Capturing Central Venous Catheterization Events in Health Record Texts*

Thomas Brox Røst[1], Christine Raaen Tvedt[2], Haldor Husby[3], Ingrid Andås Berg[4] and Øystein Nytrø[1]

***Abstract*— Infections caused by central venous catheter (CVC) use is a serious and under-reported problem. In our research we explore methods of automatically detecting CVC use from clinical documentation for quality improvement and surveillance purposes. This paper describes our initial research on this topic, where we build CVC event classifiers based on an episodes of care corpus and an annotated gold standard. After describing the available data and gold standard we then experiment with different classification algorithms and feature selection approaches. We find that even with limited data it is possible to build reasonably accurate sentence classifiers, at least for the events that are most important to us. We also find that making use of document meta information may help improve classification quality by providing additional context to a sentence. Finally, we outline some strategies on using these preliminary clinical document-centric results as a tool for future analysis and elicitation of CVC usage intervals over full patient histories.**

I. INTRODUCTION

Central venous catheters (CVC) are primarily used to administer medications and fluids and to measure central venous pressure [17]. 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 central lines in medical treatment is indispensable for many patients but is also exposing them for risk of infection and consequently increased morbidity and mortality [13]. 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 [6]. 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 [17] and the mortality rate may be as high as 25% [3].

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 [4]. Even though CVC usage is common we do not know enough about the prevalence of CVC use, CVC-related infections and the associated patient injuries. Surveillance regimes and adverse event detection are the preferred approaches to increase quality of care and is mainly performed in intensive care units. Surveillance regimes requires considerable manual labor, do not give any 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.

II. 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 to estimate 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 Norwegian language. The work of Penz et al. [15] 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.

* This work was supported by EVICARE (Evidence-based care processes: Integrating knowledge in clinical information systems), NRC project 193022

[1] Dept. Computer Science, Norwegian Univ. of Science and Technology
[2] Lovisenberg Diaconal University College
[3] Institute of Clinical Medicine, University of Oslo
[4] KNOWIT

III. RELATED WORK

A study by Hojsak et al. [8] 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. De Bruin et al. [5] applied a fuzzy logic-based system to generate rules for early detection of CVC-related infections. Trick et al. [18] 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. [15] 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. Bates et al. [1] and Govindan et al. [7] 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.

IV. METHODS

A. Data

Approximately 800 episodes of care with a total of 45,614 clinical notes were selected from the DIPS EHR database of the Akershus University hospital (Ahus). The initial design called for extraction of records for episodes involving CVC (at least 50 episodes), and CVC with BSI (50 episodes) in order to develop an annotation guide, ontologies and an experimental setup. We then aimed to extract a similar set of episodes from another hospital department for validation, annotation and initial classifier development. A third phase would extract 1,000 episodes for partial annotation and classifier training and validation. We assumed that we could use surgical intervention codes (NOMESKO Classification of Surgical Procedures), but a full search revealed that this particular coding was not reliable. Many patients would be transferred to, or from, the hospital, so the surgical procedures would be invisible in local records. It was then proposed to use quarterly prevalence surveys, where the CVC state of patients was recorded manually. The study population was patients present during one of the four survey dates [12] at the hospital. The survey collects specific indicators, e.g. use of antibiotics, presence of CVC and known infections for each patient. All record notes in the ongoing episode of care were included for each patient.

Husby [9] and Berg [2] did major efforts in order to to retrieve, clean and perform initial classifier experiments. To acquire a cohort with a sufficiently large number of episodes involving patients with known CVC, the following procedure was used to select patient episodes, and thus record notes:

1) For six quarterly prevalence survey days, all health record notes for all patients with CVC were extracted. The identity of episodes or patients, or actual survey findings, were unknown to researchers and not represented in the record.
2) For a seventh prevalence survey day, all health record notes for all inpatients in the most relevant departments were included. This was to give us a representative set of other similar patients, not necessarily having CVC at the prevalence survey date.

The extraction and annotation process resulted in a corpus consisting of notes summarized in table I.

TABLE I
CORPUS OVERVIEW: EPISODES AND NOTES

		Notes		
Survey	Episodes	Total	[a]Inspected	Annotated
1	44	2,708	2,708	377
2	28	2,883	2,883	432
3	14	1,369	1,369	165
4	23	1,595	1,595	190
5	57	2,808	2,804	341
6	22	2,147	2,147	289
7	631	[b]32,104	[c]8,668	[d]951
Totals	[e]819	45,614	22,174	2,745

[a] Read, but no relevant annotations
[b] All manually searched for content potentially relevant for CVC
[c] Positive search results, manually inspected
[d] True positives
[e] Some episodes are counted more than once, because they last longer than 3 months

Each episode of care would contain all clinical notes, including nursing notes, surgical notes, medical notes written by physicians and laboratory examinations. The high number of clinical notes for each patient ensured that notes both with and without CVC-related events were included in the study. A requirement for selection of records was a care period of more than three days for each patient.

The Norwegian Regional Committees of Medical Research Ethics (REK) approved our research plan and objectives. To preserve patient anonymity, personally identifying information was removed from the data by automated methods. An exception from confidentiality was given since the researchers could potentially indirectly recognize persons known to them privately. Only named members of the EVICARE project were given access to the data. All project participants had to sign a non-disclosure agreement. The collected data was stored on a secure offline local network with restricted access both physically and electronically. All accessing researchers had to keep a log with date, time, period of access and their name. Only system administrators had access to the physical server and only the researchers had physical access to the network and connected computers.

B. Annotation

The clinical notes were manually read and annotated in order to make a corpus of data with CVC events. A nurse with special competence in infection control was responsible for the annotation. The authors, the annotator, and a domain expert in natural language processing together defined the set of CVC-related event annotations, shown in Table II. The annotator used the Brat rapid annotation tool [16]. The annotation labels were intended to form a generalization hierarchy, e.g. "CVC" being a more general type of CVC than "Hickmann".

A hospital employee converted the notes from rich text format in cooperation with DIPS, the EHR provider. Each note had a file name with a unique serial number, an anonymous, but unique, patient number, an episode of care number, the document type and the date when the document was written. Each annotated note file had an accompanying annotation file. In total, 22,175 notes were manually inspected and read, all possible notes for survey days $1-6$, while only those containing "interesting" phrases were read for survey day 7 (see table I). In all, 4,533 notes were opened in Brat while 2,745 were annotated.

TABLE II
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.
Remcvc	CVC has been removed.
PlanRemvcvc	Removal of CVC has been planned.
Symptom	Statements indicating that there may be a blood system infection (BSI).
Sepsis	Sentence containing the word “sepsis”.
Device: CVC, Hickmann, VAP, other	Type of CVC.
Site: JugularVein, SubclavicanVein, Femoralis	Site of the vein for CVC insertion.
Possiblecvc	Sentences where CVC is discussed without mentioning the word “CVC”.

The annotations had word-level granularity, meaning that an event was defined as one or more adjoining words (sometimes complete sentences) that contained sufficient information as to why the annotator considered this an event.

C. Data Analysis

The EHR system from which the notes were collected did to some extent enforce structured documentation, in particular for the nursing notes. For the most frequent note types we built simple regular expression-based parsers that would recognize this structure. This provided us with metadata that would tell us if the extracted text belonged to a certain section, was part of a symptom description, and so on. The only purpose for this was to have a note representation that included information on which part of the note any given text belonged to, and not to do deeper parsing in the traditional sense. We knew this could be of relevance since Husby [9] already noted that 10% of the nursing notes with structured content would have CVC-related content in the section about "Skin, tissue and wounds".

After several iterations of testing and refinement on the original 45,614 notes 65 were discarded due to parsing errors. In addition, 1,892 notes were not included because they were decided to be not relevant for the experiment. Examples of such note types are letters to the patient or to various institutions. This left us with a total of 43,657 remaining notes. We considered removing infrequently occurring notes but decided against this as it could potentially affect the length of the episode of care.

The parsed notes were grouped according to the episode of care they belonged to and sorted in ascending order. A known problem was that several notes were duplicated. Every time a clinical note is reopened in the origin EHR system, a new clinical note is generated. This may happen if e.g. a nurse wants to check the status of a patient without actually updating the patient record. We wanted to remove these notes so as to avoid unnecessary repetitions of content. To do so we added a post-processing step where we sequentially compared the textual contents of each note with the previous note of the same type. If they were equal the current note was removed from the episode. After removing 851 duplicate notes a total of 42,806 notes remained.

A similar concern is if a nurse adds information to a previously written note, thus generating a new note with some additional information. We suspected that this may happen but did not know how prevalent this was. For time reasons we did not do any analysis and possible correction for this issue but it may be relevant for future work.

After cleaning the data we ended up with 778 patient histories containing 122 different types of clinical notes. Table III shows the frequency of the 10 most common note types in the data set. Some of the note types were quite sparse: a total of 50 note types occurred 5 times or less. As we can see, nursing notes were by far the most prevalent.

TABLE III
NOTE TYPES, TRANSLATED

Note Type	Count
Somatic nurse note (care, plan, evaluation)	28,265
Somatic physician note	6,641
Intensive nurse note (care, plan, evaluation)	1,830
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

Only a limited number of the available notes were anno-

tated. Table IV shows the 10 most annotated note types. It also shows the number of notes where actual CVC-related annotations were made. Of the notes remaining after cleaning the data, a total of 4,056 notes were read and 564 of those were annotated.

TABLE IV
ANNOTATED NOTE TYPES, TRANSLATED

Annotated Note Type	Total	Annotated
Somatic nurse note (Care, plan, evaluation)	2,942	380
Somatic physician note	660	105
Intensive nurse note (care, plan, evaluation)	137	16
Somatic nurse ward transfer note	51	2
Somatic nurse ward admission 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

Table V shows how the annotation classes are distributed over the annotated notes. Some of the classes are quite sparse while CVC care is the most common event type. The different number of CVC insertion and removal events is because the patient, as seen in our data set, may arrive with an already present CVC or leave without removing it.

TABLE V
ANNOTATION COUNT

Note Type	Count
Carecvc	349
Symptom	123
PlanInscvc	82
Inscvc	63
PlanCarecvc	54
Remcvc	50
CVC	37
PossibleCVC	35
Sepsis	32
PlanRemcvc	22
JugularVein	19
Hickman	13
SubclavicanVein	6

The mean number of clinical notes per episode of care was 55 while the median was 34. Figure 1 shows the distribution of all episodes of care in the data set. The longest recorded episode had a total of 643 notes.

In terms of episode duration the mean was 29 days and the median was 13 days. The longest episode of care lasted 361 days. Figure 2 gives an overview of the episode of care duration in days.

We manually inspected some of the longest episodes. In most cases, the data appeared to be valid and with sound medical reasons for the long hospitalizations. We knew that

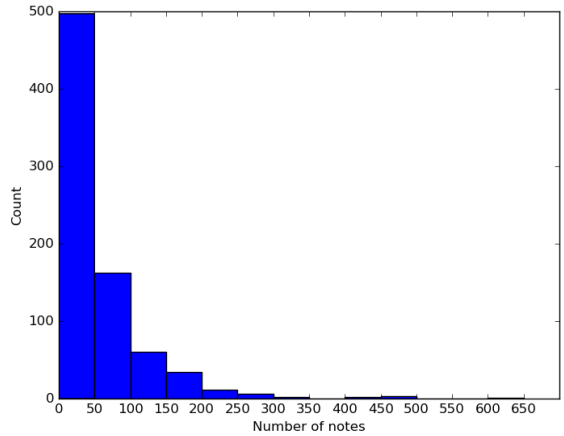

Fig. 1. Episode of care length (notes)

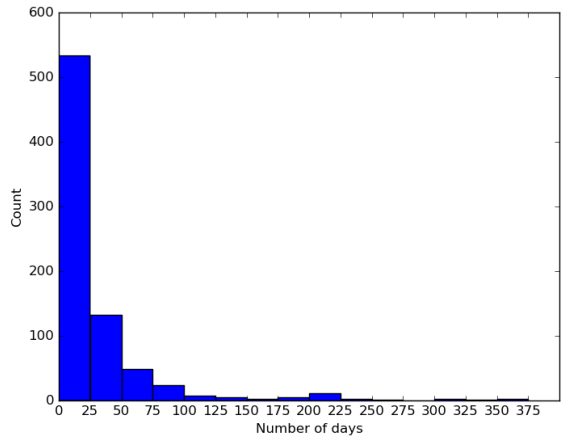

Fig. 2. Episode of care length (days)

the data set would contain descriptions of serious CVC-related infections so this made sense. In the case of the longest episode, which according to the note timestamps lasted 361 days, the actual hospitalization period from admission to discharge was just a fortnight. However, 11 months after discharge a single clinical note with a standardized report to the national cancer registry was appended. We suspect that similar occurrences may happen in other episodes, thereby artificially inflating their total duration. For this project we did not take any steps to rectify this problem, given that research on episodes of care was not a primary concern, and it was accordingly left as an issue for future work. The actual episode length was not important for any of our results.

As can be seen in Table II, some of the annotation classes are quite sparse. Also, semantically speaking, some of the classes are quite similar, such as all the CVC care and use classes. We also found that the Sepsis and Symptom classes may be used even if the occurrence is unrelated to the use of CVC. We decided that getting the exact class

right was not necessary for detecting the prevalence and duration of CVC usage, as we were more interested in the events that signify a transition between using and not using CVC. For this reason We chose to discard the Sepsis and Symptom observations and map the remaining classes into four aggregate classes (labels included in parentheses): *Plan* (PlanInscvc), *Ins* (Inscvc), *Use* (Carecvc, PlanCarecvc, CVC, PossibleCVC, PlanRemcvc, JugularVein, Hickman, SubclavicanVein) and *Rem* (Remcvc). These would give us the information we need for future reasoning over the start and end points of CVC usage intervals. Table VI shows the distribution of the new aggregate classes. While still imbalanced, the sparsest classes have been removed or subsumed. As expected, usage classes are most common.

TABLE VI
AGGREGATE ANNOTATION COUNT

Note Type	Count
Plan	82
Ins	63
Use	535
Rem	50

Figure 3 shows the relative frequency of aggregate annotation class use across the most commonly used note types, while Figure 4 shows the same, although this time for the different sections of the somatic nursing notes. For such notes, only sections with annotations were included. The numbers in parentheses indicate the total number of observations. We see that there are documents and sections where some of the sparser aggregate classes occur with higher relative frequency than could be expected. However, for some document types and section types there is not enough data to draw any conclusions.

V. EXPERIMENT

The goal of our experiment was to investigate the possibility of building classifiers to reveal whether or not a clinical note mapped to one or more of our four aggregate classes. This is a first necessary step towards later CVC usage interval analysis. Since our annotations told us the exact words that indicated CVC events, we redefined the goal to perform sentence rather than document classification: Given a complete sentence, the classifier should tell us if this sentence describes no CVC use (*None*), CVC planning (*Plan*), CVC insertion (*Ins*), CVC use (*Use*), or CVC removal (*Rem*).

We used the Python NLTK Natural Language Toolkit [11] for the initial preparation of the clinical notes. Since we were doing sentence classification it was critical to split the raw notes into proper sentences. We built a sentence tokenizer using the NLTK Punkt Sentence Tokenizer, which can be trained to perform unsupervised sentence boundary detection [10]. As is always the case with clinical documentation, there were a lot of abbreviations and spelling errors, some of which had to be added explicitly to the tokenizer for it to perform with sufficient accuracy. After several trial runs followed by manual review and corrections, we had a tokenizer that performed well enough on our source material. Note that all sentences are in Norwegian and no translation was done before applying language processing tools.

As mentioned before, we also parsed some of the notes into a semi-structured format, so that we knew e.g. the section type a sentence would occur in, the role of the author, the hospital department, and so on. We deemed the section information to be most relevant for our purposes, as it would give us additional information about the context of a particular sentence. For the nursing notes, nurses were required to document according to a fixed document structure under 12 different headings. Example headings are "Communication/Senses", "Breathing/Circulation", "Pain/Sleep/Rest/Well-being", and "Skin/Tissue/Wounds". For the most part the documentation seemed to follow this structure quite well. Typically only a small subset of the sections would be used. Using this information we converted each note into a JSON data structure where each sentence was associated with a section identifier (or a "general" section for note types other than nursing notes). We also added all other available meta information about the note.

The next step was to train our classifiers. For all classification experiments we used the Python scikit-learn library [14], as it is a well-established, stable and reasonably efficient data analysis toolkit with lots of batteries included. We selected 34,810 sentences that had been through annotation (out of a total of 344,563 sentences in the full data set) of which 640 were annotated. (For practical reasons we only had access to a subset of the annotated data for this initial experiment.) Since the annotated data set is quite small, we decided on using 4-fold cross-validation. In experiments with and without fold stratification, we generally found that stratified folds provided better overall classification performance as this would somewhat rectify the class imbalance problem. Since this is a multiclass classification problem we went for a one versus all classification approach.

Each sentence was converted to a tf-idf representation with scikit's `TfidfVectorizer`, using sublinear tf scaling and a `max_df` setting of 0.5. The latter setting removes highly-frequent words and is an alternative to using stop word lists. The only pre-processing technique we applied was to convert numbers to a generic token. `TfidfVectorizer` would anyway provide basic pre-processing such as tokenization, conversion to lowercase, and punctuation handling. We decided against using stemming, since we wanted to differentiate between verb tenses. For instance, the verb tense used when discussing a planned insertion could potentially be different from that used for a performed insertion. We did not do any analysis of the impact of word tense in our results, but it may be relevant in further research. Negation is another important aspect; here we took the easy approach of letting n-grams pick up on simple negative constructs.

We experimented with different n-gram settings and ultimately settled on using 1- to 3-grams for all experiments as

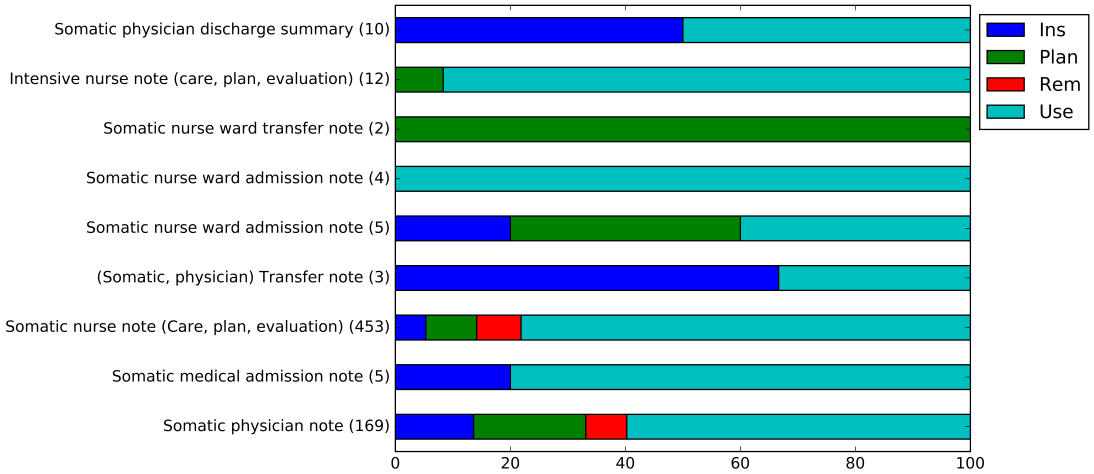

Fig. 3. Aggregate class use per document type

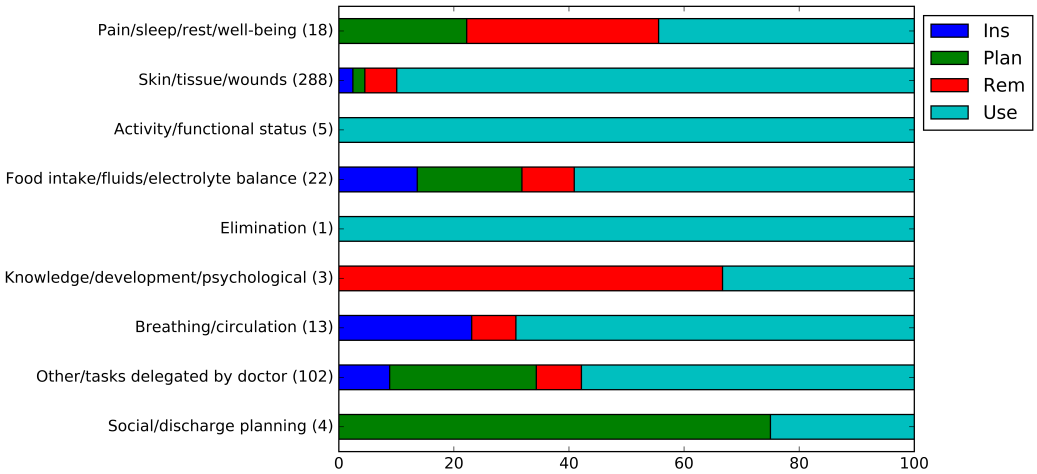

Fig. 4. Aggregate class use per nursing note section type

this seemed to yield the best results. Given the terseness of clinical language and the scarcity of training data it made sense to also allow for single-word features.

For the classification algorithm we started out with a selection of common algorithm implementations in scikit-learn, using default or recommended settings. We did an initial experiment where we varied the number of features used, just to get a feel for the effect this would have and the performance of the various algorithms. Figure 5 shows the F1 score for the majority class, *Use*, while Figure 6 shows the same experiment for the sparsest class, *Rem*. For *Use*, the best performing algorithms (`linear_svc_l1`, `linear_svc_l2`, and `ridge`) benefit from using as many features as possible. For *Rem*, the same pattern holds although here there are other algorithms that both perform better and prefer a limited set of features. Since the *Use* class is arguably the most important, we decided to do stick with the `linear_svc_l1` classifier for the rest of the experiments since this classifier appeared to give the best results for this class. This is the scikit linear kernel support vector machine implementation in `LinearSVC` with parameters `loss=squared_hinge`, `penalty=l1`, `dual=False`, and `tol=1e-3`).

Figure 7 shows the same experiment, although this time with the F1-score for all 5 classes, including the *None* class. We see that prediction quality is decidedly poorer for the sparser classes and it is reasonable to expect that more training data would help. While not directly comparable, results for the *Use* class are in the same ballpark as the results reported by Penz et al. [15] in their adverse CVC event detection.

Feature selection was done using using the scikit-learn `SelectKBest` univariate feature selector with a chi-squared statistical test. In practice, this selects the desired number of highest scoring features. On manual inspection the top features did indeed seem relevant to the domain and classification targets. The selected features for a 20-feature trial experiment are shown in Table VII, translated from Norwegian to English. As expected, given the high number of nursing notes, many of the features are closely associated

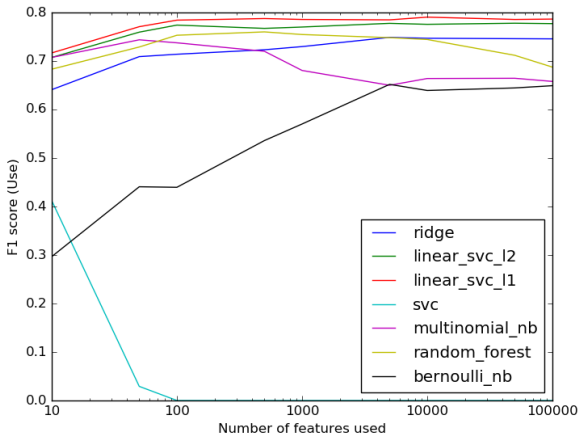

Fig. 5. F1 vs. number of features (Use)

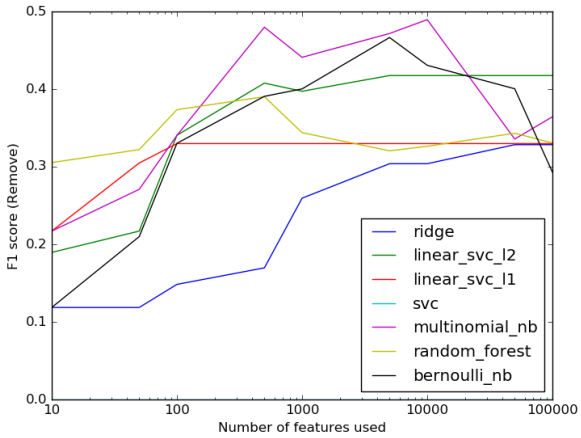

Fig. 6. F1 vs. number of features (Rem)

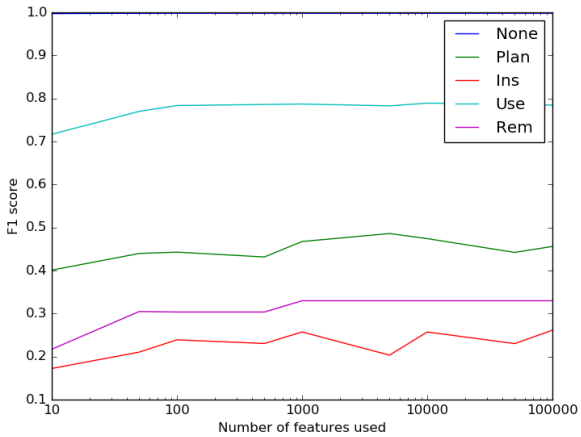

Fig. 7. F1 vs. number of features (all classes)

with nursing tasks, such as the removal of sutures.

TABLE VII
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	

Given the information in Figures 3 and 4, we also wanted to investigate if the aforementioned section information could have an impact on classifier performance. We were also interested in seeing if information about the note type was useful for sentence interpretation. To do so we used the scikit `FeatureUnion` functionality which creates combined feature vectors from several sources. For each sentence we then had a combined feature vector constructed from the sentence bag-of-words vector as well as section and note type vectors. It was possible to adjust the weighting of the vector data sources but we opted for giving each source equal weight. Table VIII shows the F1, precision, and recall results for each class with three different experiment setups: Just the sentence information, sentence + section information, and sentence + section + note type information.

TABLE VIII
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 in boldface indicate the highest scores for the given class. We see that the addition of section features tend to boost prediction quality while adding document type features on top has a negative effect, except for the *Ins* class. Upon further inspection this actually made sense since documentation of insertion is almost always recorded in the anesthesiology records.

VI. CONCLUSION

We have presented the results of our preliminary experiments in classifying CVC-related events from several different types of clinical notes. Even though the initial gold standard corpus was somewhat limited the results are still promising and there are indications that more training data will benefit the overall prediction quality, particularly for the more uncommon event types. The variety in clinical note types, many of them written for very different purposes

and with very dissimilar intentions, was one of the more interesting aspects of the experiment. Our attempts to include document meta information as classifier features did indicate that this is an approach worthy of further pursuit. Since the corpus is partitioned into complete (at least for the period of admission to the given department) episodes of care it will be of similar interest to see if previous documented treatment history can be exploited to improve classification accuracy.

Maximizing classification performance was not a priority for the work described in this paper, not the least given the scarcity of available training data. Nonetheless, classification quality is likely to have an impact on our future reasoning on CVC interval length. For this reason we plan to explore approaches such as convolutional neural networks, which have been shown to sometimes work surprisingly well even with a small gold standard data set.

The overall purpose of our research is to get better estimates of CVC use in hospitals. The work described in this paper is a stepping stone towards this goal. In future experiments we will use the classifiers to add an additional layer of event information on individual clinical notes in episodes of care. We will then try to identify the critical transitions between CVC usage states: From planning to insertion, care, and finally removal. This comes with its own set of challenges, in particular with how to align the information available to us (time-stamped, semi-structured clinical notes) with the reality of what actually happened to the patient. The variety of documentation and documentation purposes makes the mapping between reality and documentation particularly challenging. As an example, nursing notes may mostly describe events that just happened or will happen very soon, i.e. within a very narrow time slot, while physician notes may be more reflective, summarizing what brought the patient to his or her current state or outlining a long-term treatment plan. While textual event descriptions themselves are relatively straightforward to detect, giving them the right interpretation may prove critical when reasoning over the patient's transition between different treatment and health states.

ACKNOWLEDGMENT

The reported research was approved the regional ethical committee (Reference 2010/338) and supported by EVI-CARE (Evidence-based care processes: Integrating knowledge in clinical information systems), NRC project 193022.

REFERENCES

[1] David W. Bates, R. Scott Evans, Harvey Murff, Peter D. Stetson, Lisa Pizziferri, and George Hripcsak. Detecting adverse events using information technology. *Journal of the American Medical Informatics Association*, 10(2):115–128, 2003.

[2] Ingrid Andås Berg. Automated annotation of events related to central venous catheterization in norwegian clinical notes. Master's thesis, Norwegian University of Science and Technology, 2014.

[3] Christian Brun-Buisson. New technologies and infection control practices to prevent intravascular catheter-related infections. *American journal of respiratory and critical care medicine*, 164(9):1557–1558, 2001.

[4] Elaine R Cohen, Joe Feinglass, Jeffrey H Barsuk, Cynthia Barnard, Anna O'Donnell, William C McGaghie, and Diane B Wayne. Cost savings from reduced catheter-related bloodstream infection after simulation-based education for residents in a medical intensive care unit. *Simul Healthc*, 5(2):98–102, 2010.

[5] Jeroen S De Bruin, Alexander Blacky, and Klaus-Peter Adlassnig. Assessing the clinical uses of fuzzy detection results in the automated detection of cvc-related infections: a preliminary report. *Stud health technol inform*, 180:579–583, 2012.

[6] Stephen Fletcher. Catheter-related bloodstream infection. *Contin Educ Anaesth Crit Care Pain*, 5(2):49–51, 2005.

[7] M Govindan, AD Van Citters, EC Nelson, J Kelly-Cummings, and G Suresh. Automated detection of harm in healthcare with information technology: a systematic review. *Qual Saf Health Care*, 19(5):e11, 2010.

[8] Iva Hojsak, Helena Strizić, Zrinjka Mišak, Irena Rimac, Gordana Bukovina, Hrvoje Prlić, and Sanja Kolaček. Central venous catheter related sepsis in children on parenteral nutrition: a 21-year single-center experience. *Clin Nutr*, 31(5):672–675, 2012.

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

[10] Tibor Kiss and Jan Strunk. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525, 2006.

[11] Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics*, pages 63–70, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[12] Hege Line Løwer, Hanne-Merete Eriksen, Preben Aavitsland, and Finn Egil Skjeldestad. Methodology of the norwegian surveillance system for healthcare-associated infections: the value of a mandatory system, automated data collection, and active postdischarge surveillance. *Am J Infect Control*, 41(7):591–596, 2013.

[13] L McKibben, T Horan, JI Tokars, G Fowler, DM Cardo, ML Pearson, PJ Brennan, and Infection Control Practices Advisory Committee Heathcare. Guidance on public reporting of healthcare-associated infections: recommendations of the healthcare infection control practices advisory committee. *Am J Infect Control*, 33(4):217–226, 2005.

[14] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

[15] Janet F E Penz, Adam B Wilcox, and John F Hurdle. Automated identification of adverse events related to central venous catheters. *J Biomed Inform*, 40(2):174–182, 2007.

[16] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujui. Brat: a web-based tool for nlp-assisted text annotation. In *EACL '12 Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, 2012.

[17] Robert W Taylor and Ashok V Palagiri. Central venous catheterization. *Crit Care Med*, 35(5):1390–1396, 2007.

[18] William E Trick, Wendy W Chapman, Mary F Wisniewski, Brian J Peterson, Steven L Solomon, and Robert A Weinstein. Electronic interpretation of chest radiograph reports to detect central venous catheters. *Infect Control Hosp Epidemiol*, 24(12):950–954, 2003.