Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4577–4585
Marseille, 20-25 June 2022

Modeling Dutch Medical Texts for Detecting Functional Categories and Levels of COVID-19 Patients

Jenia Kim*, Stella Verkijk*, Edwin Geleijn†, Marike van der Leeden†, Carel Meskers†, Caroline Meskers†, Sabina van der Veen◇, Piek Vossen*, Guy Widdershoven◇
*Computational Linguistics and Text Mining Lab, Faculty of Humanities, Vrije Universiteit Amsterdam
jenka@protonmail.com, stellaverkijk@outlook.com, p.t.j.m.vossen@vu.nl
†Department of Rehabilitation Medicine, Amsterdam University Medical Centers
{e.geleijn, m.vanderleeden, c.meskers, c.j.w.meskers}@amsterdamumc.nl
◇Department of Ethics, Law and Humanities, Amsterdam University Medical Centers
{s.vanderveen, g.widdershoven}@amsterdamumc.nl

Abstract

Electronic Health Records contain a lot of information in natural language that is not expressed in the structured clinical data. Especially in the case of new diseases such as COVID-19, this information is crucial to get a better understanding of patient recovery patterns and factors that may play a role in it. However, the language in these records is very different from standard language and generic natural language processing tools cannot easily be applied out-of-the-box. In this paper, we present a fine-tuned Dutch language model specifically developed for the language in these health records that can determine the functional level of patients according to a standard coding framework from the World Health Organization. We provide evidence that our classification performs at a sufficient level (F1-score above 80% for the main categories and error rates of less than 1 level on a 5-point Likert scale for levels) to generate patient recovery patterns that can be used to analyse factors that contribute to the rehabilitation of COVID-19 patients and to predict individual patient recovery of functioning.

Keywords: medical text mining, electronic health records, Dutch language models, functional level classification, COVID-19

1. Introduction

Electronic Health Records (EHRs) contain a wealth of unexplored information. Some of it is captured in the structured clinical data on patients (e.g. diagnosis codes, lab results); other aspects are not recorded as structured data, but rather appear in free-text form. This unstructured data may be essential, especially in the case of new diseases such as COVID-19 on which we have little knowledge. Natural Language Processing (NLP) has a great potential to mine the free-text in EHRs for significant patterns and insights, helping us to learn more about the disease and its effects.

One important aspect of health and well-being is captured by *functional status*, i.e. the individual's ability to engage in different activities and social roles. For example, in the case of COVID-19, it is known that almost 80% of hospitalized patients report at least one persistent symptom (such as fatigue, sleep difficulty, anxiety etc.) six months after discharge (Huang et al., 2021). This might adversely impact their functioning, including their ability to work, to exercise, to maintain a healthy body weight, etc. Deployment of appropriate and personalized rehabilitation plans for these patients requires insight into how their functioning develops over time, so that relevant patterns and predictors can be associated.

The description of functioning is standardized in a WHO's framework of codes called the International Classification of Functioning, Disability and Health (ICF) (World Health Organization, 2001).[1] ICF is a hierarchical structure of codes that organizes human activity into various categories (e.g. speaking, walking, dressing). For each category, the levels of functioning/activity/participation are described by numeric qualifiers that indicate the extent of capacity or impairment (e.g. mild impairment, severe difficulty).

Unlike diagnosis codes, ICF codes are not systematically captured in the structured clinical data. Rather, the functional status of a patient is often described in the free text of EHRs (Maritz et al., 2017). These clinical notes are short texts whose purpose is to communicate the status of a patient between different healthcare professionals. As such they contain a mixture of general and specialised language. An example of such a note is a nurse's report on what a patient has eaten on a specific day, whether they were able to get out of the bed, whether they have taken their medication, etc.

At the Amsterdam University Medical Centers, over 10 million notes from EHRs were collected since 2017 and recently all records on COVID-19 diagnosed patients treated in the Amsterdam academic hospitals became available.[2] This massive text resource provides enormous possibilities to build language models and text classification systems that can derive structured information from EHRs and provide valuable insights about the functioning of patients, in relation to new diseases such as COVID-19, or to any other disease.

In this paper, we describe the language technology that

[1] WHO-ICF online

[2] The medical ethics committee approved use of EHRs for the purpose of this research provided patient privacy is secured.

can mine clinical notes for insights about the functional status of patients. Specifically, we present an automated coding of natural language descriptions of functioning in EHRs into standardized categories according to ICF. As the language in the EHRs deviates in many aspects from standard Dutch, we relied on the transformer Language Model (LM) 'MedRoberta.nl' that was built from scratch on millions of Dutch hospital notes from EHRs (Verkijk and Vossen, 2022). Verkijk and Vossen showed that this domain specific model performs better than general language models such as BERTje (de Vries et al., 2019) and RobBERT (Delobelle et al., 2020) on tasks within the Dutch medical domain.

In the present work, the MedRoberta.nl model was fine-tuned with ICF labels to capture 9 different functional categories such as the emotional status of patients, their ability to walk, breathe, eat and concentrate; the specific levels of functioning in each of these categories are captured as well. We demonstrate that our classifier performs sufficiently well to create functional recovery patterns of patients over time. The classification output can help find typical recovery patterns and correlate these to types of patients and/or treatments.

Our contributions are the following:

1. We present freely-available classifiers to assign ICF functional levels to Dutch EHRs.
2. We demonstrate that EHR text can be interpreted according to standardized categories, testing the WHO principle behind ICF.
3. We provide an instrument for analysing patient recovery patterns in relation to treatment protocols, especially relevant for new diseases like COVID-19 on which a lot is unknown and fast insights are needed.
4. Our approach can be applied to other languages, which will help the understanding of patient functioning in an international context.

In this paper, we relate our work to the state-of-the-art and explain the design of our language model in Section 2. We describe the annotation process in Section 3 and the resulting annotated dataset in Section 4. In Section 5 we describe the fine-tuning of the language model for assigning functioning categories and levels to sentences, and we report on the performance of the fine-tuned models. In Section 6 we illustrate how our system can be used to generate meaningful patterns of functioning and recovery. Finally, we conclude and discuss future options in Section 7.

2. Related work

2.1. Medical domain Language models

The purpose of EHRs is to efficiently communicate important information on the status of a patient from one health professional to the other. As such, the text has properties that are very different from texts in Wikipedia or news archives, which standard LMs for Dutch are trained on. For example, typos are very common, as notes are taken quickly. A model trained on edited Wikipedia texts and news might not capture that different versions of the same word with typos and spelling mistakes refer to the same meaning/sense. Also, because of efficiency, the language in hospital notes sometimes adheres to different syntax than general language. Examples are shorter sentences, missing pronouns or articles, but also completely novel constructions, like 'De stemming **imponeert** normofoor, met een normaal modulerend affect' (*Mood **impresses** normophore, with normal modulating affect*). Furthermore, there are major differences in terminology and word meaning distribution in medical text compared to general Dutch. Apart from a much higher frequency of medical terms, words appear that are never used or seen as ungrammatical in general Dutch (see Appendix for some examples).

Several studies demonstrate that domain-specific models give better results at tasks within their domain than general models (Beltagy et al., 2019; Huang et al., 2019; Lee et al., 2020; Chalkidis et al., 2020; Gu et al., 2020). These studies also experimented with different ways to build domain-specific models. For example, whether to train the model from scratch, i.e. initialize it with low random weights that need to be adapted, or to extend pre-training on an existing model, i.e. take trained weights from an existing model as a starting point as a way of transfer learning.

Chalkidis et al. (2020) experimented with training from scratch as well as with extending pre-training on BERT, and showed that it depends on the downstream task which model performs better. Both BioBERT (Lee et al., 2020) and ClinicalBERT (Huang et al., 2019) were made by extending pre-training on BERT, showing improved performance on downstream tasks in the biomedical domain. However, Gu et al. (2020) built a new domain-specific language model for the biomedical domain by pre-training from scratch, and show improved performance on most biomedical downstream tasks compared to both BioBERT and ClinicalBERT. They argue that transfer learning by extending pre-training on a general language model is only preferable when there is not much domain-specific data to pre-train on. They also state that a big advantage of training from scratch is the fact that a domain-specific vocabulary can be used during pre-training. The SciBERT model (Beltagy et al., 2019) was also trained from scratch and the authors experimented with using either the general vocabulary from BERT or their own domain-specific vocabulary during pre-training, concluding that using a domain-specific vocabulary indeed had a positive effect on the final performance of the model.

To the best of our knowledge, the only Dutch language model for the medical domain is the Roberta

model MedRoBERTa.nl (Verkijk and Vossen, 2022), which is used in this study. This model was developed from scratch (random initialisation) on nearly 10 million notes from EHRs with a specialized vocabulary and was shown to have better performance than standard Dutch models such as BERTje (de Vries et al., 2019) and RobBERT (Delobelle et al., 2020). Verkijk and Vossen (2022) performed intrinsic as well as extrinsic evaluations on in-domain tasks, evaluating the raw model before fine-tuning on its zero-shot performance on a similarity task as well as on a downstream classification task, after fine-tuning the model.

MedRoBERTa.nl is a unique model that suits our task of assigning ICF categories to free text from EHRs. Other models, such as English medical models trained on carefully edited medical publications from PubMed, are less optimal for this challenging classification task. They are not used to the varied and non-standardized ways of writing in clinical notes by a large variety of caretakers with different backgrounds, which, as we have shown, results in deviant language use.

2.2. Automated ICF Coding

The task of mapping natural language descriptions of functional status to the ICF framework is often referred to as 'ICF coding'. The task consists of two main components: assigning an ICF category and assigning a qualifier that indicates the level of functioning within this category. For example, the sentence *The patient was able to eat independently* is coded as *d550.00*, where *d550* refers to the Eating category and *00* is a qualifier that indicates that there is no difficulty or impairment.

To the best of our knowledge, the only existing work that addresses both components of the task is Kukafka et al. (2006). They present a rule-based system that "translates" free text to standard target concepts, and then assigns ICF codes to the concepts. They focus on 5 ICF categories; the data on which the system is trained and evaluated consists of 250 rehabilitation discharge summaries, selected from 10 relevant diagnoses. The system's performance is similar to human annotation; it performs better than non-expert human coders, but not as well as expert human coders. For the qualifier assignment task, the areas under the ROC curves for the experts, non-experts, and NLP system are 0.85, 0.79 and 0.82, respectively.[3]

Other work on the task focuses on the assignment of an ICF category, without addressing the qualifier aspect. Focusing on the Mobility chapter of the ICF (which includes about 15-20 ICF categories such as d415 Maintaining body position and d450 Walking), Thieu et al. (2021) train a named-entity-recognition (NER) model, where the entities are: Mobility (e.g. *Patient able to ambulate 40 ft. with rolling walker*) and the nested sub-entities Action (*ambulate*), Assistance (*with rolling walker*) and Quantification (*40 ft.*). They annotate 400 physical therapy notes (14,000 entities) and experiment with CRF, RNN, and fine-tuning BERT (Devlin et al., 2019) and BioBERT (Lee et al., 2020). Their best performing model is an ensemble that aggregates the outputs of all classifiers, which achieves an average F1-score of 0.85.

Newman-Griffis et al. (2021) extend the work of Thieu et al. (2021) on Mobility to two additional ICF chapters: Self-care and Domestic life. The input to their models are so-called 'activity mentions', i.e. phrases that discuss functional status information (like the Mobility entity of Thieu et al. (2021)); the task of the classifier is to assign the correct ICF code to each mention. In their work, the extraction of what they call 'activity mentions' from the free text is viewed as a separate task, and the performance of a pipeline that combines both tasks is not discussed. Our models, on the other hand, take raw clinical notes as input and do not involve a preliminary extraction of 'activity mentions'. Similarly to Kukafka et al. (2006), our work focuses on automated assignment of both an ICF category and the level of functioning (qualifier). Our method – fine-tuning a pre-trained language model – has also been applied by Thieu et al. (2021); however, we do not treat the task as a token-level NER, but rather as a sentence-level multi-label classification, followed by regression. In addition, we train and evaluate our models on all available types of clinical notes, rather than selecting specific diagnoses (like Kukafka et al. (2006)) or specific note types (like Thieu et al. (2021)).

3. Annotation

ICF code	Category	Abbrev.	Functioning levels scale
b1300	Energy level	ENR	0-4
b140	Attention functions	ATT	0-4
b152	Emotional functions	STM	0-4
b440	Respiration functions	ADM	0-4
b455	Exercise tolerance functions	INS	0-5
b530	Weight maintenance functions	MBW	0-4
d450	Walking	FAC	0-5
d550	Eating	ETN	0-4
d840-d859	Work and employment	BER	0-4

Table 1: Overview of the ICF categories in the project

Our work focuses on automated coding of 9 ICF categories, which were chosen due to their relevance to recovery from COVID-19; see Table 1. For each category, the levels of functioning (qualifiers) are defined on a scale of 0-4 or 0-5; the levels indicate the extent of functioning or disability, where 4 or 5 indicates that there is no problem or limitation, and 0 indicates a total disability (in this category).[4] The exact definitions of the scales can be found in the annotation guidelines.

[3] Kukafka et al. (2006) do not report performance on the task of assigning an ICF category.

[4] Our scales are reversed compared to the generic ICF qualifiers: in the ICF scale, 0 indicates no difficulty/impairment, while in our scale 0 indicates complete disability. Further, two of our categories have a 0-5 scale, while ICF scales are always 0-4.

3.1. Guidelines and procedure

The annotation of clinical notes from EHRs with ICF labels was performed by six native Dutch-speaking (para)medical students. The annotators were initially trained for the task by the core project team, consisting of healthcare professionals and NLP experts. In addition, throughout the entire annotation period, weekly mentoring sessions were held in which questions and difficult examples were discussed among the annotators and the core project team.

The annotation was conducted using the INCEpTION software (Klie et al., 2018), which was installed locally on a secure server within the Amsterdam UMC firewall, according to the criteria of the medical ethics committee and the EU GDPR legislation; this setup ensures that the sensitive patient data do not leave the hospital's virtual environment.

Annotation guidelines were created by the core project team, consisting of healthcare professionals and NLP experts. Before the production phase with the six annotators started, the guidelines were tested on a sample of notes by the healthcare professionals from the core team. Based on their experience, a few rounds of discussion-update-annotation occurred, resulting in the final version that was given to the annotators. Based on the first week of annotation, the guidelines were updated one last time; after that, no significant changes to the guidelines were made. The full version of the final guidelines in Dutch, as well as an abbreviated version in English, are available on the project's GitHub.[5]

The annotation consists of assigning a label to a phrase that describes one of the 9 ICF categories and another label to a phrase that describes the level of functioning. For example, in the sentence *Concentratie is nog wel iets verminderd* (Concentration is still slightly diminished), the word *concentratie* (concentration) is marked with the category label ATT (Attention) and the phrase *iets verminderd* (slightly diminished) is marked with the level label att-3, which indicates a mild functioning problem.

In addition to the category and level labels, the annotators were instructed to assign a 'disregard' label to notes that should be completely excluded from the dataset (e.g. notes about children under 12 years old), and assign a 'background' or 'target' label to sentences that discuss past or future functional status.

3.2. Inter-annotator agreement

The inter-annotator agreement is calculated based on 35 clinical notes that were annotated by all 6 annotators. The agreement is calculated on a sentence-level; the measure for the category labels is mean pairwise F1-score (as described by Hripcsak and Rothschild (2005)), and the measure for the level labels is mean pairwise MAE (mean absolute error). The results are shown in Table 2.

[5]https://github.com/cltl/a-proof-zonmw/tree/main/resources/annotation_guidelines

For the category labeling, the F1-scores range from 0.34 to 0.78. Whereas the higher scores are reasonable to good, the lower scores – specifically for INS (Exercise tolerance), BER (Work and employment) and ETN (Eating) – are too low. For INS, the difficulty to decide which sentences are relevant has been explicitly expressed by the annotators throughout the process. Specifically, it was not clear to them whether mentions of activities (e.g. that a patient plays football once a week) should be labeled as INS. In addition, there is some overlap between the INS category and the FAC (Walking), ADM (Respiration) and ENR (Energy level) categories; this was also often mentioned as a cause for confusion. For example, the sentence *Was zelf naar de WC gelopen, en was daarna uitgeput* (Walked to the WC by herself and was exhausted afterwards) describes walking (the patient can walk independently) and energy level (the patient is fatigued after a short walk); however, it was unclear whether it should also be considered as INS (the patient cannot tolerate a short walk, i.e. can tolerate only sitting activities).

For BER, it is hard to draw conclusions from the mean agreement score since there were very few examples of this category in the 35 analyzed notes; in total the annotators labeled only 1-5 sentences with this category. Because of the small number of examples, the agreement scores vary greatly between pairs: from 0 to 0.8. For ETN, the source of the low agreement is not entirely clear. This category was not mentioned by the annotators as especially confusing, and the examples that we sampled can only be explained by an oversight; e.g. the sentence *Abdominaal en voeding: Sondevoeding weer herstart* (Abdominal and nutrition: Tube feeding restarted), which was labeled as ETN only by 3 out of the 6 annotators, explicitly mentions tube feeding, which was defined as an ETN indication in the annotation guidelines.

Importantly, almost all observed disagreement for all 9 categories follows from confusion with the no label category. Out of the 206 sentences which were assigned a label by at least one annotator, only 8 sentences contain a confusion between categories. This shows that the annotations mainly differ in coverage and not in interpretation. The 8 examples that do contain "real" confusion include (a) confusion between ETN (Eating) and MBW (Weight maintenance), e.g. in the sentence *Dhr heeft een goede voedingtoestand* (Mr. has a good nutritional status), (b) confusion between ATT (Attention) and ENR (Energy level), e.g. in the sentence *Valt tijdens onderzoek steeds in slaap* (Always falls asleep during examination), (c) confusion between INS and ENR, e.g. in the sentence *Dit kostte dhr veel energie* (This took a lot of energy (the previous sentence describes getting out of the bed)), and (d) confusion between INS and FAC (Walking), e.g. in the sentence *Kon met veel hulp staan onder begeleiding van twee personen* (Could stand with a lot of help, under supervision of two people).

label	measure	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM
category	F1-score	.64	.58	.42	.66	.45	.78	.34	.62	.57
level	MAE	.25	.32	.38	.39	.28	.17	.30	.32	.31

Table 2: Inter-annotator agreement on category and level labels

The MAE scores for the level labels are all far below 1, both for the 0-4 scales and the 0-5 scales. This can be seen as a good score suggesting that when annotators agree on the relevance of a category, they also agree on the level.

4. Data

4.1. Data selection for annotation

The data that was made available for our study consists of millions of clinical notes from the EHRs of the Amsterdam UMC, dating from 2017 to 2020. This included both patients with COVID-19 as well as all kinds of other patients. This is important to make sure that our classifiers are not biased towards a specific type of patients in which a specific recovery patterns dominates. We balanced the data selection across COVID-19 and non-COVID-19 patients.

Each week, a batch for annotation was selected from this large dataset; the batch could be configured with a few parameters, as detailed below. The goal of this selection method was to collect a sufficient number of positive examples well distributed over the 9 categories.

The first parameter is the type of clinical notes to be sampled. EHR's contain a big variety of notes, including e.g. 'consultation', 'progress report', 'letter', 'discharge instructions', etc. Since specific types of notes might be more or less relevant for specific categories, this parameter can be used to influence the distribution of the labels in a batch.

The second parameter concerns a keyword-based search. There was a concern that random sampling would not yield enough positive examples as most texts are primarily not about the functional status. Therefore, a list of keywords for each category was compiled; for each batch, a certain proportion of notes was selected based on a keyword search and the rest were sampled randomly. The keyword search could be configured to focus on one or more specific categories.

The third parameter is the proportion of notes that belong to COVID-19 patients. The 2020 data was split into two subsets: notes of patients with a COVID-19 diagnosis, and notes of patients that do not have a COVID-19 diagnosis. Since the project specifically focused on COVID-19 patients, it was important to have good representation of this population in the data. However, for the purpose of training the classifiers, a diverse sample that includes a wide variety of functional levels was needed. Therefore, it was undesirable to sample only COVID-19 data, which is probably biased towards specific categories and levels.

In the first batches, the notes were sampled from all note types, 50% of the notes were related to COVID-19 patients, and 80% of the notes were selected with keywords (of all 9 categories). After the first 4 weeks of annotation, the annotated data was analyzed, with the following findings: (a) the ADM category is very dominant (41% of the labels), especially in the COVID-19 data (49% of the labels), (b) the ATT, BER and MBW categories are very rare (2-4% of the labels).

Based on these findings, in the following weeks we experimented with different configurations of the parameters, aiming to balance the distribution of the categories in our data. These experiments were partially successful, but the final dataset still suffers from some imbalance, as detailed in subsection 4.2.

To assess the effect of the keyword-based data selection procedure we compared notes that were selected with keywords with notes that were selected randomly. Both types of notes contain the same percentage of sentences with labels; this means that the keyword-based selection does not actually contribute to obtaining more labeled sentences (i.e. positive examples), which was the initial goal of implementing this method. However, the method does seem to help with obtaining more labels for certain categories, specifically ATT, MBW and STM; the proportion of labels for these three categories is slightly higher in keyword-selected notes than in randomly-selected notes.

4.2. Annotated data

In total, about 6,000 clinical notes were annotated[6] (not including any annotations collected during the pilot phase of the project). 10% were marked 'disregard' and therefore removed from the final dataset (disregarded notes include, for example, notes about children under 12 years old, listings of medications, etc.). The non-disregard notes contain in total about 286,000 sentences; 5% of these sentences contain at least one category label. This means that about 15,000 sentences with category labels were obtained in the current annotation effort.

Despite the attempts to balance the distribution of the labels, as described in subsection 4.1, ADM is still very dominant in the dataset, while ATT and BER are rare. The distribution of the levels within each category is also not completely balanced. For example, for ENR levels 1 and 2 (severe and moderate functional problem, respectively) are dominant, while for FAC level 4 (mild problem) is dominant. For certain categories, the distribution of levels in the COVID-19 data is different from the non-COVID subsets; for example, in the COVID data, there are a lot more 0 and 1 levels (complete and severe problem) for both ADM and INS.

[6]For the 35 notes that were used for the inter-annotator agreement, one annotated version was randomly selected.

4.3. Training, development and test sets

	Total number sentences	Total number notes
train	239,153	6,821
dev	21,742	431
test	22,082	431
total	282,977	7,683

Table 3: Total number of sentences and notes in the training, development and test sets

The annotated data described in subsection 4.2 – both the positive examples (i.e. sentences with labels) and the negative examples (i.e. sentences without labels) – were split into a training set (80% of the sentences), a development set (10% of the sentences), and a test set (10% of the sentences). The development set was used for evaluation during the intermediate experiments; the test set was used for the final evaluation described in section 5.

After the split, the following additional steps were applied: (a) sentences that are labeled as background/target (and do not contain any other labels) were removed from the training set (n=7,548), and (b) positive examples from the pilot phase of the project were added to the training set (n=4,410); in the pilot only 4 out of the 9 categories were annotated: BER, FAC, INS, STM. These actions were performed based on intermediate experiments, which are not discussed here due to space limitations. This resulted in the final datasets shown in Table 3, which were used for training and evaluating the models. In total, the datasets contain about 283,000 sentences belonging to about 7,600 different clinical notes.

		train	dev	test	total
Categories	ADM	4,988	411	775	6,174
	ATT	247	22	39	308
	BER	486	29	54	569
	ENR	989	105	160	1,254
	ETN	2,420	225	382	3,027
	FAC	2,489	119	253	2,861
	INS	1,967	127	287	2,381
	MBW	755	96	125	976
	STM	3,390	147	181	3,718
Levels	ADM	5,233	440	421	6,094
	ATT	251	23	32	306
	BER	216	29	26	271
	ENR	1,005	107	100	1,212
	ETN	2,491	236	183	2,910
	FAC	1,086	124	139	1,349
	INS	1,104	132	136	1,372
	MBW	766	98	60	924
	STM	1,420	148	155	1,723

Table 4: Positive examples (sentences) per category

Table 4 shows the number of positive examples (sentences) per category, both for the category classification model and the levels regression models. For the category model, we have 6,174 positive examples of ADM, 3,718 examples of STM (about half of which originate from the pilot annotations), 3,027 examples of ETN, 2,861 examples of FAC (about half of which originate from the pilot annotations), 2,381 examples of INS (about half of which originate from the pilot annotations), 1,254 examples of ENR, 976 examples of MBW, 569 examples of BER (about half of which originate from the pilot annotations), and 308 examples of ATT.

For the levels models, the pilot annotations were not added because changes in the functioning scales were introduced after the pilot; therefore, the numbers of positive examples for BER, FAC, INS and STM are lower.

5. Classification Models

Medical researchers are interested in the functional status of patients on a note-level rather than a sentence level, since more than one sentence can express the functioning of a patient at a certain moment in time. To accommodate this requirement, we built a NLP pipeline that takes as input a clinical note in Dutch and outputs a note-level ICF functioning score for each of the 9 categories (if relevant).

The first step in the pipeline is anonymizing the note (in order to make it compatible with MedRoBERTa.nl) and splitting it into sentences, which is done with spaCy[7]. The sentences are then sent to the multi-label category classification model, which assigns between 0-9 categories to each sentence; for example, the sentence *Loopt, eet, drinkt, geen dyspnoe* (Walks, eats, drinks, no dyspnea) is assigned 3 categories by the classifier: ADM, ETN, FAC.

Next, sentences that were labeled with a specific category are sent to the category-specific regression model that assigns them a functioning level. For example, the abovementioned sentence goes to 3 regression models – ADM level, ETN level, FAC level – and gets a score from each, e.g. ADM level 4.1, ETN level 3.8, FAC level 4.6.

Finally, all the sentences belonging to the same note are aggregated and a note-level score for each category is calculated. The note-level score is the mean of all the sentence-level scores; for example, if a note contains 3 sentences with ADM levels, the ADM level of the note is the mean of the ADM levels of the sentences.

In the following subsections, we present the classification and regression models in the pipeline and evaluate their performance.

5.1. Category Classification Model

5.1.1. Method

For the detection of the categories mentioned in a sentence, we fine-tuned the MedRoBERTa.nl model (Verkijk and Vossen, 2022) for the task of multi-label classification. This was implemented with the Python library Simple Transformers[8]; the default hyperparameters values were applied: AdamW optimizer, learning rate 4e-5, one training epoch, and a batch size of 8.

		P	R	F1	support
sents	ADM	.98	.49	.66	775
	ATT	.98	.41	.58	39
	BER	.56	.29	.35	54
	ENR	.96	.57	.72	160
	ETN	.92	.49	.63	382
	FAC	.84	.71	.76	253
	INS	.89	.26	.41	287
	MBW	.79	.62	.70	125
	STM	.70	.75	.72	181
notes	ADM	1.0	.89	.94	231
	ATT	1.0	.56	.71	27
	BER	.66	.44	.50	34
	ENR	.96	.70	.81	92
	ETN	.95	.72	.82	165
	FAC	.84	.89	.86	95
	INS	.95	.46	.61	116
	MBW	.87	.87	.87	64
	STM	.80	.87	.84	94

Table 5: Category classification: evaluation on test set, sentence-level and note-level

5.1.2. Results

Table 5 shows the precision, recall and F1-score results per category, both on a note-level and on a sentence-level.[9] The performance for ADM, ENR, ETN, FAC, MBW and STM is good, with a note-level F1-score above 0.8, which is comparable to the results reported by Thieu et al. (2021). The categories ATT, BER and INS, on the other hand, perform poorly, especially in terms of their recall.

For ATT and BER, the low performance is likely to be related to insufficient training examples; the results are very unstable across the two trained models, indicating that a consistent pattern was not identified. For INS, the problem is likely to be with the quality, rather than the quantity, of the training examples. As mentioned in subsection 3.2, the annotators have indicated that this category was difficult to annotate, which is also reflected in a low inter-annotator agreement (F1-score 0.34). The model has difficulty with identifying INS examples mainly because the gold labels for this category are inconsistent.

The results on a sentence-level (the original classification unit) are lower in comparison to the note-level, but the same trends hold. For the 6 categories that do perform well, the precision is relatively high (0.7-0.98) while the recall is relatively low (0.49-0.75), which means that there are many false negatives but not many false positives. The fact that on a note-level the recall recovers to above 0.7 for all 6 categories can be probably attributed to two factors: (a) there is more than one sentence in a note that discusses a specific category and the model manages to detect at least one of these sentences, and/or (b) due to incorrect sentence segmentation (which is not uncommon in clinical notes because of non-standard punctuation), the gold label and the model's label might end up in different segments of the same sentence, which is regarded as an error on a sentence-level but is not problematic on a note-level.

In terms of confusion, the predominant error that the model makes is misclassification into the 'no label' category. The confusion between the different categories is very minimal; the only categories for which there are more than two examples of confusion in the test set are: confusion between ETN and MBW (5 sentences), confusion between INS and FAC (5 sentences), confusion between INS and BER (5 sentences). The first two types of confusion were observed in the gold labels as well, as discussed in subsection 3.2. The third type of confusion (INS and BER) can be explained by the fact that exercise tolerance is sometimes expressed through the ability to work or to travel to work. For example, the fact that someone cycles to work every day is indicative of their level of exercise tolerance; such sentences are annotated as INS, but the model sometimes labels them as BER because of the work-related vocabulary.

Comparing the sentence-level F1-scores in Table 5 to the inter-annotator agreement F1-scores in Table 2, we see that the model's performance on the category assignment task is quite comparable to the human performance.

5.2. Levels Regression Models

5.2.1. Method

For the task of assigning a functioning level to a sentence, we fine-tuned the MedRoBERTa.nl model (Verkijk and Vossen, 2022) for a regression task, which generates a continuous output. A separate regression model for each of the 9 categories was trained. Similarly to the category model, this was implemented with the Simple Transformers library, using the default hyperparameters values.

5.2.2. Results

		MAE	MSE	RMSE	support
sents	ADM	.48	.55	.74	421
	ATT	.99	1.35	1.16	32
	BER	1.56	3.06	1.75	26
	ENR	.48	.49	.70	100
	ETN	.59	.65	.81	183
	FAC*	.70	.91	.95	139
	INS*	.69	.80	.89	136
	MBW	.81	.83	.91	60
	STM	.76	1.03	1.01	155
notes	ADM	.37	.34	.58	200
	ATT	1.03	1.47	1.21	21
	BER	1.49	2.85	1.69	22
	ENR	.43	.42	.65	70
	ETN	.50	.47	.68	123
	FAC*	.66	.93	.96	79
	INS*	.61	.64	.80	74
	MBW	.60	.56	.75	41
	STM	.68	.87	.93	84

Table 6: Levels classification: evaluation on test set, sentence-level and note-level

Table 6 shows the mean absolute error (MAE), mean squared error (MSE) and root mean squared error

[7] https://spacy.io/

[8] https://simpletransformers.ai/

[9] The reported metrics are the mean of two fine-tuned models, identical except for the random initialization.

(RMSE) for each category, calculated both on a note-level (the meaningful unit for the healthcare professionals) and a sentence-level (the original classification unit).[10] The MAE is below 1 for all categories, which is comparable to Kukafka et al. (2006), except for ATT and BER. The lower performance for ATT and BER can be explained by the small number of training examples (about 200 sentences for each, see Table 4).
Comparing the sentence-level MAE in Table 6 to the inter-annotator agreement MAE in Table 2, we see that the model's performance does not reach the human level on this task. To improve the performance, we believe that the models could significantly benefit from additional training data.

6. Example Recovery Trajectory

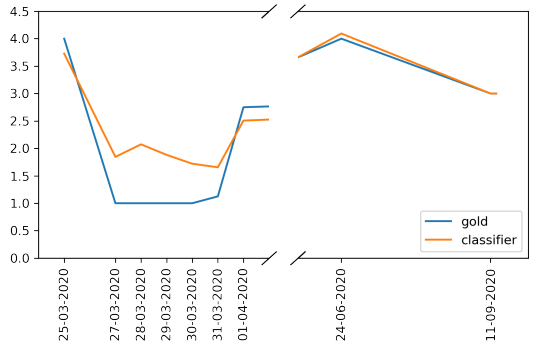

Figure 1: ADM levels of a COVID-19 patient over time

Our NLP pipeline can be used by medical researchers to analyze how the functional status of types of patients develop over time. To illustrate such a use-case, we show the trajectory of respiration functions (ADM) recovery for one COVID-19 patient from our dataset. Figure 1 shows the patient's ADM levels during her hospitalization (from 25-03-2020 to 01-04-2020), and at two outpatient consultations (3 and 5 months after discharge). The gold labels assigned by the annotator are shown in blue, and the labels predicted by the classifier are shown in orange.
The dates shown on the x axis are the actual data points, i.e. dates for which we have notes; these data points are connected by linear interpolation. During the hospitalization, multiple notes per day are available, and the level shown in the graph is the average score across same-day notes. All in all, this patient has 43 clinical notes; 6 notes of type 'letter' were excluded from the current analysis, since letters summarize previous situations, so the date of the note does not reflect the current functional status. Out of the 37 analyzed notes, 28 had a gold ADM label and 29 had a predicted ADM label.
As evident from the graph, the functional trajectory generated from the classifier's predictions is very similar to the gold trajectory. On the first day of hospitalization (25-03), the patient does not have a problem with her respiration (gold: 4, classifier: 3.7), between 27-03 and 31-03 she has a severe respiration problem (gold daily mean: 1-1.1, classifier daily mean: 1.7-2), on the discharge day (01-04) the respiration level recovers to a moderate/mild problem (gold: 2.8, classifier: 2.5). On the first follow-up (24-06) there is no problem with the respiration (gold: 4, classifier: 4.1), and on the second follow-up (11-09) there is a mild problem (gold: 3, classifier: 3). The classifier accurately predicts the higher end of the scale and the overall pattern; the only divergence is that for the days when the patient is at the lower end of the scale, the classifier's predictions are somewhat higher than the gold labels (but still within the same level, i.e. between 1 and 2).
This example shows that our pipeline's performance is sufficient to generate meaningful and reliable trajectories of functional development over time, at least for some of the studied ICF categories. Applying a similar methodology on a big scale can provide medical researchers with a time- and cost-efficient way to get data-driven insights about functioning, disability and recovery.

7. Conclusions

We described a fine-tuned Dutch language model for the medical domain that assigns functional level classifications to Electronic Health Records of COVID-19 patients. Our models are freely available on `https://huggingface.co/cltl` and the code to obtain the data, train and test the models is available on `https://github.com/cltl/a-proof-zonmw`. We showed that our classifier has sufficient performance to generate potentially reliable patient recovery patterns that can be used to search for factors that impact recovery.
In future research, we investigate how to further improve the performance of the classifiers and how well they perform on other medical communication outside the hospital context, such as reports from physiotherapists, dietitians, geriatric rehabilitation centres, general practitioners or personal reporting by patients. Additionally, our method can be applied to other ICF categories and across different languages so that the standardized ICF classification can help to analyse and guide medical practice across the globe, as is the goal of ICF by the World Health Organization.
Finally, we will further investigate how our models can be used within a clinical research context to predict patient's recovery patterns as time lines. Currently the records are classified sentence by sentence, without taking into account the context or the preceding information of a patient.

8. Acknowledgements

This research was partially funded by the NWO Spinoza Project assigned to Piek Vossen (project number SPI 63-260), the Corona Research Fund (project number 2007793 - COVID 19 Textmining) and the

[10] Categories marked with * have 0-5 scales, the rest 0-4.

NWO-ZonMW project Effectiveness of allied healthcare in patients recovering from COVID-19 (project number 80-87700-98-001, Gerichte Ronde COVID Paramedische Zorg). We express our gratitude to the annotators: Stella Avelli, Luca Bos, Joey Katsburg, Jasper Opsomer, Hannah van der Pas, Nienke Swartjes, Quinten Vervaart, and to our technical students Quirine Smit, Gianluca Truda and Bruna Aguiar Guedes. Finally, we want to thank the external medical experts for their comments and advise: Hinke Kruizenga, Edith Cup, and Ron van Heerde.

9. Appendix

Examples 1-3 below show usage of words and expressions in Dutch medical notes. Each word or word combination in bold is not used or does not convey an intended meaning in general Dutch.

1 'Ligt verder in bed en **loopt op** voor po-stoel en toilet.' (*Lies in bed and **walks "up"** for pot chair and toilet.)*

2 'Inspanningsgebonden dyspneu en thoracale druk **verdacht voor** angina pectoris(...)' (*Exercise-related dyspnoea and thoracic pressure **suspected of** angina pectoris(...)*)

3 'Patiënt met **possibele** pulmonale aspergillus met oplopend galactomannan onder vori mono' (*Patient with **possible** pulmonary aspergillus with ascending galactomannan under vori mono*)

10. Bibliographical References

Hripcsak, G. and Rothschild, A. S. (2005). Agreement, the F-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298.

Huang, C., Huang, L., Wang, Y., Li, X., Ren, L., Gu, X., Kang, L., Guo, L., Liu, M., Zhou, X., et al. (2021). 6-month consequences of COVID-19 in patients discharged from hospital: a cohort study. *The Lancet*, 397(10270):220–232.

Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., and Gurevych, I. (2018). The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9.

Kukafka, R., Bales, M. E., Burkhardt, A., and Friedman, C. (2006). Human and automated coding of rehabilitation discharge summaries according to the International Classification of Functioning, Disability, and Health. *Journal of the American Medical Informatics Association*, 13(5):508–515.

Maritz, R., Aronsky, D., and Prodinger, B. (2017). The international classification of functioning, disability and health (icf) in electronic health records. *Applied clinical informatics*, 8(03):964–980.

Newman-Griffis, D., Camacho Maldonado, J., Ho, P.-S., Sacco, M., Jimenez Silva, R., Porcino, J., and Chan, L. (2021). Linking free text documentation of functioning and disability to the ICF with Natural Language Processing. *Frontiers in Rehabilitation Sciences*, page 67.

Thieu, T., Maldonado, J. C., Ho, P.-S., Ding, M., Marr, A., Brandt, D., Newman-Griffis, D., Zirikly, A., Chan, L., and Rasch, E. (2021). A comprehensive study of mobility functioning information in clinical notes: entity hierarchy, corpus annotation, and sequence labeling. *International Journal of Medical Informatics*, 147:104351.

World Health Organization. (2001). *International Classification of Functioning, Disability, and Health*. World Health Organization, Geneva.

11. Language Resource References

Beltagy, I., Lo, K., and Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In *EMNLP/IJCNLP*.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. *ArXiv*, abs/2010.02559.

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., and Nissim, M. (2019). BERTje: A Dutch BERT model. *ArXiv*, abs/1912.09582.

Delobelle, P., Winters, T., and Berendt, B. (2020). RobBERT: a Dutch RoBERTa-based language model. In *EMNLP*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2020). Domain-specific language model pretraining for biomedical natural language processing. *ArXiv*, abs/2007.15779.

Huang, K., Altosaar, J., and Ranganath, R. (2019). ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). BioBERT: a pretrained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234 – 1240.

Verkijk, S. and Vossen, P. (2022). MedRoBERTa.nl: A language model for Dutch electronic health records. *Computational Linguistics in the Netherlands Journal*, 11.