Identifying Catheter-Related Events Through Sentence Classification

Thomas Brox Røst, Christine Raaen Tvedt, Haldor Husby,
Ingrid Andås Berg, Øystein Nytrø

June 22, 2020

Contents

1 Abstract

Infections caused by central venous catheter (CVC) use is a serious and under-reported problem in healthcare. The CVC is almost ubiquitous in critical care because it enables fast circulatory monitoring and central administration of medication and nutrition. However, the CVC exposes the patient to a risk of blood-stream infections (BSI). Explicit documentation of normal CVC usage and exposure is sparse and indirect in the health record. For a clinician, CVC presence is simple to infer from record statements about procedures, plans and results related to CVC. In order to capture evidence about CVC-related risk of infections and complications, it is important to develop computerized tools that can estimate individual patient days of CVC exposure retrospectively for large cohorts of patients. Towards that objective, we have developed methods for learning classifiers for statements about CVC-related events occurring in the textual health record. This includes developing and testing an annotation ontology of events and indicators, annotation guidelines, a gold standard of annotated clinical records selected from a corpus of complete health records for more 800 episodes of care and collecting alternate health register evidence for validation purposes. This paper describes the available data and gold standard, feature selection approaches and our experiments with different classification algorithms. We find that even with limited data it is possible to build reasonably accurate sentence classifiers for the most important events. We also find that making use of document meta information helps improve classification quality by providing additional context to a sentence. Finally, we outline some strategies on using our results for future analysis and reasoning about CVC usage intervals and CVC exposure over individual patient trajectories.

2 Introduction

The use of intravenous cannulation is a very common procedure when a patient is treated in a hospital setting. The most common type of cannulation is performed with peripheral venous catheters (PVCs), since peripheral veins are readily accessible for insertion (Mermel, 2017). Central venous catheters (CVC) are primarily used to administer medications and fluids and to measure central venous pressure (Taylor and Palagiri, 2007). They typically consist of a tube that is inserted into one of the central veins of a patient. How long a patient is in need of a CVC varies from a couple of days to several months. The use of CVC in medical treatment is indispensable and life-saving for many patients but also exposes them for risk of infection and consequently increased morbidity and mortality (McKibben et al., 2005). Bacteria that are colonised on the catheter may cause a catheter-related bloodstream infection (CRBSI). For the first 3-4 days of CVC usage the risk is low (Fletcher, 2005). As the number of CVC usage days increases, so does the risk of CRBSI. This is a severe complication of CVC usage and may lead to hospital-acquired sepsis and in worst case death. More than 15 % of patients experience one or more complications during CVC insertion or maintenance (Taylor and Palagiri, 2007). Common complications in addition to catheter-related infections include arterial puncture, hematoma, pneumothorax, and venous thrombosis. Of these, catheter-related infections and venous thrombosis are often deadly. In some cases the mortality rate may be as high as 25 % (Brun-Buisson, 2001). Even though CVC usage is common we do not know enough about the prevalence and duration of CVC use, CVC-related infections and the associated patient injuries (Wong et al., 2018).

CVC-related infections are risky for the affected patient and costly to treat, often leading to prolonged hospitalization. A 2008 study of CRBSI in an intensive care unit found that each CRBSI event added approximately USD 82,000 in extra costs and 14 additional hospital days (Cohen et al., 2010). In a 2002 study of healthcare-associated infections in U.S. hospitals, the highest death rates were associated with bloodstream infections in intensive care units. Of a total of 81,942 infections, 25 % of these had death as an outcome.

Surveillance regimes and adverse event detection are the preferred approaches to increased quality of care and is mainly performed in intensive care units. These regimes require considerable manual labor, do not give much clinical effect, and may not be applicable in all hospital wards. In Norway, quarterly prevalence surveys are used to describe the current state of all hospitalized patients, but are not sufficient for estimating risk related to days of CVC usage. Ideally, we would like to use retrospective patient data to derive a precise risk ratio of CBRSI per CVC-day, and thus gain more detailed knowledge about an important patient safety indicator. In turn, this can guide better practice related to central-line catheter usage.

3 Objectives

In this paper we describe our research on automated retrospective capture of CVC-related events from a data set of annotated clinical notes. The project was performed in collaboration with researchers at Akershus University Hospital (Ahus). From their experience, there is insufficient knowledge about prevalence and duration of CVC use for patients in Norwegian hospitals. The duration of CVC use (number of CVC days) is an important prerequisite for estimating the risk of CRBSI, and a first step towards targeted quality improvement work. It is also desirable to have better data on CVC insertion and removal events, without relying on explicit coding.

Our approach was to manually annotate the content of clinical notes with CVC-related events and states and then train machine learning classifiers on the annotated data set. Identifying events such as CVC insertion, care and removal can contribute to a faster and more accessible overview of the occurrence and duration of CVC usage. It can also provide improved monitoring of CVC-related bloodstream infections, thus contributing to patient safety. Moreover, detecting CVC placement can also be of use when performing risk evaluations.

To our knowledge, using machine learning and natural language processing for detecting CVC-related events has not been done previously on clinical notes in the Norwegian language. The work of Penz et al. (2007) on English-language clinical notes is similar but relies on a semi-automated approach and was targeted towards adverse events. Our focus is on CVC exposure time in general, and more specifically individualized risk assessment. This CVC-specific work is part of more general research on capturing episodes and exposure in health records.

4 Related work

In a systematic review of 200 studies related to bloodstream infections and intravascular devices, Maki et al. (2006) found that CVCs have far higher incidence rates than peripheral intravenous catheters and midline catheters. A study by Hojsak et al. (2012) investigated the rate of CVC-related sepsis for patients on parenteral nutrition. They found that CVC was used on average 243.9 days per patient. Because of septic episodes 12.8 % of the used catheters were removed. The importance of intervention and monitoring of catheter use was demonstrated by Pronovost et al. (2010). For a total of 300,310 catheter days their Keystone ICU quality improvement project saw a mean and median reduction of CRBSI from 7.7 and 2.7 to 1.3 and 0 over a 16-18 month period. Bruin et al. (2012) applied a fuzzy logic-based system to generate rules for early detection of CVC-related infections. Trick et al. (2003) evaluated the ability of the SymText natural language processing (NLP) system to find mentions of CVC in chest radiograph reports. SymText yielded a sensitivity of 95.8 % and a specificity of 98.7 % when compared with human interpretation. Penz et al. (2007) compared the performance of an NLP program (MedLEE) and a phrase-matching algorithm in detecting CVC-related adverse events from clinical records. They found that phrase matching was a sensitive but non-specific method while the NLP program was less sensitive but significantly more specific. Combining the methods gave an acceptable sensitivity (72.0 %) and specificity (80.1 %). Another interesting finding was that incomplete or inaccurate clinical notes hampered all methods, including manual chart review. In a 2014 study by Michelson et al. (2014), text mining methods were found to be very effective in detecting different types of surgical site infections (SSIs). They did not consider CRBSIs specifically but their system was able to identify 100 % of infection cases detected by regular surveillance as well as 37 cases not previously identified. Bates et al. (2003) and Govindan et al. (2010) both give comprehensive overviews of various adverse event detection approaches. In general, there is pervasive research on retrospective NLP analysis of health data for many other purposes, though this falls outside of the scope of this paper.

5 Methods

5.1 Data

To perform our experiments we needed to build a dataset from a cohort of patients with CVC exposure. Our study design specified that we needed the patient records for both patients with CVC use as well as patients with both CVC and BSI. Following that we would acquire a larger set of reasonably similar patient records where CVC may or may not have been used. Rather than extracting the full patient record for a given patient we narrowed the data requirements down to a single episode of care. Our definition of an episode of care corresponds to the one used in the Norwegian specialist healthcare patient register NPR [1]. For an episode of care, we acquired the continuous set of clinical notes from the hospital. The patient may have had other contacts with healthcare providers, documented in separate records, but we did not collect those records from the same period. All episodes were to be selected from the DIPS EHR database of Akershus University Hospital (Ahus). Ahus is a tertiary-level university hospital that often receives and transfers patients to other hospitals. By only selecting complete episodes, i.e.

[1] An episode of care is a period where the patient receives care and treatment from one institution for one health problem. An episode may be an outpatient visit, a day visit or a hospitalization, potentially with interspersed leaves. An episode designates activity, not only treatment.

including a concrete initial admission and final discharge, we largely avoided truncated episodes.

We decided to only extract the text in the clinical notes themselves and not any accompanying structured data. One reason was that we did not consider any of the available structured information directly useful for our purposes. There are specific NCSP (NOMESKO Classification of Surgical Procedures) surgical codes, such as PYGC00 ("Insertion of central venous catheter") and related codes, that we could use both for corpus selection and as a classification feature. This code could have been recorded in the structured part of the EHR and then possibly reported to national registries. However, when searching for this and related codes in the EHR, the number of results returned was much too low to be realistic. This was not unexpected: Part of the rationale for this study was that structured reporting of CVC use was lacking. Generally, ubiquitous procedures are not counted or documented separately if they are obvious or implicit in more comprehensive procedures. Moreover, reimbursement calculation (DRG coding) for intensive care patients tries to model severity and complexity, and CVC usage is not a distinguishing feature. In addition, our view of the patient was somewhat limited. Many patients would be transferred to or from the hospital which meant that the relevant surgical coding may not have been visible in the records available to us. The lack of CVC-related structured coding is also known from other research. In a paper on CVC adverse event detection, Penz et al. (2007) found that the unstructured text was the best source for finding patients with CVC.

Given that the coding could not aid corpus selection, we decided to make use of prevalence surveys instead. In Norway, all hospitals must perform two annual surveys on infections and the use of antibiotics. In such a survey the CVC state, in addition to several other parameters such as known infections, is recorded manually at a given date and time. Fortunately, this is done four times a year at Ahus and we thus decided to base our corpus selection on patients present on one of the four survey dates (Løwer et al., 2013) at the hospital. We needed a corpus containing a sufficiently large number of patients with CVC and decided on the following selection criteria:

1. For six quarterly prevalence survey days, all health record notes for the ongoing episode, for all patients registered as having CVC on the prevalence survey day were extracted. The identity of episodes or patients, or actual survey findings, were unknown to researchers and not represented explicitly in the record.

2. For a seventh prevalence survey day, complete episode health record notes for all inpatients in the most relevant departments were included.

> This was to give us a representative set of similar patients, not necessarily having CVC at the prevalence survey date. Still, these patients could be expected to have many similar traits and findings and be subject to (peripheral and urinary) catheters.

We required the episode length to be at least four days. The rationale for this lower bound on episode length was to increase the total volume of the corpus. Some episodes spanned more than one prevalence surveys, but duplicates were removed in the final corpus. We could not identify if unique patients gave rise to more than one distinct episode but this was irrelevant for our study. Following this approach we ended up with a corpus which is summarized in Table 1.

Table 1: Corpus overview: Episodes and notes

			Notes	
Survey	Episodes	Notes	[2]Inspected	Annotated
1	44	2708	2708	377
2	28	2883	2883	432
3	14	1369	1369	165
4	23	1595	1595	190
5	57	2808	2804	341
6	22	2147	2147	289
7	631	[3]32104	[4]8668	[5]951
Totals	[6]819	45614	22174	2745

As mentioned we would extract all clinical notes available to us for each selected episode of care, including nursing notes, surgical notes, physician notes, laboratory examinations, and more. The average number of clinical notes for each patient was high enough to ensure that at good selection of notes both with and without CVC use were included. Considerable effort was needed both to retrieve, organize and clean the data, as described by Husby (2014) and Berg (2014). Each note in our corpus was represented as a plain-text file. Since all notes in the EHR were originally in RTF format they needed to be converted to a plain-text format without losing any formatting that was relevant to the interpretation of the note. The EHR vendor did not provide built-in conversion to text so a custom solution had to be built.

[2] Read, but not annotated

[3] All manually searched for content potentially relevant for CVC

[4] Positive search results, manually inspected

[5] True positives

[6] Some episodes are counted more than once, because they last longer than 3 months

The initial corpus would contain personally identifying information (PII) about both patients, staff, family, and other related people. This meant that the research would sort under the Norwegian Medical and Health Research Act (hfo, 2008), which stipulates that the Norwegian Regional Committees of Medical Research Ethics (REK) had to be involved. The research plan and objectives, including descriptions on how PII would be handled, was submitted to the committee, which then evaluated the research ethics of the project and finally approved our application. The application stated that only named researchers in the EVICARE project who had signed non-disclosure agreements would have data access. The data would be stored on an offline restricted local network where all access would be logged with timestamps and the identification of the accessing researcher. The physical server was only accessible to system administrators.

5.2 Annotation

To identify the clinical state documented by the clinical notes in our corpus, we defined a set of CVC-related annotation labels (Table 2). This was done as a collaboration between the authors and a domain expert in natural language processing. The classes of patient states labelled were intended to form a generalization hierarchy, e.g. "CVC" being a more general type of CVC-state than "Hickmann". When applied to the text, the annotator would label one or more words that would (roughly) act as a confirming proof of a certain state, situation or event. In practice, this meant that an annotation could span everything from a single word to a complete sentence.

The classes were intended to form an ontology about events, states, devices, conditions and symptoms. However, sparsity of events and non-documented care for CVC skewed our results. Furthermore, it was a continuing challenge to separate clinically implicit patient state from textually explicit record statements when assigning labels. I.e., what a trained clinician would be able to infer about patient reality and what could be read in the text documents. For the purpose of identifying CVC-state, we had to re-interpret the labels, and this is further discussed in section 5.3.

All the notes were translated to plain text, retaining sections, section headings and cleaning punctuation and sentence-dividers; see section 6 for details. Each note was saved into a single file. Each file was named with a unique serial number, patient ID, episode of care ID, the note type, and a timestamp showing when the original clinical note was written. No other correction or parsing was applied, so the individual note would have the appearance of a well formatted clinical note. The Brat rapid annotation tool (Stenetorp et al., 2012) was set up with the designed annotation ontology, accessing one file at a time. Some test annotations were done during ontology design, but this was discarded once the annotation guideline was agreed upon and considered stable. The annotator, which is also one of the authors of this paper, was a nurse with special competence in infection control. For each processed clinical note file a corresponding annotation file was created. For each annotation, this file had a line with a local identifier, the annotation label, the start and stop character for the annotated text (referring to the original note), and the text fragment from the original note that was annotated. The following example shows what a single annotation could look like:

```
T1 RemCVC 241 277 CVC removed and tip sent for culture
```

If no annotations were made, an empty annotation file was still created; this would tell us that the file had been reviewed by the annotator but was without any annotated findings. The annotation files were named identically as the corresponding note, but given a different suffix. After the annotation process was completed we had a total of 22,174 notes. All the notes from survey days 1 to 6 were annotated. Only a quarter of the notes included after the day 7 prevalence survey was annotated. This fraction was determined by time and resources available after annotating all the notes for the episodes included because of the other 6 survey days (Table 1).

Table 2: Annotations

Annotation	Description
Carecvc	Care, observation or assessment of CVC.
PlanCarecvc	Care of CVC has not been performed, but has been booked or planned.
PlanInscvc	Admission of CVC not performed, but planned, desired or ordered for the future.
Inscvc	CVC has been inserted in the period covered by this note.
Remcvc	CVC has been removed in the period covered by this note.
PlanRemvcvc	Removal of CVC has been planned or ordered for the future.

Continued on next page

Continued from previous page

Annotation	Description
Symptom	Statements indicating that there may be a blood system infection (BSI).
Sepsis	Sentence containing the word "sepsis" or mention of similar conditions.
CVC, Hickmann, VAP, other	Labels for more or less specific type of CVC.
JugularVein, SubclavicanVein, Femoralis	Labels for site of CVC.
Possiblecvc	Sentences where CVC is discussed without implication that CVC is present.

5.3 Data Analysis

Some of the note types, in particular the nursing notes, had a distinct format. This reflected the document editing interface in the EHR system, which came with predefined templates to structure the documentation process. The nursing notes often used a template with 12 different headings. Example headings are "Communication/Senses", "Breathing/Circulation", "Pain/Sleep/Rest/Well-being", and "Skin/Tissue/Wounds". Most nursing notes would follow this template, but typically only a subset of the sections would be used. For the most frequently occurring note types where such a structure existed we built regular expression-based parsers to extract the contextual information along with the text. The assumption was that knowledge about the context of a piece of text could potentially be used as a feature to enhance its interpretation. We also knew from Husby (2014) that approximately 10 % of the nursing notes would contain CVC-related information under the "Skin/Tissue/Wounds" heading, thus making the section information a potentially valuable feature. For this project we chose not to apply any deeper linguistic analysis, such as e.g. part-of-speech analysis. We had previous experience that clinical language was often terse and grammatically incomplete. Furthermore, we did not have access to comprehensive vocabularies of clinical terms for entity recognition.

Once the parsers had been tested and refined sufficiently there were still 65 out of the original 45,614 notes that would not pass, usually because the structure had for some reason been mangled. Given the total volume of clinical notes we decided it was safe to discard these. We also chose to discard 1,892 notes that we thought were not relevant because of their note type. Examples of these were letters to the patient or to other healthcare institutions. Following this we were left with 43,657 notes. The final reduction of the corpus was to remove duplicated notes. From our initial data analysis we observed that several notes were exact duplicates where only the timestamp differed. Discussions with technical staff revealed that this was an artifact of how the EHR worked: Whenever a clinical note was reopened a new note would be generated, even if no changes were made. This could happen when a nurse opens a document for editing, but only read it. A similar situation of semi-duplication occurs when a document is edited incrementally. This creates a new note only slightly different from the previous one. We observed some cases where this happened but did not do any analysis of how prevalent this was; this could be relevant for future work. After removing duplicates, the final corpus size was 42,806 clinical notes. Another corpus reduction task we considered was to discard notes with infrequently occurring note types. Ultimately we decided against this as it would potentially affect the episode of care length.

Once we had extracted the text along with associated meta information we grouped all the note data according to the episode of care. The final processed corpus contained 778 episodes of care with 122 different types of clinical notes. Table 3 shows the most frequently occurring note types. The nursing notes were by far the most common note types. This made sense given that nurses are working three shifts and have a need to communicate throughout the day for continuity of care. For 50 of the 122 note types there were less than 5 note examples, making this a fairly long-tailed distribution. While e.g. somatic nursing notes are subdivided into "care", "plan" and "evaluation", the table aggregates this type for compactness. However, we treated these different nursing notes as separate in the analysis.

As shown in Table 1, approximately 50 % of the notes were inspected, of which 2,745 received annotations. The 10 most annotated note types are shown in Table 4. The rightmost column shows the number of notes where actual CVC annotations were made, i.e. not just empty annotation files. Of the remaining notes, 4,056 were read and 564 had annotations.

Table 5 shows how the annotation classes are distributed over the annotated notes. The most common class is CVC care (including observation and assessment), which makes sense given that this is an action likely to be performed during a nurse visit. Note that the number of CVC insertion and removal annotations differ. This can be explained by the CVC already being present when the patient arrives at the hospital or not being removed before leaving or being transferred. It may also be the case that documentation

Table 3: Note types, translated

Note Type	Count
Somatic nurse note (care, plan, evaluation)	28265
Somatic physician note	6641
Intensive nurse note (care, plan, evaluation)	1830
Somatic physician discharge summary	727
Somatic nurse ward admission note	596
Somatic medical admission note	574
Somatic nurse ward transfer note	426
Somatic nurse reception note	415
Somatic nurse summary	305
Somatic physician discharge note	183

Table 4: Annotated note types, translated

Annotated Note Type	Total	Annotated
Somatic nurse note (care, plan, evaluation)	2942	380
Somatic physician note	660	105
Intensive nurse note (care, plan, evaluation)	137	16
Somatic nurse ward transfer note	51	2
Somatic nurse ward admisson note	18	4
Somatic physician discharge summary	17	8
Somatic medical admission note	16	4
(Somatic, physician) Transfer note	16	3
Palliative note	16	0
Somatic nurse ward admission note	14	5

is missing or incomplete, although this is less likely given the seriousness of the procedure. Another possibility is that the CVC spans more than one episode of care. A further complication is that more than one CVC may be present—we found cases of up to three CVCs being present—and inserted at different times, but removed together.

Table 5: Annotation count

Annotation	Description
Carecvc	349
Symptom	123
PlanInscvc	82
Inscvc	63
PlanCarecvc	54
Remcvc	50
CVC	37
Possiblecvc	35
Sepsis	32
PlanRemvcvc	22
JugularVein	19
Hickman	13
SubclavicanVein	6

In Figure 1 we see the distribution of the number of notes in each episode of care. The mean number is 55 while the median is 34. The longest episode of care in terms of the number of notes had 643 notes. The similar statistic for episode of care duration in number of hospitalization days is shown in Figure 2. Here the mean was 29 days and the median 13. This reflects all the episodes which involve patients with CVC.

After inspecting some of the longest episodes it turned out that there were mostly sound medical reasons behind the long hospitalizations. In many ways this was expected, given that CVC use is often associated with serious medical conditions. There were, however, exceptions. In the episode with the longest duration, which lasted 361 days, it turned out that the actual admission period was approximately a fortnight. Almost a year after the discharge a single clinical note was tacked onto the episode, containing a standardized report to the national cancer registry. These deviations were also likely to occur for other episodes, so some care was needed if the admission period was to be used as e.g. a feature. We decided not to consider this as a problem since the episode length was not used in our experiments.

Table 4 shows that most of the annotation events are very sparse. This was a challenge, given that sparse classes is a common problem in machine learning. To alleviate this we decided to make use of the intended generalization hierarchy of annotation event classes. In terms of semantics, classes such as CVC care and use are fundamentally quite similar, meaning it is probably safe to group them together into a common class. Besides, for our research purposes it was not necessary to exactly predict the given annotation labels: Our interest was in the CVC usage prevalence and duration, which means that the main goal was to detect the transitions between having and not having CVC. Accordingly, we decided to create four aggregate classes from the initial fifteen: *Plan* (PlanInscvc), *Ins* (Inscvc), *Use* (Carecvc, PlanCarecvc, CVC, PossibleCVC, PlanRemcvc, JugularVein, Hickman, SubclavicanVein) and *Rem* (Remcvc). Note that planning removal of a CVC implies that the CVC is present. The reasoning behind these classes were that they should be sufficent to support our future attempts to infer periods of continuous CVC use. Note that the Sepsis and Symptom classes were discarded for now, even though they represent a substantial number of annotations. This was done because the Sepsis and Symptom labels were often used in situations that were unrelated to actual CVC use and could as such be a source of confusion to the classifiers. Table 6 shows the final distribution of our new aggregate classes. There is still some imbalance although to a smaller extent than before. As expected, tha majority of samples are in the *Use* class.

Table 6: Aggregate annotation count

Note type	Count
Plan	82
Ins	63
Use	535
Rem	50

Finally, Figure 3 shows the aggregate annotation class frequency relative to the most common note types and Figure 4 shows the same information relative to the different sections in the somatic nursing notes (excluding sections without annotations). The numbers in parentheses show the total number of observations. For some documents and sections some of the sparser aggregate classes occur with a relatively high frequency.

6 Experiments

In order to find evidence of CVC use we decided to build text classifiers that would, given clinical notes as input, make predictions as to whether or not one of our previously mentioned aggregate annotation classes should apply. In practice, the output classes would then be no CVC use (*None*), CVC planning (*Plan*), CVC insertion (*Ins*), CVC use (*Use*), and CVC removal (*Rem*). This would give us a foundation for later prediction of CVC usage intervals. Rather than classifying the whole note, we instead opted for classifying sentences given that the annotations were granular enough to attribute them to a particular sentence. This would also make it easier to use section information as an additional feature.

Our tool of choice for cleaning up the section notes and converting them into sentences was the Python NLTK Natural Language Toolkit (Loper and Bird, 2002). To perform sentence splitting with sufficient quality we used the NLTK Punkt Sentence Tokenizer. This tokenizer could be trained with our clinical notes as input data to perform unsupervised sentence boundary detection (Kiss and Strunk, 2006). We found that it was easily confused by abbreviations and spelling errors, both of which are common in clinical notes. To alleviate this we had to manually add said errors and abbreviations to the tokenizer, thus gradually improving its quality. After several iterations of manual review and corrections we found that the tokenizer yielded sufficient although not perfect quality on our source material. The fact that the source language was Norwegian did not pose any problems, so no translation or other modifications was necessary for the sentence splitting to work as intended.

As mentioned, the nursing notes largely followed templates with fixed section headers, author roles, hospital department and other information. We were particularly interested in the section information but also the other available information. After extracting the sentences from each note the resulting information was placed into a JSON data structure where each sentence was associated with relevant meta information, including the section header. If no section header information was available, as was the case for notes other than nursing notes, the sentences were given a "general" section header label.

To train our sentence classifiers we chose to use the Python scikit-learn library (Pedregosa et al., 2012). This is a well-established and efficient machine learning and data analysis toolkit which provided the functionality we required for this experiment. Data pre-processing yielded a total of 344,563 sentences. From these we selected 34,810 sentences that had been through the annotation process, out of which 640 had actual annotations. From a machine learning point of view this can be considered a fairly small data set, so we decided on using 4-fold cross validation rather than the more common 10-fold approach. Given the highly imbalanced data set (most sentences belonged to the *None* class) we considered whether or not stratification would make sense. Experiments both with and without fold stratification indicated that stratified folds slightly alleviated the class imbalance problem and provided overall better classification performance. Accordingly, we settled on stratified folds. Our task was a multiclass classification problem and we decided to take the one-versus-all approach in this experiment.

Using the scikit `TfidfVectorizer` we gave each sentence in the training data set a tf-idf representation, using sublinear tf scaling and a `max_df` parameter setting of 0.5. This choice would remove frequently occurring words and was an alternative to techniques such as removing stop words. Another inherent feature of this vectorizer was that it performed automatic tokenization, lowercase conversion, and punctuation handling, thus providing basic text pre-processing functionality. In addition to this we also converted numbers to a generic number token, this to reduce the variability in the text and on the assumption that the actual numeric values had limited value for our classification task. Stemming was considered but since we wanted to preserve verb tenses we decided against this. An example of a case where verb tense could make a difference would be the discussion of a planned CVC insertion versus an actual insertion. For this reason we set up a separate experiment to investigate the effect of stemming. Handling of negation is another common challenge in natural language processing tasks. We did not make any efforts towards explicitly handling negation, assuming instead that the use of n-gram models would enable the classifiers to differentiate between negated and non-negated concepts. As for n-gram models, we experimented with different n-gram dimensions and their impacts on classifier performance. In the end we settled on using 1- to 3-grams for all experiments as this combination seemed to provide the best results. The use of unigrams was partly motivated by the terseness of clinical language; single-word features could make a difference as single-word sentences were known to exist.

We selected a set of common algorithm implementations in scikit-learn using the default or recommended settings as the initial parameters. For the first experiment we wanted to see how the number of features used would affect the performance of the selected algorithms on the majority class *Use* and the minority class *Rem*. A key aspect when limiting the number of features used is how features are selected. We decided to use the scikit-learn `SelectKBest` univariate feature selector with a chi-squared statistical test for scoring. This selector scores the features according to the chosen scoring function and returns the desired number of features. Manual inspection of the top features showed that the chosen features made sense given the context and our domain knowledge. For example, direct references to CVC or various catheter types, were highly ranked. Also, many features were closely associated with nursing tasks, e.g. the removal of sutures. This was reasonable since we had a large number of nursing notes in our data set.

Table 7 shows an example of the 20 highest ranked features, translated from Norwegian to English, in a trial experiment on all 34,810 sentences.

Table 7: Highest scoring features

cvc	cvc day	cvc care
removed sutures	removed sutures from	from hickman catheter
given cvc	cvc was inserted	received new cvc
have been inserted	hickman	hickman catheter
new cvc	disc cvc	discontinuing cvc
discontinued cvc day	discontinued cvc	care
sutures from	sutures from hickman	

In Figure 5 we see the balanced F_1 score for the *Use* class for the chosen algorithms while Figure 6 shows the same experiment for the sparse *Rem* class. For both classes the performance of the `linear_svc_l1`, `linear_svc_l2` and `ridge` algorithms improves with the number of features. For the *Rem* class there are better performing algorithms that seem to perform well with a limited number of features. In terms of priorities, we decided that optimizing performance for the *Use* class should be our primary experiment objective. Having a well-performing CVC usage classifier would not only be beneficial for the purpose of counting days of CVC use but would also make good use of the more prevalent *Use*-related annotations in our data set. For these reasons we opted to use the `linear_svc_l1` and `linear_svc_l2` algorithms for the remainder of our experiments. These algorithms are the scikit-learn implementations of a linear kernel support vector machine (SVM) with parameters `loss=squared_hinge`, `penalty=l1` (or `l2`), `dual=False` and `tol=1e-3`. The `l1` and `l2` influences the sparsity of the internal coefficient vectors.

We repeated the number of features experiment although this time only using the `linear_svc_l1` algorithm. Figure 7 shows the F_1-score for all 5 classes. As could be expected, the prediction performance is lower for the classes with less training data. Coincidentally the predictive quality of the *Use* classifier is similar to the results seen in the adverse CVC event detection by Penz et al. (2007), although a direct comparison can not be made.

The next experiment sought to evaluate if inluding sentence section information (see Figures 3 and 4) and note type as features could improve classifier performance. The simplest way to achieve this was to use the scikit-learn `FeatureUnion` functionality which combines different feature sources into a unified feature vector. Applying this on each sentence would give us a combined feature vector that relied on both the standard bag-of-words features as well as additional section and note type features. We chose to give each feature source equal weight rather than weighting some of them as more important than others. We defined three experiment setups with different feature source combinations: sentence, sentence + section, and sentence + section + note type. Table 8 shows the results, where F_1, precision, and recall are given for each setup.

Table 8: Experiments combining sentence and note type information

	Sen			Sen/Sec			Sen/Sec/Not		
Class	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
None	**99.8**	**99.9**	**99.8**	**99.8**	99.8	**99.8**	**99.8**	99.8	**99.8**
Plan	63.4	38.2	47.5	**69.2**	**41.1**	**51.2**	66.9	38.1	48.4
Ins	47.5	18.1	24.6	**50.0**	18.1	25.5	**50.0**	**19.8**	**27.5**
Use	74.0	84.5	78.9	**74.3**	**85.4**	**79.5**	73.9	85.0	79.1
Rem	**81.3**	22.4	35.0	**81.3**	**26.6**	**39.1**	**81.3**	22.4	35.0

The numbers shown in bold are the highest scores for the given class. Most noteworthy is that adding section features has a positive effect on prediction quality while note type has a negative effect. The one exception for the latter is the *Ins* (insertion) class. A manual inspection revealed that documentation of CVC insertion was almost always found in the anesthesiology record note type, so in that way it made sense that including note type information would have a positive impact.

In our final experiment we wanted to investigate the effect of stemming on classifier performance. The effect of pre-processing techniques such as stemming may be highly dependent on e.g. the text domain and the language used (Uysal and Gunal, 2014). Using `linear_svc_l2` we ran an experiment with and without stemming, otherwise using all available text features, 4-fold cross-validation and no sentence or note type features. For stemming we used the Norwegian Snowball stemmer that is bundled with NLTK. The results can be seen in Table 9.

Table 9: Experiments with and without stemming

For the most common class, *Use*, stemming has a slightly negative although negligible effect. For the sparsest class, *Rem*, there is however a

	Without stemming			With stemming		
Class	Pr	Re	F1	Pr	Re	F1
None	**99.7**	**99.9**	**99.8**	**99.7**	**99.9**	**99.8**
Plan	80.0	46.1	57.9	**82.9**	**48.7**	**60.7**
Ins	**48.1**	**34.3**	**38.5**	**48.1**	**34.3**	**38.5**
Use	**77.5**	**78.5**	**78.0**	**77.5**	77.9	77.7
Rem	63.2	31.8	41.7	**66.5**	**38.4**	**48.5**

marked improvement with stemming. A similar effect is seen with *Plan*, which is also quite sparse. For *Ins* there is no difference. A possible explanation is that a potential benefit from differentiating between e.g. verb tenses is outweighed by the dimensionality reduction of the feature space that stemming provides in our quite small data set. A similar effect for another non-English language is seen in e.g. Torunoglu et al. (2011), where stemming was found to be beneficial for small training sets.

7 Conclusion

We found that even with limited training data it is still possible to predict CVC use events from sentences in clinical notes with adequate precision and recall. This gives us a foundation for later inference of CVC usage periods, thus allowing us to get better estimates for the number of days that CVC has been in use. It seems likely that additional training data will improve classifier performance; in particular, it would be useful with better performance for insertion and removal events. Another interesting finding was that using sentence context information would provide an additional performance boost. It can reasonably be assumed that more accurate classifiers for the sparser events will lead to less ambiguity when attempting to map CVC use intervals from discrete CVC events.

There are several avenues for further research on this topic, most importantly the aforementioned CVC usage period prediction and day count. In addition there are many possible approaches towards strengthening the event prediction foundation. Given the sparsity of training data, one interesting option would be to see how convolutional neural networks perform, given that they have been shown to sometimes work well even with limited training data sets. Another option is to expand our notion of sentence context to also include data from previous clinical notes, thus providing even more background that may aid the classifiers. The key here is probably to find a representation of previous events, treatment and patient background that is both at a high-enough level to be useful but also not overly simplistic.

Another observation is on the difficulty of extracting text from EHR systems and the fact that exporting options are often quite limited. While the importance of secondary use of clinical data is increasingly recognized (Meystre et al., 2017), there are often many practical obstacles towards accessing such data. The trend among EHR vendors is somewhat towards e.g. interoperability and API access, but often only for structured content. For research on unstructured clinical text, it is also necessary that the text is available in a format useful for export and that elements of text structure and the usage context are not lost during the export process. In particular, text as part of forms lose their meaning unless the specific form is also available for text processing.

The end goal of this project is as mentioned to improve our knowledge of the prevalence and duration of CVC use in hospitals. The work described in this paper is preliminary and can be considered a means towards this end. A key element is to be able to accurately identify transitions between CVC use states: from planning to insertion, care during use, and removal. When doing so it is important to recognize the difference between the actions that were originally applied to the patient and how these were ultimately documented. There are multiple aspects that must be taken into account, not least given the variety of note types. For example, a nursing note will typically describe actions and observations from the current 8-hour shift and which are relevant for the next shift. These notes are mostly descriptive, and will also be written shortly after the described events took place. On the other hand, a discharge note will summarize a wider variety of events that took place over a longer period of time. It may also be more reflective and also outline plans for further treatment. Mapping descriptions in the clinical notes as accurately as possible to the points on the timeline where they actually took or will take place is critical for getting an accurate CVC use day count.

8 Acknowledgements

The reported research was approved by the regional ethical committee (Reference 2010/338) and supported by EVICARE (Evidence-based care processes: Integrating knowledge in clinical information systems) (NRC project 193022). To a lesser extent, the BigMed project (NRC project 259055) also contributed. This article is an extended version of an article accepted at BIBM 2018 Røst et al. (2018).

9 Figures

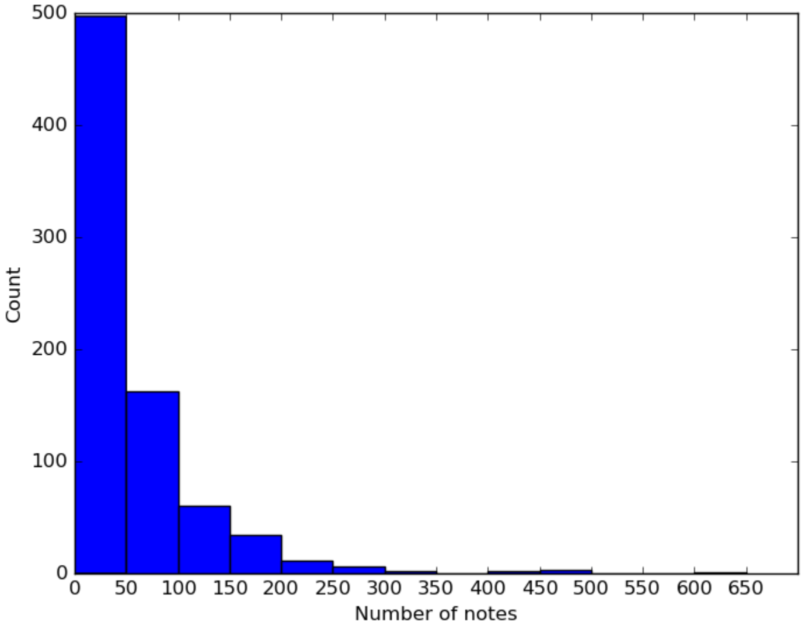

Figure 1: Episode of care length (notes)

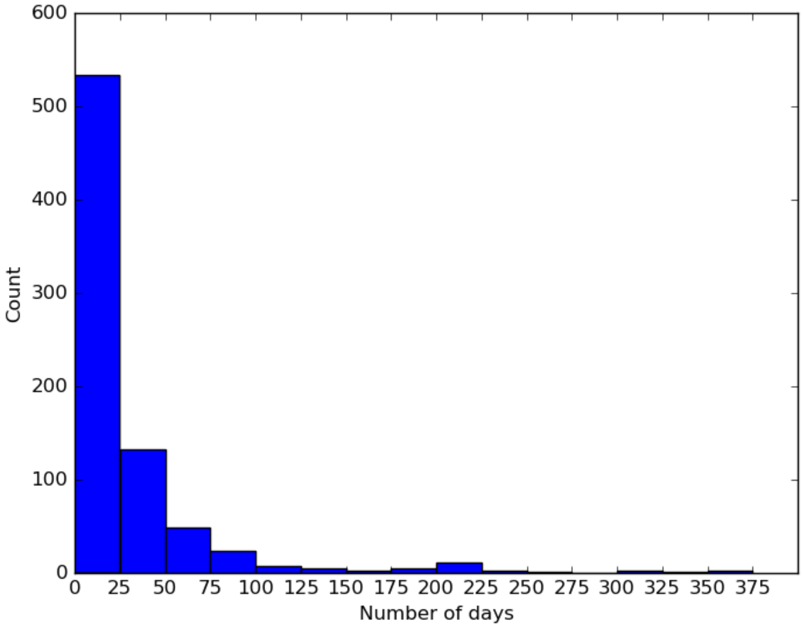

Figure 2: Episode of care length (days)

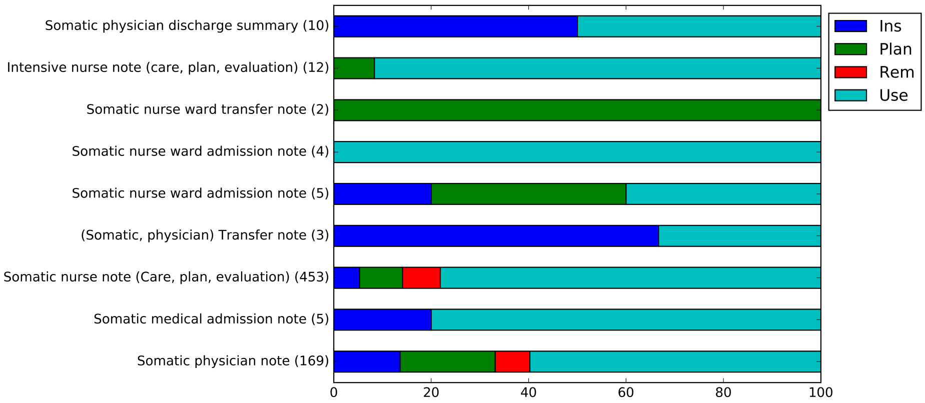

Figure 3: Aggregate class use per document type

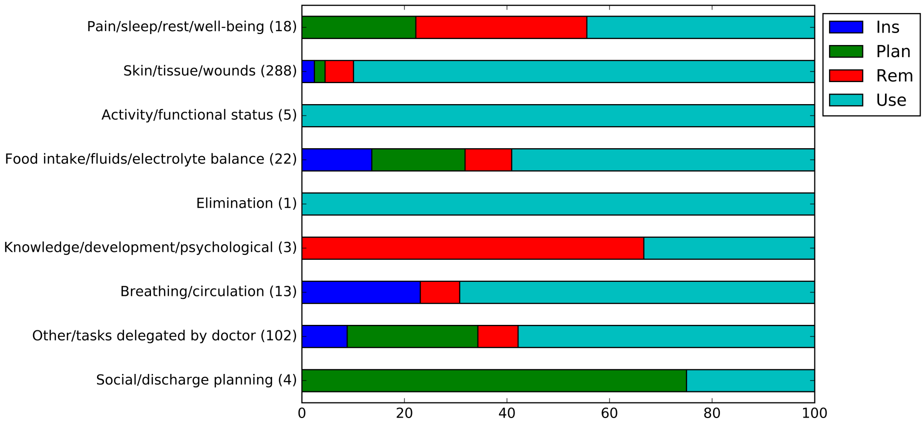

Figure 4: Aggregate class use per nursing note section type

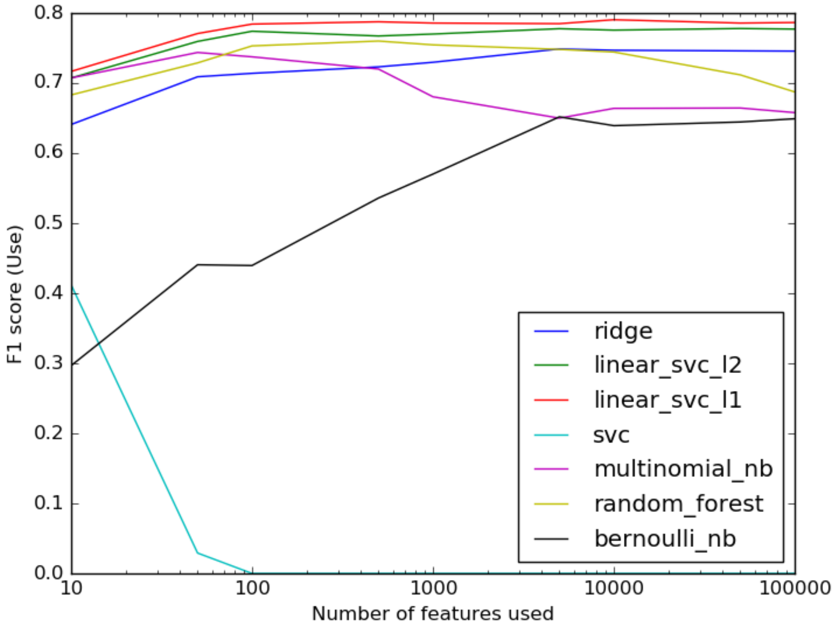

Figure 5: F1 vs. number of features (Use)

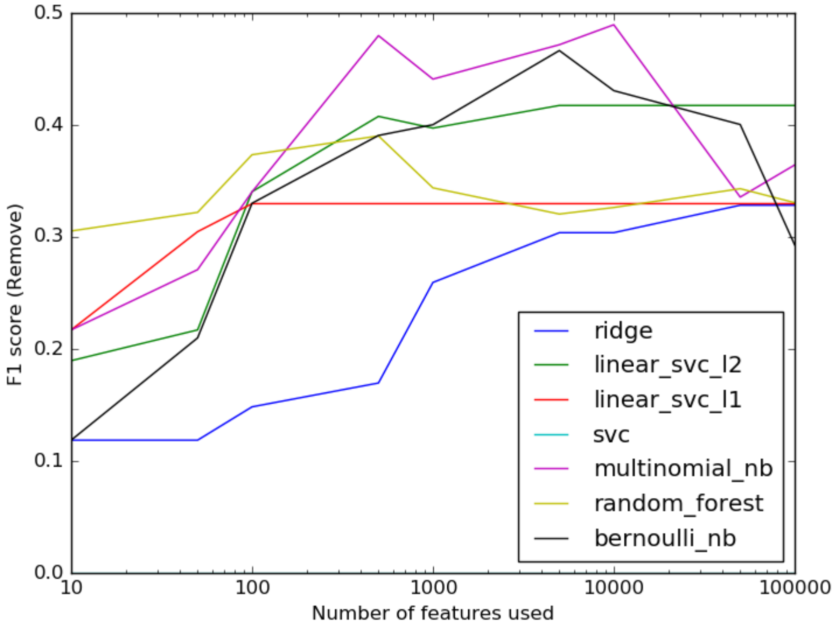

Figure 6: F1 vs. number of features (Rem)

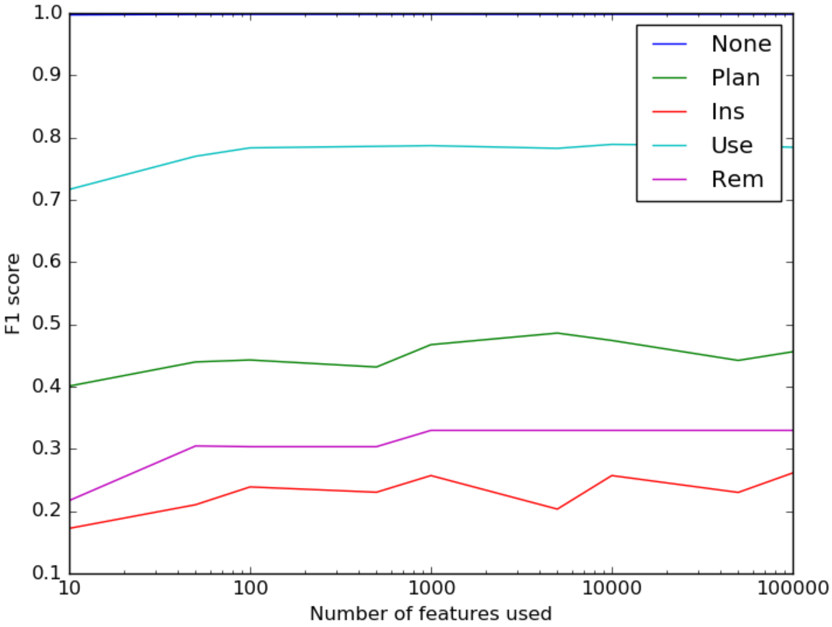

Figure 7: F1 vs. number of features (all classes)

10 Bibliography

References

(2008). Lov 20. juni 2008 nr. 44 om medisinsk og helsefaglig forskning.

Bates, D. W., Evans, R. S., Murff, H., Stetson, P. D., Pizziferri, L., and Hripcsak, G. (2003). Detecting adverse events using information technology. *Journal of the American Medical Informatics Association*, 10(2):115–128.

Berg, I. A. (2014). Automated annotation of events related to central venous catheterization in norwegian clinical notes. Master's thesis, Norwegian University of Science and Technology.

Bruin, J. S. D., Blacky, A., and Adlassnig-Peter, K.-P. (2012). Assessing the clinical uses of fuzzy detection results in the automated detection of cvc-related infections: a preliminary report. *Studies in Health Technology and Informatics*, 180(Quality of Life through Quality of Information):579–583.

Brun-Buisson, C. (2001). New technologies and infection control practices to prevent intravascular catheter-related infections. *American Journal of Respiratory and Critical Care Medicine*, 164(9):1557–1558.

Cohen, E. R., Feinglass, J., Barsuk, J. H., Barnard, C., O'Donnell, A., McGaghie, W. C., and Wayne, D. B. (2010). Cost savings from reduced catheter-related bloodstream infection after simulation-based education for residents in a medical intensive care unit. *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare*, 5(2):98–102.

Fletcher, S. (2005). Catheter-related bloodstream infection. *Continuing Education in Anaesthesia Critical Care & Pain*, 5(2):49–51.

Govindan, M., Citters, A. D. V., Nelson, E. C., Kelly-Cummings, J., and Suresh, G. (2010). Automated detection of harm in healthcare with information technology: a systematic review. *BMJ Quality & Safety*, 19(5):e11–e11.

Hojsak, I., Strizić, H., Mišak, Z., Rimac, I., Bukovina, G., Prlić, H., and Kolaček, S. (2012). Central venous catheter related sepsis in children on parenteral nutrition: a 21-year single-center experience. *Clinical Nutrition*, 31(5):672–675.

Husby, H. (2014). Klassifisering av sykepleiejournalen - kan kunnskap om sykepleiedokumenter forbedre gjenkjenning av hendelser knyttet til sentralvenekateterisering? Master's thesis, Norwegian University of Science and Technology.

Kiss, T. and Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Loper, E. and Bird, S. (2002). Nltk. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics.*

Løwer, H. L., Eriksen, H.-M., Aavitsland, P., and Skjeldestad, F. E. (2013). Methodology of the norwegian surveillance system for healthcare-associated infections: the value of a mandatory system, automated data collection, and active postdischarge surveillance. *American Journal of Infection Control*, 41(7):591–596.

Maki, D. G., Kluger, D. M., and Crnich, C. J. (2006). The risk of bloodstream infection in adults with different intravascular devices: a systematic review of 200 published prospective studies. *Mayo Clinic Proceedings*, 81(9):1159–1171.

McKibben, L., Horan, T., Tokars, J. I., Fowler, G., Cardo, D. M., Pearson, M. L., and Brennan, P. J. (2005). Guidance on public reporting of healthcare-associated infections: Recommendations of the healthcare infection control practices advisory committee. *American Journal of Infection Control*, 33(4):217–226.

Mermel, L. A. (2017). Short-term peripheral venous catheter-related bloodstream infections: a systematic review. *Clinical Infectious Diseases*, 65(10):1757–1762.

Meystre, S. M., Lovis, C., Bürkle, T., Tognola, G., Budrionis, A., and Lehmann, C. U. (2017). Clinical data reuse or secondary use: Current status and potential future progress. *Yearbook of Medical Informatics*, 26(01):38–52.

Michelson, J. D., Pariseau, J. S., and Paganelli, W. C. (2014). Assessing surgical site infection risk factors using electronic medical records and text mining. *American Journal of Infection Control*, 42(3):333–336.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2012). Scikit-learn: Machine learning in python. *CoRR*.

Penz, J. F., Wilcox, A. B., and Hurdle, J. F. (2007). Automated identification of adverse events related to central venous catheters. *Journal of Biomedical Informatics*, 40(2):174–182.

Pronovost, P. J., Goeschel, C. A., Colantuoni, E., Watson, S., Lubomski, L. H., Berenholtz, S. M., Thompson, D. A., Sinopoli, D. J., Cosgrove, S., Sexton, J. B., Marsteller, J. A., Hyzy, R. C., Welsh, R., Posa, P., Schumacher, K., and Needham, D. (2010). Sustaining reductions in catheter related bloodstream infections in michigan intensive care units: Observational study. *BMJ*, 340(feb04 1):c309–c309.

Røst, T. B., Tvedt, C. R., Husby, H., Berg, I. A., and Nytrø, Ø. (2018). Capturing central venous catheterization events in health record texts. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 488–495.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsuji, J. (2012). Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Taylor, R. W. and Palagiri, A. V. (2007). Central venous catheterization. *Critical Care Medicine*, 35(5):1390–1396.

Torunoglu, D., Cakirman, E., Ganiz, M. C., Akyokus, S., and Gurbuz, M. Z. (2011). Analysis of preprocessing methods on classification of turkish texts. In *2011 International Symposium on Innovations in Intelligent Systems and Applications*, pages 112–117.

Trick, W. E., Chapman, W. W., Wisniewski, M. F., Peterson, B. J., Solomon, S. L., and Weinstein, R. A. (2003). Electronic interpretation of chest radiograph reports to detect central venous catheters. *Infect Control Hosp Epidemiol*, 24(12):950–954.

Uysal, A. K. and Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management*, 50(1):104–112.

Wong, A. V., Arora, N., Olusanya, O., Sharif, B., Lundin, R. M., Dhadda, A., Clarke, S., Siviter, R., Argent, M., Denton, G., Dennis, A., Day, A., Szakmany, T., and group, T. F. I. C. N. A. P. I.-. (2018). Insertion rates and complications of central lines in the uk population: A pilot study. *Journal of the Intensive Care Society*, 19(1):19–25.