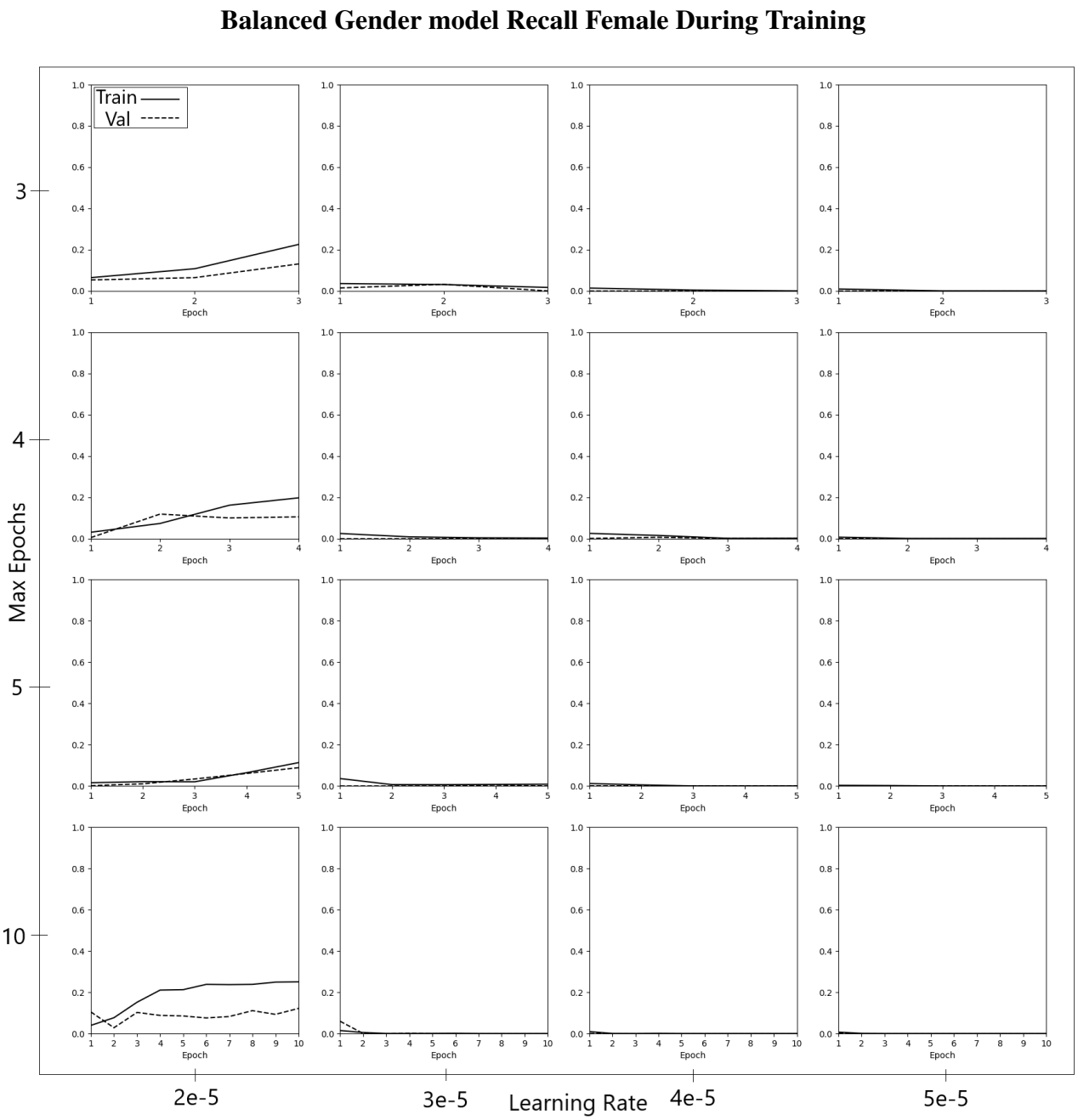


Figure 17: The average training and validation Recall for female high calcium scores per learning rate and max epochs combination. Average taken over 10 training runs.

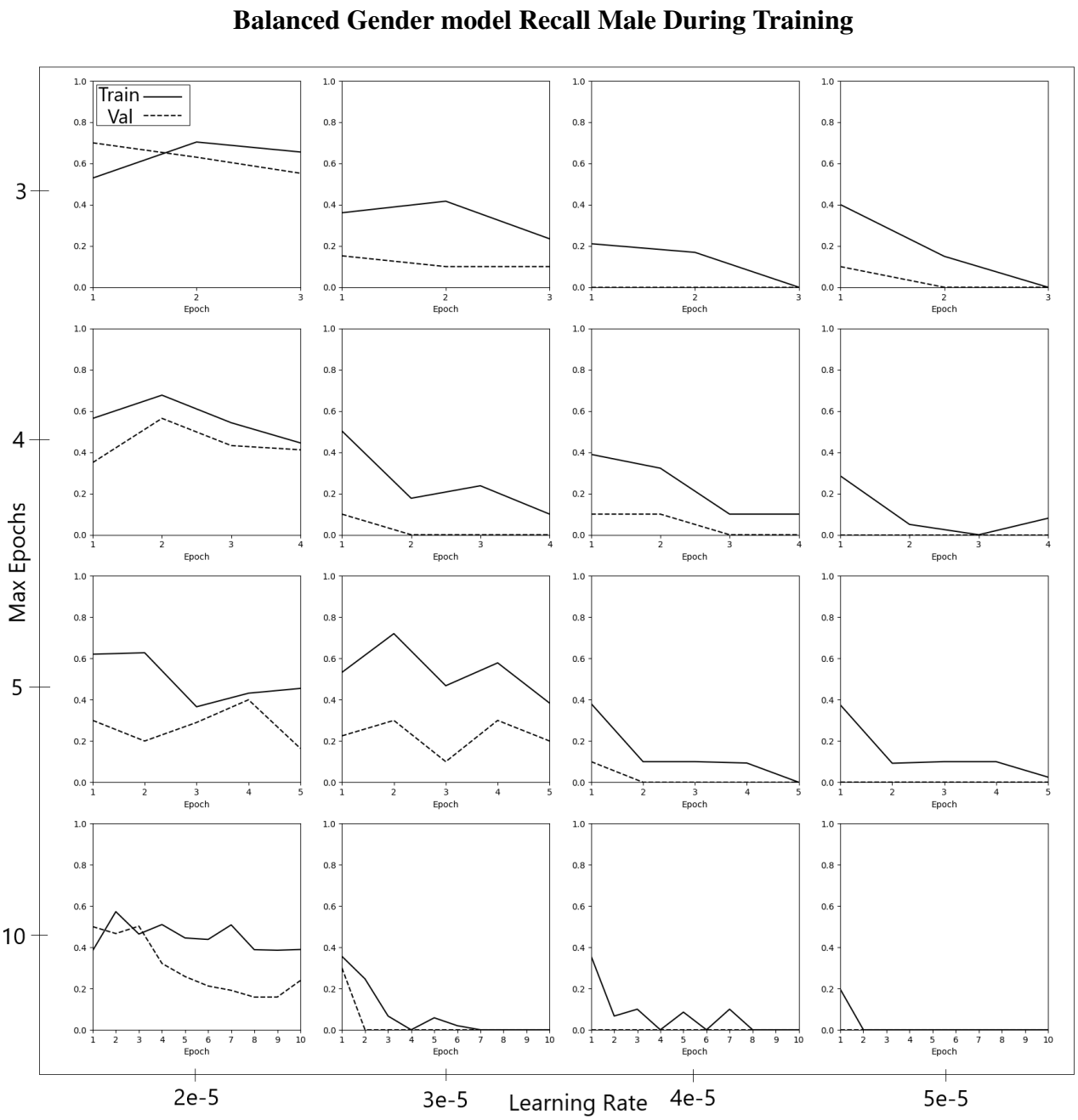


Figure 18: The average training and validation Recall for male high calcium scores per learning rate and max epochs combination. Average taken over 10 training runs.

Appendix D: Results Patient Based Model

Training curves

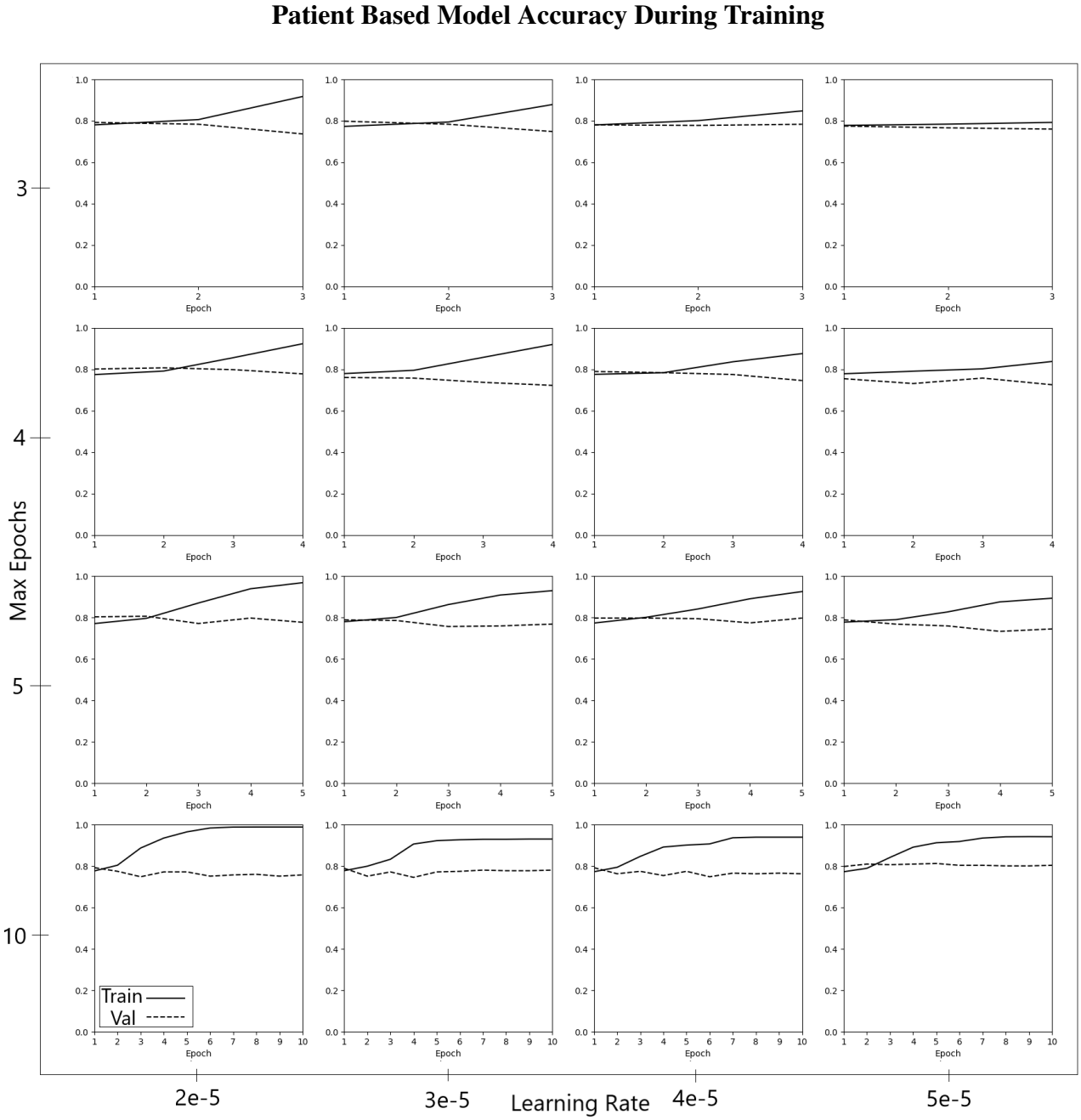


Figure 19: The average training and validation Accuracy per learning rate and max epochs combination. Average taken over 10 training runs.

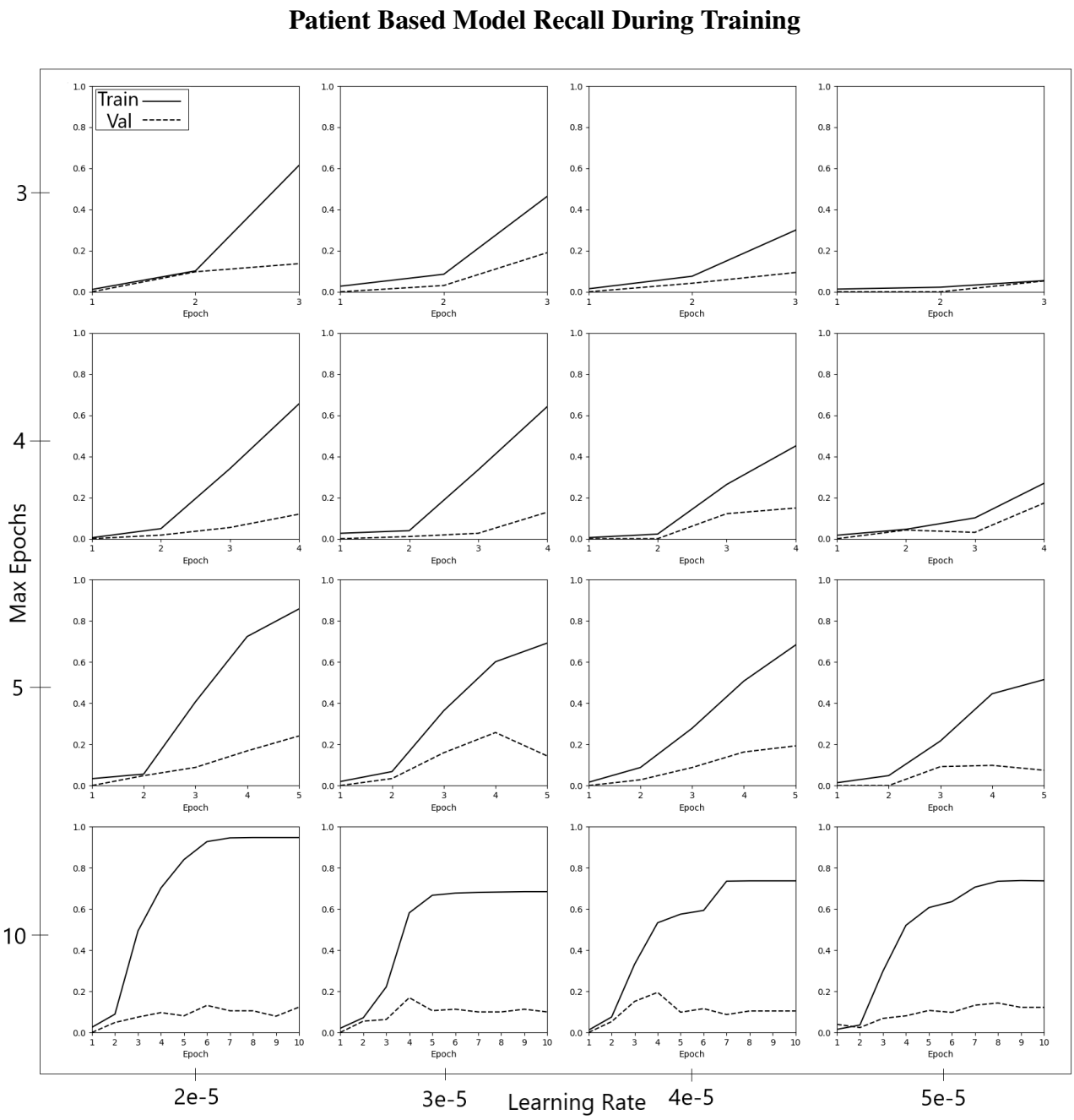


Figure 20: The average training and validation Recall per learning rate and max epochs combination. Average taken over 10 training runs.

Appendix E: Results Question Based Model

Training curves

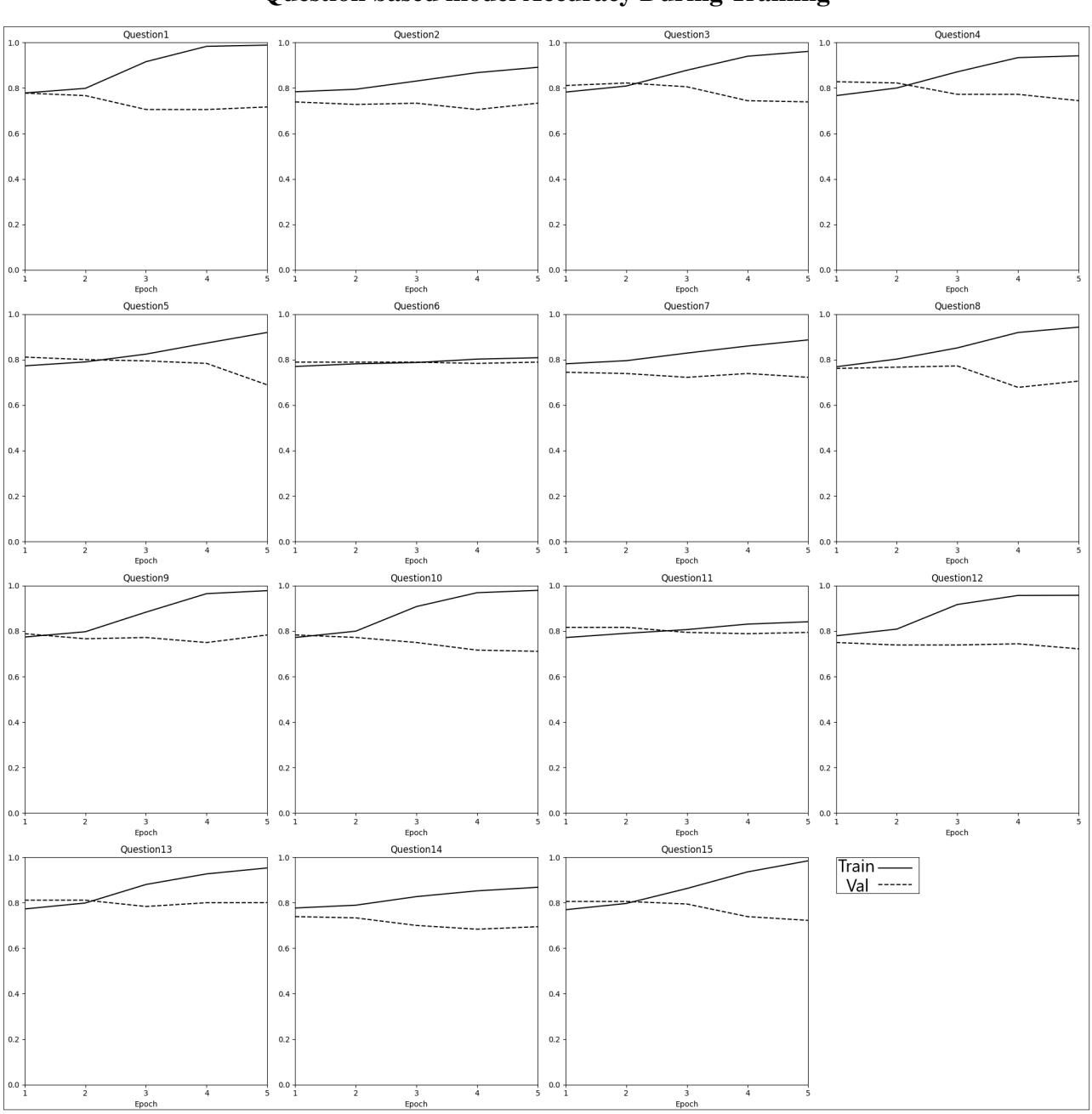


Figure 21: The average training and validation Recall for the question based model for each question. Average taken over 10 training runs.

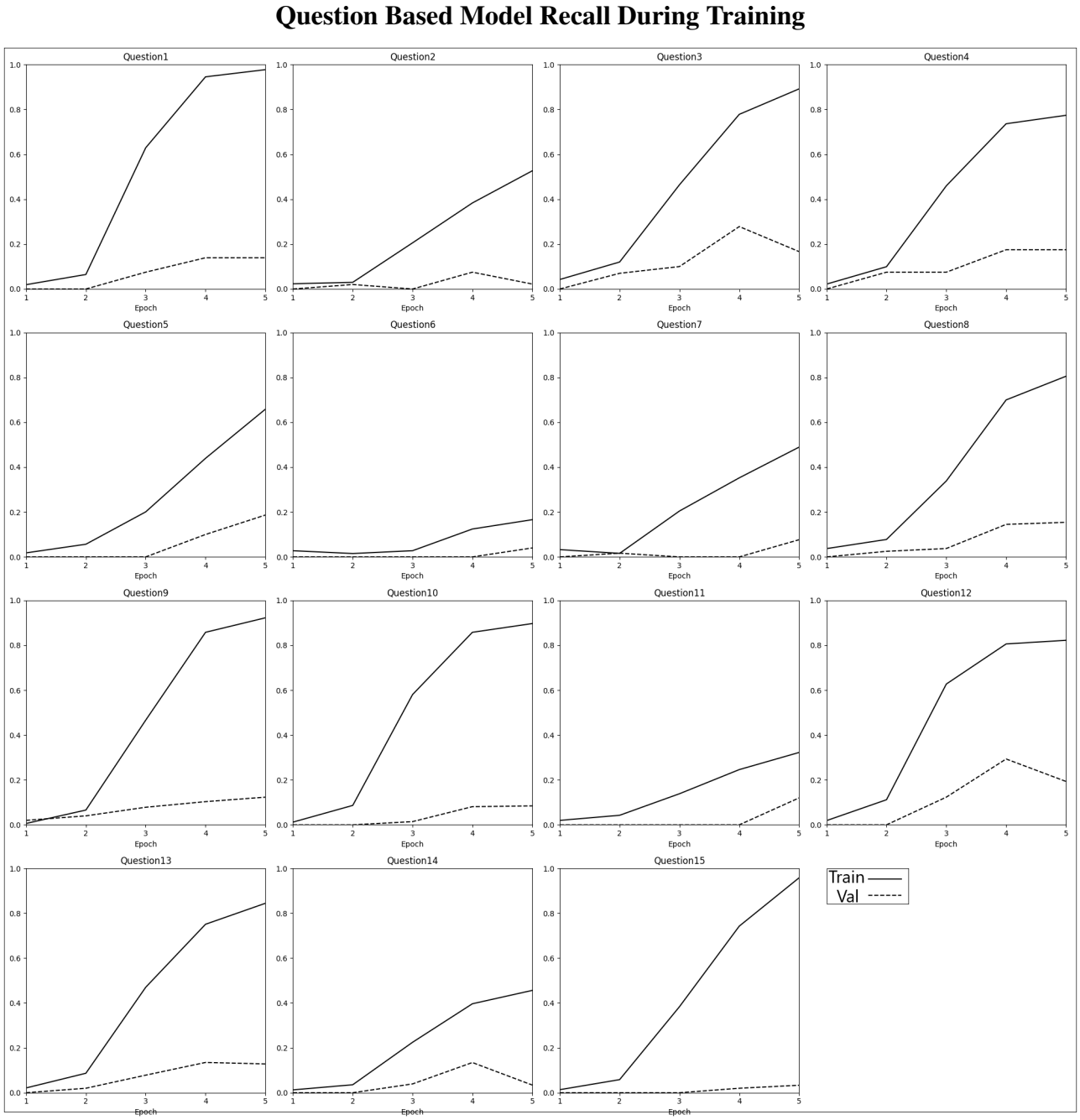


Figure 22: The average training and validation Recall for the question based model for each question. Average taken over 10 training runs.