Benchmark for Evaluation of Danish Clinical Word Embeddings

Martin S. Laursen[*], University of Southern Denmark, Odense, Denmark msla@mmmi.sdu.dk
Jannik S. Pedersen[*], University of Southern Denmark, Odense, Denmark jasp@mmmi.sdu.dk
Pernille Just Vinholt, Odense University Hospital, Denmark pernille.vinholt@rsyd.dk
Rasmus Søgaard Hansen, Odense University Hospital, Denmark rasmus.sogaard.hansen@rsyd.dk
Thiusius Rajeeth Savarimuthu, University of Southern Denmark, Odense, Denmark trs@mmmi.sdu.dk

Abstract In natural language processing, benchmarks are used to track progress and identify useful models. Currently, no benchmark for Danish clinical word embeddings exists. This paper describes the development of a Danish benchmark for clinical word embeddings. The clinical benchmark consists of ten datasets: eight intrinsic and two extrinsic. Moreover, we evaluate word embeddings trained on text from the clinical domain, general practitioner domain and general domain on the established benchmark. All the intrinsic tasks of the benchmark are publicly available[1].

1 Introduction

Word embeddings are real-valued vectors that are trained to represent words based on the context in which they appear. Based on the distributional hypothesis (Harris, 1954), which suggests that words with similar contexts have similar meaning, embeddings of semantically similar words are expected to appear close to each other in vector space.

Since their introduction, word embeddings have been ubiquitous in natural language processing (NLP) due to their ability to represent word meaning. Typically, word embeddings are trained on a general text corpus such as Wikipedia. Afterwards, word embeddings are used as stand-alone features or as input to neural networks to perform a wide variety of NLP tasks such as text classification, named entity recognition (NER) and machine translation.

In specialized domains, such as the clinical, word embeddings are also widely used to e.g. extract information from electronic health records (EHRs). However, the text in clinical EHRs differs significantly from the general domain. Clinical EHRs include rare words, domain specific abbreviations and a mix of languages (for example Latin, English and Danish). The text is often non-narrative and very concise, free of syntactic rules, sometimes consisting of a sequence of keywords. Moreover, it contains many spelling errors, and the semantic meaning of words can differ from that of the general domain (Leaman et al., 2015). In the clinical domain, word embeddings are, therefore, often trained on an in-domain corpus to better capture the vocabulary and the semantic meaning of words. After being trained on an in-domain corpus, they are used for e.g. clinical NER, International Classification of Diseases coding, clinical event detection, de-identification and patient similarity estimation with improved performance over general word embeddings (Zhao et al., 2018; Wang et al., 2018; Chen et al., 2019).

For evaluating word embeddings, two different methods are typically used: intrinsic and extrinsic evaluation (Wang et al., 2019c). In intrinsic evaluation, word embeddings are evaluated based on their inherent information, e.g. by exploring the syntactic or semantic relationship between words. In extrinsic evaluation, word embeddings are evaluated based on their ability to solve a downstream task, e.g. by using them as input to a neural network. While word embeddings can be evaluated using extrinsic benchmarks by holding the network architecture fixed while varying the set of word embeddings, intrinsic benchmarks provide an intermediate evaluation of the embeddings' properties before being used as input to a larger system. This supports the need for intrinsic evaluation.

Word embeddings for the general domain are publicly available in many languages (Grave et al., 2018). However, publicly available embeddings for the clinical domain are scarce (Khattak et al., 2019). This is

[*] Both authors contributed equally to this paper.
[1] www.github.com/jannikskytt/DaClinWordEmbeddings

most likely due to strict regulations around clinical data which contain sensitive information making them unsuitable for sharing. Therefore, researchers in clinical NLP are often forced to create their own word embeddings in order not to expose sensitive information (Abdalla et al., 2020).

Clinical intrinsic benchmark datasets do not necessarily contain sensitive information and can, in that case, be shared openly, benefitting researchers producing clinical word embeddings. For the English language, both intrinsic and extrinsic benchmarks exist, e.g. University of Minnesota Medical Residents Similarity / Relatedness Set (UMNSRS) (Pakhomov et al., 2010) for word similarity and relatedness, and BLUE (Peng et al., 2019), which includes both clinical and biomedical datasets, for extrinsic evaluation. For Danish, though, no clinical benchmark exists.

In this paper, we introduce a clinical word embedding benchmark for the Danish language. Moreover, we produce clinical word embeddings and use the benchmark to compare them to embeddings trained on the general domain and embeddings trained on the general practitioner (GP) domain.

The benchmark is specifically constructed to evaluate static word embeddings such as GloVe (Pennington et al., 2014), Continuous Bag-of-Words (Mikolov et al., 2013a), Skip-gram (Mikolov et al., 2013a) and FastText (Bojanowski et al., 2017). It is therefore not suitable for evaluation of contextual word embeddings produced by transformer models like BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) and ELECTRA (Clark et al., 2019).

Although transformer models achieve state-of-the-art results (Wang et al., 2019b; Wang et al., 2019a), static word embeddings are still useful as input to NLP pipelines. Some advantages are that they require less compute to train and at inference time, and they work better on limited data (Peng et al., 2021). This is relevant for research within specialized domains, such as clinical NLP, where researchers must often train their own word embeddings on limited data and the hardware to train and run a transformer model is not necessarily available. Static word embeddings are also relevant in time-critical tasks in clinical practice such as expanding single-word searches in the EHR using the nearest neighbors of the search term. Expanding single-word searches is especially relevant in the clinical domain where many different terms can be used about the same basic symptom or disease. Another advantage is their ease of use for medical doctors (MDs) and clinical researchers who are not machine learning scientists compared to contextual word embeddings.

The remainder of this paper first introduces the benchmark including the methods for creating each intrinsic and extrinsic dataset. It then describes the training methods of the produced Danish clinical word embeddings and those from the general and GP domains which they will be benchmarked against. Finally, the benchmark results are presented and discussed.

2 Establishing Benchmark

The benchmark consists of an intrinsic and extrinsic part. In this paper, intrinsic performance is evaluated based on the quality of the semantic and syntactic inherent information using analogy and similarity tasks. We produce datasets for three different intrinsic evaluation methods: analogy tasks, similarity and relatedness tasks, and an equality task.

Analogy tasks, introduced by Mikolov et al. (2013b), take tuples of four words (A, B, C and D) and evaluate 'what is to C as B is to A' by selecting the nearest neighbour to the calculated vector in the embedding space, excluding the words forming the analogy:

$$\overrightarrow{C} + \overrightarrow{B} - \overrightarrow{A} = \overrightarrow{D}$$

If the nearest neighbor to the calculated vector is D, the analogy task is correct. The task is evaluated on the percentage of correct predictions in the dataset.

Similar to Pennington et al. (2014), the similarity tasks take a tuple of two words and their similarity score, in our case, produced by one or more MDs. The similarity score of a word pair is compared to the cosine similarity of the pair's word embeddings. The cosine similarity is calculated as:

$$similarity(\overrightarrow{v}, \overrightarrow{u}) = \frac{\overrightarrow{v} \cdot \overrightarrow{u}}{\|\overrightarrow{v}\| \cdot \|\overrightarrow{u}\|}$$

The correlation between MD scores and cosine similarities for the dataset of word pairs is evaluated using the Spearman's rank correlation coefficient. Relatedness tasks are identical to similarity tasks except the MDs produce a relatedness score instead of a similarity score. Relatedness refers to one word calling to mind another word (e.g., needle–thread), while similarity reflects the degree of semantic feature overlap between words (e.g., whale–dolphin) (Pakhomov et al., 2010).

Equality tasks take a tuple of two terms with the exact same meaning. As the similarity score of a pair is 1 for a perfect match, the objective is maximization of the cosine similarity between terms. The task is evaluated as the mean of the cosine similarities for all pairs in the dataset.

The extrinsic part consists of two different text classification tasks in the clinical domain with the word embeddings as input. The quality of the word embeddings is evaluated based on the evaluation metric of the classification task.

An overview of all datasets can be seen in Table 1.

Task	Description	Example
Intrinsic tasks		
Clinical analogy	Evaluate "what is to C as B is to A"	$\overrightarrow{joint} + \overrightarrow{colonoscopy} - \overrightarrow{colon} = \overrightarrow{arthroscopy}$
Clinical similarity UMNSRS similarity UMNSRS relatedness	Compare the human similarity/relatedness score of a word pair to the cosine similarity of the pair's word embeddings	uterus, cervix
Clinical abbreviation equality	Compare the similarity of a word and its abbreviation	cm, centimeter
Verb, adjective, and noun inflection analogy datasets	Evaluate "what is to C as B is to A"	$\overrightarrow{adjusted} + \overrightarrow{remove} - \overrightarrow{removed} = \overrightarrow{adjust}$
Extrinsic tasks		
Bleeding classification	Classify a paragraph as either positive or negative for bleeding	15-year-old girl hospitalized with bleeding tendency and anemia symptoms
Hospital department classification	Classify a paragraph into one of six hospital departments	Clinical contact. Prepared by clinic. Conclusion and plan: As agreed and as a follow-up to the note on 10.2.99, I have contacted pt. However, pt. is hospitalized due to …

Table 1: Overview of the datasets used in the benchmark. Examples are translated to English.

2.1 Intrinsic Datasets

The intrinsic part consists of the following semantic tasks: clinical analogy, clinical similarity, clinical abbreviation equality, and UMNSRS similarity and relatedness; and the following syntactic tasks: verb inflection analogy, adjective inflection analogy, and noun inflection analogy. The intrinsic syntactic tasks are evaluating the syntactic properties of word embeddings in general rather than specifically for clinical use cases. As good clinical word embeddings must also contain syntactic information, the syntactic tasks are constructed to specifically evaluate the inherent syntactiv information on words from the clinical domain.

The development of each intrinsic task consisted of 1) selecting the terms to use for the task and 2) creating the evaluation dataset. This is described for each task below. All intrinsic datasets are supplied in the supplementary material.

2.1.1 Clinical Analogy Dataset

Two MDs, in agreement, created 41 distinct clinical analogies such as (translated from Danish)

$$\overrightarrow{colonoscopy} - \overrightarrow{colon} = \overrightarrow{arthroscopy} - \overrightarrow{joint}$$

where the word pairs on each side of the equation have the same one-to-one relationship. For the example above with the one-to-one relationship 'is telescopic examination of', it means that colonoscopy is a telescopic examination of only the colon, and that the colon has only one telescopic examination: a colonoscopy. Some other common relationships were 'treats', 'is indicator for', 'is disease in anatomy', 'is test for', 'is examination of', 'leads to' and 'is symptom of'. We relaxed the one-to-one relationship condition in a few cases: if for example a symptom is predominant for one disease but also minorly associated with another, we accepted the word pair. We augmented each distinct analogy to form four analogies by changing the order of the words inside the word pairs and by changing the order of the word pairs. This means that, for the analogy example above, we predicted each of 'colonoscopy', 'colon', 'arthroscopy', and 'joint' from the remaining three words. We performed this augmentation because the analogy tasks are based on evaluating the nearest neighbour to the calculated vector. Since the surrounding embedding space for each of the four calculated vectors may vary in distance to neighbours, the result may vary depending on which of the four words is predicted.

The clinical analogy dataset consists of 164 analogies.

2.1.2 Clinical Similarity Dataset

For the clinical similarity dataset, we predefined the following goals for achieving a diverse set of word pairs:

1. The selected words should be of different categories, e.g. they should not all be diseases.
2. The selected words should appear with varying frequency in clinical EHRs.
3. Word pairs should be matched within and across the categories and frequencies.
4. Words should not be selected based on an existing clinical EHR database because it could introduce bias to the dataset, e.g. the frequency of

words in our clinical EHR database might differ from other databases.

To achieve this, we predefined five clinical categories: anatomy, symptom/finding, disease, treatment, and diagnostic; and three frequency categories indicating how frequently a word appears in clinical EHRs: infrequent, occasional, and frequent. Then, two MDs selected words from a reference work on internal medicine (Schaffalitzky de Muckadell et al., 2009) by turning to approximately every fifth page, randomly selecting words, and subjectively assigning them categories until all three frequency categories per five clinical categories had 36 words each. This generated a total of 108 words per clinical category and 540 words overall.

We defined 270 word pairs by pairing 36 words from each clinical category with 36 words from the same category and 36 words evenly distributed on the four other clinical categories. We opted to use more words per group for intra-category-pairs than inter-category-pairs because we expected it would decrease the overrepresentation of pairs with low similarity. The pairings were distributed evenly across frequency categories. Finally, to further decrease overrepresentation of pairs with low similarity, the MDs subjectively defined 19 extra pairs with high similarity by pairing any two words from the word pool, resulting in a total of 289 word pairs.

Ten MDs with 2 to 17 years (mean: 7.5 years) of clinical experience used between 17 and 45 minutes (mean: 30.5 minutes) to rate the 289 pairs. Nine MDs had clinical biochemistry as speciality and one had pathology. The pairs were rated for similarity on a scale from 0 to 6 with 0 being lowest similarity and 6 being highest similarity. It was emphasized that the MDs should rate for similarity and not relatedness. If a word pair was unknown to the MDs, they did not rate it. One pair was rated by eight MDs and the rest were rated by at least nine. The similarity score for each pair is the mean rating. The mean ratings span from 0 to 6 with a minimum similarity score of 0.3, a mean of 1.1, and a maximum of 5.4. The standard deviations range from 0.3 to 1.6 with a mean of 0.7.

2.1.3 Clinical Abbreviation Equality Dataset

A list of 319 clinical abbreviations and their corresponding words was collected from online sources (supplementary material). Only abbreviations of single words were collected to simplify the evaluation of word embeddings, which usually represent single words. Ambiguous abbreviations and the abbreviations deemed unlikely to appear in clinical EHRs by an MD were removed. For example, the abbreviation 'all' is ambiguous because it could both mean 'allergy' or 'acute lymphocytic leukemia'. The final dataset comprises 195 abbreviation–word pairs with the same meaning.

2.1.4 UMNSRS Similarity and Relatedness Datasets

The UMNSRS consists of 566 English term pairs rated for semantic similarity and 587 for semantic relatedness on a continuous scale from 0 to 1600. One MD translated the datasets into Danish. Pairs consisting of a term that translates into a multi-word expression were removed. As were terms that do not exist in Danish, for example a non-traded drug. In cases where a Danish counterpart drug exists, for example 'betalaktam' for 'cefoxitin', this term was used as a translation. The Danish translation of the UMNSRS consists of 528 similarity pairs and 557 relatedness pairs.

2.1.5 Verb Inflection Analogy Dataset

A list of all verbs was extracted from the Danish orthographic dictionary (Danish Language Council, 2012). One MD selected verbs from the list that were deemed would occasionally or frequently occur in a clinical EHR. Next, verbs were conjugated in the following inflections: infinitive, present/future (same form in Danish), past tense, and present/past perfect. If a verb did not exist in all four inflections or had the same form in multiple inflections, it was removed from the list as it would cause analogy tasks involving the zero-vector. The final list contained 92 words, each in four inflections.

For each verb, six types of inflection pairs were made, for example infinitive–past, by pairing each inflection with the three other inflections. Next, we randomly combined each verb with 20 other verbs, evenly distributed on types of inflection pairs except for the remainder after equal division. This produced 1,840 analogies like the following of type infinitive–past (translated from Danish):

$$\overrightarrow{remove} - \overrightarrow{removed} = \overrightarrow{adjust} - \overrightarrow{adjusted}$$

2.1.6 Adjective Inflection Analogy Dataset

The same method as described for the verb inflection analogy dataset was used to develop the adjective inflection analogy dataset. Adjectives were declined in the following inflections: common positive, neuter positive, plural positive, comparative and superlative. The final list contained 43 words, each in five inflections.

For each adjective, we made seven types of inflection pairs by pairing each of the three positive inflections with comparative and superlative and finally, the comparative with the superlative.

We combined each adjective with all other adjectives to produce 1,806 analogies.

2.1.7 Noun Inflection Analogy Dataset

We created a list from the 180 frequent words from the combined five clinical categories of the clinical similarity dataset. We removed words which were not nouns and declined the remaining in the following inflections: indefinite singular, definite singular, indefinite plural and definite plural. If a noun did not exist in all four inflections or had the same form in multiple inflections, it was removed from the list. The final list contained 138 words, each in four inflections. For each noun, we made six types of inflection pairs by pairing each inflection with the three other inflections. Next, we randomly combined each noun with 13 other nouns, evenly distributed on types of inflection pairs except for the remainder after equal division, to produce 1,794 analogies.

2.2 Extrinsic Datasets

The extrinsic part consists of a hospital department classification task and a bleeding classification task. All datasets were obtained according to each dataset's respective data usage policy. The datasets are described below.

2.2.1 Bleeding Classification

For the bleeding classification dataset, we used that of Pedersen et al. (2021). It consists of 9,430 training sentences, 1,178 validation sentences, and 1,178 test sentences which are evenly distributed on the two classes: 'indicates bleeding' and 'does not indicate bleeding'. The latter class consists of 50% sentences that were deemed by the MDs to be at high risk of being misinterpreted by the deep learning model. The other 50% were random negative sentences. The classification objective is to predict if a sentence indicates bleeding.

The data came from 300 EHRs corresponding to 88,477 notes from the EHR system of the Region of Southern Denmark between 2015 and 2020. The sentences were annotated by splitting the annotation of EHRs between twelve MDs.

2.2.2 Hospital Department Classification

The hospital department classification dataset was constructed without the need of human annotators by using the department associated with each note as a label. This approach is an advantage since the task of annotating clinical records is time consuming and expensive.

The hospital department classification dataset consists of 42,000 clinical EHR notes evenly distributed on the following six Odense University Hospital departments: Cardiology; Cardiac, Thoracic and Vascular Surgery; Orthopaedic Surgery; Rheumatology; Surgery; and Medical Gastrointestinal Diseases. Danish clinical EHR notes have a tree structure consisting of many generic node headlines. MDs only fill out the end-nodes manually. To avoid node headlines or text passages specific to one department making the classification a simple task, each note was preprocessed by only keeping the lowercased end-node texts. Furthermore, end-nodes which were duplicates based only on their words, disregarding all but letters, were removed across the whole dataset. The notes are between 51 and 220 tokens. The dataset contains 7,000 notes from each department in a class-balanced train:validation:test ratio of 5:1:1. The classification objective is to predict the hospital department.

3 Word Embedding Evaluation

This section describes an evaluation of word embedding models, trained on data from different domains, using the established benchmark. We make a clinical–general domain comparison using a FastText (Bojanowski et al., 2017) model as it has the best performance on Danish text according to benchmark results (Brogaard Pauli et al., 2021). We make a clinical–GP domain comparison using a GloVe (Pennington et al., 2014) model as it is the only available type of embeddings trained on Danish GP data. We describe how the benchmark can be used to show strengths and weaknesses of different word embeddings.

We trained two sets of clinical word embeddings using the FastText and GloVe methods. The embeddings were trained on 299,718 Danish EHRs from Odense University Hospital. The text was preprocessed by lowercasing and removing headlines, subheadings, phone numbers, social security numbers, emails, URLs, dates and time stamps. Samples were defined as text from the same subheading. After removal of duplicates and samples with less than 3 words, the corpus consisted of 1.4 billion tokens.

For the clinical–general domain comparison, the clinical FastText embeddings were trained with the default settings from the FastText API (www.fasttext.cc) except from a vector size of 300, 10 negative samples and 10 epochs. The hyperparameters were chosen to be able to compare the produced embeddings with the FastText word embeddings from Grave et al. (2018) pretrained on a general domain, specifically Wikipedia and Common Crawl. The FastText models can generate out-of-vocabulary (OOV) words from subwords which e.g. makes it capable of representing unknown spelling errors. For clarity, only the results without OOV generation are reported here while the results with OOV generation are found in Appendix A.

For the clinical–GP domain comparison, the clinical GloVe embeddings are 100-dimensional embeddings

trained with the default settings from the code and paper by Pennington et al. (2014) except for a min-count of 3. The hyperparameters were chosen to be able to compare with the GloVe word embeddings from Rasmussen et al. (2019) trained on 323,122 GP EHRs.

The word embedding models are benchmarked on the established intrinsic and extrinsic datasets. For each intrinsic task, we show the performance of the embeddings on the part of the evaluation dataset which is in-vocabulary (IV), ignoring the word pairs or analogies containing OOV words. We also produce the IV rate as the proportion of word pairs or analogies which are in the vocabulary of the embeddings. Additionally, Appendix B contains the IV intersection results which show the performance of the embeddings on the intersection of all embeddings' IV dataset for that task.

For the extrinsic tasks, the word embeddings are used as input to a recurrent neural network which is initialized and trained three times with the same set of standard hyperparameters. No hyperparameter tuning is performed. A bidirectional gated recurrent unit (Cho et al., 2014) with 128 units followed by a dropout layer with probability 0.3 is trained with the Adam optimizer with a learning rate of 5e-4 for a maximum of 100 epochs using early stopping. The best model, based on the validation loss, is evaluated on the test set. The test set accuracy is reported as the evaluation result.

3.1 Intrinsic Results

We present the intrinsic semantic and syntactic benchmark results.

3.1.1 Semantic Results

Table 2 shows the intrinsic semantic results. The clinical FastText embeddings achieve better performance than the general FastText embeddings on the abbreviation equality task, clinical similarity task, UMNSRS similarity task and UMNSRS relatedness task. The clinical analogy task shows different results with the general FastText embeddings performing better with an IV accuracy of 0.14 while the clinical FastText embeddings have an IV accuracy 0.05. The clinical GloVe embeddings perform better than the GP GloVe embeddings on all intrinsic semantic tasks.

The word embeddings trained on the clinical domain show the highest IV rates, followed by the GP domain and then the general domain. The two clinical models have an IV rate equal to or higher than 0.83 for all semantic tasks. The GP GloVe embeddings have IV rates between 0.57 and 0.75 while the general FastText embeddings have IV rates between 0.54 and 0.61.

Appendix C presents the correct clinical analogy predictions for all word embedding models. Moreover, Appendix D shows the results on the clinical analogy task where a prediction is considered correct if the correct term is in the top 1, 5 and 10 nearest neighbours to the calculated vector.

3.1.2 Syntactic Results

Table 3 shows the intrinsic syntactic results. The results show that the general FastText embeddings achieve better performance than the clinical FastText embeddings on all syntactic tasks with an IV accuracy of 0.69 on verbs, 0.60 on nouns and 0.41 on adjectives. The clinical FastText embeddings perform at IV accuracies of 0.28, 0.19 and 0.16, respectively. The clinical GloVe embeddings perform better than the GP GloVe embeddings on the verb and noun inflection tasks with IV accuracies of 0.21 and 0.04, and 0.09 and 0.01, respectively. The GP GloVe embeddings perform best on the adjective inflection task with an IV accuracy of 0.04 contra 0.03 for the clinical GloVe embeddings.

The clinical domain embeddings have the highest IV rates for the verb and noun inflection tasks at 0.99 and 0.39, respectively. The general FastText embeddings have the highest IV rate for the adjective inflection task at 0.65, followed by the clinical GloVe embeddings at 0.47.

3.2 Extrinsic Results

Table 4 shows the extrinsic results. For both the FastText and GloVe models, the clinical domain embeddings achieve higher performances than their respective general domain and GP domain counterparts.

4 Discussion

In this paper, we have presented the first benchmark for evaluating Danish clinical word embeddings. Although the clinical word embeddings cannot be shared due to privacy concerns, having a publicly available benchmark will allow researchers to compare and evaluate locally available clinical word embeddings. Below, we discuss the capability of the benchmark to compare word embedding performance in the clinical domain.

As the intrinsic benchmark tasks consist of words which are typically, and in some cases, exclusively, used in the clinical domain, we expected higher IV rates from clinical domain embeddings. In concurrence, the results show that the clinical word embeddings, in general, have higher IV rates than those trained on the GP and general domain. An exception is that the general FastText embeddings have the highest IV rate for the adjective inflection analogy task. One explanation could be that clinical written language does not use as many inflections of adjectives as the general. Interestingly, when comparing the GP and general word em-

	FastText (300d)	
	Clinical	**General**
Clinical analogy, accuracy (IV)	0.05 (0.88)	0.14 (0.54)
Abbreviation equality, similarity (IV)	0.53 (0.84)	0.27 (0.58)
Clinical similarity, ρ (IV)	0.64 (0.93)	0.43 (0.61)
UMNSRS similarity, ρ (IV)	0.60 (0.88)	0.30 (0.59)
UMNSRS relatedness, ρ (IV)	0.54 (0.83)	0.32 (0.56)
	GloVe (100d)	
	Clinical	**GP**
Clinical analogy, accuracy (IV)	0.08 (0.88)	0.06 (0.61)
Abbreviation equality, similarity (IV)	0.49 (0.85)	0.24 (0.57)
Clinical similarity, ρ (IV)	0.56 (0.96)	0.34 (0.75)
UMNSRS similarity, ρ (IV)	0.41 (0.89)	0.18 (0.74)
UMNSRS relatedness, ρ (IV)	0.41 (0.84)	0.21 (0.70)

Table 2: Semantic benchmark results on the in-vocabulary (IV) dataset for each task by model type (FastText, GloVe) and domain (clinical, general, general practitioner (GP)). The accuracy metric is the accuracy on the dataset. The similarity metric is the average cosine similarity on the dataset. The ρ metric is the Spearman's rank correlation coefficient on the dataset. IV rates are reported in parenthesis. We underline the best results per task by model type.

	FastText (300d)	
	Clinical	**General**
Verb inflection analogy, accuracy (IV)	0.28 (0.99)	0.69 (0.92)
Noun inflection analogy, accuracy (IV)	0.19 (0.36)	0.60 (0.13)
Adjective inflection analogy, accuracy (IV)	0.16 (0.36)	0.41 (0.65)
	GloVe (100d)	
	Clinical	**GP**
Verb inflection analogy, accuracy (IV)	0.21 (0.99)	0.09 (0.83)
Noun inflection analogy, accuracy (IV)	0.04 (0.39)	0.01 (0.18)
Adjective inflection analogy, accuracy (IV)	0.03 (0.47)	0.04 (0.25)

Table 3: Syntactic benchmark results on the in-vocabulary (IV) dataset for each task by model type (FastText, GloVe) and domain (clinical, general, general practitioner (GP)). The accuracy metric is the accuracy on the dataset. IV rates are reported in parenthesis. We underline the best results per task by model type.

beddings on the semantic tasks, the GP embeddings, in four out of five tasks, have higher IV rates but lower accuracy. This result shows that the GP embeddings have seen more clinical domain words than the general embeddings during training, but the general embeddings capture higher quality information for the words that it has seen. This could be due to the size and quality of the dataset, differences between model types or the dimensionality of the embeddings. Future work should investigate these claims further.

The benchmark shows that the clinical embeddings surpass the general and GP embeddings in all semantic tasks except for the clinical analogy task where the general FastText embeddings performed better than the clinical FastText embeddings. This discrepancy may be caused by the clinical analogy dataset only containing 164 analogies of which only 54% are IV for the general FastText model.

The general embeddings surpass the clinical embeddings on the syntactic tasks which shows that it has captured higher quality syntactic information for the words that it has seen during training. This is most likely due to Wikipedia and Common Crawl, which it was trained on, containing a higher quality of syntactic information than clinical EHRs.

The fact that the general embeddings achieve the highest IV rate on the adjective inflection task suggests that the task consists of more inflections specific to the general domain than our clinical dataset. On the contrary, clinical domain embeddings achieve the highest IV rates on the verb and noun inflection tasks which suggests that these syntactic tasks do contain inflections specific to the clinical domain.

Similar to earlier work (Zhao et al., 2018; Wang et al., 2018; Chen et al., 2019), we found that the clinical word embeddings perform better than the GP and general domain embeddings on extrinsic tasks. It is notable that for the extrinsic tasks, the GP GloVe embeddings are closer to the performance of the clinical GloVe embeddings than the general FastText embeddings are to

	FastText (300d)	
	Clinical	**General**
Bleeding classification, accuracy	<u>0.93</u>	0.84
Department classification, accuracy	<u>0.83</u>	0.65
	GloVe (100d)	
	Clinical	**GP**
Bleeding classification, accuracy	<u>0.90</u>	0.87
Department classification, accuracy	<u>0.76</u>	0.66

Table 4: Extrinsic benchmark results by model type (FastText, GloVe) and domain (clinical, general, general practitioner (GP)). We report accuracies on the class-balanced bleeding and department classification tasks using the word embeddings as input. We underline the best results per task by model type.

that of the clinical FastText embeddings. This could be explained by the fact that there is some similarity between the GP and clinical domains, both being subdomains of the healthcare domain.

Considering that the general embeddings perform well on syntactic tasks and clinical embeddings perform well on semantic and extrinsic tasks, future work should explore training word embeddings from the general FastText checkpoint on clinical data. This might provide word embeddings that better capture both clinical and general syntactic and semantic properties.

4.1 Limitations

This study compared GloVe and FastText word embeddings. While FastText performed best on some benchmarks, other word embedding methods might perform better. We leave these investigations to future work.

Future use of the presented resources relies on the assumption that the words in the intrinsic datasets also appear in the user's vocabulary. In section 2.1.2 we described how we tried to mitigate this shortcoming.

The clinical similarity dataset would benefit from including more pairs with high similarity and decreasing the mean standard deviation, e.g. by including more raters from different specialities. To alleviate MD rating disagreement, we have included in the supplementary material the clinical similarity ratings for each MD with information about the standard deviation of each word pair, which can be used to set a threshold of maximum allowed disagreement. Appendix E shows the results on the clinical similarity dataset consisting of pairs with standard deviations at or below 1.

The extrinsic department classification task might as well classify the writing styles of specific MDs in a department, thus not necessarily generalizing to other MDs. This can be remedied by having unique authors in the test split.

It is a limitation to the extrinsic results that no hyperparameter tuning was performed. Results from a model trained with a standard set of hyperparameters can rank the word embeddings but the results are not indicative of the best performance of each embedding.

We have shown a discrepancy between the clinical analogy task and all other semantic tasks. We believe it would be beneficial to include more analogies in the clinical analogy dataset as the result is based on few IV analogies.

The syntactic results suggest that the adjective inflection task consists of more inflections specific to the general domain than the clinical domain. Many of the inflections do exist in the clinical domain but it is a limitation for the evaluation of clinical word embeddings that not enough inflections are specific for the clinical domain.

The tasks were designed to evaluate static word embeddings using only single-word expressions which limits the use of the benchmark for contextual word embeddings such as tranformer models and word embeddings trained on n-grams.

It is a limitation to our benchmark that it only provides two extrinsic tasks, and in general, that there are no Danish clinical extrinsic datasets publicly available. Due to privacy concerns, we cannot publish the extrinsic datasets, but we provide a method for creating an extrinsic test that leverages already existing labels in the form of the department of the clinical note. This method does not need any labeling but still requires access to EHRs. We encourage interested researchers to contact us for the possibility of sharing the extrinsic datasets.

Future work should focus on developing more diverse extrinsic tasks such as named entity recognition, relation extraction and question answering.

5 Conclusion

In this paper, we presented a benchmark for Danish clinical word embeddings. The benchmark consists of two extrinsic tasks, five intrinsic semantic tasks and three intrinsic syntactic tasks. We developed clinical word embeddings and compared them with word embeddings trained on a general and general practitioner

domain. The benchmark showed that the word embeddings trained on clinical data performed better on the extrinsic and semantic tasks, except for the clinical analogy task. On the syntactic tasks, the FastText word embeddings trained on a general domain performed better than those trained on a clinical domain.

Acknowledgements

The authors thank Anne Bryde Alnor, Charlotte Gils, Eline Sandvig Andersen, Ida Stangerup, Jesper Dupont Ewald, Jesper Farup Revsholm, Katrine Sølling Borlund Madsen and Kristina Bjerg Appel for the rating work for the clinical similarity task.

References

Abdalla, Mohamed, Moustafa Abdalla, Graeme Hirst, and Frank Rudzicz. 2020. Exploring the privacy-preserving properties of word embeddings: Algorithmic validation study. *J Med Internet Res*, 22(7):e18055.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Brogaard Pauli, Amalie, Maria Barrett, Ophélie Lacroix, and Rasmus Hvingelby. 2021. DaNLP: An open-source toolkit for danish natural language processing. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa 2021)*.

Chen, Qingyu, Yifan Peng, and Zhiyong Lu. 2019. Biosentvec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5.

Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Clark, Kevin, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Danish Language Council. 2012. *Retskrivningsordbogen*, 4 edition. Danish Language Council. Including 8 digital issues (2017).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Harris, Zellig S. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Khattak, Faiza Khan, Serena Jeblee, Chloé Pou-Prom, Mohamed Abdalla, Christopher Meaney, and Frank Rudzicz. 2019. A survey of word embeddings for clinical text. *Journal of Biomedical Informatics*, 100:100057. Articles initially published in Journal of Biomedical Informatics: X 1-4, 2019.

Leaman, Robert, Ritu Khare, and Zhiyong Lu. 2015. Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics*, 57:28–37.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Schaffalitzky de Muckadell, Ove B., Stig Haunsø, and Hendrik Vilstrup. 2009. *Medicinsk kompendium*, 17 edition. Nyt Nordisk Forlag.

Pakhomov, Serguei, Bridget McInnes, Terrence Adam, Ying Liu, Ted Pedersen, and Melton Genevieve B. 2010. Semantic similarity and relatedness between clinical terms: An experimental study. *AMIA annual symposium proceedings*, 2010:572–576.

Pedersen, Jannik S., Martin S. Laursen, Thiusius Rajeeth Savarimuthu, Rasmus Søgaard Hansen, Anne Bryde Alnor, Kristian Voss Bjerre, Ina Mathilde Kjær, Charlotte Gils, Anne-Sofie Faarvang Thorsen,

Eline Sandvig Andersen, Cathrine Brødsgaard Nielsen, Lou-Ann Christensen Andersen, Søren Andreas Just, and Pernille Just Vinholt. 2021. Deep learning detects and visualizes bleeding events in electronic health records. *Research and Practice in Thrombosis and Haemostasis*, 5(4):e12505.

Peng, X, Y Zheng, C Lin, and A Siddharthan. 2021. Summarising historical text in modern languages. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3123–3142. Association for Computational Linguistics (ACL).

Peng, Yifan, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Rasmussen, Mathias, Nichlas Berggrein, and Leon Derzcynski. 2019. Named entity recognition and disambiguation in danish electronic health records. Master's thesis, IT University of Copenhagen.

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019a. Superglue: a stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 3266–3280.

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019b. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*.

Wang, Bin, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C.-C. Jay Kuo. 2019c. Evaluating word embedding models: methods and experimental results. *APSIPA Transactions on Signal and Information Processing*, 8:e19.

Wang, Yanshan, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. 2018. A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*, 87:12–20.

Zhao, Mengnan, Aaron J. Masino, and Christopher C. Yang. 2018. A framework for developing and evaluating word embeddings of drug-named entity. In *Proceedings of the BioNLP 2018 workshop*, pages 156–160, Melbourne, Australia. Association for Computational Linguistics.

A Benchmark Results Including Models With OOV Generation

Table 5 shows the semantic benchmark results for all models, including the FastText models with OOV generation.

Table 6 shows the syntactic benchmark results for all models, including the FastText models with OOV generation.

Table 7 shows the extrinsic benchmark results for all models, including the FastText models with OOV generation.

B In-Vocabulary Intersection Results

We report the IV intersection results which show the performance of the embeddings on the intersection of all embeddings' IV dataset for that task. We also report relative coverage (RC) for each model as the proportion that the IV words of a model constitute out of the union of all models' IV words.

Table 8 shows the semantic benchmark results on the intersection of IV datasets of all embeddings.

Table 9 shows the syntactic benchmark results on the intersection of IV datasets of all embeddings.

C Correctly Predicted Semantic Analogies

Table 10 shows the correctly predicted semantic analogies for all models, including the FastText models with OOV generation.

D Clinical Analogy Task Top N Accuracies

Table 11 shows the results on the clinical analogy task where a prediction is considered correct if the correct term is in the top 1, 5 and 10 nearest neighbours to the calculated vector.

E Results on Clinical Similarity Dataset With Standard Deviation at or Below 1

Table 12 shows the results on the clinical similarity dataset with standard deviation at or below 1.

F Supplementary Material

All datasets are txt files with tab-separated values. Each row has one word pair or analogy. Some datasets are divided into parts. A headline of a part is in all caps and introduced with ': '.

The following datasets are attached:

- Clinical analogy dataset (txt)
- Abbreviation equality dataset (txt)
- Clinical similarity dataset (txt)
- Clinical similarity SD1 dataset (txt)
- UMNSRS similarity dataset (txt)
- UMNSRS relatedness dataset (txt)
- Verb inflection analogy dataset (txt)
- Noun inflection analogy dataset (txt)
- Adjective inflection analogy dataset (txt)

The clinical similarity ratings and their standard deviations are found in file:

- Clinical similarity ratings (xlsx)

The online sources of clinical abbreviations are found in file:

- Abbreviation sources (xlsx)

	FastText (300d)			
	Clinical		General	
	No gen.	OOV gen.	No gen.	OOV gen.
Clinical analogy, acc (IV)	0.05 (0.88)	0.04 (1.0)	0.14 (0.54)	0.07 (1.0)
Abbreviation equality, sim (IV)	0.53 (0.84)	0.52 (1.0)	0.27 (0.58)	0.30 (1.0)
Clinical similarity, ρ (IV)	0.64 (0.93)	0.62 (1.0)	0.43 (0.61)	0.32 (1.0)
UMNSRS similarity, ρ (IV)	0.60 (0.88)	0.58 (1.0)	0.30 (0.59)	0.25 (1.0)
UMNSRS relatedness, ρ (IV)	0.54 (0.83)	0.54 (1.0)	0.32 (0.56)	0.27 (1.0)
	GloVe (100d)			
	Clinical (No gen.)		GP (No gen.)	
Clinical analogy, acc (IV)	0.08 (0.88)		0.06 (0.61)	
Abbreviation equality, sim (IV)	0.49 (0.85)		0.24 (0.57)	
Clinical similarity, ρ (IV)	0.56 (0.96)		0.34 (0.75)	
UMNSRS similarity, ρ (IV)	0.41 (0.89)		0.18 (0.74)	
UMNSRS relatedness, ρ (IV)	0.41 (0.84)		0.21 (0.70)	

Table 5: Semantic benchmark results by model type (FastText, GloVe), domain (clinical, general, general practitioner (GP)), and out-of-vocabulary (OOV) generation (no gen., OOV gen.). The acc metric is the accuracy on the in-vocabulary (IV) dataset. The sim metric is the average cosine similarity on the IV dataset. The ρ metric is the Spearman's rank correlation coefficient on the IV dataset. IV rates are reported in parenthesis. We underline the best results per task by model type.

	FastText (300d)			
	Clinical		General	
	No gen.	OOV gen.	No gen.	OOV gen.
Verb inflection, acc (IV)	0.28 (0.99)	0.28 (1.0)	0.69 (0.92)	0.66 (1.0)
Noun inflection, acc (IV)	0.19 (0.36)	0.11 (1.0)	0.60 (0.13)	0.20 (1.0)
Adjective inflection, acc (IV)	0.16 (0.36)	0.07 (1.0)	0.41 (0.65)	0.29 (1.0)
	GloVe (100d)			
	Clinical (No gen.)		GP (No gen.)	
Verb inflection, acc (IV)	0.21 (0.99)		0.09 (0.83)	
Noun inflection, acc (IV)	0.04 (0.39)		0.01 (0.18)	
Adjective inflection, acc (IV)	0.03 (0.47)		0.04 (0.25)	

Table 6: Syntactic benchmark results by model type (FastText, GloVe), domain (clinical, general, general practitioner (GP)), and out-of-vocabulary (OOV) generation (no gen., OOV gen.). The acc metric is the accuracy on the in-vocabulary (IV) dataset. IV rates are reported in parenthesis. We underline the best results per task by model type.

	FastText (300d)			
	Clinical		General	
	No gen.	OOV gen.	No gen.	OOV gen.
Bleeding classification, acc	0.93	0.92	0.84	0.84
Department classification, acc	0.83	0.83	0.65	0.64
	GloVe (100d)			
	Clinical (No gen.)		GP (No gen.)	
Bleeding classification, acc	0.90		0.87	
Department classification, acc	0.76		0.66	

Table 7: Extrinsic benchmark results by model type (FastText, GloVe), domain (clinical, general, general practitioner (GP)), and out-of-vocabulary (OOV) generation (no gen., OOV gen.). We report accuracies on the bleeding and department classification task using the word embeddings as input. We underline the best results per task by model type.

	FastText (300d)	
	Clinical	**General**
Clinical analogy, accuracy (RC)	0.06 (1.0)	0.14 (0.61)
Abbreviation equality, similarity (RC)	0.55 (0.97)	0.27 (0.67)
Clinical similarity, ρ (RC)	0.67 (0.98)	0.44 (0.63)
UMNSRS similarity, ρ (RC)	0.57 (0.98)	0.28 (0.66)
UMNSRS relatedness, ρ (RC)	0.52 (0.97)	0.29 (0.66)
	GloVe (100d)	
	Clinical	**GP**
Clinical analogy, accuracy (RC)	0.13 (1.0)	0.08 (0.69)
Abbreviation equality, similarity (RC)	0.60 (0.98)	0.25 (0.66)
Clinical similarity, ρ (RC)	0.60 (1.0)	0.35 (0.79)
UMNSRS similarity, ρ (RC)	0.40 (0.99)	0.15 (0.82)
UMNSRS relatedness, ρ (RC)	0.40 (0.98)	0.23 (0.82)

Table 8: Semantic benchmark results on the intersection of IV datasets for each task by model type (FastText, GloVe) and domain (clinical, general, general practitioner (GP)). The accuracy metric is the accuracy on the dataset. The similarity metric is the average cosine similarity on the dataset. The ρ metric is the Spearman's rank correlation coefficient on the dataset. Relative coverage (RC) is reported in parenthesis. We underline the best results per task by model type.

	FastText (300d)	
	Clinical	**General**
Verb inflection analogy, accuracy (RC)	0.29 (0.99)	0.71 (0.92)
Noun inflection analogy, accuracy (RC)	0.17 (0.92)	0.63 (0.54)
Adjective inflection analogy, accuracy (RC)	0.18 (0.55)	0.54 (0.98)
	GloVe (100d)	
	Clinical	**GP**
Verb inflection analogy, accuracy (RC)	0.23 (0.99)	0.09 (0.83)
Noun inflection analogy, accuracy (RC)	0.08 (0.98)	0.03 (0.64)
Adjective inflection analogy, accuracy (RC)	0.05 (0.71)	0.04 (0.38)

Table 9: Syntactic benchmark results on the intersection of IV datasets for each task by model type (FastText, GloVe) and domain (clinical, general, general practitioner (GP)). The accuracy metric is the accuracy on the dataset. Relative coverage (RC) is reported in parenthesis. We underline the best results per task by model type.

	FastText (300d)				GloVe (100d)	
	Clinical		General		Clinical	GP
	No gen.	OOV gen.	No gen.	OOV gen.	No gen.	No gen.
hoftealloplastik + knæ - knæalloplastik = hofte (hip replacement + knee - knee replacement = hip)	✓	✓	✓	✓	✓	✓
hofte + knæalloplastik - knæ = hoftealloplastik (hip + knee replacement - knee = hip replacement)	✓	✓	✓	✓	✓	✓
knæalloplastik + hofte - hoftealloplastik = knæ (knee replacement + hip - hip replacement = knee)	✓	✓	✓	✓	✓	✓
knæ + hoftealloplastik - hofte = knæalloplastik (knee + hip replacement - hip = knee replacement)	✓	✓			✓	✓
ovarier + mand - testikler = kvinde (ovaries + man - testicles = woman)	✓	✓	✓	✓	✓	✓
testikler + kvinde - ovarier = mand (testicles + woman - ovaries = man)			✓	✓		
trombocytpool+anæmi - sag-m = trombocytopeni (thrombocyte pool + anemia - sag-m = thrombocytopenia)	✓	✓				
sag-m + trombocytopeni - trombocytpool = anæmi (sag-m + thrombocytopenia - thrombocyte pool = anemia)	✓	✓			✓	
høretab + øjne - synstab = ører (hearing loss + eyes - visual obscuration = ears)			✓	✓	✓	
synstab + ører - høretab = øjne (visual obscuration + ears - hearing loss = eyes)			✓	✓		
mad + tørst - væske = sult (food + thirst - liquid = hunger)			✓	✓	✓	
milt + gastrektomi - mavesæk = splenektomi (spleen + Gastrectomy - stomach = splenectomy)			✓	✓	✓	
aids + borrelia - neuroborreliose = hiv (aids + borreliosis - neuroborreliosis = hiv)			✓	✓		
levothyroxin + hyperthyroidisme - thiamazol = hypothyroidisme (levothyroxine + lyperthyroidism - thiamazole = hypothyroidism)			✓	✓		
respirator + nyresvigt - dialyse = respirationssvigt (respirator + renal failure - dialysis = respiratory failure)			✓	✓		
virus + dyrkning - bakterie = pcr (virus + cultivation - bacteria = pcr)					✓	
tarm + hæmoptyse - lunger = melæna (intestine + hemoptysis - lung = melena)					✓	
Kreatinin + knoglemarvsfunktion - differentialtælling = nyrefunktion (creatinine + bone marrow function - differential count = renal function)					✓	
nyrefunktion + differentialtælling - knoglemarvsfunktion = kreatinin (renal function + differential count - bone marrow function = creatinine)						✓

Table 10: Overview of the correctly predicted semantic analogies by model type (FastText, GloVe), domain (clinical, general, general practitioner (GP)), and out-of-vocabulary (OOV) generation (no gen., OOV gen.). Each analogy is presented in Danish, and the English translation is parenthesis. A mark signifies a correctly predicted analogy.

	FastText (300d)			
	Clinical		General	
	No gen.	OOV gen.	No gen.	OOV gen.
Top 1, acc	0.05	0.04	<u>0.14</u>	0.07
Top 5, acc	0.10	0.09	<u>0.27</u>	0.16
Top 10, acc	0.13	0.11	<u>0.41</u>	0.28
IV rate	0.88	1.0	0.54	1.0
	GloVe (100d)			
	Clinical (No gen.)		GP (No gen.)	
Top 1, acc	<u>0.08</u>		0.06	
Top 5, acc	<u>0.15</u>		0.08	
Top 10, acc	<u>0.21</u>		0.13	
IV rate	0.88		0.61	

Table 11: Top n accuracies and IV rate on the semantic clinical analogy dataset by model type (FastText, GloVe), domain (clinical, general, general practitioner (GP)), and out-of-vocabulary (OOV) generation (no gen., OOV gen.). A prediction is considered correct if the correct term is in the top n nearest neighbours to the calculated vector. The IV rate is the proportion of word pairs or analogies which are in-vocabulary. We underline the best results by model type.

	FastText (300d)			
	Clinical		General	
	No gen.	OOV gen.	No gen.	OOV gen.
Clinical similarity SD1, ρ	<u>0.60</u>	0.57	0.44	0.35
IV rate	0.94	1.0	0.61	1.0
	GloVe (100d)			
	Clinical (No gen.)		GP (No gen.)	
Clinical similarity SD1, ρ	<u>0.54</u>		0.40	
IV rate	0.96		0.75	

Table 12: Results on the clinical similarity dataset with standard deviation at or below 1 by model type (FastText, GloVe), domain (clinical, general, general practitioner (GP)), and out-of-vocabulary (OOV) generation (no gen., OOV gen.). The ρ metric is the Spearman's rank correlation coefficient on the IV dataset. The IV rate is the proportion of word pairs or analogies which are in-vocabulary. The dataset contains 255 word pairs. We underline the best result by model type.