Iterative development of family history annotation guidelines using a synthetic corpus of clinical text

Taraka Rama* **Pål H. Brekke†** **Øystein Nytrø‡** **Lilja Øvrelid***

*University of Oslo, Department of Informatics
†Oslo University Hospital, Department of Cardiology, Center for Cardiological Innovation
‡Norwegian University of Science and Technology, Department of Computer Science
tarakark@ifi.uio.no,pabrek@ous-hf.no,nytroe@ntnu.no,liljao@ifi.uio.no

Abstract

In this article, we describe the development of annotation guidelines for family history information in Norwegian clinical text. We make use of incrementally developed synthetic clinical text describing patients' family history relating to cases of cardiac disease and present a general methodology which integrates the synthetically produced clinical statements and guideline development. We analyze inter-annotator agreement based on the developed guidelines and present results from experiments aimed at evaluating the validity and applicability of the annotated corpus using machine learning techniques. The resulting annotated corpus contains 477 sentences and 6030 tokens. Both the annotation guidelines and the annotated corpus are made freely available and as such constitutes the first publicly available resource of Norwegian clinical text.

1 Introduction

The limited availability of clinical text corpora constitutes a major challenge for the development of clinical NLP tools. Such text originates in the (electronic) health record (EHR), and access to and use of the EHR is governed by strict data privacy and health service regulations, which usually restricts secondary use and prohibits re-distribution and sharing with the larger NLP community. Among notable exceptions are anonymized health record texts published as part of the i2b2 challenges (Uzuner and Stubbs, 2015) and the CLEF corpus (Roberts et al., 2008b). For languages other than English the situation is even more difficult, and despite notable annotation efforts (Dalianis et al., 2012), the underlying corpora are largely unavailable.

Clinical texts are radically different in form and function from other biomedical texts: They are communicative, conveying information between health service providers, terse (in that the patient is implicit), and very specialized according to the role of the narrative and profession of the author (Allvin et al., 2010; Røst et al., 2008). In this work, the targeted narrative of family history corresponds to the anamnesis recorded by the cardiologist when interviewing the patient as part of a consultation. However, lacking a corpus of family history statements, we decided to develop a synthetic corpus (Lohr et al., 2018; Boag et al., 2018).

Development of most NLP tools requires manually annotated data and the design of annotation guidelines is crucial for consistent and high quality data suitable for machine learning and classification. Development of annotation guidelines is a time consuming process which in the case of clinical data often also requires access to domain experts (clinicians). The question of how to involve the clinician in the annotation process and make the best use of their domain knowledge is therefore highly relevant.

This article describes the systematic development of annotation guidelines for family history information in Norwegian clinical text. We make use of incrementally developed synthetic clinical text describing patients' family history relating to cases of cardiac diseases. The domain expert is an integral part of this methodology and generates synthetic examples that challenge the guidelines and further participates both in the annotation and development of guidelines. In doing so, the domain knowledge of the clinician informs the annotation process systematically. Measures of inter-annotator agreement is actively used to improve the annotation guideline, as well as to extend the synthetic corpus and range of annotated concepts.

In the rest of the paper, we describe the methodology for corpus generation and annotation guideline design in more detail and provide an overview of our current state of progress in the fam-

Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis (LOUHI 2018), pages 111–121
Brussels, Belgium, October 31, 2018.

ily history domain. We analyse inter-annotator agreement based on the developed guidelines and present results from experiments aimed at evaluating the validity and applicability of the purpose-made annotated corpus using machine learning.

2 Family history in clinical text

A family history is an important part of the medical record. It helps the clinician in identifying risk factors, in diagnosing conditions that have genetic components, and in identifying family members that should be offered genetic counselling or medical follow up. Specific patterns of disease or symptoms in a family suggest modes of inheritance, and could be helpful in the diagnosis of an unrecognised disease or syndrome. For example, if only men in the family are affected, one might expect an X-linked trait, or if approximately half of the offspring in a generation seem to be affected, it would suggest an autosomal dominant disease. In the cases where a pathological mutation has already been identified, the pedigree is used to plan further genetic screening or counselling. Figure 1 shows an example pedigree with a typical autosomal dominant inheritance pattern.

For some diseases, the course of events in the patient's family are important in judging the patient's own risk of serious events. In patients with hereditary hypertrophic cardiomyopathy, the European Society of Cardiology recommends using an online risk calculator to estimate a patient's 5 year risk of sudden cardiac death (SCD). Among the seven factors included in the underlying model – a strong contributor to individual risk – is a history of SCD in first degree relatives (Elliott et al., 2014).

Family histories occur as descriptive text in the EHR, but acknowledging that computational reasoning about family history have substantial benefits in research, diagnosis and decision support where many tools has been developed for interactive pedigree input (Welch et al., 2018). The underlying objective of our NLP challenge is to be able to infer the pedigree of a patient from text. However, even checking consistency of family history information represented in OWL proves to be a challenge (Stevens et al., 2014). A potential outcome of our work would be to transform statements about pedigree into tabular formats directly usable in risk calculators and for bioinformatics application like genome-wide analysis (Hiekkalinna et al., 2005).

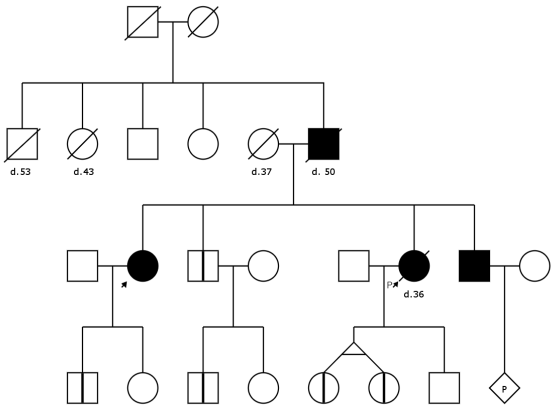

Figure 1: An example pedigree chart with a typical autosomal dominant inheritance pattern. Horizontal rows represent generations, lines represent relationships, lines of descent and sibship. Squares are male, circles female, and diamond shape is unknown gender. A symbol with a 'P' inside denotes a pregnancy. Diagonal lines through symbols denote deceased individuals and the text below their age at the time of death (eg. 'd. 43' means died when 43 years old). Filled symbols represent individuals with manifest disease, symbols with a vertical line are healthy gene carriers who may develop disease later. The small arrow denotes the current patient ("self") and the arrow with the 'P' is the proband or index patient where the genetic analysis of the family started (Bennett et al., 2008).

2.1 Previous work

There has been some previous work aimed at extracting family history information from clinical text. Bill et al. (2014) annotate 284 sentences from the publicly available MTSamples corpus of synthetically produced English clinical text for information about family members and clinical observations with some additional attributes (vital status, negation and age of death). However, they do not provide any measures of inter-annotator agreement. Polubriaginof et al. (2015) compared the information contained in structured and free-text descriptions of family history information and found that the free-text descriptions were more comprehensive.

In another work, Goryachev et al. (2008) developed a pipeline of rule based systems to detect family members and diagnosis concepts; and, then assign the family diagnosis to a specific family number. The authors run standard NLP tools such as sentence splitter and part-of-speech taggers on

discharge summary notes. The pipeline system is related to Friedlin and McDonald (2006) in only identifying diagnosis concepts that are present in standard medical dictionaries and do not perform relation extraction as performed in this paper.

Both rule based systems (Abacha and Zweigenbaum, 2011) and machine learning methods such as Roberts et al. (2008a) and Minard et al. (2011) use multi-class SVMs to perform relation extraction from clinical reports. Our work in this paper is closest to the work of Roberts et al. (2008a) who manually annotated cancer narratives for entities and relations and, then, trained and tested a one-vs-rest SVM classifier for training and testing. In this paper, we employ widely used features in general purpose named entity recognition (Hong, 2005; Miwa and Sasaki, 2014) to train the SVM models.

3 Incremental annotation guideline and synthetic corpus development

One immediate goal of this work is to develop a tool for the extraction of family history information from Norwegian clinical text. Due to the unavailability of the real health records describing family histories, we developed a methodology for annotation guideline development which makes use of an incrementally developed synthetic corpus. The textual data contained in the corpus was produced by a clinician who has extensive experience with clinical work and genetic cardiology. The data consists of statements that summarize the family history of a patient and will typically correspond to a small part of a patient journal. The descriptions were made by performing web searches for images of "autosomal dominant pedigree", and pseudo-randomly describing parts of the displayed pedigrees while assigning invented but realistic medical events. No real patient histories are reproduced, but coincidental similarities must be expected. The text does not contain any personal identification information.

The first step in a semantic annotation of text is to decide upon the entities and the relations that are interesting to extract or characterize. Biomedicine employs terminologies and classifications that may be used for annotation (Savova et al., 2010). In our domain of family history, we started with family members and relationships, and largely ignored medical conditions apart from death or known (cardiac) disease in general.

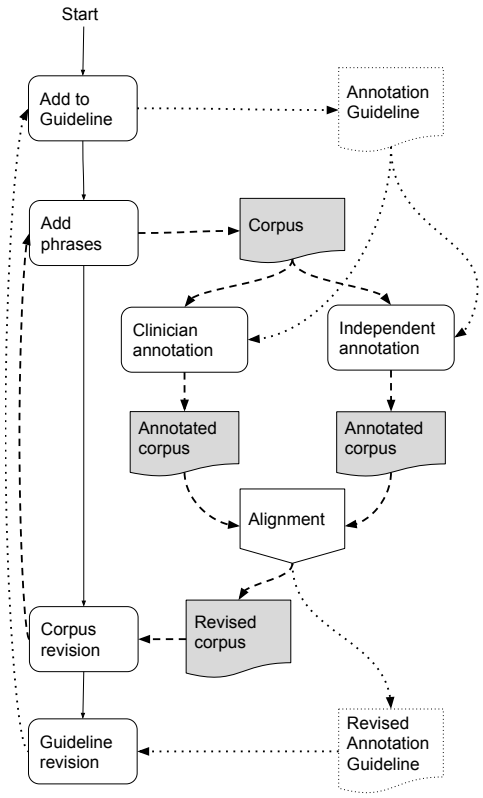

Figure 2: Incremental development of corpus and annotation guidelines

The guideline developers consisted of a clinician and three computational linguists and/or computer scientists. We usually maintained two roles: The clinician would produce a set of representative sentences and along with one of the others propose an annotation scheme for these. Then, the clinician would annotate while another independent person not involved in the design of the annotation scheme would make an *independent annotation*. The results were compared and discrepancies were recorded. We (sometimes artificially) could identify both *semantic* and *pragmatic* discrepancies. Semantic discrepancy would signify a misunderstanding of the underlying domain and required amending the ontology, whereas the pragmatic discrepancy would uncover an underspecified or incomplete annotation rule which could be further specified by adding more examples to the corpus. The drivers and amendments in this quiz-like game is shown in the table 1.

Figure 2 shows the double loops of corpus production and guideline development. As shown, the family history statements were produced iteratively. In the initial round, the clinician was

	Driver	Amendment
Guideline iteration	Semantic discrepancy	Add concept or revise guideline
Corpus Iteration	Pragmatic discrepancy	Add sentence to corpus

Table 1: Drivers and amendments guiding the development of annotation guidelines.

asked to produce a set of representative statements about SCD-related family history. Example 1 below shows a sentence from the corpus.

(1) *Indekspasienten er hans onkel på*
Index-patient is his uncle on
farssiden, som hatt hjertestans og
father's-side, who had cardiac-arrest and
fått implantert ICD.
had implanted ICD.
'The index patient is his uncle on the father's side, who had cardiac arrest and implanted ICD.

Following the initial iterations and discussions with the clinician the need to account for i) relations to groups of family members, ii) temporal statements, and iii) negation emerged. During this iteration the clinician was therefore tasked with the generation of statements that challenged the current guidelines, whilst still producing representative family statements. Example 2 shows an example sentence containing a temporal statement and example 3 shows another type of temporal statement describing the age of the family member at the time of diagnosis.

(2) *Han har kjent hjertebank de siste*
He has felt heart-palps the last
fire-fem månedene.
four-five months
'He has been feeling heart palpitations during the last four-five months'

(3) *Broren fikk diagnosen i femti-årene.*
Brother-the got diagnosis i fifty-years
'The brother was diagnosed in his fifties'

After arriving at a fairly stable set of guidelines, a large portion of the data set (320 sentences) was doubly annotated. Following this, disagreements were resolved in a round of consolidation between the annotators. The final portion of the data set (91 sentences) was then annotated doubly and the resulting inter-annotator agreement on these data sets is reported here in Section 4.5.

All annotation was performed using the Brat web-based annotation tool (Stenetorp et al., 2012). The data was manually segmented and tokenized prior to annotation.

4 Annotation guidelines

The following section presents an overview of the resulting annotation guidelines. The annotation of the corpus distinguishes semantically relevant clinical *entities* and shows how these relate to each other in the text via a set of *relations*. Figure 4 shows a graphical overview of the annotation schema, where rectangles indicate core clinical entities, ovals indicate modifier entities, and all possible relations are indicated by directed arcs.

4.1 Clinical entities

Clinical entities are marked with one of the following entity types:

- `Family` describes various family members (e.g. *onkelen* 'the uncle', *bestefar* 'grandfather').
- `Self` is used only for the patient under consideration (e.g. *pasienten* 'the patient', *hun* 'she').
- `Index` entities designate the property of being the index patient or *proband*, i.e. the first identified family member with disease *indekspasienten* 'the index patient'.
- `Condition` entities describe a range of clinical conditions such as diseases (*koronarsykdom* 'coronary disease'), diagnoses, various types of mutations, test results (*testet negativt* 'tested negative'), treatments (*hjertetransplantert* 'heart-transplanted'), and vital state (*død* 'dead', *frisk* 'healthy').
- `Event` entities describe clinical events (e.g. *hjertestans* 'cardiac arrest' and *synkope* 'syncope').

The distinction between conditions and events relate to the temporal extension of the entity described: an event is something that happens and then is over, but a condition is a prolonged state of the patient, for instance, the patient has a heart attack (`Event`), but from this point on she is considered to have heart disease (`Condition`).

In addition to the main clinical entities described above, the annotation guidelines also distinguish a set of modifier entities that further de-

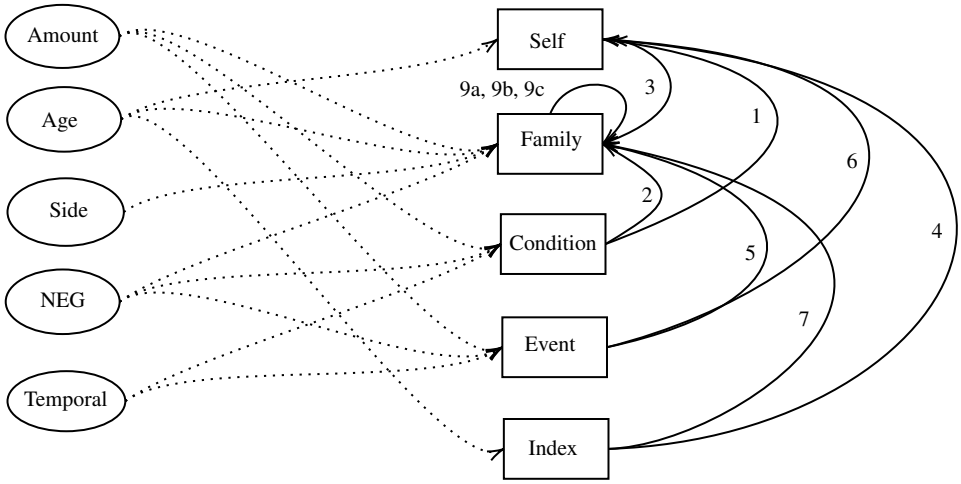

Figure 3: Schematic diagram showing the possible relations between entities. The different relations are marked with a number to avoid cluttering. `Holder`: 1, 2, 4, 5, 6, 7, 8; `Modifier`: Dotted lines; `Related_to`: 3, 9a; `Subset`: 9b; `Partner`: 9c.

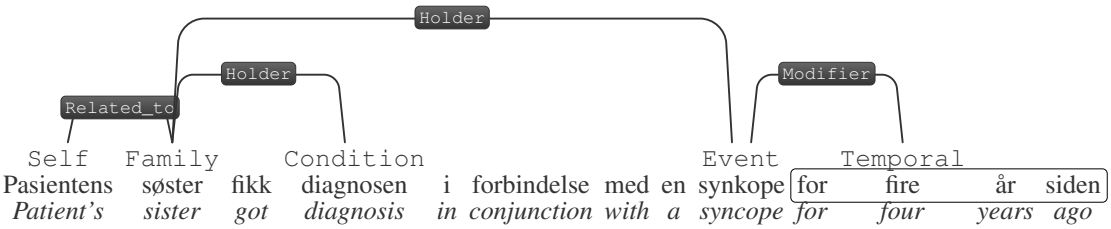

Figure 4: Annotation of clinical entities and relations for an example sentence from the corpus.

scribe the clinical entities for a number of properties that are relevant for semantic interpretation of family history information:

- `Side` entities describe the side of the family and thus modify `Family` entities (e.g. *farssiden* 'paternal side').
- `Age` entities describe the age of a family member *40 år gammel* '40 years old'.
- `Negation` entities mark lexical items that signal negation, so-called *negation cues* in the terminology of Morante and Daelemans (2012). These may be negative adverbs, such as e.g., *ikke* 'not', *aldri* 'never', or negative determiners/pronouns *ingen* 'nobody'. Note that in contrast to Morante and Daelemans (2012), we do not annotate morphological negation cues (e.g. **im**-*possible*). In this version of the guidelines, we treat negation as encompassing uncertainty. The main reason for this is that just like the presence of negation, it marks missing information in the family history.
- `Amount` modifiers describe quantifiers that describe numerical properties of clinical entities, e.g. *to* 'two', *mange* 'many'.
- `Temporal` modifiers typically position `Condition/Event` entities in time, e.g. *i sommer* 'this summer', *for tre år siden* 'three years ago'. These are similar to temporal expressions (so-called *timexes*) in previous temporal annotation schemes (Ferro et al., 2002; Saurí et al., 2006).

4.2 Family history relations

In addition to the clinical entities described above, we further annotate a number of relationships between entities in our annotation scheme. Example 4 shows a fully annotated example containing entities and their relations for an sentence from the corpus. The relations are binary undirected relations of the following types:

- `Holder` relations are always between `Condition/Event` entity on the one hand and its holder, a `Family/Self/Index` entity.

- `Modifier` relations hold between modifier entities (e.g. `Side, Negation`) and clinical entities (e.g. `Family, Condition`).
- `Related_to` relations specify relations between family members and always hold between entities of the `Family` type.
- `Subset` relations specify relations between family members, where one is a subset of the other, e.g. in statements such as *Hun har to brødre, den ene har mutasjonen* 'She has two brothers, one of them has the mutation', where *den ene* 'one of them' would be connected to the `Family` entity *brødre* 'brothers' with a `Subset`-relation.
- `Partner` relations specify relations between entities of the `Family` type, used to identify couples (husbands and wives, civil partnerships) that are able to provide offspring. The assumption is no kinship.

4.3 Span of annotations

In general, annotation should pick out the minimal span in the text which denotes the entity or property in question. This will most often be a single word (*onkel* 'uncle', *mutasjon* 'mutation') but will in some cases also include more than one word (*plutselig hjertedød* 'sudden cardiac death', *voksende hjerte* 'growing heart'). Genitive modifiers of an entity, e.g. *farens* 'father's' in *farens søster* 'the father's sister' or *Søsteren til faren* 'the sister of the father' should not be included in the annotation span. Rather, these are annotated as two separate entities related by a `Related_to` relation. The span of `Family` entities usually encompass only the family term itself (*onkel* 'uncle', *søster* 'sister'), however, when the family term is described using a pronominal element (*hun* 'she', *den ene* 'one (of them)') this should be annotated as a family entity. When both are present (*den ene broren* 'the one brother') only the family term is annotated.

Temporal expressions will often be more complex and should include both numerical expressions denoting amount (*tre* 'three', *flere* 'several'), temporal units such as month/year, as well as expressions denoting temporal ordering or duration (*i* 'in', *siden* 'since' as in *tre år siden* 'three years since', *i tre år* 'for three years'). Initial iterations of annotation showed that agreement for this category was low due to differences in annotation span. We therefore introduced the generalization that temporal annotation should make use of a replacement rule where the full constituent replaced by a temporal pronoun corresponding to English *then* is annotated. This means that unlike e.g. Ferro et al. (2002), our temporal annotations will include prepositions (e.g. *i tre år* 'for three years').

Entities	Number	Spans
Family	1704	96
Condition	681	135
Event	542	115
Self	509	–
Amount	273	9
Temporal	214	178
Negation	131	33
Age	57	34
Side	36	3
Index	7	–
Relations	**Number**	**Spans**
Holder	880	–
Modifier	687	–
Related_to	389	–
Subset	108	–
Partner	14	–

Table 2: Distribution of entities and relations in the data annotated by the clinician. The Spans column shows the number of entities that span across words. Both the entities and relations are sorted in decreasing order of number of occurrences.

4.4 Statistics

The resulting annotated corpus contains 477 sentences and 6030 tokens. In table 2 we present the distribution of the entities and relations in the corpus. We see that `Condition` and `Event` entities are fairly equally distributed in the corpus. Temporal modifiers span more than one word in a majority of cases. Whereas `Holder`-relations are the most common type of relation in the corpus, there are only 14 cases of the `Partner` relation.

4.5 Inter-Annotator Agreement

As described in Section 3, two final rounds of annotation with different second annotators (in addition to the clinician, here dubbed A1 and A2) were used to complete the annotation guidelines. We measured the inter-annotator agreement at two levels. At the first level, IAA is based on match of the entities spans and their labels. At the second

level, IAA is based on the relationship matches between the matched spans. Therefore, the relationship agreement measurement is stricter than the entity level agreement measurement. We examine token level agreement where we treat the clinician's notes as gold standard and compute the per token F-measures i.e., Precision, Recall, and F_1-score. We measure the inter-annotator agreement using micro F_1-score. The Precision, Recall, and F_1-scores of the agreement is provided in table 3.

Annotator	Precision	Recall	F_1-score
A1, 320	0.743	0.648	0.692
	0.645	0.559	0.599
A2, 91	0.821	0.797	0.809
	0.752	0.678	0.713

Table 3: Each row shows the number of sentences annotated by each annotator. The first and second rows shows the Precision, Recall, and F_1-score for entities and relations. All the results are in comparison to the texts annotated by the clinician.

We find that the round of consolidation and improvement of the guidelines was useful and improves the IAA scores for both entities and relations. When we compare the annotations of the clinician (A0) and the second additional annotator (A2), we find that there are still a number of remaining discrepancies. Some of these are what we termed semantic discrepancies above in Section 3 above, annotation decisions that require domain knowledge. For instance, in several places A2 annotates clinical conditions that are not marked by the clinician, e.g. marking *symptomer* 'symptoms' as a `Condition`. There are also examples where additional distinctions should probably be added to the guidelines, in particular with respect to annotation of temporal and negation-related information, both examples of complex annotation tasks by themselves. For instance, A2 annotates the phrase *under en flytur til Spania* 'during a flight to Spain' as Temporal, where A0 does not. With respect to negation, the distinction between negation and uncertainty causes differences in annotation spans, where A0 annotates *husker ikke* 'does not remember' as `NEG`, whereas A2 annotates only *ikke* 'not'.

5 Preliminary experiments

In this section, we perform entity classification and relation extraction experiments to verify the viability of our annotation. We train and test SVM model on the data annotated by the domain expert in five-fold cross-validation fashion. The domain expert annotated dataset has 477 sentences and we performed five-fold cross-validation to train and test our model. In all our experiments, we split the sentences into five folds and extracted entities and relations. Then, we treated each of five folds as test dataset and trained on the other four folds in an iterative fashion.

5.1 Entity detection

In this experiment, we trained and tested a linear classifier (SVM model) for entity classification. We treat entity classification as a multi-class classification problem where there are 11 classes including the `O` entity that denotes unmarked lexical units. Our model is a linear SVM model that is trained on the following features:

- Lexical: Current word, words in a context window size of 2.
- Universal POS tags: Current word, words in a context window size of 2.
- Entity tags: The two previous entity tags where the model uses the gold entity tags to train but uses the previous predicted entity tags to predict the current tag.

We also experimented with lowercasing a word and orthographic features such as prefixes and suffixes of length 3 which did not improve the performance of the SVM model. We evaluate the performance of the SVM model using weighted F_1 score to account for class imbalance. On an average, these feature templates yielded 5000 features across the five cross-validation experiments. All the Universal POS tags are obtained through the CoNLL17 Baseline model (Zeman et al., 2017) trained on the publicly available Universal Dependencies Norwegian Bokmål treebank (Øvrelid and Hohle, 2016). We used the majority class "O" as the baseline in our experiments. The results of our experiments are given in table 4. It has to be noted that these results are not comparable to the IAA scores presented in table 3, which are calculated only over entities and completely disregard the remaining tokens. Moreover, the IAA

System	Precision	Recall	F_1-score
Baseline	0.34	0.582	0.429
SVM	0.843	0.843	0.841

Table 4: The average of the weighted F_1-scores across the five folds. On an average, there are 6030 training instances and 1507 test instances.

score is computed only on parts of the annotated data whereas the SVM models are trained and tested on the whole of the data annotated by A0. The SVM model performs better than the majority class baseline model across all the measures. The SVM model made errors at distinguishing Condition entities from Event entities and Age from Temporal entities. Most of the errors occurred when the SVM model misclassified the rest of the classes as "O".

5.2 Relation extraction

In this subsection, we performed a relation detection and classification experiment. In this experiment, we treat a relation defined between exactly two entities to belong to one of the six relations where five of them are given in table 2 and the sixth relation is "No_Relation". We train and test an SVM model in a five-fold cross-validation fashion. Apart from entity labels, we experimented with increasingly complex set of features:

- Lexical: Words belonging to the entities are treated as two separate features.
- POS tags: Universal POS tags of the entities' lexical tokens as separate features.
- Dependency features: The dependency label of a entity word's incoming arc as a feature.

If a entity is spanning across multiple words, we concatenate the per-word feature and treated them as a single feature when training and testing the SVM model. The results of the experiments are given in table 5. Our results suggest that word based features themselves yield a performance which is close to the model with more complex features. Incremental inclusion of POS tags and dependency labels increases the performance of the SVM model, whereas the inclusion of predicted entity labels does not improve the performance of the SVM model. We experimented if including the gold standard labels would improve the performance of the SVM model. We find that the quality of entity labels does improve the performance of the model.

Finally, we present the confusion matrix for the best fold is presented in table 6. The SVM model makes most of the errors when it misclassifies one of the five annotated relations as "No_Relation" and vice-versa. The classifier errs when distinguishing between "Related_to", "Partner" and "Subset" relations. Finally, the classifier makes errors when distinguishing between the Norwegian indefinite determiner *en* which is unmarked and the quantifier *en*.

Features	Precision	Recall	F_1-score
Words	0.716	0.732	0.719
+POS tags	0.73	0.738	0.731
+Dependency labels	0.743	0.746	0.743
+Entity labels (Predicted)	0.743	0.745	0.743
+Entity labels (Gold)	0.771	0.767	0.768

Table 5: Average of the weighted F_1-scores on five fold cross-validation. On an average, there are 5530 training instances and 1461 test instances.

	Holder	Modifier	No_Relation	Partner	Related_to	Subset
Holder	127	1	54	0	1	0
Modifier	2	82	66	0	0	0
No_Relation	65	100	1045	1	30	15
Partner	0	0	0	2	3	0
Related_to	7	0	22	0	58	2
Subset	0	0	11	0	4	13

Table 6: Confusion matrix for the SVM model at the task of relation extraction on the best performing fold.

6 Discussion

The validity of our study is limited by using synthetic data. While the clinician producing the clinical text works in genetic cardiology, and writes similar patient histories in his clinical practice, the synthetic data can not be expected to be fully representative of real clinical notes from a large patient cohort. The analysis pertaining to the synthetic data should be thought of as an illustration of one iteration of the cycle described in 2, and the objective of the iterative process is a stepwise, guided, design of an annotation guideline in a setting where the target text data is unavailable. The same process could be used with a real corpus, where specific new examples would present challenges driving guideline development. The only difference is that in our case, a specialist produced text, instead of finding representative text.

The guideline development workflow itself may also be improved or expanded by storing a representation of the input data (the pedigree) and linking it to the resulting synthetic text description, which would allow further downstream comparison of extraction results to the actual source material.

7 Conclusion

This article has investigated the development of annotation guidelines for family history information in Norwegian by leveraging synthetically produced clinical text. Inter-annotator agreement scores show that the annotation schema can be applied fairly consistently and that it may also be generalized to unseen text using machine learning. Both the annotation guidelines and the annotated corpus will be made freely available and as such constitutes the first freely available resource of Norwegian clinical text.

In the near future, we will apply the annotation schema to real clinical texts. The family history is but a minor part of a patient record, and segmentation as shown in Bill et al. (2014) is needed. Analysis of the annotation disagreements along with the experimental results also highlighted part of the schema that will need to be further refined, e.g. the analysis of temporality and our treatment of uncertainty. We will develop the method for incremental and systematic annotation guideline development further. The method will be put to test when we iteratively improve the current guideline in order to capture real patient pedigree information from the EHR.

Acknowledgments

We are grateful to three anonymous reviewers for constructive comments on the first version of the paper. This work is funded by the Norwegian Research Council and more specifically by the BigMed project, an IKTPLUSS Lighthouse project.

References

Asma Ben Abacha and Pierre Zweigenbaum. 2011. Automatic extraction of semantic relations between medical entities: a rule based approach. *Journal of biomedical semantics*, 2(5):S4.

Helen Allvin, Elin Carlsson, Hercules Dalianis, Riitta Danielsson-Ojala, Vidas Daudaravičius, Martin Hassel, Dimitrios Kokkinakis, Heljä Lundgren-Laine, Gunnar Nilsson, Øystein Nytrø, et al. 2010. Characteristics and analysis of finnish and swedish clinical intensive care nursing narratives. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, pages 53–60. Association for Computational Linguistics.

Robin L Bennett, Kathryn Steinhaus French, Robert G Resta, and Debra Lochner Doyle. 2008. Standardized human pedigree nomenclature: update and assessment of the recommendations of the national society of genetic counselors. *Journal of genetic counseling*, 17(5):424–433.

Robert Bill, Serguei Pakhomov, Elizabeth S Chen, Tamara J Winden, Elizabeth W Carter, and Genevieve B Melton. 2014. Automated extraction of family history information from clinical notes. In *AMIA Annual Symposium Proceedings*, volume 2014, page 1709. American Medical Informatics Association.

Willie Boag, Tristan Naumann, and Peter Szolovits. 2018. Towards the creation of a large corpus of synthetically-identified clinical notes. *CoRR*, abs/1803.02728.

Hercules Dalianis, Martin Hassel, Aron Henriksson, and Maria Skeppstedt. 2012. Stockholm EPR Corpus: A Clinical Database Used to Improve Health Care. In *Proceedings of the Fourth Swedish Language Technology Conference*, pages 17–18.

Perry M Elliott, Aris Anastasakis, Michael A Borger, Martin Borggrefe, Franco Cecchi, Philippe Charron, Albert Alain Hagege, Antoine Lafont, Giuseppe Limongelli, Heiko Mahrholdt, William J McKenna, Jens Mogensen, Petros Nihoyannopoulos, Stefano Nistri, Petronella G Pieper, Burkert Pieske, Claudio Rapezzi, Frans H Rutten, Christoph Tillmanns, Hugh Watkins, Additional Contributor, Constantinos O'Mahony, ESC Committee for Practice Guidelines (CPG), Jose Luis Zamorano, Stephan Achenbach, Helmut Baumgartner, Jeroen J Bax, Héctor Bueno, Veronica Dean, Christi Deaton, Çetin Erol, Robert Fagard, Roberto Ferrari, David Hasdai, Arno W Hoes, Paulus Kirchhof, Juhani Knuuti, Philippe Kolh, Patrizio Lancellotti, Ales Linhart, Petros Nihoyannopoulos, Massimo F Piepoli, Piotr Ponikowski, Per Anton Sirnes, Juan Luis Tamargo, Michal Tendera, Adam Torbicki, William Wijns, Stephan Windecker, Document Reviewers, David Hasdai, Piotr Ponikowski, Stephan Achenbach, Fernando Alfonso, Cristina Basso, Nuno Miguel Cardim, Juan Ramón Gimeno, Stephane Heymans, Per Johan Holm, Andre Keren, Paulus Kirchhof, Philippe Kolh, Christos Lionis, Claudio Muneretto, Silvia Priori, Maria Jesus Salvador, Christian Wolpert, Jose Luis Zamorano, Matthias Frick, Farid Aliyev, Svetlana Komissarova, Georges Mairesse, Elnur Smajić, Vasil Velchev, Loizos Antoniades,

Ales Linhart, Henning Bundgaard, Tiina Heliö, Antoine Leenhardt, Hugo A Katus, George Efthymiadis, Róbert Sepp, Gunnar Thor Gunnarsson, Shemy Carasso, Alina Kerimkulova, Ginta Kamzola, Hady Skouri, Ghada Eldirsi, Ausra Kavoliuniene, Tiziana Felice, Michelle Michels, Kristina Hermann Haugaa, Radosław Lenarczyk, Dulce Brito, Eduard Apetrei, Leo Bokheria, Dragan Lovic, Robert Hatala, Pablo Garcia Pavía, Maria Eriksson, Stéphane Noble, Elizabeta Srbinovska, Murat Özdemir, Elena Nesukay, and Neha Sekhri. 2014. 2014 esc guidelines on diagnosis and management of hypertrophic cardiomyopathy: the task force for the diagnosis and management of hypertrophic cardiomyopathy of the european society of cardiology (esc). *European heart journal*, 35(39).

L. Ferro, L. Gerber, I. Mani, B. Sundheim, and G. Wilson. 2002. Instruction manual for the annotation of temporal expressions. Technical report, MITRE, Washington C3 Cen- ter, McLean, Virginia.

Jeff Friedlin and Clement J McDonald. 2006. Using a natural language processing system to extract and code family history data from admission reports. In *AMIA Annual Symposium Proceedings*, volume 2006, page 925. American Medical Informatics Association.

Sergey Goryachev, Hyeoneui Kim, and Qing Zeng-Treitler. 2008. Identification and extraction of family history information from clinical reports. In *AMIA Annual Symposium Proceedings*, volume 2008, page 247. American Medical Informatics Association.

Tero Hiekkalinna, Joseph D. Terwilliger, Sampo Sammalisto, Leena Peltonen, and Markus Perola. 2005. Autogscan: Powerful tools for automated genome-wide linkage and linkage disequilibrium analysis. *Twin Research and Human Genetics*, 8(1):16–21.

Gumwon Hong. 2005. Relation extraction using support vector machine. In *International Conference on Natural Language Processing*, pages 366–377. Springer.

Christina Lohr, Sven Buechel, and Udo Hahn. 2018. Sharing copies of synthetic clinical corpora without physical distribution — a case study to get around IPRs and privacy constraints featuring the German JSYNCC corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 1259 – 1266, Miyazaki, Japan.

Anne-Lyse Minard, Anne-Laure Ligozat, and Brigitte Grau. 2011. Multi-class svm for relation extraction from clinical reports. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 604–609.

Makoto Miwa and Yutaka Sasaki. 2014. Modeling joint entity and relation extraction with table representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1858–1869.

Roser Morante and Walter Daelemans. 2012. Conandoyle-neg: Annotation of negation cues and their scope in conan doyle stories. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. European Language Resources Association (ELRA).

Lilja Øvrelid and Petter Hohle. 2016. Universal Dependencies for Norwegian. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Fernanda Polubriaginof, Nicholas P Tatonetti, and David K Vawdrey. 2015. An assessment of family history information captured in an electronic health record. In *AMIA Annual Symposium Proceedings*, volume 2015, page 2035. American Medical Informatics Association.

Angus Roberts, Robert Gaizauskas, and Mark Hepple. 2008a. Extracting clinical relationships from patient narratives. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 10–18. Association for Computational Linguistics.

Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Andrea Setzer, and Ian Roberts. 2008b. Semantic annotation of clinical text: The clef corpus. In *Proceedings of the LREC 2008 workshop on building and evaluating resources for biomedical text mining*, pages 19–26.

Thomas Brox Røst, Ola Huseth, Øystein Nytrø, and Anders Grimsmo. 2008. Lessons from developing an annotated corpus of patient histories. *JCSE*, 2(2):162–179.

R. Saurí, J. Littman, B. Knippen, R. Gaizauskas, A. Setzer, and J. Pustejovsky. 2006. TimeML annotation guidelines version 1.2. 1. Technical report, LDC.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, pages 102 – 107, Avignon, France.

Robert Stevens, Nicolas Matentzoglu, Uli Sattler, and Margaret Stevens. 2014. A family history knowledge base in OWL 2. In *Informal Proceedings of the 3rd International Workshop on OWL Reasoner Evaluation (ORE 2014) co-located with the Vienna Summer of Logic (VSL 2014), Vienna, Austria, July 13,*

2014., volume 1207 of *CEUR Workshop Proceedings*, pages 71–76. CEUR-WS.org.

Özlem Uzuner and Amber Stubbs. 2015. Practical applications for natural language processing in clinical research: The 2014 i2b2/uthealth shared tasks. *Journal of biomedical informatics*, 58(Suppl):S1.

Brandon M. Welch, Kevin Wiley, Lance Pflieger, Rosaline Achiangia, Karen Baker, Chanita Hughes-Halbert, Heath Morrison, Joshua Schiffman, and Megan Doerr. 2018. Review and comparison of electronic patient-facing family health history tools. *Journal of Genetic Counseling*, 27(2):381–391.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Václava Kettnerová, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria dePaiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, Canada.

A Supplemental material

The annotated data and the code used in this paper is available at: https://github.com/ltgoslo/NorSynthClinical.