Using Anchors from Free Text in Electronic Health Records to Diagnose Postoperative Delirium

Karl Øyvind Mikalsen[a,b,*], Cristina Soguero-Ruiz[b,c], Kasper Jensen[d], Kristian Hindberg[a], Mads Gran[e], Arthur Revhaug[e,f,g], Rolv-Ole Lindsetmo[e,g], Stein Olav Skrøvseth[a,d], Fred Godtliebsen[a], Robert Jenssen[h,d,b]

[a]*Dept. of Mathematics and Statistics, UiT The Arctic University of Norway, Tromsø, Norway*
[b]*UiT Machine Learning Group*
[c]*Dept. of Signal Theory and Comm., Telematics and Computing, Universidad Rey Juan Carlos, Fuenlabrada, Spain*
[d]*Norwegian Centre for E-health Research, University Hospital of North Norway (UNN), Tromsø, Norway*
[e]*Dept. of Gastrointestinal Surgery, UNN, Tromsø, Norway*
[f]*Clinic for Surgery, Cancer and Women's Health, UNN, Tromsø, Norway*
[g]*Institute of Clinical Medicine, UiT, Tromsø, Norway*
[h]*Dept. of Physics and Technology, UiT, Tromsø, Norway*

Abstract

Objectives. Postoperative delirium is a common complication after major surgery among the elderly. Despite its potentially serious consequences, the complication often goes undetected and undiagnosed. In order to provide diagnosis support one could potentially exploit the information hidden in free text documents from electronic health records using data-driven clinical decision support tools. However, these tools depend on labeled training data and can be both time consuming and expensive to create.

Methods. The recent "Learning with Anchors" framework resolves this problem by transforming key observations (anchors) into labels. This is a promising framework, but it is heavily reliant on clinician's knowledge for specifying good anchor choices in order to perform well. In this paper we propose a novel method for specifying anchors from free text documents, following an exploratory data analysis approach based on clustering and data visualization techniques. We investigate the use of the new framework as a way to detect postoperative delirium.

Results. By applying the proposed method to medical data gathered from a Norwegian University Hospital, we increase the area under the precision-recall curve from 0.51 to 0.96 compared to baselines.

Conclusions. The proposed approach can be used as a framework for clinical decision support for postoperative delirium.

Keywords: Electronic Health Records, semi-supervised learning, "Learning with Anchors" framework, postoperative delirium, data-driven clinical decision support, clustering

*Corresponding author at: Department of Mathematics and Statistics, Faculty of Science and Technology, UiT – The Arctic University of Norway, N-9037 Tromsø, Norway

Email address: karl.o.mikalsen@uit.no (Karl Øyvind Mikalsen)

1. Introduction

Complications after major surgery are unfortunately not uncommon. Central nervous system dysfunction, including postoperative delirium (PD), is often seen in geriatric patients undergoing major surgery [1]. Despite its potentially serious consequences, such as an increase in length of hospitalization, morbidity, mortality, and adverse events, it is often hard to detect PD [2]. Moreover, if the complication goes undiagnosed, it could have economical consequences for the care giver, as hospitals' reimbursement rates are dependent on correct coding.

For these reasons several works have investigated risk factors and prediction of PD. Bohner et al. predicted the risk for PD among patients undergoing aortic, carotid, and peripheral vascular surgery using multivariate linear logistic regression [3]. In [4, 5], the authors predicted risk for PD after major abdominal surgery and found well-known predictors such as advanced age or ASA-score. Common for the previous studies is that only a few structured variables have been used as features for the prediction model. However, we believe that also the free text documents in the patients' electronic health records (EHRs) contain valuable information about PD that can be used for diagnosis support. In particular, nurses collect useful information about the patient health status since they observe the patients after the surgery and report about them three times every day.

Recent advances in machine learning for healthcare have shown great potential for exploiting the "hidden" information in the EHRs to provide data-driven clinical decision support, especially if large amounts of labeled data are available [6, 7, 8, 9]. In the aforementioned studies, the patients were manually labeled with and without PD. However, the labeling task could be a time consuming and expensive process [10]. To overcome this drawback Halpern et al. proposed a very promising framework, with a large number of possible applications. In this framework, which we refer to as the anchor method (AM), one can learn phenotypes and predict clinical state variables from EHR unlabeled data only by specifying a few key observations called anchors [11, 12]. An underlying assumption is that the presence of an anchor variable implies the presence of the latent label of interest. Thus, training examples for which the anchor variable is present are positive examples, while nothing can be said for the remaining examples.

If the data mainly consist of free text, a limitation with AM is that trustworthy anchors could be difficult to identify, even for clinicians. Moreover, in settings where the sample size is larger than the dimensionality ($N > d$) , the originally proposed (ridge) l_2-regularized logistic regression classifier within AM works well. It keeps all variables in the model and the coefficients of correlated variables are shrunken toward each other. However, when $d >> N$ ridge regularization is not a good choice [13].

In this paper we investigate the use of AM as a way to develop models to detect PD, and thereby being able to diagnose and code it properly. To resolve the problem of specifying reliable anchors we develop a problem specific method based on domain knowledge and exploratory data analysis using clustering and visualization techniques. Furthermore, we propose to use a different classifier in the AM framework, namely the elastic net, which forces sparsity and has been shown to provide robustness in settings where the dimensionality is higher than the sample size [14, 13]. We show that, by introducing this new methodology, AM can be successfully applied to problems where no obvious anchors exist. In particular, by applying it to clinical data gathered from a Norwegian university hospital, we show that it can be used to extract hidden information from unstructured free text and thereby provide diagnosis support for PD.

The rest of this paper is organized as follows. Section 2 describes methods, including the AM framework and our proposed anchor specification method. In section 3 data and feature

representation are described. Experiments and results are presented in Section 4. In Section 5, we discuss the results and further work. Conclusions are drawn in Section 6.

2. Methods

2.1. Background on the "Learning with Anchors" framework

AM is particularly well suited for text documents where the features can be represented using e.g. bag-of-words or medical ontologies. In the method there are two different kinds of binary variables; *observed* and *latent*. An observed variable is a variable that can be observed directly from the EHR. It could for example be the answer to a question such as *Does the word "confused" appear anywhere in some of the free text documents?* A latent variable cannot be extracted directly from the EHR and could be the answer to a higher level question such as *Does the patient have postoperative delirium?* Formulating such questions is difficult since there are so many different ways to answer them, and it could also be that answers are not documented in the EHR.

An *anchor* variable is an observed variable that can be extracted directly from the EHR and contains valuable information about the *latent* variable one wants to uncover. The anchor should satisfy two properties, 1) given that the anchor is observed, then also its latent variable is on, and 2) it is independent of all other observations, conditioned on the latent variable. The latter property states that once the value of the latent variable is known, no other observed variables provide additional information about the anchor.

Given these definitions, a description of the steps in the original AM is as follows: (1) Select data source; (2) Represent features using e.g. bag-of-words; (3) Specify anchor (for this step our proposed method can be used); (4) Extract the vector that represents the anchor from the feature matrix and use it as a label vector; (5) Train a classifier to predict whether the anchor is on or not (Elastic net can be used); (6) The trained model can be calibrated using a validation set [15]; and (7) For an unseen patient where the anchor is not observed, the model is used to predict the likelihood of the anchor being on. This scheme is illustrated in the upper part of Figure 1.

In more detail the framework is as follows. Assume that there are N patients and p observed variables. Let Y be the latent variable we want to predict for each patient. Let $\mathbf{x}^-$ represent all observed variables except for the anchor A. Assuming that we have found an anchor, A, the last three steps are as follows:

(5) Learn $P(A = 1 \,|\, \mathbf{x}^-)$ using a classifier that provides a probabilistic output.

(6) Using a validation set, K, compute

$$C = \frac{1}{|K|} \sum_{k \in K} P(A = 1 \,|\, \mathbf{x}_k^-), \tag{1}$$

where $\mathbf{x}_k^-$ is the data for patient k with the anchor removed.

(7) For an unseen patient, t, with $A = 0$, predict

$$P(Y_t = 1) = \frac{P(A = 1 \,|\, \mathbf{x}_t^-)}{C}. \tag{2}$$

If $A = 1$, $P(Y_t = 1) = 1$ because of the first property of anchors.

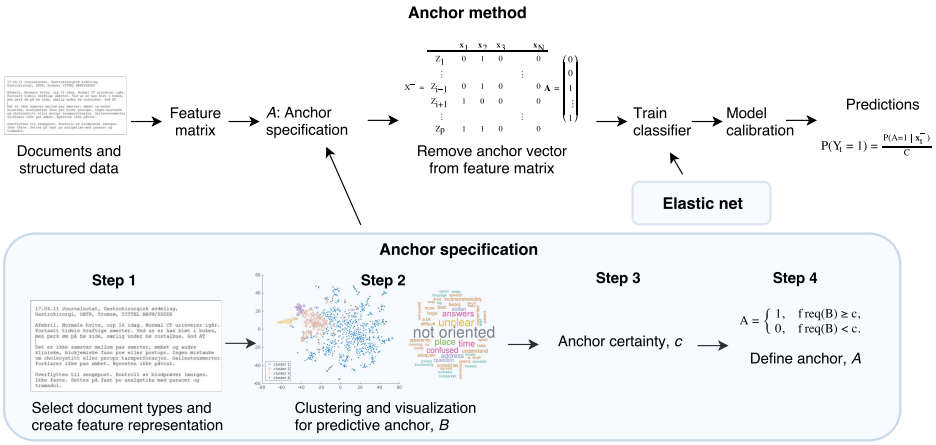


Figure 1: Schematic diagram of the method. The upper part of the figure explains how the "Learning with Anchors" framework works and the lower part illustrates the proposed anchor specification method.

2.2. *Proposed anchor framework solution*

Figure 1 illustrates how the *"Learning with Anchors" framework* and the proposed *anchor specification method* work. In the following we explain how to specify anchors using an exploratory data analysis and review the proposed classifier.

2.2.1. *Predictive anchors via exploratory analysis*

The two properties that anchors are supposed to satisfy are very strict and therefore it often turns out that it is difficult to find such anchors. However, in practice, the conditional independence property does not have to be completely satisfied [12]. On the other hand, if property 1 is relaxed, the false positive rate will automatically increase. With our proposed method, it is possible to define an anchor from free text by first searching for a *predictive* anchor – an observed variable that originally does not satisfy property 1, but by adding a certainty measure we can define a true anchor from it. This makes the AM framework applicable for a larger variety of problems.

The proposed method consists of four steps, which are explained below and is as follows.

In *step 1* one has to *identify a subset of relevant document types*, which requires domain knowledge, and create a feature representation.

In *step* 2, we *define a predictive anchor*, *B*, as a feature that is a surrogate for the latent variable of interest, and whose semantic meaning could vary in different settings in general, but restricted to the subset of relevant document types, it has a clearer meaning. We propose to use clustering to suggest predictive anchor candidates, *B*. For this reason it is important that the clustering method is robust and not sensitive to parameter choices. We therefore use the *kNN mode seeking consensus clustering* algorithm [16] (Appendix A), which has been shown to be robust on a variety of datasets. The idea with the clustering is to identify groups of patients of different health status. The visualization method t-SNE (Appendix B) is used, in combination with clinical knowledge, to further analyze the clustering results and thereby, identify groups containing patients with normal outcomes and groups of patients in worse condition. An example

of a helpful tool for this task is to plot wordclouds of the most informative words for each cluster and then let the domain experts identify predictive anchor candidates from the wordclouds.

In *step 3*, we define the *certainty*, c, of the predictive anchor, B, as the lowest frequency that makes the predictive anchor trustworthy. Frequency in this setting means the frequency across the set of documents associated with a specific patient. We note that applying a global threshold of 1 basically corresponds to saying that the predictive anchor is an anchor. If one wants to make more conservative anchors, one can use a higher global threshold to reduce the probability of obtaining false positives. However, this definition also enables the opportunity to use a locally varying certainty. For example one could apply the proposed clustering and visualization techniques to stratify the data into groups with varying certainty.

Step 4 consists of using the term frequency restricted to the subset of relevant document types of the predictive anchor candidate B and the certainty measure c to *define the anchor* A as

$$A = \begin{cases} 1, & freq(B) \geq c, \\ 0, & freq(B) < c. \end{cases} \tag{3}$$

The idea behind the procedure is that, in general, some words are not anchors when they are written in a random document, but in certain documents it could be that the words are used in special settings and therefore are more trustworthy. It is also possible that some words, that in themselves cannot be trusted as anchors, could become more certain when they appear more than once.

2.2.2. Elastic net

In AM, a classifier that provides a probabilistic output is required. Halpern et al. applied l_2-regularized logistic regression. We propose to use the *elastic net* instead since it is robust in settings where the dimension is higher than the sample size [14]. A review is given here.

For a data point, $\mathbf{x}$, with an unknown label $y \in \{0, 1\}$, the logarithm of the ratio of the posterior probabilities $P(y = 0 \,|\, \mathbf{x})$ and $P(y = 1 \,|\, \mathbf{x})$ is modeled via a linear function, $w_0 + \mathbf{w}^T\mathbf{x}$. Given a training set, $\{(\mathbf{x}_k, y_k)\}$, the parameters $w = (w_0, \mathbf{w})$ are found by maximizing a regularized log-likelihood,

$$l(w) = \sum_{k=1}^{N_0} \log P(y_k = 0 \,|\, \mathbf{x}_k^{(0)}, w) + \sum_{k=1}^{N_1} \log P(y_k = 1 \,|\, \mathbf{x}_k^{(1)}, w) - \lambda \left((1 - \alpha)\|\mathbf{w}\|_2^2 + \alpha\|\mathbf{w}\|_1 \right), \tag{4}$$

where $\lambda > 0$, $\alpha \in [0, 1]$, $\|\cdot\|_p$ is the l_p-norm and N_j is the number of data points in the jth class.

A "side-effect" of the elastic net is that it provides a ranked list of the most important features. The list can be used together with clinical knowledge to suggest new predictive anchors. One can then create a composite anchor out of the union of the individual anchors. Using multiple anchors could often be beneficial because it gives more positive examples for training.

3. Experiments and results

3.1. Data description

We wanted to use AM to detect whether a patient had developed PD or not. Hence, the latent variable of interest was $Y = hasPD$. For this particular task we explored a data set extracted from the Department of Gastrointestinal Surgery (DGS) at the University Hospital of North Norway

(UNN) from 2004 to 2012. In particular, we extracted EHRs for 7741 patients. The data include structured data such as ICD-10 codes describing the main diagnosis, age, sex, length of surgery, blood tests and health status, as well as free text from documents such as doctor's notes, radiology reports and semi-structured nurses notes. The nurses notes are semi-structured since they are formulated as questionnaires with 12 bullet points and the nurses answer the questions using free text. For each patient the nurses write at least three notes every day; morning, afternoon and evening.

A clinician (author M.G.) made a list of surgeries of interest, basically consisting of major abdominal surgeries requiring general anesthesia. Based on this, 1138 patients who potentially could suffer from PD were selected into a cohort. In AM no labels are needed, but to test the learning system, the clinician manually read the EHR for a subset consisting of 308 patients and found that 24 of them had PD after the surgery. Hence, the training set consisted of the remaining 830 unlabeled patients.

The remainder of this section is divided into two main subsections. In Subsection 3.2 we apply the proposed methodology to specify the first anchor. For the clarity of this exposition we leave some of the details for Appendix C, for example the specification of the other anchors. In Subsection 3.3 we apply AM and demonstrate the results of the methodology we have proposed.

3.2. Anchor specification using proposed method

As a first step, our clinicians suggested some words that potentially can be used as anchors; *delirium, delir, postoperative*. However, these words rarely or never occur in the EHR and cannot be used as anchors. We therefore employed our proposed method to specify anchors.

Step 1. Identification of relevant types of text documents. It was hypothesized by our clinicians that since the nurses take care of the patients continuously after the surgery, most likely information about PD would be discovered and reported by them. In particular, the bullet points in the semi-structured nurses notes related to communication/senses and knowledge/ development/ psychological are important descriptors of the mental status for the patient. Following this clinical knowledge, we chose to search for anchors in the free text only from the first two bullet points in the nurses notes.

A *term frequency - inverse patient frequency* (tf-ipf) representation was used instead of the more common *inverse document frequency* (idf) since we did not have access to each document for each patient [17]. However, the effect of the tf-ipf is the same, the value of the tf-ipf is proportional to the number of times a word appears for each patient, and is reduced by the frequency of the word for all patients. To further compensate for the redundancy in the features because of a lack of preprocessing (correlation between misspelled and correctly spelled words, etc.) principal component analysis (PCA) [18] was used to reduce the dimensionality. Based on a plot of the eigenvalues, we decided to use the 20 dimensions corresponding to the 20 top ranked eigenvalues. We notice that it is possible to compute both the tf-ipf and PCA feature representation also for new unseen patients.

Step 2. Identification of a predictive anchor. The kNN mode seeking consensus clustering algorithm was run for the 830 patients in the training set. Based on the dendrogram [19], the number of clusters was automatically chosen to 4. A low dimensional embedding of the data was created using t-SNE and the resulting mapping is shown in Figure 2. The different colors and markers represent the different clusters. This figure verifies that the clustering results are reasonable; nearby points in the two dimensional space are clustered together. Table 1 provides a summary

of the clustering results and more details are provided in Appendix C. Cluster 4 seems to contain patients with normal, positive outcomes. In cluster 1 words like *confused, unclear, disoriented* dominate, whereas in cluster 3 the theme is sedation and sedation drugs. In cluster 2, many of the most frequent words are related to speech and communication.

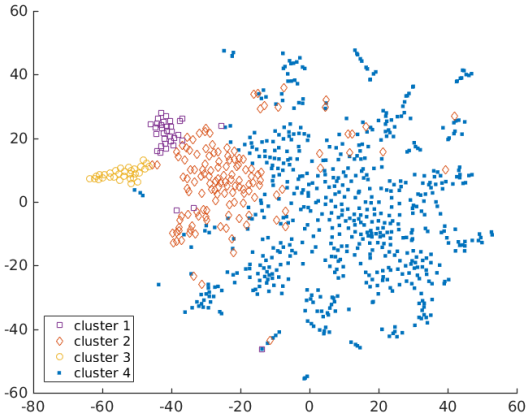


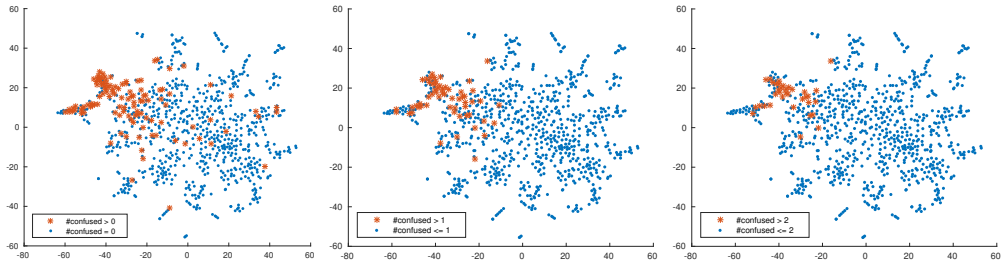
Figure 2: Locations of the four clusters in the t-SNE map, obtained using the kNN mode seeking consensus clustering algorithm.

Cluster	# of patients	Marker/color in Figure 2	Keywords
1	34	Purple squares	*Disoriented* and *confused*
2	134	Red diamonds	*Adequate* and *communicates*
3	31	Yellow circles	*Sedated*
4	631	Blue dots	*Good mood* and *nothing to report*

Table 1: Summary of clustering results. The table shows the number of patients belonging to each cluster, the marker and color representing the cluster in the t-SNE map and certain keywords describing the different clusters.

The fact that most of the high-frequent words in cluster 1 are words describing a patient's mental status, e.g. *disoriented, unclear, confused, messes* (see Figure C.4 in Appendix C), indicates that it is natural to search for anchor candidates in this cluster. Clinicians suggested to use *confused* as the most evident word. Hence, we considered it as our first predictive anchor.

Step 3. Certainty assessment. Figures 3a-3c show the location in the two dimensional t-SNE map of the patients with different frequencies of the word *confused* in their nurses notes. We see that *confused* also appears for some patients in the cluster containing "normal" patients (cluster 4), but for many of these patients only once. Figure 3c shows that patients that have a frequency of at least three for *confused* are concentrated around cluster 1 and 3. Higher frequency probably means that several nurses made the same observation more times. Hence, it is reasonable to assume that higher frequency means higher certainty. An underlying cluster assumption is that patients that belong to the same cluster are similar, and therefore one could argue that if *confused* appears for a patient that belong to cluster 1 or 3 only once, then it is probably not noise since the patient is supposed to be similar to patients for whom the word appear with a higher frequency.



(a) (b) (c)

Figure 3: The red stars show the location in the t-SNE map of the patients for whom the word *confused* appear in their nurses notes. In (a) it appears at least once, in (b) at least twice and in (c) at least three times.

Cluster 2 and 4 are larger and have higher variance. Some observations of *confused* in these clusters could be treated as noise and we therefore following clinicians' input defined the certainty measure, c_1, as

$$c_1 = \begin{cases} 2, & \text{if the patient belongs to cluster 1 or 3} \\ 3, & \text{if the patient belongs to cluster 2 or 4.} \end{cases}$$

Step 4. Definition of anchor. The anchor $A_1 = confused*$ was then defined according to Eq. (3). The * means that the certainty measure is considered to define the anchor. Note that we probably could have chosen $c_1 = 1$ for patients in cluster 1 and 3 as well, but we rather want false negatives than false positives, and therefore make a more conservative choice for the certainty.

3.3. Classification based on specified anchors

3.3.1. Feature representation for classifier

A bag-of-words (BoW) model was used to represent the presence or absence of each different word that appeared in the clinical narrative [20]. Stop words and words that appeared for fewer than five patients were removed. The structured data, gender, type of surgery and some ICD-10 codes, were represented as booleans. Age was discretized into two intervals; older or younger than 65 years, since the literature emphasizes that especially patients older than 65 years have higher risk of getting PD [21, 2, 1]. For the American Society of Anesthesiologists (ASA) physical state grade, Scholz et al. [22] showed that for a score of at least three is a risk factor for PD. We therefore made a boolean by putting a threshold at three. In total this resulted in 20949 different features.

3.3.2. Evaluation of proposed method

The R-package *glmnet* [23] was used to run the elastic net logistic regression. The regularization parameter λ was chosen using 10 fold cross-validation. We could also choose the other regularization parameter α using cross-validation. However, to ensure that we did not see the effect of different types of regularizations when comparing to baselines we chose $\alpha = 0.5$. To incorporate that our prior belief is that each variable is equally important, we ensured that the penalty applied equally to all variables by standardizing the binary variables to zero mean and standard deviation one [13]. We chose to measure performance using the area under the

	Anchor	*confused*	*confused x3*	*confused+*	A_1
AM - Elastic	AUC-PR	0.507	0.707	0.503	**0.803**
	95% CI	(0.351, 0.652)	(0.541, 0.856)	(0.360, 0.637)	(0.633, 0.918)

Table 2: Area under the PR-curve (AUC-PR) for three baselines and the proposed method. The two first anchors are chosen from all documents, whereas the two last one are chosen from the nurses notes (D). 95 % confidence intervals are shown in parenthesis.

precision-recall curve (AUC-PR) because it captures the performance over the entire operating range and has been shown to work well on imbalanced data [24]. For this measure only the ordering of the scores is needed and therefore it was not necessary to tune the calibration coefficient. 95 % confidence intervals (CIs) were evaluated using 100 bootstrap samples from the test set [25].

3.3.3. Demonstrating the effect of exploratory anchor selection

Section 3.2 introduced a text-based method for exploring anchors from EHRs using clinical knowledge, basically creating labels for a classifier (see Figure 1). Here we demonstrate the effect of this exploratory anchor selection procedure by comparing to baselines where we applied AM with anchors not specified using the proposed method. To isolate the effect of the proposed anchor specification method we used the elastic net with $\alpha = 0.5$ also for the baseline. The effect of the classifier choice will be demonstrated in a later subsection.

The first baseline we compared to was AM with the anchor *confused*, where *confused* was specified by naively letting all patients where the word confused appeared in some of their documents have an anchor. We also applied AM to the anchor *confused x3*, which was defined such that it is on if confused appeared at least three times in any of the documents. To demonstrate that it is not only a matter of choosing the correct document types, we compared to yet another baseline; we applied AM to the anchor *confused+*, which is on only if confused is observed in the free text only from the first two bullet points in the nurses notes.

Table 2 shows AUC-PR values and 95%-CIs obtained using the baselines and AM with the anchor A_1. We see that with the anchor A_1 an AUC-PR value of 0.803 was obtained, which is a considerable increase compared to the baselines.

By comparing to different baselines, we have now isolated the effects of 1) specifying the anchor only from the free text only from the first two bullet points in the nurses notes, and 2) specifying the anchor using our proposed methodology. We have shown that both steps are necessary to obtain a reasonably good performance.

3.3.4. Demonstrating the effect of document selection in feature representation for classifier

Clinical knowledge was used to suggest that anchor selection should come from the first two bullet points in the nurses notes. However, it was also hypothesized that the nurses notes likely is the most important data source for identifying information about PD. Surgical operation notes, doctor's notes, radiology reports, etc., will probably introduce more noise than relevant information. We therefore used clinical knowledge to reduce the number of data sources for the classifier to only structured data and free text from the nurses notes, which reduced the number of features to 2008. With this approach the AUC-PR value increased from 0.803 to 0.838, 95% CI $(0.694, 0.930)$, with the anchor A_1. The CI is wide, but at least we see that the AUC-PR did not decrease.

3.3.5. Demonstrating the effect of adding more anchors and classifier choice

The elastic net outputs a ranked list of the most important features, which potentially could contain suggestions of new predictive anchors. Table 3 shows the ranked features provided by AM when A_1 was used as anchor (second column). Based on the ranking and clinical knowledge, we added the word *disoriented* as a predictive anchor.

Rank	A_1	A_2	A_3	A_4
1	disoriented	unclear	haloperidol	perceive
2	unclear	eye contact	messes	messes
3	clear	responds	responds	responds
4	case	picking	perceive	picking
5	bed	hands	indistinct	indistinct
6	messy	indistinct	agitated	understand
7	visions	sleep	remembers	agitated
8	eyes	messy	understand	opens
9	called	messes	hallucinated	hallucinated
10	fall	bring	messy	messy
⋮				
24	haloperidol	ASA score[1]	forgets	incomprehensible

Table 3: Lists of the top ranked features obtained using elastic net logistic regression with the anchors A_1 =*confused**, $A_2 = \{A_1$, *disoriented**$\}$, $A_3 = \{A_1, A_2$, *unclear** $\}$ and $A_4 = \{A_1, A_2, A_3$, *haloperidol** $\}$, respectively, as labels.

Using the same certainty measure as for *confused* we defined the anchor *disoriented** according to Eq. (3) and created a composite anchor, A_2, as the union of *confused** and *disoriented**. Table 4 shows that AM with the anchor A_2 gave an AUC-PR value of 0.925, which is a considerable improvement.

Based on the ranking in the third column in Table 3 and clinical knowledge we added the word *unclear* as a predictive anchor. We defined the composite anchor, A_3, as the union of *confused**, *disoriented** and *unclear**. Table 4 shows that using A_3 we obtained an AUC-PR value of 0.964, which is a large improvement. Similarly, we created the anchor A_4 using the predictive anchor *haloperidol*. However, the AUC-PR value of 0.962 is very similar to the result obtained using the anchor A_3.

We see that the list of the top ranked features obtained using four anchors contains words like *messes, picking, indistinct, understand, agitated, hallucinated, visions* and *incomprehensible*. These words are definitely related to the mental status and potentially we could continue to add more anchors. However, we decided to not add more anchors because these candidates were not predictive enough and/or ambiguous.

As we mentioned above, since the sample size is lower than the dimensionality, we chose to use the elastic net. We compared to l_2 regularization by computing AUC-PR values and 95% CIs using the anchors A_1, A_2, A_3 and A_4. Table 4 shows that the elastic net is clearly beneficial. For example, for the anchor A_1 using l_2 regularization an AUC-PR value of 0.692 was obtained, whereas using the elastic net we got 0.838.

We also compared to a supervised baseline where we trained a classifier (elastic net) on the test set using 5-fold cross-validation. Mean AUC-PR and standard errors were calculated using

[1]The only structured variable that appeared among the 25 top ranked variables.

	Elastic net		l_2-regularization	
	AUC-PR	95% CI	AUC-PR	95% CI
A_1	0.838	(0.694, 0.930)	0.692	(0.555, 0.844)
A_2	0.925	(0.851, 0.975)	0.815	(0.658, 0.916)
A_3	0.964	(0.911, 0.993)	0.910	(0.817, 0.975)
A_4	0.962	(0.923, 0.996)	0.915	(0.827, 0.998)
Supervised baseline	0.770	(0.652, 0.888)	0.580	(0.469, 0.691)

Table 4: AUC-PR values obtained by adding more anchors. In the columns to the right we have also shown AUC-PR values obtained using l_2 regularized logistic regression as the classifier in AM.

bootstrap (creating 100 different 5-folds). Table 4 shows that with this a approach an AUC-PR value of 0.770 was obtained, considerably lower than for AM with two or more anchors.

4. Discussion

The proposed method is not fully automatic, it still requires some manual work. Therefore a natural question to ask is whether one actually gains something in terms of reduced labor intensity compared to manual label annotation. However, then one should keep in mind that while the latter must be done individually (e.g. by retrospectively reading the EHR for each patient one wants to label), in the former the manual work is done once and for all. Hence, the time spent on anchor annotation is actually not comparable to manual label annotation, and the difference becomes larger the larger the dataset is. We also want to emphasize that the proposed method is not fully generalizable to all diagnostic challenges. That being stated, it is easy to find other clinically interesting problems, both in retro- and prospective settings, where the method is applicable. One example is to use this method to pre-operatively identify malnourished patients [26]. In this case the notes regarding nutritional status would be particularly relevant. We also believe that the method is transferable to *predicting* patients at risk for post-operative complications. Potentially the method can be used in more general text-based settings, not necessarily in a clinical application.

4.1. Limitations and further work

AM falls into the classical PU-learning setting where one assumes that only the unlabeled dataset, U, is contaminated, whereas the positive set, P, is assumed to not contain false positives. In our approach we adapted the way of choosing the set, P, such that this assumption is not broken. However, recently, approaches where one assumes that also P can be contaminated, have been proposed [27, 28]. The main ingredient in these methods is to use resampling on P to provide robustness against false positives. In [29] Claesen et al. showed that this approach can be used to predict whether a patient will start glucose-lowering pharmacotherapy. It will be interesting to use the anchors as proposed by Halpern et al. such that P is contaminated and thereafter applying an approach similar to the robust ensemble SVM, proposed in [28], in further work.

There are of course many challenges related to the unstructured text we have available [30, 31]. Often the time spent on entering text into the EHRs is limited. A document could for example be a dictate of a conversation during a consultation. In other cases information could be recorded on an audio-recorder and then transcribed by a secretary at a later time. For these

reasons incomplete sentences and typos are more common in medical text than in usual published text. In addition, there are words that contain digits, medical short forms and acronyms. Another challenge, special to Norwegian medical text, is related to the fact that there are two official languages in Norway and that a relatively large fraction of the employees at UNN are from other countries in Scandinavia. Some of them write in their own language, others have learned some Norwegian and therefore text written by them could be a mixture of several languages. We could have done more natural language processing to compensate for these challenges, but would have required a lot of effort since all the text mining software that is developed for English language does not exist for Norwegian language. However, there is ongoing work in our group trying to introduce less noisy conceptual features based on medical ontologies [32]. Since the AM framework do not make any assumption on how the features are represented, these can be included in further work.

Another limitation of our work is the quality of the gold standards. The clinicians created the gold standard of PD based on actual information in the EHR. Diagnosing PD was in part based on a consciousness assessment tool, the Observational Scale of Level of Arousal (OSLA) [33, 34], as the EHR lacked sufficient data to use standardized delirium screening instruments. Hence, there is a risk that the gold standard could be biased.

Finally, we want to mention that in this work we have demonstrated the effects of the proposed methodology on a medium-sized dataset. The focus has been on diagnosing PD. However, in future work we would like to even more investigate the generalization abilities on bigger datasets and other problems. In particular, we will look at the problem of pre-operatively identifying and predicting malnourished patients at UNN.

5. Conclusion

We have adapted the "Learning with Anchors" framework to medical data gathered from a Norwegian University Hospital. We introduced a new method for specifying anchors, providing the opportunity to obtain a labeled training set without manual label annotation. The importance of the proposed method was demonstrated on task where the aim was to detect postoperative delirium. By creating the labels in naive way we got an area under the PR-curve (AUC-PR) of 0.51, whereas by introducing our suggested improvements and adaptations we got an AUC-PR value of 0.96. We believe that the method potentially can be used in other clinical problems as well as in a more general text-based settings, not necessarily related with healthcare.

Conflict of interest

The authors have no conflict of interest regarding the study.

Acknowledgements

This work was partially funded by the Norwegian Research Council FRIPRO grant no. 239844 on developing the *Next Generation Learning Machines* and IKTPLUSS grant no. 270738 *Deep Learning for Health*. Cristina Soguero-Ruiz is partially supported by projectS TEC2016-75361-R from Spanish Government and by Project DTS17/00158 from Institute of Health Carlos III (Spain).

6. References

[1] S. Deiner, J. H. Silverstein, Postoperative delirium and cognitive dysfunction, BJA: British Journal of Anaesthesia 103, suppl. 1 (2009) i41–i46. doi:10.1093/bja/aep291.

[2] S. K. Inouye, T. Robinson, C. Blaum, J. Busby-Whitehead, M. Boustani, A. Chalian, S. Deiner, D. Fick, L. Hutchison, J. Johanning, M. Katlic, J. Kempton, M. Kennedy, E. Kimchi, C. Ko, J. Leung, M. Mattison, S. Mohanty, A. Nana, D. Needham, K. Neufeld, H. Richter, Postoperative delirium in older adults: Best practice statement from the american geriatrics society, Journal of the American College of Surgeons 220 (2) (2015) 136–148. doi:10.1016/j.jamcollsurg.2014.10.019.

[3] H. Böhner, T. C. Hummel, U. Habel, C. Miller, S. Reinbott, Q. Yang, A. Gabriel, R. Friedrichs, E. E. Müller, C. Ohmann, et al., Predicting delirium after vascular surgery, Ann Surg 238 (2003) 149–156.

[4] Y. Morimoto, M. Yoshimura, K. Utada, K. Setoyama, M. Matsumoto, T. Sakabe, Prediction of postoperative delirium after abdominal surgery in the elderly, Journal of Anesthesia 23 (1) (2009) 51–56. doi:10.1007/s00540-008-0688-1.

[5] J. W. Raats, W. A. van Eijsden, R. M. P. H. Crolla, E. W. Steyerberg, L. van der Laan, Risk factors and outcomes for postoperative delirium after major surgery in elderly patients, PLOS ONE 10 (8) (2015) 1–12. doi:10.1371/journal.pone.0136071.

[6] C. Soguero-Ruiz, K. Hindberg, I. Mora-Jiménez, J. L. Rojo-Álvarez, S. O. Skrøvseth, F. Godtliebsen, K. Mortensen, A. Revhaug, R.-O. Lindsetmo, K. M. Augestad, R. Jenssen, Predicting colorectal surgical complications using heterogeneous clinical data and kernel methods, Journal of Biomedical Informatics 61 (2016) 87–96. doi:10.1016/j.jbi.2016.03.008.

[7] S. Huang, P. LePendu, S. Iyer, M. Tai-Seale, D. Carrell, N. H. Shah, Toward personalizing treatment for depression: predicting diagnosis and severity, JAMIA 21 (6) (2014) 1069–1075. doi:10.1136/amiajnl-2014-002733.

[8] S. Dua, U. R. Acharya, P. Dua, Machine Learning in Healthcare Informatics, Vol. 56, Springer, 2014.

[9] P. Yadav, M. Steinbach, V. Kumar, G. Simon, Mining electronic health records (EHR): A survey, Technical Report.

[10] C. Shivade, P. Raghavan, E. Fosler-Lussier, P. J. Embi, N. Elhadad, S. B. Johnson, A. M. Lai, A review of approaches to identifying patient phenotype cohorts using electronic health records, Journal of the American Medical Informatics Association 21 (2) (2013) 221–230.

[11] Y. Halpern, Y. Choi, H. Steven, D. Sontag, Using anchors to estimate clinical state without labeled data, AMIA Annual Symposium Proceedings (2014) 606–615.

[12] Y. Halpern, S. Horng, Y. Choi, D. Sontag, Electronic medical record phenotyping using the anchor and learn framework, Journal of the American Medical Informatics Associationdoi:10.1093/jamia/ocw011.

[13] T. J. Hastie, R. J. Tibshirani, J. H. Friedman, The elements of statistical learning : data mining, inference, and prediction, Springer series in statistics, Springer, New York, 2009, autres impressions : 2011 (corr.), 2013 (7e corr.).

[14] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society, Series B 67 (2005) 301–320. doi:10.1111/j.1467-9868.2005.00503.x.

[15] C. Elkan, K. Noto, Learning classifiers from only positive and unlabeled data, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008, 2008, pp. 213–220. doi:10.1145/1401890.1401920.

[16] J. N. Myhre, K. Ø. Mikalsen, S. Løkse, R. Jenssen, Consensus clustering using kNN mode seeking, in: Image Analysis, Springer, 2015, pp. 175–186. doi:10.1007/978-3-319-19665-7_15.

[17] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information processing & management 24 (5) (1988) 513–523.

[18] I. Jolliffe, Principal component analysis, Wiley Online Library, 2002.

[19] J. C. Gower, G. J. S. Ross, Minimum spanning trees and single linkage cluster analysis, Journal of the Royal Statistical Society. Series C (Applied Statistics) 18 (1). doi:10.2307/2346439.

[20] C. Soguero-Ruiz, K. Hindberg, J. Rojo-Alvarez, S. O. Skrøvseth, F. Godtliebsen, K. E. Mortensen, A. Revhaug, R.-O. Lindsetmo, K. M. Augestad, R. Jenssen, Support vector feature selection for early detection of anastomosis leakage from bag-of-words in electronic health records., IEEE journal of biomedical and health informatics 20 (5) (2016) 1404–1415. doi:10.1109/JBHI.2014.2361688.

[21] T. N. Robinson, B. Eiseman, Postoperative delirium in the elderly: Diagnosis and management., Clinical Interventions in Aging 3 (2) (2008) 351–355.

[22] A. F. M. Scholz, C. Oldroyd, K. McCarthy, T. J. Quinn, J. Hewitt, Systematic review and meta-analysis of risk factors for postoperative delirium among older patients undergoing gastrointestinal surgery, British Journal of Surgery 103 (2) (2016) e21–e28. doi:10.1002/bjs.10062.

[23] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, Journal of Statistical Software 33 (1) (2010) 1–22. doi:10.1145/1273496.1273501.

[24] J. Davis, M. Goadrich, The relationship between precision-recall and ROC curves, in: Proceedings of the 23rd International Conference on Machine Learning, ICML '06, ACM, 2006, pp. 233–240. doi:10.1145/1143844.1143874.
[25] B. Efron, The bootstrap and modern statistics, Journal of the American Statistical Association 95 (452) (2000) 1293–1296.
[26] N. Ward, Nutrition support to patients undergoing gastrointestinal surgery, Nutrition Journal 2 (1) (2003) 1–5. doi:10.1186/1475-2891-2-18.
[27] F. Mordelet, J.-P. Vert, A bagging SVM to learn from positive and unlabeled examples, Pattern Recognition Letters 37 (2014) 201–209. doi:10.1016/j.patrec.2013.06.010.
[28] M. Claesen, F. D. Smet, J. A. Suykens, B. D. Moor, A robust ensemble approach to learn from positive and unlabeled data using SVM base models, Neurocomputing 160 (2015) 73–84. doi:10.1016/j.neucom.2014.10.081.
[29] M. Claesen, F. D. Smet, P. Gillard, C. Mathieu, B. D. Moor, Building classifiers to predict the start of glucose-lowering pharmacotherapy using Belgian health expenditure data, CoRR abs/1504.07389.
[30] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, J. F. Hurdle, Extracting information from textual documents in the electronic health record: a review of recent research., Yearbook of medical informatics (2008) 128–144.
[31] P. Jensen, L. Jensen, S. Brunak, Mining electronic health records: towards better research applications and clinical care., Nature Reviews. Genetics 13 (6) (2012) 395–405. doi:10.1038/nrg3208.
[32] C. Soguero-Ruiz, L. Lechuga-Suarez, I. Mora-Jiménez, J. Ramos-Lopez, O. Barquero-Perez, A. Garcia-Alberola, J. L. Rojo-Alvarez, Ontology for heart rate turbulence domain from the conceptual model of snomed-ct, IEEE Transactions on Biomedical Engineering 60 (7) (2013) 1825–1833.
[33] B. Neerland, M. Ahmed, L. Watne, K. Hov, T. Wyller, New consciousness scale for delirium., Tidsskrift for den Norske lægeforening: tidsskrift for praktisk medicin, ny række 134 (2) (2014) 150.
[34] Z. Tieges, A. McGrath, R. J. Hall, A. M. MacLullich, Abnormal level of arousal as a predictor of delirium and inattention: an exploratory study, The American Journal of Geriatric Psychiatry 21 (12) (2013) 1244–1253.
[35] W. L. Koontz, P. M. Narendra, K. Fukunaga, A graph-theoretic approach to nonparametric cluster analysis, Computers, IEEE Transactions on 100 (9) (1976) 936–944. doi:710.1109/TC.1976.1674719.
[36] R. P. Duin, A. L. Fred, M. Loog, E. Pekalska, Mode seeking clustering by kNN and mean shift evaluated, in: Structural, Syntactic, and Statistical Pattern Recognition, Springer, 2012, pp. 51–59.
[37] Y. Cheng, Mean shift, mode seeking, and clustering, Pattern Analysis and Machine Intelligence, IEEE Transactions on 17 (8) (1995) 790–799. doi:10.1109/34.400568.
[38] D. Comaniciu, P. Meer, Mean shift: A robust approach toward feature space analysis, Pattern Analysis and Machine Intelligence, IEEE Transactions on 24 (5) (2002) 603–619. doi:10.1109/34.1000236.
[39] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, Journal of Machine Learning Research 9 (2008) 2579–2605.

Appendix A. kNN mode seeking consensus clustering

In this section we give a brief description of the clustering method used for anchor specification. The clustering method belongs to the consensus framework, meaning that the same *kNN-mode seeking* algorithm is applied many times with a random k-parameter to a resampled version of the dataset each time. The kNN mode seeking algorithm [35, 36] is a density based algorithm, similar to mean-shift [37, 38], but the kernel density estimates are replaced by k-nearest neighbors (kNN) density estimates. This algorithm is used in each iteration in the consensus clustering. A detailed description of the framework is given in Algorithm 1. An advantage with this method is that there are no critical parameter choices such as number of clusters, bandwidth parameters, etc.

To assign cluster labels to new patients cannot be done using the kNN mode seeking consensus clustering algorithm since there exist no out-of-sample mapping. However, since the clustering algorithm is based on a k-nearest neighbours search, one could assign cluster labels to new data points using a kNN classifier [13].

Algorithm 1 Consensus clustering using kNN mode seeking

Input Dataset X, range of k-values K, subsampling rate p and number of clustering trials M.

Initialize I and S as $\mathbf{0}_{N\times N}$
for each clustering trial **do**
    Draw a random k^* from K.
    Draw a random sample of size pN, X^*, from X.
    For each pair of data points in X^* update the counter matrix I by $I_{ij} = I_{ij} + 1$, where (i, j) are the indices of the data points in X.
    Use kNN mode seeking with parameter k^* to obtain a clustering of X^*.
    For each pair of data points in X^*, (i, j), that belong to the same cluster, update S by $S_{ij} = S_{ij} + 1$.
end for
Normalize the consensus matrix, $\mathcal{S}$, by dividing element-wise by the counter matrix; $\mathcal{S}_{ij} = \frac{S_{ij}}{I_{ij}}$
Create a dendrogram using average linkage.
Obtain the final clustering by selecting the cluster configuration with the longest lifetime.

Output Clustering C of X.

Appendix B. t-distributed stochastic neighbor embedding (t-SNE)

The t-SNE algorithm is one of the most well-established techniques for visualizing high-dimensional data in two or three dimensions. It has shown robustness and has become the state-of-the-art visualization method for many different data types [39]. The algorithm has the property that it creates a single map that reveals structure in the data at many different scales. The objective in this algorithm, which consists of two main stages, is to map points, $\mathbf{x} \in \mathbb{R}^p$, in a high dimension, p, to a low dimension d, $\mathbf{v} \in \mathbb{R}^d$ [39]. Firstly, one estimates a joint probability distribution, $p_{ij} = \frac{p_{j|i}+p_{i|j}}{2N}$, in the original, high-dimensional space over each pair of data points using a Gaussian kernel

$$p_{j|i} = \frac{e^{-\frac{1}{2\sigma_i^2}\|\mathbf{x}_i-\mathbf{x}_j\|^2}}{\sum_{k\neq i} e^{-\frac{1}{2\sigma_i^2}\|\mathbf{x}_i-\mathbf{x}_k\|^2}}. \tag{B.1}$$

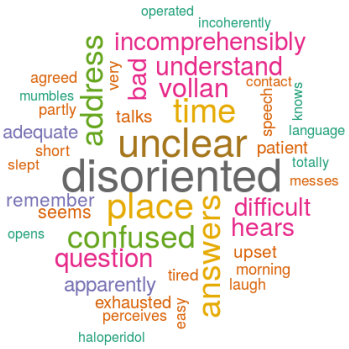
Hence, p_{ij} represents the similarity between the data points $\mathbf{x}_i$ and $\mathbf{x}_j$. Secondly, the heavy-tailed Student t-distribution with one degree of freedom is used to model similarities in the low-dimensional space as

$$q_{ij} = \frac{(1 + \|\mathbf{v}_i - \mathbf{v}_j\|^2)^{-1}}{\sum_{k\neq l}(1 + \|\mathbf{v}_k - \mathbf{v}_l\|^2)^{-1}}. \tag{B.2}$$

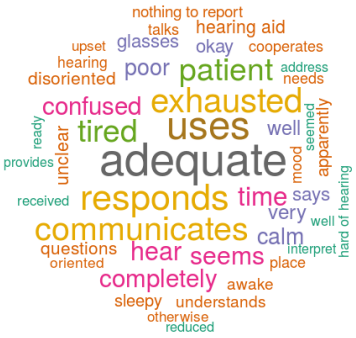
Then, the locations of the points $\mathbf{v}_i$ are found by minimizing the Kullback–Leibler divergence, $KL(P\,\|\,Q) = \sum_{i\neq j} p_{ij}\log(p_{ij}q_{ij}^{-1})$, using gradient descent. P and Q are the joint probability distributions over all data points in the high- and low-dimensional space, respectively.

Appendix C. Anchor specification

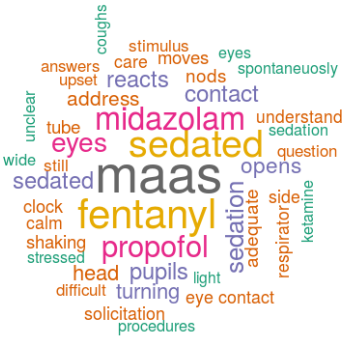
In addition to the wordclouds (Figure C.4) and the t-SNE map shown in Figures 3a-3c, Table C.5 contains information related to the word *confused* that was used to assess the certainty of this predictive anchor. For example Table C.5 shows that for 35% of the patients in cluster 1 the frequency of *confused* is at least three, whereas for 71% of the patients the frequency is at least one.



(a) Cluster 1.



(b) Cluster 2.



(c) Cluster 3.

(d) Cluster 4.

Figure C.4: By applying the clustering procedure as described in Section 3.2 to the training data four clusters were obtained. In this figure we have shown the most important features in each cluster. The size of each word corresponds to their relative tf-ipf values.

Appendix C.1. Adding more anchors

As we described in Section 3.3.5, we used the ranking provided by AM with the anchor A_1 and clinical knowledge, to add the word *disoriented* as a predictive anchor. By looking at the wordcloud in Figure C.4 and Table C.6, we observe that the top ranked word *disoriented* also is very frequent in cluster 1. This is another reason for using *disoriented* as the next predictive anchor.

The semantic meanings of disoriented and confused are quite similar. Moreover, the t-SNE plot (not shown here) of the patients with the word *disoriented* in their nurses notes is very

Frequency	1	2	3	4
Cluster 1	0.7059	0.5000	0.3529	0.2647
Cluster 2	0.3507	0.2090	0.1343	0.0970
Cluster 3	0.4839	0.2903	0.1290	0.0323
Cluster 4	0.0349	0.0063	0.0016	0
Overall	0.1301	0.0699	0.0422	0.0277

Table C.5: Fraction of patients in each cluster for whom the word *confused* appeared at least 1,2,3 and 4 times, respectively, in their nurses notes.

Frequency	1	2	3	4
Cluster 1	0.9117	0.7941	0.6470	0.5882
Cluster 2	0.2910	0.2089	0.0970	0.0522
Cluster 3	0.5161	0.2903	0.2258	0.1935
Cluster 4	0.0285	0.0031	0	0
Overall	0.1253	0.0795	0.0506	0.0397

Table C.6: Fraction of patients in each cluster for whom the word *disoriented* appeared at least 1,2,3 and 4 times, respectively, in their nurses notes.

similar to the t-SNE plot corresponding to *confused* shown in Figure 3. Therefore we decided to use (almost) same certainty measure for these two predictive anchors,

$$c_2 = \begin{cases} 2, & \text{if the patient belongs to cluster 1 or 3, or other} \\ & \text{predictive anchors appear at least twice.} \\ 3, & \text{otherwise.} \end{cases}$$

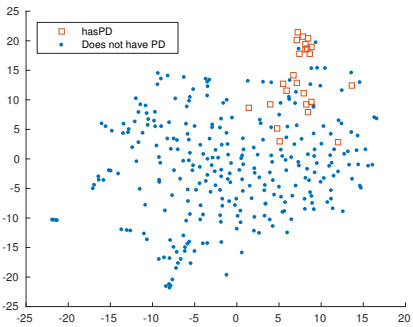
We defined the anchor *disoriented** according to Eq. (3) and created a composite anchor, A_2, as the union of *confused** and *disoriented**.

The two other anchors, *unclear** and *haloperidol**, were added in a very similar fashion. From them we defined the composite anchor, A_3, as the union of *confused**, *disoriented** and *unclear** and A_4 as the union of all four.

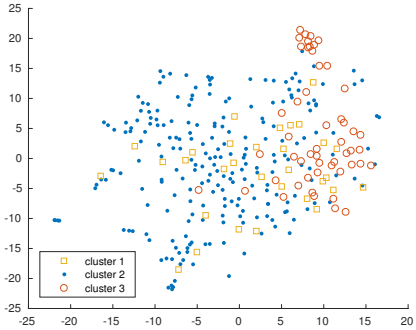
Appendix D. Classification based on clustering of test set

By looking at the clustering results shown as word clouds in Figure C.4 it seems like cluster 1 contains many words related to PD and this might indicate that doing classification only based on the clustering results could solve the problem we have considered in this paper. We investigate this further here.

We applied the clustering algorithm to the labeled test set alone and obtained three clusters. A t-SNE mapping of the data in two dimensions is shown in Figure D.5. By looking at the high-frequent words in the different clusters, we also in this case found a cluster containing many words related to the mental status of the patient. Based on these results we classified all patients in cluster 3 as *hasPD* and got an AUC-PR value of 0.456 with a 95 % CI (0.436, 0.483). These results are not very convincing and we conclude that it is meaningful to apply the AM for this problem.



(a) Locations of the patients with PD in a two dimensional t-SNE map. The red squares correspond to patients that have PD, and the blue dots to patients that do not have PD.



(b) Locations of the three clusters in a two dimensional t-SNE map. Yellow squares correspond to patients that belong to cluster 1, red circles to patients in cluster 2, blue dots to patients in cluster 3.

Figure D.5: Plots of the t-SNE mapping of the test set.