From Hospital to Home: Predicting Post-Discharge Patient Functioning with NLP-based extracted ICF ratings

Stan Frinking

Supervisors:
Sanne Abeln (UU), Edwin Geleijn (AUMC), Marike van der Leeden (AUMC)

Second Reader: Albert Gatt (UU)

July 12, 2024

Abstract

The transition from hospital to home is a critical period for patients often accompanied by complications that can lead to readmission. Effective post-discharge monitoring is essential to enhance patient recovery and reduce healthcare costs. Despite existing monitoring methods, quantitatively analyzing the reports from these methods remains challenging. Recent advancements in Natural Language Processing (NLP) offer new opportunities to extract meaningful information from clinical texts. This thesis explores the use of an NLP model specifically designed to classify and rate mentions of the International Classification of Functioning, Disability, and Health (ICF) from clinical notes to predict post-discharge functioning levels. The research goals are defined in two research questions:

1. How does a professional's judgment of a patient's functioning as extracted from unstructured clinical data correlate with the patient's own perception of their functioning?

2. How effectively can NLP-extracted ICF functioning levels from in-hospital patient notes predict post-discharge functioning as documented in follow-up notes?

The first experiment assesses the correlation between the ICF classifier model's outputs on clinical notes and patients' self-reported functioning levels. We found that for each of the relevant ICF classes except Weight Maintenance there is at least some data representation that yields a significant correlation with the self-reported values. However, discrepancies were noted in specific classes due to infrequent labeling and the nature of the questions. The second experiment evaluates the predictive value of in-hospital functioning levels on post-discharge functioning levels. The constructed prediction models were able to predict post-discharge functioning with a comparable error as the researchers that developed the ICF models reported as performance metrics for their models. This thesis shows that while the extracted functioning data has some predictive value for post-discharge functioning, more efforts are needed to evaluate the used ICF models further to better understand its implications.

Contents

List of Common Abbreviations

NLP - Natural Language Processing

AUMC - Amsterdam University Medical Centers

ICF - International Classification of Functioning, Disability and Health

LSTM - Long Short-term Memory

BERT - Bidirectional Encoder Representations from Transformers

ADM - Respiratory Functioning

STM - Emotional/Mood-related Functioning

ENR - Energy-related Functioning

MBW - Weight Maintenance

Chapter 1

Introduction

1.1 The challenge of post-discharge monitoring

Many patients encounter a variety of problems in the first weeks after being discharged from the hospital (Bull et al., 2000; Cole, 2001; Hyde et al., 2000; Mistiaen et al., 1999; Parker et al., 2002; Shepperd et al., 2004). Patients with post-discharge complications are more likely to be readmitted to the hospital, and early identification of these complications can significantly reduce readmission rates (Tevis and Kennedy, 2014). Effective post-discharge monitoring and support can enhance a patient's capacity for self-care while reducing hospital readmission risk (Leppin et al., 2014).

Inadequate transitional care processes can lead to impaired functional statuses, increased mortality rates, and greater healthcare costs (Burke et al., 2016). McIlvennan et al. (2015) note that preventing readmissions by even 10% could save the United States Medicare system $1 billion annually. In the Netherlands, about 1.2 million people have surgery every year. Between 3-16% of these patients face complications after their surgeries (World Alliance for Patient Safety, 2009; Haynes et al., 2009).

Despite clear benefits, monitoring post-discharge complications can be very challenging. During hospitalization, patients receive frequent assessments of vital signs and symptoms, allowing healthcare providers to respond promptly to changes in patient functioning. After discharge, however, the care burden partially shifts to the patient and their families. Many struggle to handle the complex and evolving demands of managing their conditions without significant support (Brandberg et al., 2021).

Most hospitals already employ post-discharge monitoring methods, including constructing a post-discharge plan of care, documenting discharge summaries, regular telephone follow-ups, and scheduling post-discharge checkup appointments. Mistiaen and Poot (2006) found that telephone follow-ups are not only appreciated by patients; they also facilitate information exchange, symptom management, early recognition of complications, and quality aftercare services. However, the study also described a wide variety of approaches and outcomes for the telephone follow-ups, complicating consistent and automatic analysis.

Similarly, studies have shown that patients who attend their follow-up appointment have a lower chance of being readmitted in 30 days (Coppa et al., 2021; Chang et al., 2012). Patients with follow-up appointments with physicians that have access to discharge summaries also show a trend toward reduced readmissions compared to those without (Walraven et al., 2002).

Before surgery, patients go through a Preoperative Screening (POS) to check their health risks

[1]. Studies have found that lifestyle choices such as smoking, excessive alcohol consumption, and lack of exercise can increase the likelihood of complications and delay recovery. Improving these lifestyle factors before surgery has been shown to lead to better post-operative outcomes.

While these post-discharge monitoring methods have in common that they decrease a patient's chance of being readmitted, they are also of qualitative nature. They involve a physician's professional assessment and are recorded in an unstructured linguistic manner, making additional quantitative analysis challenging.

1.2 Using NLP to understand clinical notes

Recent advancements in Natural Language Processing (NLP) have opened new possibilities for utilizing these unstructured data sources. Techniques such as word embeddings, recurrent neural networks, and Transformers have significantly enhanced our ability to extract and analyze complex medical information from clinical notes.

Word embeddings like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) marked the first step in understanding the relationships between words, enabling models to capture semantic similarities. With recurrent neural networks (RNNs), particularly LSTMs (Hochreiter and Schmidhuber, 1997), models were able to retain and interpret the contextual meaning of words across longer text sequences, providing more accurate analyses. Transformer models (Vaswani et al., 2017) built on these developments by incorporating an attention mechanism which prioritizes important parts of the text while allowing computations to be parallelized and by using positional embeddings to maintain sequence order without sequential processing constraints.

Transformers have shown particular promise in healthcare, where they are used to predict medical diagnoses, analyze patient outcomes, and stratify patient risks based on vast amounts of unstructured clinical notes. In the context of this transformative capability, the ICF classification model developed by Kim et al. (2022) at Amsterdam University Medical Centers (AUMC) is a relevant example. This model, leveraging a Transformer architecture, was specifically trained to predict relevant classes from the WHO's International Classification of Functioning, Health, and Disability (ICF), focusing on conditions critical for COVID-19 related research such as exercise tolerance, mobility, and respiratory functions.

The ICF classification model operates by analyzing each clinical note to determine the presence of specific ICF classes and assessing the qualifier level of mentioned conditions. This process effectively transforms unstructured textual data into a structured format of classes and severity levels, facilitating trend analysis of patient functioning over time. The model's ability to track these trends not only during hospitalization but also through post-discharge follow-ups—using notes from telephone follow-ups and checkup appointments—makes it a powerful tool for quantitatively analyzing the results from continuous patient monitoring. In this project, the ICF classifier models are used to construct functioning trends over time to predict functioning levels after discharge.

1.3 The difference between a patient's and a professional's perception of functioning

The ICF classification model leverages clinical notes annotated by healthcare professionals, reflecting professional judgments of patient functioning during and post-hospitalization. However,

[1]Richtlijn "Het Preoperatieve Traject". Netherlands Society of Anesthesiologists, 2010

discrepancies often exist between clinician-assessed and patient self-reported functional statuses, as evidenced in various studies (Feuering et al., 2014; Sabbag et al., 2012). These discrepancies highlight the potential for bias or misunderstanding in professional assessments or patient self-evaluations.

This thesis not only aims to explore correlations between functioning statuses recorded during hospitalization and those post-discharge but also intends to examine the alignment between the ICF classification results and patient self-reports. Such comparisons are crucial for validating the clinical utility of the ICF model and enhancing the accuracy of patient care strategies. Since this is purely observational and not interventional, we do not expect this alignment to be perfect. The constructs to which we compare the ICF classification results were not originally intended for this purpose, and discrepancies between timing of self-reporting and professional reporting, as well as differences in what is being measured, are expected to affect the potential alignment.

In 2021, a small-scale study at the AUMC hospital in Amsterdam was conducted using questionnaires to map patients' rehabilitation progress after elective surgery; the 'Beter Voorbereid' (Be Prepared) application (Velde et al., 2021). This study aimed to understand recovery by asking specific questions related to patients' daily activities and lifestyle changes. These questions focused on:

Preoperative risk assessment (on lifestyle risk factors)

Preoperative patient education (on increasing physical activity, increasing muscle strength, smoking cessation, alcohol intake reduction, and dietary changes)

Preoperative exercise therapy for high-risk patients (prehabilitation).

The aim of this study was to provide insights into whether tailored lifestyle advice to patients undergoing surgery is feasible and effective in improving postoperative functional recovery. They assessed this by randomizing the patients among a control and intervention group, with the intervention being given advice to reduce lifestyle risk factors. Both groups were asked to fill in the questionnaires to keep track of the self-reported functional rehabilitation.

While the "Beter Voorbereid" application showed potential in terms of usability and changing risk behavior prior to major surgery, no preliminary effect of the app on functional recovery was found. However, in order to test these effects, the research group gathered a lot of data on the functional recoveries of patients after elective surgery in general. This is still very valuable data that could be used for correlational and predictive analyses. Detailed information about the dataset will be discussed in the methodology section, outlining how these elements contribute to validating the ICF classification outcomes.

Recognizing these discrepancies between a professional's assessment and a patient's own experiences is pivotal, as it underscores the complexity of accurately assessing patient functioning from clinical notes alone. This thesis aims to bridge this gap by employing advanced NLP techniques to analyze and predict functioning levels, comparing these predictions with patient self-reports to validate the accuracy and utility of the ICF classification model.

1.4 Research objectives and questions

In the introduction of this thesis, we discussed the relevance of post-discharge monitoring, the limitations of textual reporting making them hard to quantitatively analyze, and how NLP can potentially structure this information for analysis purposes. We introduced a previously developed

NLP model that extracts ICF class mentions and severity levels and we discussed how this can be used to predict post-discharge functioning. Furthermore, we also discussed how we can clinically validate the utility of the output by this model by making a comparison to existing self-reported functioning data.

This thesis explores the relationship between the functioning levels reported in clinical notes during hospitalization and those documented in post-discharge notes. It aims to assess the predictive value of in-hospital functioning trends on post-discharge outcomes, enhancing the efficacy of post-discharge monitoring. By identifying functioning classes that may require heightened attention due to predicted risks, this research seeks to improve continuity of care and support targeted interventions, ultimately aiming to improve patient recovery trajectories.

As such, the primary research question is:

> How effectively can NLP-extracted ICF functioning levels from in-hospital patient notes predict post-discharge functioning as documented in follow-up notes?

The methodology section will go in-depth on the construction of the ICF classifier models and how methodological choices might affect the quality of downstream tasks such as the one introduced in this thesis. While these choices can affect the generalizability of the results, since the main research question is about predicting post-discharge ICF results based on the ICF results during hospitalization, any systematic mistakes the model makes will be present both during hospitalization as well as after discharge. As such, we hypothesize that this experiment can show good results.

Additionally, we want to explore to what extent the extracted ICF classes relate to a patient's own perception of their functioning. This is formulated as the secondary research question:

> How does a professional's judgment of a patient's functioning as extracted from unstructured clinical data correlate with the patient's own perception of their functioning?

Contrary to the hypothesis for research question 1, here any systematic inconsistencies in the model can affect the potential upper bound quality of the analysis, so we hypothesize that this experiment could warrant suboptimal results.

Additionally, since this project is a continuation of previous research done at the AUMC hospital, we will discuss obstacles and potential solutions that are connected to downstream modeling tasks. Since this is not the main research objective, it will be discussed throughout the thesis and especially in the discussion section. This might not be academically as relevant as the experiments, but it will be useful for hospitals to circumvent any of the challenges we encountered during this research.

Chapter 2

Literature Review

This literature review explores the role of Natural Language Processing (NLP) in healthcare, particularly through its integration with predictive modeling to enhance patient care and outcomes. NLP sits at the intersection of computer science, artificial intelligence, and linguistics, providing tools to decode the complexities of human language. In healthcare, these tools offer unprecedented opportunities to extract, structure, and analyze clinical data from unstructured text sources like patient records and notes. This review not only outlines the evolution of relevant NLP techniques but also illustrates their current applications in healthcare and how they support predictive modeling to better anticipate patient needs and improve treatment outcomes post-discharge.

2.1 Natural Language Processing

Natural Language Processing (NLP) enables powerful interactions between computers and human language by employing various computational techniques. Fundamental to NLP are methods like tokenization, which breaks text into manageable pieces for analysis, and embeddings, which translate these pieces into numerical forms that capture their underlying meanings. Furthermore, advancements in sequence models like Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs) are pivotal for understanding and predicting textual context, making them valuable in handling complex linguistic data in medical documents.

2.1.1 Word Embeddings

Word embeddings represent a significant advancement in the way machines understand human language. Embeddings map words into a continuous vector space where semantic similarities between words are reflected by their closeness in the space. This concept was popularized by breakthrough models such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) which efficiently encode linguistic patterns and relationships from large text corpora.

Word2Vec, developed by Mikolov et al. (2013) at Google, uses shallow neural networks to produce embeddings that capture complex word relationships such as analogies, demonstrating that certain linguistic intuitions can be represented algebraically. It offers two architectures: Skip-Gram and Continuous Bag of Words (CBOW). Skip-Gram predicts surrounding context words given a target word, whereas CBOW predicts a target word from a bag (unordered list) of context words.

GloVe, developed by researchers at Stanford Pennington et al. (2014), extends this idea by effectively leveraging global word-to-word co-occurrence statistics from a corpus to yield embed-

dings. It constructs a co-occurrence matrix (of size words x context) that counts how often things appear in context with each other. The model then learns embeddings by approximating this matrix through least squares fitting, effectively capturing global statistics of the corpus in the embeddings.

Representing text in terms of word embeddings has several benefits. First of all, it reduces the dimensionality of the input because they transform sparse high-dimensional vectors (one-hot encoding) into dense lower-dimensional spaces. As explained above, word embeddings also capture contextual similarity. An embedding for the word "cat" would be more similar to the embedding of "dog" than the embedding of "car" because both words are expected to appear in a similar context, namely that of "pet" even when the words "cat" and "car" are orthographically less different. Finally, once embeddings are trained, they can be used across different tasks without needing retraining, accounting for an increase in the transferability of the trained concepts.

An example of this unique capability is given by Sousa et al. (2020), who explored in different languages how words are represented in word embedding vectors. The example used in all languages is the one of the embedding of the word "king". If you subtract the embedding for "man" from the one for "king", you essentially subtract the gender association. If you then add the embedding for "woman", you shift the gender association to a female context, which brings the resulting vector closer to the one for "queen". This demonstrates how word embeddings encapsulate the meaning of words. While this was not the first research exploring this concept, it includes a comparison between multiple languages, and shows that the same logic applies for all language explored, which shows the generalizability of this concept to other target languages.

2.1.2 Sequence Modeling: From RNNs to LSTMs

Word embeddings have shown to be a powerful representation of language that can be used by computers due to their numerical nature. Several different machine learning models can use these representations, among which are Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs).

RNNs are fundamental in natural language processing for their capability to handle sequential input, such as text or speech. RNNs process sequences by maintaining a memory of previous inputs, allowing them to make predictions based on what they have processed so far. However, RNNs often face challenges with long-term dependencies within text due to vanishing gradients, where the contribution of information decays geometrically over time, making it hard to maintain long-range dependencies.

To overcome these limitations, LSTMs were introduced. LSTMs enhance the basic RNN architecture with a complex system of gates that regulate the flow of information. These gates control the extent to which new input should change the memory and the extent to which the current content of memory should contribute to the output. By selectively updating and retrieving information, LSTMs can maintain long-term dependencies, making them highly effective for tasks that require understanding over longer contexts, such as document classification or conversational models.

2.1.3 Transformers: Revolutionizing Contextual Understanding

The development of Transformers marked a significant evolution in how models process textual data. Unlike RNNs and LSTMs, which process data sequentially, Transformers use self-attention

mechanisms and positional encoding to analyze all words in the input data simultaneously. This architecture allows them to capture contextual relationships between word embeddings in a sentence regardless of their positional distances. Because of this parallel processing capability, Transformers are not only more efficient in handling longer texts but also significantly faster to train. The ability to attend to all parts of the input simultaneously makes Transformers especially powerful for complex NLP tasks that require a nuanced understanding of language context.

2.1.4 BERT: Bidirectional Context Tuning

Building on the success of Transformers, BERT (Bidirectional Encoder Representations from Transformers) introduced a novel approach to training language models (Devlin et al., 2019). BERT is trained on a large corpus of text and then fine-tuned for specific tasks. The bidirectional encoder allows it to understand the context from both left and right sides of a token simultaneously. This deep bidirectional training is particularly effective for tasks that require a precise understanding of language context, such as question answering and language inference. BERT has significantly advanced the state-of-the-art in NLP, providing a new framework for building more nuanced and effective models.

2.1.5 MedRoBERTa

In the same year that BERT was published, Liu et al. (2019) replicated the study and found that BERT was significantly undertrained and would perform even better if the training procedure was optimized. They aggregated their proposed results in a new model, RoBERTa, the Robustly Optimized BERT Approach.

Building on the framework of RoBERTa, the MedRoBERTa model was specifically trained to address the nuances of Dutch clinical data. Developed through a collaboration between the AUMC hospital and VU university, MedRoBERTa was subjected to both transfer learning and training from scratch methodologies, with the latter showing better performance (Verkijk and Vossen, 2021). The model was further fine-tuned to predict the presence four specific classes from the ICF classifier model—Walking, Exercise Tolerance, Emotional Functions, and Work and Employment—to externally validate it and compare various training setups and hyperparameters.

2.1.6 ICF Classifier Model

The next step in the construction of the ICF classifier model is to take the MedRoBERTa pretrained model and finetune it on the required classification and regression tasks (Kim et al., 2022). Since the ICF classifier's purpose was to aid in the identification of COVID patients and their complications, the selection of the ICF classes to use was based on classes related to COVID. As such, a selection of 9 classes was made: Respiration functions, Attention functions, Work and employment, Energy, Eating, Walking, Exercise tolerance, Weight management and Emotional functions.

The International Classification of Functioning, Disability, and Health (ICF), developed by the World Health Organization (WHO), provides a standard language and framework for the description of health and health-related states. The ICF is structured around the following components:

- **Body Functions and Structures:** Physiological functions of body systems and anatomical parts of the body.
- **Activities and Participation:** Execution of tasks or involvement in life situations.

- **Environmental Factors:** Physical, social, and attitudinal environment in which people live and conduct their lives.
- **Personal Factors:** Personal influences on functioning and disability not classified in the ICF but considered in assessments (World Health Organization, 2001, 2013).

Paragraph 3.1 in the Methodology chapter discusses the methodological setup of the ICF classifier project in more detail.

Transitioning from the specific application of the ICF classifier model to the broader context of predictive modeling in healthcare, we can see how integrating sophisticated NLP techniques with predictive analytics can enhance patient care. Predictive modeling plays a critical role in transforming reactive medical practices into proactive health management. This approach allows for better forecasting of patient outcomes, management of post-discharge care, and improvement of rehabilitation strategies.

2.2 Predictive Modeling in Healthcare

Predictive modeling in healthcare plays a critical role in transforming reactive medical practices into proactive health management. This section explores how predictive analytics aids in forecasting patient outcomes, managing post-discharge care, and improving rehabilitation strategies.

2.2.1 Overview of Predictive Modeling Techniques in Healthcare

Predictive modeling utilizes historical data and machine learning techniques to forecast future outcomes. In healthcare, this involves predicting patient risks, disease progression, and potential readmission, which can significantly enhance clinical decision-making and resource allocation. By anticipating future health events, healthcare providers can implement preventative measures, tailor treatments to individual needs, and improve overall patient care quality. Rajkomar et al. (2018) demonstrated that deep learning models could predict a range of clinical outcomes from electronic health records with high accuracy, outperforming traditional models.

Traditional statistical models have been foundational in healthcare predictions but often struggle with the complex non-linear relationships inherent in medical data. These models typically require explicit programming for each variable and interaction, limiting their adaptability and scalability. Furthermore, they struggle with large datasets, high dimensionality, and missing or unstructured data, which are common in medical records. Inouye (2001) illustrates the effectiveness of simpler, more adaptable predictive models in clinical settings, especially for stroke rehabilitation outcomes.

2.2.2 Post-Discharge Monitoring

Traditional methods for post-discharge monitoring, while essential, often fall short due to their reactive nature and inability to capture subtle signs of patient deterioration. The integration of advanced monitoring technologies, such as TeleHealth Electronic Monitoring (THEM) systems, offers a proactive approach. For instance, a study by Mousa et al. (2019) demonstrated the effectiveness of THEM in reducing post-operative complications and improving patient outcomes through real-time data transmission and monitoring.

Another innovative approach is the use of Interactive Voice Response Systems (IVRS), which have been shown to significantly improve patient safety following hospital discharge by enabling better monitoring and quicker response to patient needs (Forster and van Walraven, 2007).

Despite these technological advances, there are barriers to effective post-discharge monitoring and communication. A qualitative study by Brajcich et al. (2021) identified key barriers such as discrepancies in technology access, literacy, and the need for enhanced education on post-discharge care (Brajcich et al., 2021). Addressing these barriers is crucial for the successful implementation of post-discharge monitoring technologies.

The future of post-discharge monitoring looks promising with the advent of virtual healthcare solutions, particularly for chronic conditions like heart failure. These solutions offer comprehensive monitoring capabilities that significantly enhance patient management outside the hospital setting, reducing rehospitalization rates and improving quality of life (Lee et al., 2023).

2.2.3 Predictive Models for Rehabilitation and Readmission

Predictive modeling in rehabilitation is pivotal for optimizing patient recovery trajectories. These models use historical data on treatment outcomes and individual patient characteristics to predict and customize therapy regimens. Such models are instrumental in determining the most effective interventions for each patient, thereby enhancing the potential for successful recovery. For example, research by Whiteneck et al. (2012) illustrates that the volume and type of treatment received, combined with specific patient characteristics, can significantly influence rehabilitation outcomes. These findings underscore the capability of predictive models to tailor rehabilitation efforts to individual needs, thereby improving the effectiveness of therapeutic interventions (Whiteneck et al., 2012).

Models predicting the risk of hospital readmission play a crucial role in post-discharge care, particularly for patients at high risk of readmission. These models assess a range of factors such as medical history, the specifics of treatment received, and conditions of the post-discharge environment to estimate the likelihood of a patient returning to the hospital. Shirkavand et al. (2023) demonstrate how transformer-based models can effectively predict postoperative complications, a key factor in readmissions. Their work highlights the potential of advanced predictive models to not only foresee but also mitigate the risk of readmission, thus significantly contributing to improved patient care and reducing the burden on healthcare facilities (Shirkavand et al., 2023).

The integration of predictive models in healthcare settings, particularly in the domains of rehabilitation and readmission risk management, offers substantial benefits. These models enhance the ability of healthcare providers to make data-driven decisions that are critical for patient care post-discharge. By effectively predicting and addressing potential complications and optimizing rehabilitation processes, predictive models serve as a cornerstone of proactive healthcare strategies, aiming to reduce readmissions and improve overall patient outcomes.

2.3 Integrating NLP with Predictive Modeling in Healthcare

This section examines the intersection of NLP and predictive modeling in healthcare, highlighting how NLP technologies are instrumental in enhancing the predictive capabilities of healthcare systems. It discusses the role of NLP in transforming unstructured clinical data into structured analyzable formats that significantly improve the accuracy and effectiveness of predictive models.

NLP is pivotal in extracting meaningful information from unstructured clinical text such as patient records, doctor's notes, and medical transcripts. By employing techniques like named entity recognition, sentiment analysis, and topic modeling, NLP helps in identifying critical information that forms the basis for predictive analysis (Wu et al., 2015; Kabaev et al., 2023; Huddar et al., 2016). These NLP processes convert unstructured text into structured data that can be easily analyzed by predictive algorithms, enhancing their ability to forecast health outcomes based on comprehensive data sets.

Lineback et al. (2021) tested three NLP techniques for feature selection in predicting unplanned readmission within 30 days of discharge. They constructed patient-word co-occurrence matrices for both unigrams and bigrams in the clinical notes. They tested predictive models using differently aggregated data from the clinical notes, such as tf-idf metrics calculated from the co-occurrence matrices, the same co-occurrence matrix with Principal Component Analysis (PCA) applied to extract the principal components that describe the most variance in the dataset, and finally a Word2Vec word embedding model to represent the data in the notes.

These numerical representations were used in combination with different machine learning models such as logistic regression, Naïve Bayes modeling, support vector machines, random forests, gradient boosting machines, and extreme gradient boosting techniques (XGBoost). Additionally, they employed a method called stacking, where they combined the generated outputs of a base ML model with the existing set of features to train a second model. They found that this stacked classifier ensemble model performed best.

Wang et al. (2022) evaluated multiple similar research papers where ML was used for the prediction of readmission risks. They identified strengths and weaknesses of these employed techniques, as well as the strengths of different types of features, rebalancing techniques, the interpretability of different models, and the challenges of taking these solutions into production. They also discussed how embeddings can effectively capture high-dimensional features and mentioned attempts at constructing a word embedding model similar to Word2Vec: Med2Vec.

Huddar et al. (2016) used NLP-extracted features from clinical notes to predict complications for patients in critical care. They identified that the mention of a medical term can have different meanings depending on the part of the note it was present in. For example, the word "blood" can have different meanings depending on whether it's part of the note describing cardiovascular complications or the genitourinary part of the note. They created a supervised topic modeling system using Latent Dirichlet Allocation (LDA) techniques. The target labels were extracted from the text using a rule-based system after identifying a standardized structure in their dataset, which was selected from the open-source MIMIC-III dataset. This renders their approach very specific to their dataset and hinders the generalizability of the approach. Huddar et al. combined the text-extracted features with structured data using a multi-view learning method, which essentially means they created co-occurrence matrices for the structured data and the extracted text features. This feature representation was used to train multiple machine learning models: logistic regression, support vector machines, decision trees, AdaBoost, and Random Forest.

The last publication that outlines the relevance of this thesis is written by Murff et al. (2011): *Automated Identification of Postoperative Complications Within an Electronic Medical Record Using Natural Language Processing*. While this publication is from before recent advancements in AI and NLP in particular, it is very close to the research objective described in this thesis. This paper describes the use of rule-based search queries to map clinical notes to SNOMED-CT concepts (Systematized Nomenclature of Medicine Clinical Terms). They compared how these extracted concepts differ in their predictivity of postoperative complications compared to traditional patient safety indicators assigned to the dataset (VASQIP) by previous studies.

In conclusion, while there is a lot of research in the topic of representing unstructured textual clinical data in a way that it can be used as features in predictive ML models, it is also apparent that most applied methods are relatively traditional machine learning methods that generally do not capture the nuanced non-linear relationships between complex medical terms. The approach described in the next chapter will differ from those methods as it employs a data source that was preprocessed using complex deep neural networks.

Chapter 3

Methodology

This thesis consists of two experiments. The first experiment assesses the correlation between the previously developed ICF classifier model's output on clinical notes and patients' self-reported functioning levels. The second experiment evaluates the predictive value of in-hospital functioning levels on post-discharge functioning. Since this research project continues on work done previously at the AUMC, this section starts by explaining the methodology and data of the previous work: the ICF classifier models. Additionally, it explains the chosen methodology for the rest of the thesis.

3.1 Overview Methodology

This thesis consists of two experiments:

Experiment 1

The first experiment serves as a response to the research question: *How does a professional's judgment of a patient's functioning as extracted from unstructured clinical data correlate with the patient's own perception of their functioning?*

We attempt to answer this question by plotting the self-reported functioning data from the BV questionnaires against an average of the extracted ICF labels from the clinical notes written by healthcare professionals. We apply a statistical test to see if these two datasets appear to be correlated and therefore describing the same functioning. The aim of this experiment is to validate the usability of the ICF classifier models in a healthcare environment.

Experiment 1 involves the Beter Voorbereid data which consists of five questionnaires administered at different time points: prior to hospitalization, and one, three, six, and twelve weeks post-discharge. Due to patient variability in completing these questionnaires, mapping responses to specific times and corresponding clinical notes is challenging. We selected relevant questions for each ICF class and manually extracted the corresponding time periods.

Three scenarios were identified, each requiring a specific approach. They are described in table 3.1

Scenario	Approach
Lowest Score in Specific Period	Selecting the lowest ICF level predicted during a specified period for weight management questions.
Day-Specific Questions	Average the 2 closest ratings.
Repeated Day-Specific or Period-Specific Questions	Weighted average in same period. Analyzing trends over time using delta changes in ratings.

Table 3.1: Three identified scenarios based on questionnaire constructs and the approach to select corresponding ICF data

Kendall's Tau-b is used to analyze the relationship between ICF outputs and self-reported data because of the ordinal nature of the self-reported data.

Experiment 2

The aim of the second experiment is to explore to what extent this now validated ICF model output can be used for predictive modeling. We use the ICF model output assigned to notes during a patient's hospitalization as features for a model that predicts an average ICF rating assigned to post-discharge notes. This experiment is the basis for the answer to the main research question be posed in chapter 1: How effectively can NLP-extracted ICF functioning levels from in-hospital patient notes predict post-discharge functioning as documented in follow-up notes?

For experiment 2, we selected 1,037 patients, focusing on the ICF ratings assigned to follow-up appointments around six weeks post-discharge. We test two methods of aggregating the target labels: a weighted average approach that targets all ICF predictions between four and eight weeks post-discharge and a nearest two neighbors approach that only looks at the two closest measurements to the 6 week target.

Given the variability in patient lengths of stay, LSTM models were chosen for their ability to handle variable-length sequences and capture temporal patterns. Missing values are represented as a separate feature to preserve information, and structured data from the EHR (age, gender, BMI, etc.) is included to test whether those features affect model performance.

We employ Mean Absolute Error (MAE) to evaluate model performance. We compare the performance of models with and without ICF ratings (baseline test) to evaluate the added predictive value of the ICF data. In an error analysis we will identify model mistakes and areas for improvement.

3.2 ICF Classifier

This section outlines the methodology that was employed for the construction of the ICF classifier models, integrating insights from several research efforts conducted previously at the AUMC

hospital and Vrije Universiteit Amsterdam (Verkijk and Vossen, 2021; Kim et al., 2022). The development of the ICF classifier models involved multiple phases: the construction of a specialized medical language model, gathering and annotating data, and the development of the classifier. We added the important methodological aspects of these previous works to the methodology of this thesis because they directly affect the results we gathered.

3.2.1 Language Model Construction

The initial step in developing the ICF classifier involved creating a language model tailored to Dutch medical texts. The MedRoBERTa.nl model was specifically developed for this purpose (Verkijk and Vossen, 2021).

The research team conducted an experiment to determine the most effective approach for training the language model. They compared two methodologies:

> **Training from Scratch:** The MedRoBERTa.nl model was trained from the ground up using millions of Dutch clinical notes. This approach involved initializing the model with random weights and employing a specialized vocabulary designed to capture the nuances of medical terminology.
>
> **Pretrained Model with Frozen Layers:** Alternatively, the team used a general Dutch language model RobBERT and extended its training on the medical corpus. In this method, certain layers of the pretrained model were frozen, meaning their parameters were not updated during training, while other layers were fine-tuned on the medical text data.

There were two different evaluation methods employed to assess the quality of the language model. The first one, referred to as intrinsic validation, assessed the quality of the model by comparing similarity metrics between texts as computed through the word embeddings with a professional's similarity assessment. It is important to note that text similarity was defined in this research with respect to ICF classes. Verkijk and Vossen (2021) used a set of triples of phrases from clinical notes that may or may not contain mentions of specific ICF classes. One of the three sentences was defined as being most dissimilar to the others. The triples were categorized into one of four groups reported in 3.2.

Description
All sentences belonging in the same ICF domain but two of them contain overlapping keywords.
Two sentences from the same ICF domain without overlapping keywords, one from another domain.
Two sentences from the same ICF domain with overlapping keywords, one from another domain.
All sentences from the same ICF domain but only one differs in terms of the assigned qualifier level.

Table 3.2: Four different triple sets used for extrinsic evaluation of the Language Model MedRoBERTa.nl

The extrinsic evaluation method included two tasks: a scaled-down version of the ICF classification task that contained 4 classes, and a general Named Entity Recognition (NER) task on a

pre-existing dataset [1].

The results from the intrinsic evaluation showed that the transfer learning approach on the RobBERT architecture performs best, albeit by a few percent. The researchers defined the best model as the one with the highest accuracy but did not account for sample size by not employing a statistical test [2].

For the extrinsic evaluation, a p-value showing a significant relation was reported but only for the ICF class where the model trained from scratch outperformed the pre-existing models: Walking (FAC). It was not mentioned in the publication based on which statistical test this was computed nor what assumptions were tested. The other three ICF categories were identified with a higher accuracy by the non-finetuned model BERTje (De Vries et al., 2019).

3.2.2 Data Collection and Annotation

This section explains the data collection, annotation, and evaluation processes for training the ICF classifier models that were trained previously at the AUMC hospital (Kim et al., 2022). It focuses on the methodology the researchers used, highlighting how these considerations impact our research and downstream modeling tasks. Chapter 2.1.6 discussed the academic relevance and background of the ICF classes; here we address the methodological considerations.

The process of mapping natural language descriptions from clinical notes into the ICF framework is traditionally known as 'ICF coding'. This task involves assigning an ICF category and a qualifier to each relevant mention within the clinical texts. For example, "The patient was able to eat independently" would be coded as "d550.00" where "d550" denotes the 'Eating' category and "00" signifies no impairment. For ICF annotation task, the researchers reversed the original ICF scale, with 0 signifying the lowest functioning status and 4 or 5 denoting no impairment. In this thesis, we use the same reversed scale for analysis, where higher levels mean better functioning.

The annotation process was conducted by six native Dutch-speaking (para)medical students to ensure accurate understanding and interpretation of the clinical data. These annotators identified mentions of the selected ICF classes within clinical notes and assigned the appropriate qualifier levels. Each ICF class features five qualifier levels (0 to 4) except 'Walking' and 'Exercise Tolerance' which have six levels to cover a broader spectrum of functionality.

To ensure high annotation accuracy, the annotators received initial training from a core team of healthcare professionals and NLP experts. This team provided essential background and technical knowledge. Additionally, weekly mentoring sessions were held throughout the annotation period to address ongoing challenges and complex cases, allowing the annotators to continuously learn and adapt.

The annotation process involved two primary tasks: identifying phrases that describe an ICF category and labeling phrases that depict the level of functioning. For example, in the sentence "Concentratie is nog wel iets verminderd" (Concentration is still slightly diminished), "concentratie" (concentration) is labeled with the category code ATT (Attention). Simultaneously, "iets verminderd" (slightly diminished) is labeled with the level code att-3, indicating a mild functioning

[1] The distribution of sentences across training and test sets was reported and this shows a discrepancy. The difference in the amount of sentences per note was significant between the test and training set and since no details were shared on how those were gathered, it's difficult to assess what effect this might have on downstream tasks (chi-squared test with df = 7 shows a p-value of 2×10^{-273}).

[2] Based on the support values provided in the report, we recalculated the difference between models while contributing for sample size and found none of the models statistically outperformed the others: p = 0.5, chi = 44.37, df = 45).

problem. Which qualifier level to assign to which class was outlined in the annotation guidelines. These levels were of categorical nature. For instance, for Respiratory Functioning (ADM), a qualifier level of 0 means the patient was intubated for additional oxygen.

An important aspect of the annotation protocol is the exclusion of mentions referring to past or predicted future functioning. This ensures that the model focuses solely on the current state of functioning, avoiding extracting mentions in historical or predictive contexts. Consequently, during model evaluation, the computed performance metrics do not account for how the model handles mentions of functioning from other timeframes [3].

To ensure the accuracy and consistency of the annotations, an inter-annotator agreement (IAA) was calculated. This process involves multiple annotators working on the same texts, which allows for a comparative analysis of their annotations. By evaluating the similarity in annotations across different annotators, the IAA provides a quantitative measure of annotation reliability.

For the ICF classifier model, two types of IAA metrics were employed:

F1-score: Used for the annotation task for identifying classes. The F1-score is the harmonic mean of precision and recall, providing a balance between the completeness and accuracy of the classification.

Mean Absolute Error (MAE): Used to assess quality of assigned qualifier levels. MAE measures the average magnitude of errors in the annotations, giving insights into the typical deviation seen in qualifier assignments[4].

A total of 206 sentences from 35 clinical notes were selected as an overlapping set for the computation of IAA [5]. This subset was used to assess both the class allocation and qualifier level allocation accuracy, though it is not specified whether the same sentences were used for both calculations. Table 3.3 below shows the IAA metrics.

ICF Class	F1-score	MAE
FAC: Mobility	0.78	0.17
ADM: Respiratory Functions	0.64	0.25
ENR: Energy	0.66	0.39
STM: Mood	0.57	0.31
MBW: Weight Maintenance	0.62	0.32
BER: Work	0.42	0.38
ATT: Attention	0.58	0.32

Table 3.3: Inter-Annotator Agreement Metrics as reported by Kim et al. (2022)

This table shows that while there is some agreement between annotators, many of the classes were inconsistently annotated, especially INS, ETN, BER, and ATT. Consequently, this inconsis-

[3]While these decisions make sense for a research project in isolation, this does complicate how the model can be used for downstream tasks. There is no consistent way to filter out past and future references to ICF class mentions from the clinical notes and since the model was never evaluated on these type of notes, it is uncertain how the models handle these events. As such, this uncertainty gets carried over to the downstream tasks.

[4]To in terpret any metric that relies solely on the size of the error, it is important to also take into account the total range of the potential error. This is not specifically mentioned in table 3.3. All class levels range from 0-4 except for FAC and INS, which can range from 0-5.

[5]It is challenging to assess the representativeness of these 35 selected notes from the total set of 6000 notes. To best interpret the results from this research, it would be beneficial to assess representativeness of the selection by comparing it to a truly random selection. Additionally, distributional and variance information of the datasets is essential to interpret these metrics.

tently annotated data is used during evaluation, which places an upper bound on the maximum performance of subsequent models and downstream tasks.

Given the expectation that relevant ICF classes would appear infrequently in the clinical notes, the annotation strategy did not rely solely on random sampling. Instead, to increase the efficiency of finding relevant cases, a keyword search was implemented. This method was designed to elevate the frequency at which certain ICF classes were found, particularly those that are typically low-frequency such as Attention, Weight Management, and Emotional Functions. Interestingly, comparative analysis between the random samples and keyword-based samples indicated a similar total presence of relevant classes across the datasets, though it did increase the presence in the aforementioned low-frequency categories.

Overall, the annotation effort covered approximately 6000 clinical notes, which included over 286,000 sentences. Approximately 10% of these notes were marked as 'disregard'. Of the remaining notes, 5%—or about 15,000 sentences—contained at least one relevant ICF label. However, the distribution of different qualifier levels among these labeled sentences was not detailed in the report. Table 3.4 below shows the distribution of labeled positive classes across the dataset.

		Train	Dev	Test	Support
Categories	Respiratory (ADM)	4,988	411	775	6,174
	Attention (ATT)	247	22	39	308
	Work and Employment (BER)	486	29	54	569
	Energy (ENR)	989	105	160	1,254
	Eating (ETN)	2,420	225	382	3,027
	Mobility (FAC)	2,489	119	253	2,861
	Exercise Tolerance (INS)	1,967	127	287	2,381
	Weight Management (MBW)	755	96	125	976
	Mood (STM)	3,390	147	181	3,718
Levels	Respiratory functioning (ADM)	5,233	440	421	6,094
	Attention (ATT)	251	23	32	306
	Work and Employment (BER)	216	29	26	271
	Energy (ENR)	1,005	107	100	1,212
	Eating (ETN)	2,491	236	183	2,910
	Mobility (FAC)	1,086	124	139	1,349
	Exercise Tolerance (INS)	1,104	132	136	1,372
	Weight Management (MBW)	766	98	60	924
	Mood (STM)	1,420	148	155	1,723

Table 3.4: Distribution of Categories and Levels across Train set, Dev set and Test set. Including support levels for each category. (Kim et al., 2022)

The table shows how some of the labels appear more frequently in the dataset, creating a more substantial dataset for the model to converge on. The infrequency of certain labels in the training set affects how well the model can converge on the task, and the infrequency of these labels in the development and test sets affect the confidence interval of any of the computed performance metrics.

3.2.3 Model Training and Evaluation

The annotated dataset was divided into training (80%), validation (10%), and test (10%) sets to prepare for the model training phase. Two distinct tasks were set for the MedRoBERTa model:

> **Multilabel Classification Task:** A single instance of MedRoBERTa was fine-tuned to classify sentences based on the presence of any of the 9 relevant ICF labels.
>
> **Regression Tasks for Qualifier Levels:** Separate instances of MedRoBERTa—one for each ICF class—were fine-tuned to perform regression on qualifier levels, adjusting for the degree of impairment or functioning as specified by the annotations.

For both tasks, default hyperparameters were used without modifications, focusing on sentence-level predictions.

The outcomes of these training tasks are detailed in the following tables 3.5 and 3.6, which present the performance metrics for both the classification and regression models as they were reported in the publication (Kim et al., 2022). It is important to note that each modeling task was validated individually on gold data. This approach does not account for potential interdependencies or transferability of errors between the two tasks. For instance, the regression models only evaluate sentences flagged by the classification model, implicitly assuming perfect accuracy in the initial classification stage, which could skew the validation results.

Moreover, since the annotation task was performed on a token level, the class mention and the qualifier level can appear in different places in the text. To split the clinical notes into sentences, a SpaCy package (Honnibal and Montani, 2017) was used, which contains a rule-based system to split sentences. However, due to the non-standard writing style in clinical notes, SpaCy is regularly expected to incorrectly split sentences into chunks. Since the article did not report on any inspection on the sentence splitting task, it is unknown how many class mentions and qualifier levels got separated in the process. Since the proposed system only sends a sentence and the context sentences to the regressor model if and only if the class mention was identified by the multi-label ICF classifier model, it is uncertain how incorrectly split sentences are evaluated in the current pipeline.

Results for both training tasks are reported in the below tables 3.5 and 3.6.

		P	R	F1	Support
Sents	ADM	0.98	0.49	0.66	775
	ATT	0.98	0.41	0.58	39
	BER	0.56	0.29	0.35	54
	ENR	0.96	0.57	0.72	160
	ETN	0.92	0.49	0.63	382
	FAC	0.84	0.71	0.76	253
	INS	0.89	0.26	0.41	287
	MBW	0.79	0.26	0.40	125
	STM	0.70	0.75	0.72	181
Notes	ADM	1.0	0.89	0.94	231
	ATT	1.0	0.56	0.71	27
	BER	0.66	0.44	0.50	34
	ENR	0.96	0.87	0.91	92
	ETN	0.95	0.72	0.82	165
	FAC	0.84	0.89	0.86	95
	INS	0.95	0.46	0.61	116
	MBW	0.87	0.87	0.87	64
	STM	0.80	0.87	0.84	94

Table 3.5: Category classification: evaluation (Precision - P, Recall - R and harmonic mean of P and R - F1) on test set for sentence-level (sents) and note-level (notes). Including support levels for each category (Kim et al., 2022)

		MAE	MSE	RMSE	Support
Sents	ADM	0.48	0.55	0.74	421
	ATT	0.99	1.35	1.16	32
	BER	1.56	3.06	1.75	26
	ENR	0.48	0.49	0.70	100
	ETN	0.59	0.65	0.81	183
	FAC*	0.70	0.91	0.95	139
	INS*	0.69	0.80	0.89	136
	MBW	0.81	0.83	0.91	60
	STM	0.76	1.03	1.01	155
Notes	ADM	0.37	0.34	0.58	200
	ATT	1.03	1.47	1.21	21
	BER	1.49	2.85	1.69	22
	ENR	0.43	0.42	0.65	70
	ETN	0.50	0.47	0.68	123
	FAC*	0.66	0.93	0.96	79
	INS*	0.61	0.64	0.80	74
	MBW	0.60	0.56	0.75	41
	STM	0.68	0.87	0.93	84

Table 3.6: Levels regression: evaluation (Mean Absolute Error - MAE, Mean Squared Error - MSE, Root Mean Squared Error - RMSE) on test set for sentence-level (sents) and note-level (notes). Including support levels for each category (Kim et al., 2022)

Table 3.5 highlights that certain classes, notably ATT (Attention) and BER (Work and Employment), have low support values, indicating a small number of positive samples for evaluation. Reporting on the sample size alongside precision and recall metrics is crucial for assessing the statistical significance of these performance metrics.

Table 3.6 illustrates the use of Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) to evaluate the regression models. Although the annotation of qualifier levels followed the ICF practical guide with discrete categories, using MAE, MSE, and RMSE assumes continuous data. For instance, a prediction of 3.2 with a target of 4 incurs a lower loss than a prediction of 2.8, even though both round to the same discrete category. This raises challenges in assessing the significance of decimal predictions. Therefore, we recommend retraining and evaluating the model using discrete levels to compare methodologies in a follow-up research. In order to facilitate intuitive comparisons of error rates between the models proposed in this thesis and the ICF classifier models used as input, we employ the same performance metric.

It is also important to note that performance metrics improve when the model is evaluated on a note level. Although the model was not retrained for note-level prediction, sentence-level predictions were averaged out, as well were the annotations. This averaging seems to improve output quality but reduces sample size which increases the standard error. Averaging predicted and annotated values may lose original resolution and information provided by outliers.

Evaluating at the note level instead of the sentence level also rebalances the dataset. For instance, ADM sentences appear 775 times in the 22,802 test set sentences (3.3%), but in 231 out of 431 notes (54%). This can inflate precision and recall by increasing the proportion of positive samples, thereby reducing the impact of false positives and negatives. The improved performance metrics may reflect this rebalance rather than actual model improvement. Without the exact

distribution of true and false positives and negatives, the resampling effect on the performance metrics is challenging to quantify.

The improvement in performance metrics might indicate that the original class imbalance affected model performance. In a follow-up experiment, retraining the model on a sentence-level classification and regression task using a rebalanced dataset could clarify the impact of the original class imbalance.

Evaluating the methodology and results of the ICF research is crucial for two reasons. First, it helps identify methodological decisions that might affect previous and current research outcomes. Second, it allows us to determine which ICF classes are most relevant to focus on and which may generate labels of inadequate quality.

Due to the low inter-annotator agreement (IAA) values for the INS functioning class, poor F1 scores on the test set, and the inconsistency of the test set results being higher than the IAA scores[6], we decided to exclude this class from all further analyses. As for the other ICF classes, while the IAA metrics for those classes are only slightly better, for now we decide to keep those in the analysis.

3.3 Experiment 1: Correlation Analysis between Beter Voorbereid Data and ICF Outputs

The aim of experiment 1 is to assess the correlation between the Beter Voorbereid questionnaire responses and the ICF classifier models output on notes written around the same date as the questionnaire was filled in. This is especially important because there were methodological decisions made by the previous researchers that complicate downstream modeling tasks, such as not evaluating the models in pipeline, not providing enough distributional information on their datasets and evaluating the ordinal discrete levels on a continuous scale. These decisions make it more difficult to assess what the generated labels really mean in the context of functioning over time. Since the focus of this project is to explore to what extent the ICF labels during hospitalization can confidently predict what ICF labels would have been given to post-discharge clinical notes, we need to quantify what these labels mean for a patient. We do this by comparing the ICF output with existing questionnaire-based self-reported patient functioning data that was gathered in the Beter Voorbereid (BV) project.

We investigate whether there is a correlation between the self-reported functioning levels in the Beter Voorbereid Dataset and the ICF outputs. This serves as an external validation, exploring how the model relates to external datasets on the same topic. This experiment, referred to as "Experiment 1," is a prerequisite for the predictive modeling task.

3.3.1 Data Description Beter Voorbereid

The BV project consists of five questionnaires. The first is a baseline measurement filled in prior to hospitalization. The subsequent questionnaires were administered one week, three weeks, six weeks, and twelve weeks after discharge, respectively. However, patients had the flexibility to decide when to complete these questionnaires. Additionally, patients often took several days to complete each

[6]While it is not impossible for a model to reach higher performance scores than the IAA score for that class, it would still be useful to explore the effect of inconsistent annotations on the model training and evaluation further. One way to test the effect of different annotators on the model performance is to stratify the test set on the annotators and compare the model output quality between annotators.

questionnaire, resulting in varying start and end dates. This variability complicates mapping the questionnaire responses to specific moments in time and consequently to corresponding clinical notes. Since most questionnaires focus on the post-discharge period, which coincides with less frequent entries in the Electronic Health Records (EHR), this data mismatch further complicates the analysis.

Since this dataset was originally gathered for different purposes, the constructs measured in the Beter Voorbereid questionnaires do not directly correspond to ICF classes. Some ICF classes have multiple related questions, while others have none. Therefore, to accurately relate the ICF labels to the BV questions, we need to analyze each individually. Tables 3.7 and 3.8 below shows the relevant questions for each ICF class except INS.

ICF Class	**Relevant Question**	**Answer Options**
FAC: Mobility	"Please select one option for each of the groups in the list below that contains a sentence that best describes your health situation today. Mobility:"	-I don't have difficulties walking. -I have some difficulties walking. -I am bedridden.
ADM: Respiratory Functions	"Did you - during your hospitalization - experience any issues with your lungs (such as coughing and/or shortness of breath)?"	-Not at all -A little -Moderately -Quite a lot -Extremely much
ENR: Energy	"How would you rate your average fatigue?" (Hoe zou u gemiddeld uw vermoeidheid beoordelen?)	-No fatigue. -Light fatigue. -Medium fatigue. -Heavy fatigue. -Extremely heavy fatigue.
STM: Mood	"Please select one option for each of the groups in the list below that contains a sentence that best describes your health situation today. Mood:" (Wilt u bij iedere groep in de lijst hieronder één mogelijkheid aanklikken voor de zin die het best past bij uw eigen gezondheidstoestand van vandaag. Stemming:)	-I am not anxious or feeling down. -I am somewhat anxious or feeling down. -I am very anxious or feeling down.

Table 3.7: Relevant Questions and Answer Options for Each ICF Class (Part 1)

ICF Class	Relevant Question	Answer Options
MBW: Weight Maintenance	"Did you unintentionally lose more than 3kg weight in the last month?" "Did you unintentionally lose more than 6kg in the last 6 months?"	-Yes -No
ETN: Eating	None	None
BER: Work	A dynamic set of questions according to the PROMISS Itembank "Social Score" (Cella et al., 2007)	A T-score ranging from 0 to 100.
ATT: Attention	"Did you generally feel confused during your hospitalization?"	-Never -Rarely -Sometimes -Quite often -Extremely often

Table 3.8: Relevant Questions and Answer Options for Each ICF Class (Part 2)

The ICF category ETN Eating (ETN) did not have any corresponding questions in the BV project, so we also ignored this class for further analysis.

	Number Participants
Started as participant	144
Stopped prematurely	49
Had no clinical notes post-discharge	21
Remaining number of participants	**74**

Table 3.9: Number of Participants in the BV study

Table 3.9 reports on the amount of partipants that took part in the BV study. A total of 144 patients participated in the Beter Voorbereid project. 49 did not complete the project for various reasons. Of the 95 patients who finished, 21 did not have any clinical notes in the relevant post-discharge period. Therefore, 74 patients were selected for Experiment 1.

As explained previously and shown in table 3.7 and 3.8, the questions in the Beter Voorbereid questionnaires do not always refer to the same period. Therefore, for each question, we manually extracted the relevant time period. For the questionnaire dates, we chose the completion date as the anchor point. Additionally, because we do not always have a relevant clinical note with the correct class mention on the exact date the question was asked, we averaged the two nearest neighboring ratings around the relevant date for all cases where questions refer to a single moment in time. Table 3.10 shows the extracted timeframes for each question and the questionnaire number, indicated with T0 to T4, in which the question appeared. It only includes the 5 ICF classes we

considered usable for this thesis, because we identified those as best-performing and of highest quality. We illustrate this decision in section 4.1: Data Analysis.

ICF Class	Relevant Timeframes	Questionnaire Number (indicating time point)
FAC: Mobility	Day of question + 2-week window	T0, T1, T2, T3, T4
ADM: Respiratory Functions	Hospitalization period	T1
ENR: Energy	Day of question + 2-week window	T0, T1, T2, T3, T4
STM: Mood	Day of question + 2-week window	T0, T1, T2, T3, T4
MBW: Weight Maintenance	1 month prior to day of question; 6 months prior to day of question	T0, T4

Table 3.10: Extracted Timeframes for Relevant Questions and the questionnaire numbers - indicated with T0 - T4 - in which they appear

3.3.2 Statistical Correlation Metrics and Data Aggregation

To effectively analyze the correlation between the ICF results and the questionnaire responses, we must account for differences in periods, frequencies, and constructs measured by the questions. We have identified three distinct cases, each requiring a specific approach:

Case 1: Lowest Score in Specific Period

Questions about weight management pertain to any experience of losing a specified amount of weight within a certain period (e.g., 3kg/6kg, 1 month/6 months). A patient might have lost the weight at some point during the period but regained it later. The response would still be "yes" as the weight loss happened at some point. Therefore, for this construct, we select the lowest ICF class predicted for any of the notes during the relevant period.

Case 2: Hospital Period Questions

One of the questions in the questionnaire was about respiratory functioning, and this question was only present at timepoint T1, while inquiring the patient about their experience during hospitalization. This is a single question referring to a period in time. Therefore, we aggregate the ICF ratings into a single rating by taking an average of all relevant ICF ratings during hospitalization.

Case 3: Repeated Day-Specific or Period-Specific Questions

The last case we identified is one where the same question is being repeated throughout the BV study, once for each questionnaire. The questions either refer to the day itself or to a period (i.e. last week). For the question about the same day, we aggregate the ICF ratings by taking the average of two nearest ratings and for the questions about a period in time, we aggregate the ICF information by taking an average over the whole period.

In addition to identifying the methods to aggregate ICF qualifier levels into a single rating that represents the same functioning level as the questionnaire responses, it is also essential to

look at the BV questions themselves and compare them with the ICF classes. This would help us hypothesize about the relationship between the BV rating and the ICF ratings, which is required for finding a suitable measure of association to quantify this relationship. Whether the BV constructs measure the same functioning element as the ICF classifier models do, is different for each of the ICF classes. For example, for Weight Maintenance (MBW) the annotation guidelines are in table 3.11.

level	**generic qualifier**	**Weight Maintenance (MBW)**
4	no problem	Healthy weight, no unintentional weight loss or gain, SNAQ 0 or 1.
3	mild problem	Some unintentional weight loss or gain, or lost a lot of weight but gained some of it back afterwards.
2	moderate problem	Moderate unintentional weight loss or gain (more than 3 kg in the last month), SNAQ 2.
1	severe problem	Severe unintentional weight loss or gain (more than 6 kg in the last 6 months), SNAQ $\geq$ 3.
0	complete problem	Severe unintentional weight loss or gain (more than 6 kg in the last 6 months) and admitted to ICU.

Table 3.11: Description of MBW Levels and Generic Qualifiers

Qualifier levels 0 and 1 directly match with the Weight Maintenance questions in the BV project when it comes to weight loss, but it also includes unintentional weight gain, which the BV questions do not: "Did you unintentionally lose more than 3kg weight in the last month?" and "Did you unintentionally lose more than 6kg in the last 6 months?"

While the BV question about weight loss and the ICF class about Weight Maintenance are measuring almost the same constructs, we can see a different relationship for Respiratory functioning (ADM). The question from the questionnaire is:

> "Did you - during your hospitalization - experience any issues with your lungs (such as coughing and/or shortness of breath)?".

The possible responses were: Not at all, A little, Moderately, Quite a lot, Extremely much. Table 3.12 shows the description of the ICF qualifier levels for ADM.

level	**Respiratory functioning (ADM)**
4	No problem with respiration, and/or respiratory rate is normal (EWS: 9-20).
3	Shortness of breath in exercise (saturation $\geq$90), and/or respiratory rate is slightly increased (EWS: 21-30).
2	Shortness of breath in rest (saturation $\geq$90), and/or respiratory rate is fairly increased (EWS: 31-35).
1	Needs oxygen at rest or during exercise (saturation <90), and/or respiratory rate >35.
0	Mechanical ventilation is needed.

Table 3.12: Description of ADM Levels

We can see that these constructs do not completely overlap. The BV question is referring to coughing and shortness of breath, which corresponds with ADM qualifier levels 2 and 3. However, a patient who is intubated or is supplied with oxygen, might not experience a shortness of breath due to the intervention. This means that the relationship between the BV responses and the ICF labels is not linear for all classes. In other words, while both the BV and the ICF labels are scaled ordinally and not rationally, due to the mismatch in measured constructs, we cannot assume that the relationship is monotonic. In appendix .1 we shared the full annotation guidelines for all relevant classes.

Since the BV questionnaire responses are always on an ordinal scale and the ICF qualifier levels can be handled as either continuous data (using the qualifier levels as assigned by the ICF classifier models) or as ordinal data (rounding the values so they reflect the original ordinal ICF scale), we need to use a correlation metric that can handle data of ordinal nature. Spearman Rank Correlation is an option, but as we explained, we do not assume all relationships to be monotonic, which is assumed when using Spearman. Additionally, our ordinal scales are relatively small, with anywhere between 2 and 5 values. Khamis (2008) argues that for ordinal datasets with a small amount of levels, Spearman rank correlation is not sufficient, since it does not handle ties in the ranks well. Instead, Kendall's Tau b is preffered, this measure of association is appropriate when one variable is ordinal and the other one is ordinal or continuous. Therefore, we use Kendall's Tau b for all three idenfitied cases.

Kendall's Tau scores range from -1 to 1. 0 indicates no association and Kendall's Tau values toward -1 and 1 indicate a stronger negative or positive association respectively.

In section 3.2.3, we discussed how the qualifier assignment by the ICF classifier models is evaluated using the MAE, which assumes the data is on a rational measurement level. Because experiment 1 involves using correlational tests to determine the strength of the relationship between the datasets and since the BV questionnaire responses are also on a ordinal discrete level instead of a continuous measurement level, we round each ICF value to its nearest whole number, to reflect the original scale of the data.

This results in experiment 1 consisting of 3 different cases, each aggregating the ICF ratings in a unique way to reflect the relationship between measurable constructs. Each analysis consists of ICF ratings being used on a continuous scale and on an ordinal scale. Additionally, in case 3, where the relevant questionnaire questions are repeated throughout the BV study, we also compare the change between BV rating and ICF ratings between each point in time, which would quantify the relationship of the trends as described by both measures. We use Kendall's Tau b for each of these analyses.

3.4 Experiment 2: Predictive modeling of ICF ratings during hospitalization on expected ICF ratings after discharge.

The objective of the second experiment is to predict the expected ICF classes for clinical notes post-discharge based on ICF predictions during hospitalization. This involves developing a predictive model that leverages the ICF ratings assigned to patient notes during their hospital stay to forecast the ICF classes that would likely be assigned to notes created after discharge.

3.4.1 Model Choice and details: LSTM

Previously, we explained that there is a variability in patients' lengths of stay. This variability impacts our modeling approach, particularly concerning the variable lengths of input sequences. Each patient's hospital stay consists of a different number of clinical notes, resulting in sequences of varying lengths. This section delves into the methods to handle variable input lengths and justifies our choice of the Long Short-Term Memory (LSTM) model for this task.

3.3.2.1 Handling variable sequence lengths

When dealing with sequences of variable lengths, several strategies can be considered:

Dimensionality Reduction: Methods such as Principal Component Analysis (PCA) can reduce the dimensionality of the input data by transforming the original variables into a smaller set of uncorrelated variables, called principal components, which still capture most of the data variance.

Embeddings: Creating embeddings for the ICF outputs helps represent the notes in a fixed-dimensional space. This involves learning a mapping from the high-dimensional space of the original data to a lower-dimensional space. One common approach is using a neural network model trained on an unsupervised task, such as predicting ICF model output and the type of note based on input features. Alternatively, a next-rating prediction task can be used, where the network starts with random weights and learns to predict the target of the next ICF ratings, updating the representations in the last hidden layer. These embeddings capture meaningful patterns and relationships in the data.

Alternatively, we could use a model that handles variable-length sequences directly. Recurrent Neural Networks, such as LSTM models, can process each note in the sequence step-by-step without the need for padding or truncation, making them ideal for our dataset where each patient's sequence of notes varies in length.

Additionally, the LSTM architecture allows for generating predictions even on incomplete sequences, aligning with the hospital's interest in obtaining predictions partway through a patient's hospitalization. While this is not the primary evaluation goal, it provides additional utility and flexibility. Moreover, LSTMs excel at capturing temporal patterns in sequential data, which is crucial for understanding the progression of patient health over time. Therefore, we expect that using LSTMs will enhance our ability to forecast ICF ratings post-discharge.

3.3.2.2 Missing value representation

For this research, we compared several feature representation methods to handle variable lengths. However, in addition to the fact that each patient has a different amount of clinical notes entries in the EHR, not all entries contain ICF class mentions. As a result, input sequences per patient often contain empty values. Missing ICF values can have different meanings depending on the type of note from which the value is missing, for example, a note without a mention of the patient's mobility can mean nothing is wrong (i.e. if the note was written by the physiotherapist) or it can mean the patient's mobility was not relevant (i.e. a note written by a dietitian). Ignoring notes with missing values or imputing these values risks losing important information. However, simply representing a missing value as 5 could mislead the model into interpreting it as an actual ICF qualifier level of 5, meaning there is no impairment. Therefore, we model the presence of an ICF qualifier level binarily as a separate feature and we denote the missing value itself as 5, which corresponds to no issues in the functioning domain.

For example, if a patient has three consecutive notes in the EHR with the following FAC ratings: 0, missing, 2, this would be represented as:

Ratings: [0, 5, 2]

Presence: [0, 1, 0]

Resulting in the total vector being: [0, 5, 2, 0, 1, 0]. Here, the second sequence indicates the presence of missing values.

3.3.2.3 Structured Data

In addition to the extracted ICF class ratings, we include structured data from the EHR to improve model performance. These features are age, gender, BMI, number of diagnoses, ICU stay requirement, and total surgery duration, which are relevant indicators of a patient's overall health and recovery trajectory.

Unlike ICF ratings, missing values in these structured features are treated as genuine absences. Continuous features such as age, BMI, and number of diagnoses are imputed with the median value, while categorical features such as gender and ICU stay requirement are imputed with the most frequent label.

3.3.2.4 Baseline

To evaluate the added predictive value of ICF ratings during hospitalization, we compare the performance of models with and without these ratings. The model without the ICF ratings from the *hospital set* is considered the baseline model and it includes only structured data features (age, gender, BMI, number of diagnoses, ICU stay requirement, and surgery duration). By comparing this baseline with a model that also includes ICF ratings, we can assess the impact of the ICF data on predictive performance.

If the model with ICF ratings outperforms the baseline, it would indicate a predictive relationship between in-hospital ICF ratings and post-discharge functioning. This comparison is crucial for validating our hypothesis that functioning during hospitalization, as captured by the ICF models, can predict post-discharge outcomes.

3.3.2.5 Model choice: post-discharge functioning prediction

The aim of this thesis project is to explore how information about a patient's functioning throughout hospitalization can be used to predict the expected functioning levels after being discharged from the hospital. If we can confidently predict how a patient rehabilitates after a surgery, we can target interventions for patients with worse expected functioning. This section explains the model architecture decision for the purpose of building predictive models that utilize functioning data throughout hospitalization to predict post-discharge functioning.

In section 3.4.1, we discussed that RNN models, such as LSTM, can handle input sequences of variable lengths, making them a viable choice. In section 3.3.2, we highlighted that ICF constructs do not always relate linearly to the BV questions. A significant benefit of using Neural Networks, such as LSTM models, is their ability to capture these non-linear relationships through non-linear activation functions and gating mechanisms.

There are multiple ML methods that can address non-linear relationships. However, traditional ML methods like random forests and SVMs cannot handle variable-length input sequences.

Additionally, methods such as linear regression models cannot capture non-linear relationships effectively. Although traditional ML techniques offer interpretability, the interpretability of the ICF classifier output itself is not quantified, complicating the interpretability of downstream tasks. While GRU, another RNN architecture, can be used, LSTMs are more effective at capturing long dependencies, which is crucial given the average sequence length of 70 to 85 notes per patient.

Transformer models share the benefits of LSTMs but require padding for variable-length sequences. While Transformers handle padding through masked tokens, sequences longer than the maximum input size need complex techniques like truncation or rolling windows to fit into the input layer. Additionally, due to their dense attention layers, Transformers require larger, more varied datasets to converge. Given our dataset size (300-500 patients per subset), LSTMs are more appropriate.

A key practical reason for choosing LSTM models is our collaboration with AUMC hospital, which aims to integrate the developed models within their EHR software. While Transformer models are faster for training and inference due to parallel computations, they require specialized Tensor GPUs, which are not available in the EPIC EHR software. Therefore, LSTMs, which are computationally less expensive and can run on standard hardware, are a more practical choice for deployment in a healthcare environment.

The LSTM architecture includes several parameterizable values: number of layers, batch size, learning rate, optimizer choice, dropout rate, activation functions, and initialization schemes. Due to the size of this project and the multitude of experiments, we are not able to search the hyperparameter space for the best set of parameters.

In conclusion, we chose the LSTM architecture for its ability to handle variable-length sequences, capture non-linear relationships, and meet practical implementation constraints. We will explore different data representation methods and combinations of structured and unstructured data. Additionally, we will compare the results with a baseline model that excludes the ICF ratings to evaluate the added predictive value of these ratings.

3.3.2.6 Model Training

For the training setup of the model, we split the data into a training, development and test set following a 60-20-20 split. Since this experiment includes constructing multiple models for each of the 5 classes, with different input feature representations, we cannot employ cross-validation during these experiments due to technical and time-related limitations.

In table 4.2 we show that, for each of the ICF classes, there are approximately 300 to 550 positive target labels, meaning there are anywhere between 500 and 700 samples per ICF class without the corresponding ICF rating in the target period. We represent these missing values as qualifier level 5, meaning no impairment. However, since this results in level 5 being over-represented in the target labels, we rebalance the dataset by removing some of the negative samples so it contains as many items as the positive samples. We only rebalance the set of imputed target labels, if a post-discharge note was assigned a qualifier level 5 naturally, we always keep them in the dataset.

Ideally, we use a grid search to find the optimal combination of (hyper)parameters that allows the model to best generalize on the data. However, due to the number of different model input representations we want to explore, this would greatly increase training time. It would also complicate a direct comparison between the experiments, since it would result in many more models to compare. Therefore, we use the same default settings for each of the trained models in Experiment 2. Table 3.13 contains all (hyper)parameters we used in this experiment.

Hyperparameter	Value
Activation Function	tanh
Recurrent Activation	sigmoid
Input Layer Shape	(None, # of features)
Layer 1 Dimensions	64
Layer 2 Dimensions	32
Output Layer Activation	linear
Optimizer	Adam
Learning Rate	0.001
Loss Function	MAE
Batch Size	8
Number of Epochs	20
Dropout Rate	0
Weight Initialization	random

Table 3.13: Model Hyperparameters

3.4.2 Model Evaluation

Model evaluation is a critical component of this research, ensuring that our predictive model performs accurately and reliably. The next sections describe all methods we use to evaluate the final models.

3.4.2.1 Evaluation Metric

The prediction models we propose in this thesis aim to predict post-discharge functioning of patients based on their functioning data during hospitalization. Both the input of the model and the target label are based on the output by the previously developed ICF classifier models. While ICF qualifier levels are originally ordinally scaled discrete levels, the researchers who developed these models chose to use Mean Absolute Error (MAE) as a loss function and as the evaluation metric. This essentially got rid of the discrete levels, because the model evaluates an output of 1.45 as opposed to a target of 2 as an error rate of 0.55 and an output of 2.45 as opposed to the target of 2 as a slightly better error of 0.45. In reality, 1.45 would correspond to the ICF qualifier level 1 and 2.45 relates to qualifier level 2.

While we believe using a classification error metric is better suitable to evaluate a predicted ICF level, to be able to directly compare our proposed models against the ICF classifier models, we employ the MAE score for model evaluation and as a loss function instead. The MAE scores as reported for the ICF classifier models in Kim et al. (2022) are essentially the average expected errors, so as long as the MAE scores of the prediction model are of the same magnitude, those predictions are within the expected error margins of the assigned ICF labels. We would consider this to be a positive result.

3.4.2.2 Error Analysis

To gain deeper insights into the model's performance and identify areas for improvement, we will conduct an error analysis. This involves examining the types of mistakes the model makes. For each ICF class, we randomly select examples of correct and incorrect predictions. We also anonymize

the phrases referring to those classes to be able to report on them as well. This analysis will provide deeper insights into the model's performance and help identify areas for improvement.

By combining these evaluation metrics and error analysis, we can comprehensively assess the model's performance and ensure that future improvements are meaningful.

The methodology outlined in this chapter provides the foundation for analyzing the correlation between ICF outputs and self-reported functioning, and predicting post-discharge ICF classes. The next chapter will present the results of these experiments.

Chapter 4

Results

This chapter presents the findings from the two experiments that are introduced in the previous chapters. The first experiment assesses the correlation between the ICF classifier model's outputs on clinical notes and patients' self-reported functioning levels from the Beter Voorbereid (BV) project. The second experiment evaluates the predictive value of in-hospital ICF ratings on post-discharge ICF ratings, utilizing data from oncological surgery patients. For both experiments, we compare results between analyzing the ICF scores as continuous values and analyzing ICF scores as ordinal categorical values.

4.1 Data Analysis

In chapter 3 we evaluated the methodology applied by the researchers who developed the Language Model (Verkijk and Vossen, 2021) and the ICF classifier models (Kim et al., 2022). Based on the Inter-Annotator Agreement they reported for the ICF classed, we identified that the class of Exercise Tolerance (INS) is not a suitable class to use for downstream modeling tasks.

We also compared the questions in the BV set with the ICF classes, to identify the questions that most closely resemble what the ICF classes are measuring. For the Eating (ETN) ICF class we found no corresponding question, therefore we also ignore that class in future analysis.

In this section, we continue this analysis by computing distributional statistics of the output of the ICF classifier models on the notes from the participating patients in the BV study. Based on these statistics we can further analyse which ICF class mentions appear frequently enough to be suitable for correlational analysis. Since the correlational analysis serves as the validation of the usability of the ICF classifier models, if we decide to exclude a class for experiment 1, we also exclude it for experiment 2.

4.1.1 Data Description ICF Classifier Output on Patients from the Beter Voorbereid Project

Table 4.1 shows the average amount of ICF class mentions found in the clinical notes for all patients, including notes from hospitalization until 8 weeks post-discharge.

ICF Class	Amount of Notes containing Class Mentions per Patient (Average)
FAC: Mobility	30
ADM: Respiratory Functions	53
ENR: Energy	10
STM: Mood	19
MBW: Weight Maintenance	16
BER: Work	5
ATT: Attention	3

Table 4.1: Average Amount of notes containing ICF Class Mentions per Patient

Since the BER (work) and ATT (attention) classes appear infrequently, they cannot be used for the correlation analysis. Therefore, only the following classes are considered: FAC (mobility), ADM (respiratory functioning), ENR (energy), STM (mood), MBW (weight maintenance).

In addition to looking at the frequency of notes containing the relevant ICF class mentions, it is also essential to look at the distribution of qualifier levels. This is useful to test assumptions to determine a suitable statistical test, but it also helps us reason about the MAE evaluation scores. 4.2 until 4.5 show histograms per ICF class containing the total distribution of assigned levels across all 74 patients.

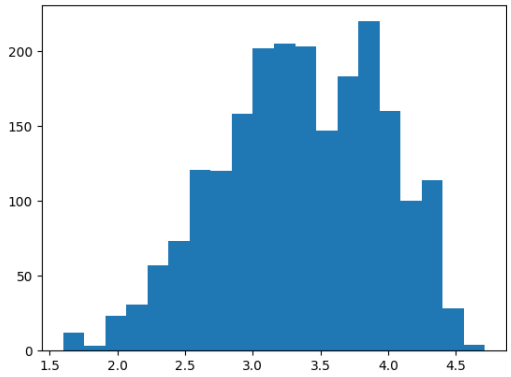


Figure 4.1: Distribution of Mobility (FAC) class mention qualifier levels across all patients

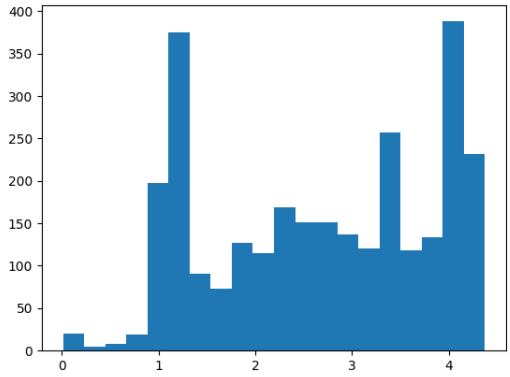


Figure 4.2: Distribution of Respiratory functioning (ADM) class mention qualifier levels across all patients

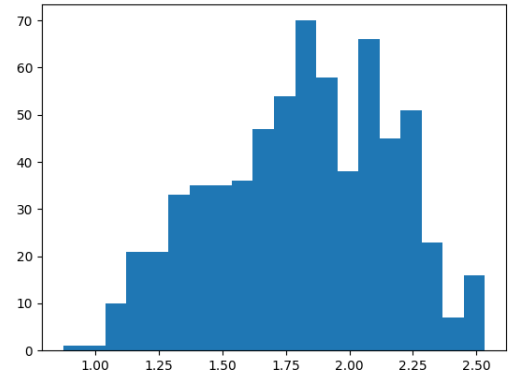


Figure 4.3: Distribution of Energy (ENR) class mention qualifier levels across all patients

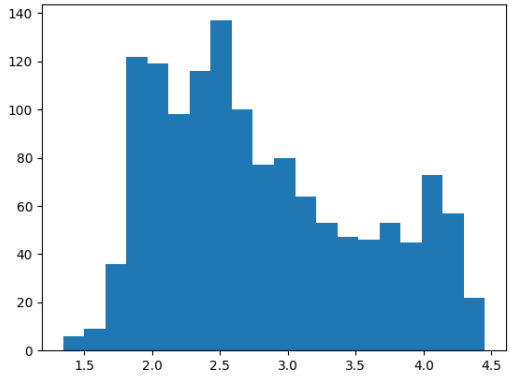


Figure 4.4: Distribution of Mood (STM) class mention qualifier levels across all patients

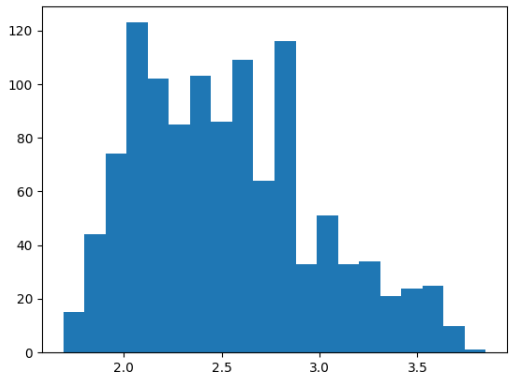


Figure 4.5: Distribution of Weight Maintenance (MBW) class mention qualifier levels across all patients

These histograms clearly show that the distribution of levels varies across classes. FAC appears to follow a near-normal distribution while ADM and STM show more skewed distributions. In paragraph 3.2.3, we noted the absence of distributional information about qualifier levels in previous publications. While one of the reports contains pie charts displaying the relative distribution of the levels, it is important to have the exact values in order to compare error metrics between ICF classes. If this information had been available, we could have compared both distributions to better understand dataset differences. Significant distribution differences might indicate a lack of transferability from the training environment to a production environment. Additionally, comparing the distribution of qualifier levels from the annotated and predicted data could provide more insight into model performance than an error-rate performance metric can on its own.

Using the ICF classifier output, we can graph the functioning over time for each patient in figure 4.6. Since FAC qualifier levels range from 0 to 5 instead of 0 to 4, we rescaled all other values to the same scale.

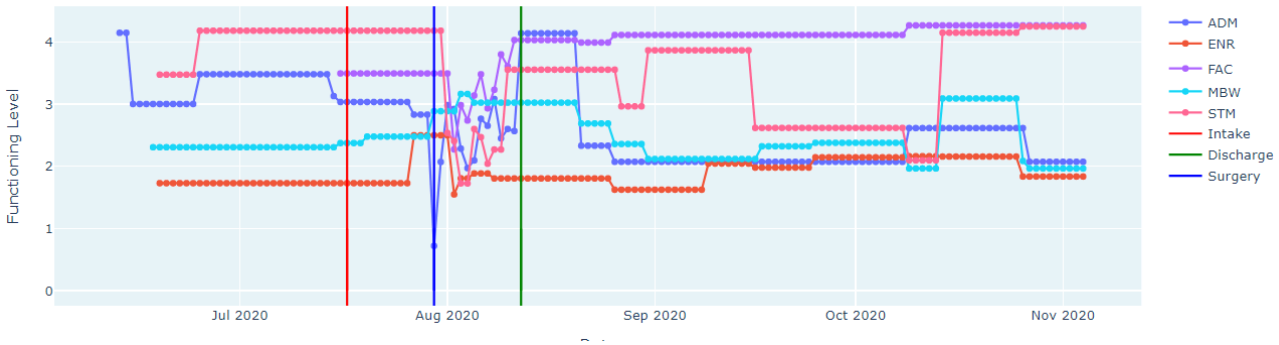


Figure 4.6: Functioning Over Time for a Sample Patient, a colored line per ICF class and vertical lines for the Intake (Red), Surgery (Blue) and Discharge (Green) dates (ADM = respiratory functioning, ENR = energy, FAC = mobility, MBW = weight maintenance, STM = mood)

Due to the low frequency of ICF mentions appearing in general, many days do not contain an ICF label. For visualization purposes in this graph, we filled those in by carrying over the last observation. Therefore, any change in ratings reflects a note with a newly identified ICF class mention. If a line does not change over time, that means there were no new observations. The vertical lines denote the intake date, surgery date, and the discharge date respectively. It is clear from the graph that between surgery and discharge, most notes with new ICF-related information are written, which is expected because monitoring is much more precise and frequent during hospitalization.

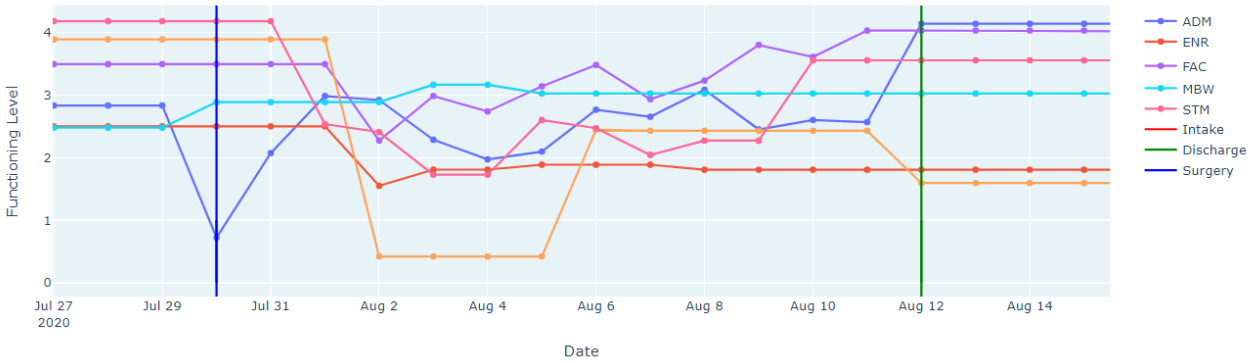


Figure 4.7: Zoomed in Functioning Over Time for a Sample Patient, a colored line per ICF class and vertical lines for the Surgery (Blue) and Discharge (Green) dates (ADM = respiratory functioning, ENR = energy, FAC = mobility, MBW = weight maintenance, STM = mood)

When zooming in on the period between surgery and discharge, as can be seen in figure 4.7, we can see the functioning over time in more detail. For example, ADM (respiratory functioning) is assigned an average qualifier level of less than 1 on July 30th. By July 31st, it is up to slightly over 2, and it continues to rise until the first of August. We identified three notes from July 30th with ADM qualifiers levels assigned to them, the relevant phrases from those notes are:

> "RESP: Extubated at 5 PM. Acceptable gasses with 3L O2 on the nasal cannula."
> Dutch: (RESP: Om 17uur gedetubeerd. Acceptabel gas met 3L O2 op de neusbril)
> ADM: 1.06

> "Based on the decreased Tiffeneau index (FEV1/VC), there is an expiratory obstruction. Perioperative details: A: double-lumen tube on the left, good position --> switch to single-lumen tube at the end of the operation B: tolerates one-lung ventilation well (left lung collapsed). [...] Saturation 100% under 12L O2, ventilation 16/min over Ambu bag."
> Dutch:)Op basis van de verlaagde tiffeneau index (FEV1/VC) is er sprake van een expiratoire obstructie. Bijzonderheden per-operatief: A: dubbellumen tube links, goede positie --> wissel naar single lumen tube op einde OK B: verdraagt 1-long beademing goed (linkerlong gecollabeerd) [...] sat. 100% onder 12L O2, beademing 16/min over ambuballon.)
> ADM: 0.40

> "COPD with reduced lung function. Management per problem: 1. reduce sedation and extubation [...] [DATE: 30 July] Addendum by [NAME DOCTOR] [DATE: 31 July]: Clear and awake. Extubated. [DATE: 17 July] VC is normal. FEV1 is moderately reduced (60-69%). Based on the decreased Tiffeneau index (FEV1/VC), there is an expiratory obstruction. [...] [DATE: 30 July] Respiration: Ventilation/O2 therapy."
> Dutch: (COPD met vermiderde longfunctie Beleid per probleem: 1. sedatie afbouwen en detubatie [...] [DATE: 30 July] Addendum door [NAME DOCTOR] [DATE: 31 July]: Goed helder wakker. Gedetubeerd. [DATE: 17 July] VC is normaal. FEV1 is matig verlaagd (60-69%). Op basis van de verlaagde tiffeneau index (FEV1/VC) is er sprake van een expiratoire obstructie. [...] [DATE: 30 July] Respiratie: Beademing/O2-therapie)
> ADM: 1.5

The first two notes got assigned 1.06 and 0.40 ADM qualifier levels respectively. This seems to reflect the note content, where the first note describes extubation and the second one describes intubation. Both notes were written on the day of surgery, but the first one is a Plan of Care type of EHR entry, while the second one is a postoperative anesthesia handover note. However, the third example also shows how the identified ICF classmentions in an EHR entry on a specific date do not always necessarily reflect the functioning of a patient on that day. In this case, the note seems to be continuously expanded on, which is mentioned in the note with "Addendum". Additionally, the note also is appended with a previous note, which describes the situation on the day of hospitalization. Consequently, the functioning status that is extracted from this note does not reflect this patient's functioning at that time, and it decreases the average qualifier level assigned for respiratory functioning for that patient on that day

This is a complicating factor in most research that focuses on extracting information from clinical notes: a note does not always reflect the date of creation. The ICF model extracts class mentions - the mention of a patient's functioning - it does not extract events. The EHR from a patient that is intubated and consequently is assigned an ADM score of 0, can have multiple mentions of the intubation in different notes in the EHR. However, if a patient has 10 ADM=0 ratings or just a single one, we cannot disambiguate the two, we do not know whether a patient with 10 intubation mentions was experiencing respiratory issues for a longer period than a patient with a single mention did.

That being said, if the appended older notes or the chance of a note being expanded on through means of an addendum, are expected to appear in the dataset relatively consistently, then this results in a latent lagging variable affecting all scores evenly. For example, the addition of information that reflects the functioning status on the date of admission is expected to be present in other progress or department handover notes as well, which would result in the same misplaced qualifier level affecting the average score of a note across hospitalization.

However, if there are more notes for a patient on a certain date, the effect the misplaced qualifier levels have on the average score changes, which means that averaging out scores for a given day is likely not the best approach. Also, there is a possibility that the type of the note (e.g. progress note, plan of care, telephone appointment) affects the chance of such an addendum appearing, which is why we expect that adding the note type as a feature in experiment 2 would benefit the model. For example: if mobility functioning is not mentioned in a neurosurgical report, it might mean this was not worth mentioning while mobility could still have been impaired. However, if a physiotherapist does not mention mobility in their notes, this probably means there was no issue with mobility, since we expect this doctor to report on this class more frequently.

We can see that the graph for ADM increases a day after, on July 31st, to 2.07. This correctly reflects the fact that the patient's intubation was ended. In total, 8 different EHR entries were assigned ADM scores on that day. These range from 0.40 to 4.02. The note that was assigned 0.40 is a progress report that contains an exact copy of the postoperative anesthesia handover note from the day before, hence the exact same score. However, since the patient also has a few notes describing the increased respiratory functioning, the average score still increased, which correctly reflects the content of the notes. As we can see, the score keeps increasing until August first. While this is a small qualitative inspection of the ICF output, it shows how the lagging latent variable affects the score, but also how over time the scores still display a trend that fits the functioning development of the patient.

Finally, the ADM line decreases again before increasing to a high level prior to discharge. When we inspect the note ADM ratings on the local minimum on August 4th (average ADM = 1.96), we can see multiple ratings ranging from 1.30 to 2.76. The highest rating is assigned to the following

phrases:

> "Respiration rate 22 — spO2 94% [...] Xth consistent with the differential diagnosis of pneumonia, however, given the decreasing infection markers and absence of fever, no treatment is necessary."
> Dutch: (Ademhal. 22 — spO2 94% [...] Xth conform dd pneumonie echter nu gezien dalend infectielab en geen koorts niet behandelen.)
> ADM: 2.76

The lowest rate (ADM = 1.30) and the next lowest (ADM = 1.65) were assigned to the following phrases:

> "Resp: Still requiring oxygen."
> Dutch: (Resp: Nog zuurstofbehoeftig)

> "Question: pneumonia?"
> Dutch: (Vraagstelling: pneumonie?)

This local minimum corresponds with a decrease in respiratory functioning, which is indicated by still requiring oxygen and the question whether the patient might be suffering from pneumonia. The highest rating is given to the mention of the O2 saturation and the fact that treating pneumonia is being discontinued. These examples show that the overall trend of the assigned ICF labels seem to reflect the note content, but on an individual note level we cannot easily conclude that an ICF label reflects the functioning at that moment in time.

4.1.2 Data Description: selected patients and ICF classifier output on the notes

Healthcare professionals that were part of this research identified that post-surgery patients typically have a follow-up appointment around six weeks after discharge, making the ICF ratings assigned to these reports a good target for prediction. Since the exact timing of these appointments can vary, we use a weighted average approach targeting ICF predictions between four and eight weeks post-discharge, with the highest weight given to the six-week mark. This method ensures alignment with the expected follow-up period. We refer to the set of post-discharge notes in the four to eight-week period as the '*discharge set*' and the notes written during hospitalization as the *'hospital set*' throughout this section.

Since the BV project focused mainly on post-surgical oncological patients, experiment 2 also uses data from oncological surgery patients. Unlike the previous experiment, which was limited by the small sample size, this experiment benefits from a larger dataset, allowing the employment of complex ML methods such as neural networks.

Out of 1380 patients selected for experiment 2, 1037 had at least one EHR entry in the 4 to 8 week period post-discharge, but not all post-discharge notes contained ICF class mentions. The table below lists the number of patients with at least one class mention in the discharge set for each relevant ICF class, along with the average number of notes during hospitalization and post-discharge periods. For example, only 329 of the 1037 patients were assigned at least one

ADM rating in the discharge set, meaning that the remaining 708 patients' EHR did not contain a post-discharge note with an ADM rating. Additionally, the table below also describes the average number of notes in the hospital and discharge sets as well as the number of notes that contain the relevant ICF rating. For the total dataset, on average, a patient has 68 notes in the EHR in the hospital set, and 16 in the discharge set.

ICF Class	**#patients**	**#notes *hospital set***	**#notes *discharge set***	**#notes *hospital set* - notes with class-mention**	**#notes *discharge set* - notes with class-mention**
Respiratory functioning (ADM)	329	79	40*	26	11
Energy (ENR)	302	75	36*	7	3
Mobility (FAC)	366	85*	36*	17	5
Weight Maintenance (MBW)	541	84*	24*	6	3
Mood (STM)	348	79	35*	10	5
Whole dataset	1037	68	16	-	-

Table 4.2: Average amount of notes per ICF category for patients with at least one relevant class mention in their post-discharge notes (* denotes statistical significance according to paired t-test as compared with the total set)

Comparing the average number of notes without class mentions during hospitalization and during the post-discharge period with the averages for all patients (68 and 16 notes for hospital and discharge set respectively) shows how the selection of these categories affects the distribution of the notes. For FAC and MBW, we can see that the averages during hospitalization are significantly higher than the overall average (using the paired t-test and compared with the total population). For all classes it holds that the number of notes in the discharge set is significantly higher than the overall population. This significance can be interpreted as, for each of these subsets, the number of notes in the post-discharge set cannot be explained if the subset had been a random sample of the population. Filtering on patients with notes containing relevant classes in the discharge set significantly alters the distribution of those notes.

Additionally, we can observe that the size of the difference between the overall population hospital set size is smaller than that of the discharge set. Where all subset's hospital set sizes range from 75 to 85 as opposed to the 68 overall average, the discharge set sizes differ much more, ranging from 24-40 as opposed to the average of 16. And where FAC and MBW were the only significantly different subsets for the hospital period, all subsets were significantly different for the discharge set (paired t-test due to the subsets being sampled from the population, hence no independence).

By comparing the average number of notes per ICF class with the average number of notes that contain the relevant class mention, we can get a better understanding of how frequent the class mentions appear. For instance, while patients with MBW mentions in the discharge set have on average 84 notes in the hospital set, only 6 of those on average contain the relevant MBW label (9.5%). This is a different percentage than if we take a look at the ADM class, where the 329 patients with at least one ADM mention in the discharge set have 79 notes on average during hospitalization, out of which 26 contain the relevant mention (32.9%).

We observed that patients in all of the subsets contain - on average - more notes during hospitalization than in the total population. Table 4.3 explains this observation by reporting the

average length of stay and the average number of notes per day. The average length of stay for all patients, regardless of whether ICF mentions are present in their discharge set is 12.4 days, and with 68 notes on average per patient in this period, that amounts to 5.5 notes per day on average.

ICF Class	**Average Length of Stay**	**Average amount of notes per day in *hospital set***
Respiratory functioning (ADM)	12.5	6.3
Energy (ENR)	11.7	6.4
Mobility (FAC)	13.7	6.2
Weight Maintenance (MBW)	12.9	6.5
Mood (STM)	13.0	6.1
Whole dataset	12.4	5.5

Table 4.3: Average Length of Stay and Average Amount of Notes per during hospitalization for patients whose post-discharge note contains at least a single relevant ICF class mention

To interpret these values, we take ADM as an example. Out of the 1073 patients selected for this experiment, 329 patients have post-discharge notes with at least 1 ADM mention. These patients have on average 79 EHR entries during their hospitalization as opposed to the 68 EHR entries we found for the total dataset (4.2). This greater amount of EHR entries can be explained in two ways: either they were hospitalized for a longer period or - on average - more notes are written for this patient per day. By looking at 4.3 we can observe that it is in fact the latter; patients with at least one ADM mention in their post-discharge notes get - on average - 6.3 new EHR entries per day, as opposed to the overall average of 5.5.

The table shows that patients with FAC, STM and MBW functioning mentions in the discharge set are hospitalized anywhere between 0.6 and 1.3 days longer than the overall population. For all classes it holds that patient with a mention of said ICF class in their post-discharge notes received - on average - more new EHR entries per day of hospitalization than the overall population. This suggests that both the length of stay and the number of notes per day have predictive value for determining whether these patients will have functioning mentions in their post-discharge notes. To leverage this insight, we will include the length of stay and the frequency of note entries as features in our predictive models by encoding the days since hospitalization into the vector representation. By incorporating these variables, we aim to enhance the model's ability to accurately forecast ICF ratings post-discharge.

4.2 Experiment 1: Correlation Analysis Between ICF Outputs and Self-Reported Functioning

The objective of the first experiment is to validate the ICF classifier model by examining its correlation with patient-reported outcomes. This serves as an external validation of the model's outputs, ensuring that the extracted ICF ratings are reflective of the patients' actual experiences as reported in the BV questionnaires. In chapter 3: Methodology, we identified three different relationships between the BV questions and the ICF output, for all of these we will employ different analyses.

As detailed in the methodology section, the BV project comprises five questionnaires administered at different stages relative to the patient's hospitalization. These questionnaires measure various aspects of patient functioning and well-being. The dataset used for this experiment includes responses from 74 patients who completed the project and had corresponding clinical notes with ICF ratings.

To analyze the correlation between the ICF outputs and the BV questionnaire responses, we employed different strategies tailored to the nature of the questions. Three primary scenarios were identified and detailed in table 3.1.

For each of these cases, we will provide data visualisations per ICF class and corresponding BV questionnaire responses. We include the original exact ICF values on a continuous scale, but we analyse the results by converting these measurements to an ordinal scale. We use these graphs to reason about potential correlations and end each section with a table reporting on the Spearman Rank correlation and - where applicable - Kendall's Tau metric.

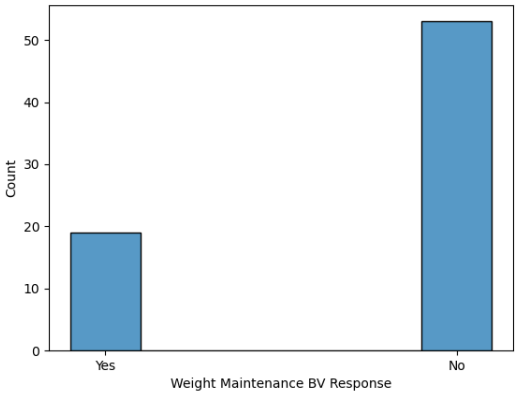
4.2.1 Results Experiment 1 - Case 1: Weight Management

The Beter Voorbereid questions related to weight management were:

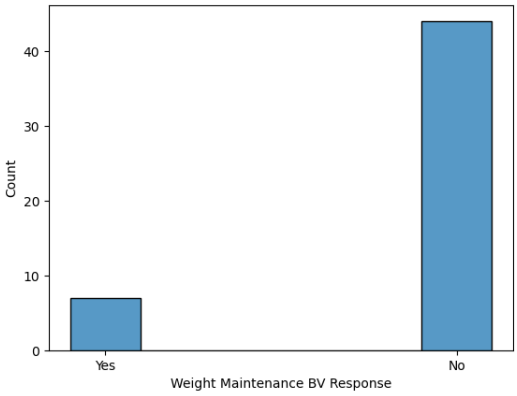
> "Did you unintentionally lose more than 3kg in the last month?"
> "Did you unintentionally lose more than 6kg in the last 6 months?"

These questions were present in the intake questionnaire (T0) and the last questionnaire (T4) (see: table 3.10). The possible responses were "Yes" (encoded as 1) and "No" (encoded as 2). The ICF values are on the same scale, with higher values indicating better Weight Management functioning.

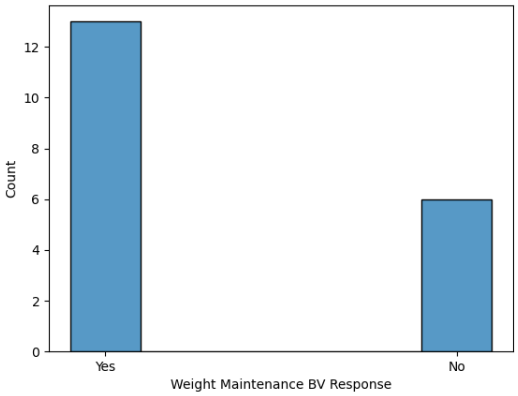
Before we can interpret the relationship between the ICF ratings and the BV responses, we need to inspect the distribution of the BV responses. Especially because not every patient responded to both questions and in order to reason about the plots that visualize the correlations this is important information



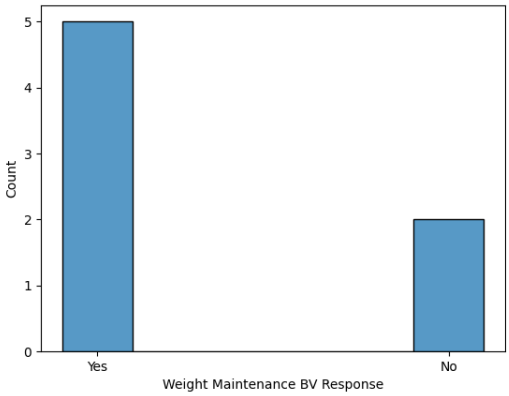
(a) MBW BV responses: 3KG loss - timepoint T0



(b) MBW BV responses: 3KG loss - timepoint T4



(c) MBW BV responses: 6KG loss - timepoint T0



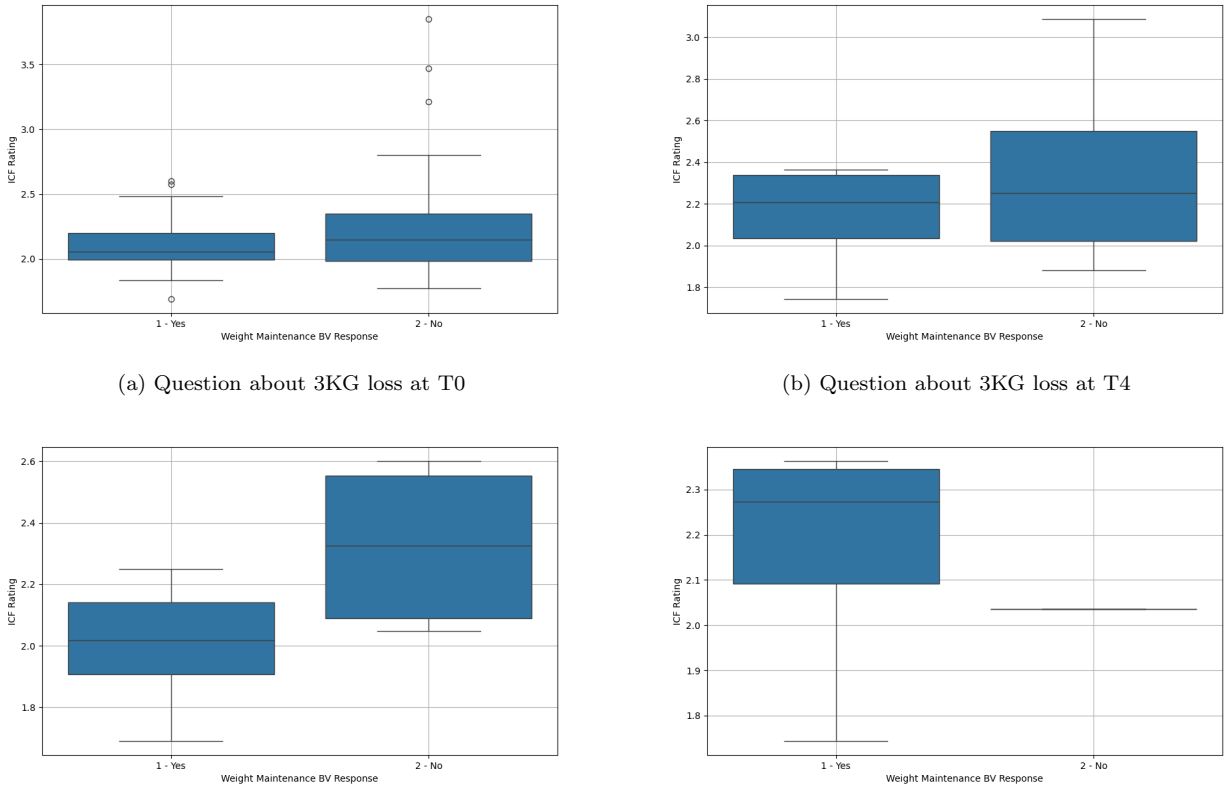
(d) MBW BV responses: 6KG loss - timepoint T4

Figure 4.8: Histograms MBW BV responses for both weights (3KG, 6KG) and timepoints (T0, T4)

The main observation in figure 4.8c and 4.8d is that many patients did not respond to the second question about losing 6kg in the last six months when comparing with the question about losing 3kg in 1 month (figure 4.8a and 4.8b). When they did respond, "yes" was a more common answer.

We identified the following relevant time periods: 1 month and 6 months. As explained in chapter 3, this question refers to the worst weight management functioning in the time period, regardless of better functioning at other times. Therefore, we compare the self-rated values to the minimum values in the MBW ICF class.

Figure 4.9 contains box plots for each BV weight question, showing the distribution of paired ICF ratings.



(a) Question about 3KG loss at T0

(b) Question about 3KG loss at T4

(c) Question about 6KG loss at T0

(d) Question about 6KG loss at T4

Figure 4.9: Box plots of ICF MBW ratings for each BV weight loss question

Graphs 4.9a and 4.9b show that the distribution of ICF ratings does not change a lot depending on whether the BV response was "yes" or "no". For "no", the median is slightly higher and the interquartile range is slightly bigger, but the overall distribution does not shift a lot between both responses. For graph 4.9c and 4.9d, the shift is slightly bigger, but there are also less datapoints. Especially for figure 4.9d this is the case, as we saw in figure 4.8d previously: only 2 participants responded with "no" to this question and 5 with "yes".

For each of the plots, it holds that most ICF ratings were between 2 and 3, regardless of the question response. In section 3.3.2, table 3.11 shows that MBW qualifier level 2 is described as "Moderate unintentional weight loss or gain (more than 3kg in the last month)." Since this question in 4.9a and 4.9b measure the exact same construct as ICF qualifier level 2 does, and the distributions look very identical for both question responses, we can conclude that the ICF values likely do not correlate with the BV question responses.

In table 4.4 we reported on the Kendall's Tau correlation coefficient, for each weight-related question, except for the last one (6KG loss in the last 6 months at timepoint T4). This is because we only have 7 datapoints for that question, which is not enough to compute statistical significance for. The table shows that the only statistically significant correlation was found for the BV question about having lost 6kg in 6 months at T0. A statistically significant correlation means that we have enough evidence to reject the null hypothesis that there is no correlation between the variables. However there are not as many respondents for this question as for the questions about 3KG, and based on the distribution of replies, we believe this was more biased towards patients who would have replied with "No" to the question. Therefore, we do not believe these question responses are

representative of what the patients really experienced.

In addition to evaluating a potential correlation while using the exact ICF ratings, we also evaluate it using the rounded ICF values. This essentially converts the data on a continuous scale back to the original ordinal scale the ICF system was build for. We display this data using contingency tables (4.5) and we also report on the Kendall's Tau correlation metric.

	3KG - 1 Month - T0	**3KG - 1 Month - T4**	**6KG - 6 Months - T0**	**6KG - 6 Months - T4**
Kendall's Tau	Coef: 0.07, p = 0.5	Coef: 0.11, p = 0.4	**Coef: 0.50, p = 0.01**	X

Table 4.4: Kendall's Tau Correlation Coeffecient (Coef) for exact ICF values vs BV responses for each Weight Maintenance question

Self-Ratings / ICF Ratings	**3KG T0**			**3KG T4**			**6KG T0**			**6KG T4**		
	Yes	**No**	**Tot.**	**Yes**	**No**	**Tot.**	**Yes**	**No**	**Tot.**	**Yes**	**No**	**Tot.**
MBW = 1	0	0	0	0	0	0	0	0	0	0	0	0
MBW = 2	16	35	51	5	21	26	12	4	16	5	2	7
MBW = 3	2	5	7	0	12	12	0	2	2	0	0	0
MBW = 4	0	1	1	0	0	0	0	0	0	0	0	0
Total	18	41	59	5	33	38	12	6	18	5	2	7

Table 4.5: Combined Contingency Tables of ICF Ratings vs BV Ratings for All Different Weights (3KG and 6 KG) and Timepoints (T0 and T4). Tot. is abbreviation for Total

Table 4.5 contains the data when comparing the rounded ICF values with the BV questionnaire responses, essentially converting the ICF qualifier levels back to their original ordinal scale. We display this data in a contingency data and we also calculate a Kendall Tau Correlation coefficient for each question. This also shows that for the questions about losing 6kg weight, we have fewer responses. We can also see that there were no ICF MBW qualifier levels of 0 assigned to the notes and only a single time did MBW level 4 appear. The latter is likely because we selected the minimum rating in each relevant period.

	3KG 1 Month T0	**3KG 1 Month T4**	**6KG 6 Months T0**	**6KG 6 Months T4**
Kendall's Tau	Coef: 0.05, p = 0.7	Coef: 0.26, p = 0.1	**Coef: 0.5, p = 0.04**	X

Table 4.6: Kendall's Tau Correlation Coeffecient (Coef) for rounded ICF values vs BV responses for each Weight Maintenance question

As evident from the graphs and contingency tables (4.5 and as reported on in table 4.6, we did not find a strong statistically significant correlation. The only statistically significant correlation was for 6kg loss in 6 months at T0. However, similarly to table 4.4 we believe this is a result of

a bias introduced by the patients not replying as much to this question and if they do, they are more likely to reply with "yes".

In conclusion, the relative size of the "yes" answers to the weight loss (encoded as 1) increased while the expected minimum value did as well, increasing the correlation even if the question and ICF classes are not necessarily related. We expect that if we were to encode all patients who did not respond to this question with value 2 (indicating they did not lose the specified amount of weight), this correlation would fade away as well.

4.2.2 Results Experiment 1 - Case 2: Respiratory Functioning

For case 2, which contains the Hospital Period Question, we identified the following class: ADM - respiratory functioning. This question was only asked a single time, at timepoint T1.

The related question from the BV questionnaires is:

> Did you - during your hospitalization - experience any issues with your lungs (such as coughing and/or shortness of breath)?

The possible responses were: "Extremely much" (1), "Quite a lot" (2), "Moderately" (3), "A little" (4) and "Not at all" (5)

Figure 4.10 shows the distribution of BV responses for this question.

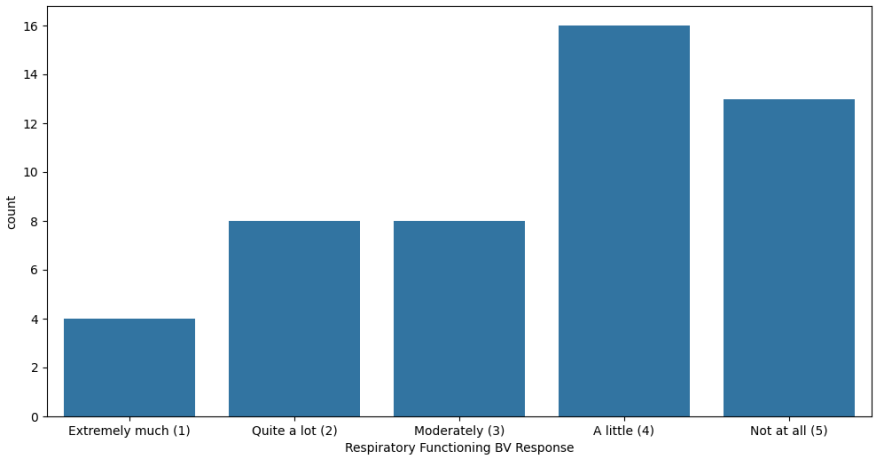


Figure 4.10: Histogram ADM BV Question Responses

As explained in section 3.3.2, the measurable construct from the questionnaires differs from the ICF class. As such, we do not expect the minimum value in this period to relate to the question response, as we did for Weight Management. For example, if a patient was intubated during hospitalization, the minimum rating should be 0. However, this does not necessarily mean the patient experienced respiratory problems or coughing.

Figure 4.11 shows the relationship between the average ICF rating during hospitalization and the BV responses for the same period using exact ICF values and table 4.7 using the rounded values.

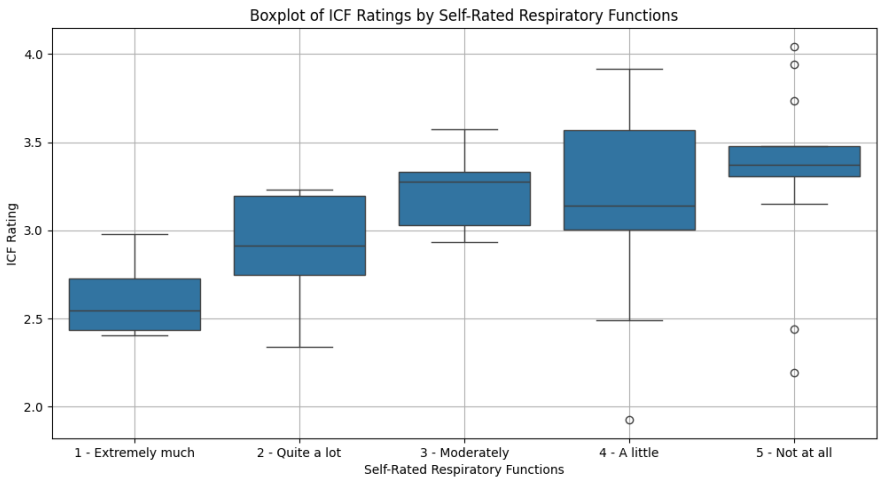


Figure 4.11: Box plot of ICF ADM ratings for each BV Respiratory Functioning question response

ICF Ratings / Self-Ratings	**1**	**2**	**3**	**4**	**5**	Total
ADM = 1	0	0	0	0	0	0
ADM = 2	2	1	0	2	2	7
ADM = 3	2	7	7	9	8	33
ADM = 4	0	0	1	5	3	9
Total	4	8	8	16	13	49

Table 4.7: Contingency Table of ADM ICF Ratings vs Self-Ratings, with for self-ratings: 1 = Extremely much, 2 = Quite a lot, 3 = Moderately, 4 = A little, 5 = Not at all

For both the case where ICF ratings are rounded and the case where they are used as their exact value, we report Kendall's Tau Correlation coefficients in table 4.8:

	Correlation
Kendall's Tau (unrounded)	**Coef: 0.36,** $p \ll 0.0001$
Kendall's Tau (rounded)	Coef: 0.20, p = 0.12

Table 4.8: Kendall's Tau Correlation Coeffecient (Coef) for exact ICF values and for rounded ICF values vs BV responses for the respiratory functioning question

When using the exact ICF values prior to aggregating them for the relevant period, we find a statistically significant correlation. We do not find such a correlation when using the rounded values.

4.2.3 Results Experiment 1 - Case 3: Repeated Day-Specific Questions

The third case we identified involves BV questions about the day itself or a recent period, asked multiple times across the research period. This applies to the following classes: Mobility (FAC), Energy (ENR), and Mood (STM). Since we have measurements at multiple points in time, we can

consider both direct correlations at each timepoint and the changes between each questionnaire and their relevant time periods. To compute the latter, we plot the delta change of ratings between each timepoint. Additionally, to evaluate average trends over time, we will plot the average rating for all patients over time for both ICF and BV ratings into a single figure. While this does not quantify the relationship, it helps visualize how average functioning over time is reflected in these ratings.

Mobility (FAC)

The mobility-related BV question was:

> "Please select one option for each of the groups in the list below that contains a sentence that best describes your health situation today. Mobility:"

With the following possible responses and their encodings in below graphs and tables: "I don't have difficulties walking" (3), "I have some difficulties walking" (2) and "I am bedridden" (1).

Figure 4.12 shows the distribution of the question answers.

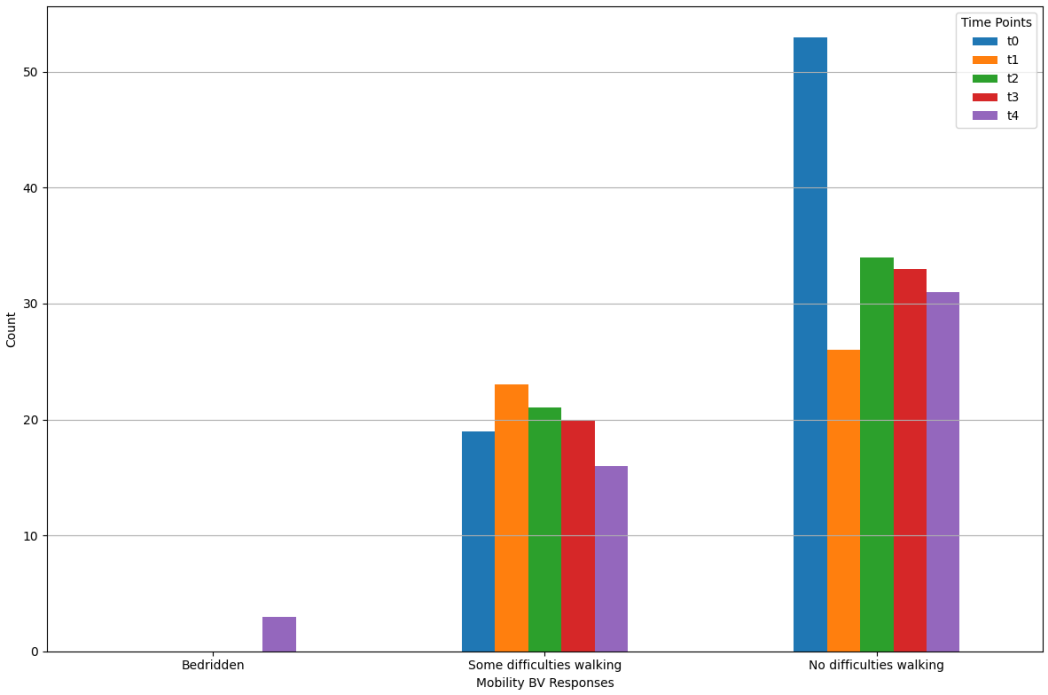


Figure 4.12: Histogram of Mobilty (FAC) BV Question Responses for each timepoint (t0 to t4) seperately

It is apparent that almost no one responded with answer option 1: I am bedridden. We can also see that most people rated their mobility with "I don't have difficulties walking" at timepoint T0, but did so less frequently for the subsequent timepoints.

Figures 4.13 and 4.14 show the distribution of the ICF ratings for each BV Mobility question response and the relationship between changes in ratings between subsequent points in time. The same data for rounded ICF values can be found in table 4.9 for the pairwise comparison ratings

and table 4.10 for the delta change between timepoints. For both analyses, we computed Kendall's Tau Correlation coefficient again in 4.11.

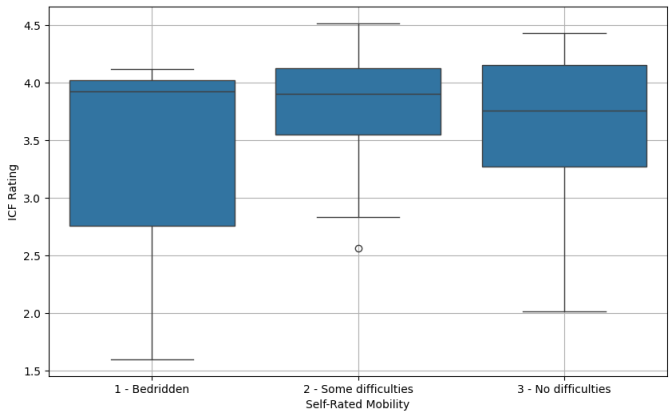


Figure 4.13: Box plot of ICF FAC ratings for each BV Mobility question response

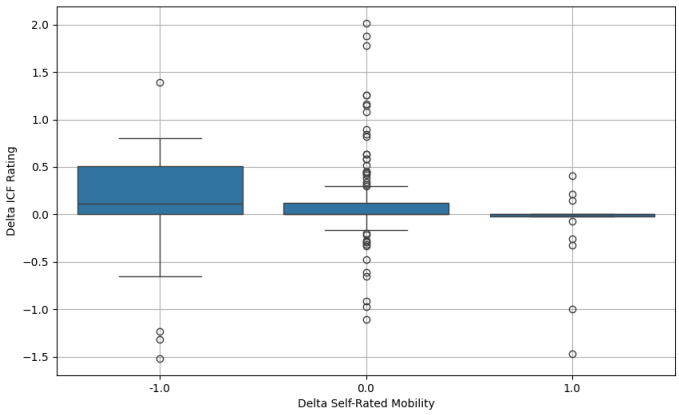


Figure 4.14: Box plot of Delta FAC ratings for each BV Mobility question response

ICF Ratings / Self-Ratings	**Bedridden**	**Some Difficulties Walking**	**No Difficulties Walking**	Total
FAC = 1	0	0	0	0
FAC = 2	1	0	7	8
FAC = 3	0	12	61	73
FAC = 4	2	77	109	188
Total	3	89	177	269

Table 4.9: Contingency Table of FAC ICF Ratings vs Self-Ratings

In table 4.9, we see that two patients who rated their mobility as 1 (bedridden) received an ICF rating of 4 during the same period, indicating almost no issues with mobility, only requiring verbal assistance. This is an interesting case to observe in detail, which we do in the Error Analysis chapter. Figure 4.14 plots same data, but instead of using the exact values, it used the delta change from T-1 to T. This helps us understand the relationships between our the change in ratings in our datasets. We also included the contingency table for the delta change between the rounded, discrete values.

ICF Ratings / Self-Ratings	**-1**	**0**	**1**	Total
-1	4	6	4	14
0	16	115	18	149
1	0	17	5	22
2	0	3	1	4
Total	20	141	28	189

Table 4.10: Contingency Table of FAC delta ICF Ratings between subsequent timepoints vs delta Self-Ratings

To interpret figure 4.14, a diagonal line would indicate a perfect relationship. We can see that if the self-rated delta value is 0, indicating no change in rating between timepoints, then the delta ICF rating ranges between -1 and +2, centering around 0. An ICF delta of -1 when the delta BV rating equals 0 indicates that the ICF rating decreased with 1 in the same period that the self-rated BV value did not change. Table 4.10 clearly shows how, for most delta changes, the delta value equals to 0. This means that between two subsequent timepoints, the rating did not change. We can see that for 16 delta ICF ratings of 0, the self-rating decreased with 1. For 18 delta ICF ratings of 0, the BV score increased with 1. However, for most cases, if the ICF rating did not change between two measurements, the same happened to the BV score.

	Pairwise	Delta change
Kendall's Tau (unrounded)	Coef: 0.06, p = 0.25	**Coef: -0.16, p = 0.01**
Kendall's Tau (rounded)	**Coef: 0.18, p = 0.003**	Coef: -0.12, p = 0.08

Table 4.11: Kendall's Tau Correlation Coeffecient (Coef) for exact ICF values and for rounded ICF values vs BV responses for the mobility question for both the pairwise comparison as the delta change between timepoints

Table 4.11 shows that the Kendall's Tau correlation found a significant correlation for the pairwise comparison when using the rounded ICF ratings. For the delta change between subsequent timepoints, however, the significant correlation was only found for the unrounded ICF values. Additionally, the coefficient is negative, indicating a negative correlation, which is contrary to expectations. This means that if one variable increases for a patient between subsequent questionnaires, the other variable is likely to decrease a little bit.

Figures 4.15 and 4.16 show the average rating over time and the average change over time.

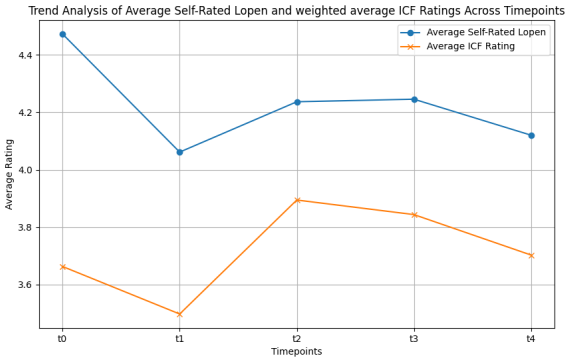


Figure 4.15: Overall Trend FAC ratings

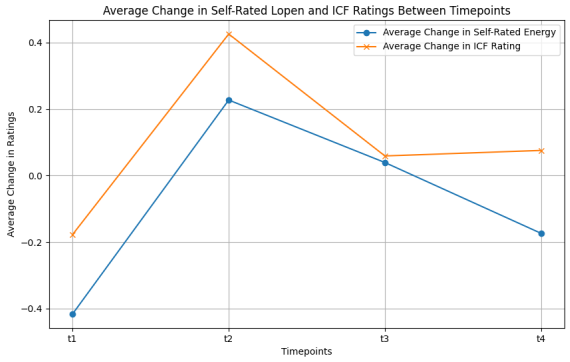


Figure 4.16: Overall Trend delta FAC ratings

Figure 4.15 clearly shows that while there is an absolute distance between ratings, with ICF ratings generally being lower than BV ratings, they develop similarly over time. When interpreting this graph with respect to the Kendall's Tau correlation we found earlier, it is important to consider that this graph is averaging the values for all patients at each timepoint. The above analysis is agnostic to timepoints, it groups all pairs of ratings together. Additionally, by averaging out all values, we lose some of the resolution provided by a pairwise comparison. This figure mainly serves to display the overall functioning curve a patient goes through, it is not a validation of the correlation we calculated.

Figure 4.16 shows the average change in values between subsequent points in time. It is, in a sense a derivative of figure 4.15 and therefore, it mainly serves to show the speed at which ratings change. We can see that when the ICF rating change is negative, the self-rated change is also negative, except for T4, which indicates the change from T3 to T4.

Energy (ENR)

The question for the Energy class is:

In the past 7 days, how would you rate your fatigue on average?

With the following possible responses: "Extremely heavy fatigue" (1), "Heavy fatigue" (2), "Medium fatigue" (3), "Light fatigue" (4), "No fatigue" (5).

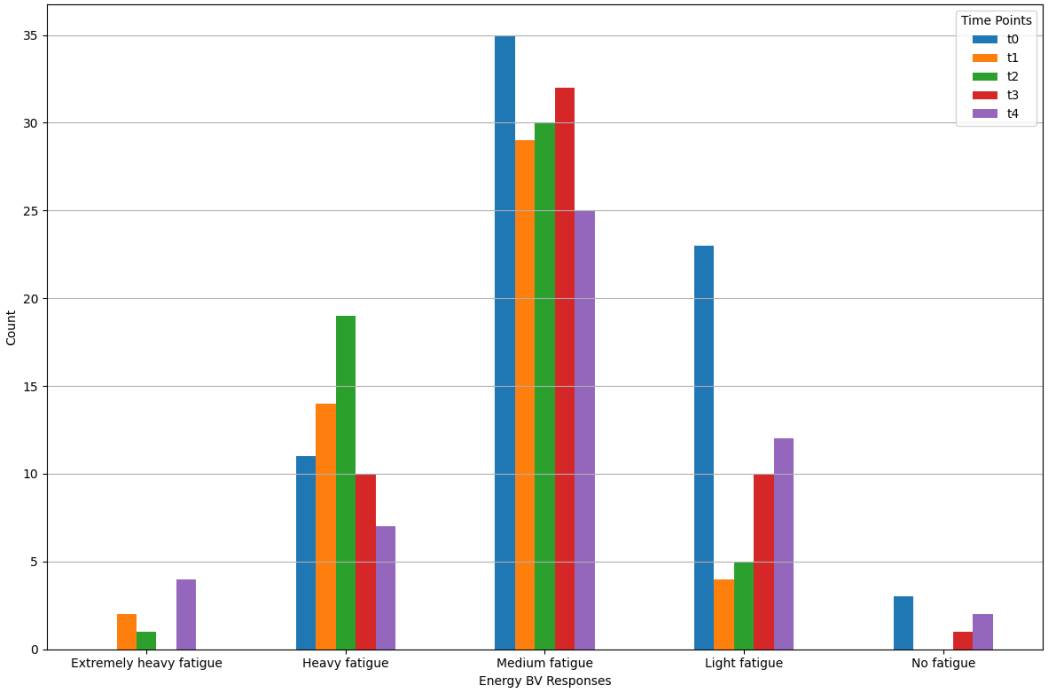


Figure 4.17: Histogram of Energy (ENR) BV Question Responses for each timepoint (t0 to t4) seperately

In figure 4.17 we can see how most people refered to their own energy as "Medium fatigue", however, there are also observations of patient with worse an better functioning. It is essential to look at this distribution in combination with the distribution of extract ICF ENR labels, as reported in 4.3, since that shows clear differences.

Since the question inquires about the patient's fatigue over the past 7 days, we aggregated the ICF levels by averaging the ratings for the same period and plotted them in a box plot 4.18. The first observation we make when reviewing this plot, is that BV rating 5 - No fatigue is not represented here. This is because in all 6 cases a participant replied with this option, no ICF ENR mentions were found in any of the notes in the last week. In figure 4.3 we showed the distibution of all ICF Energy levels extracted from the clinical notes. This shows that no ENR level above 2.5 was ever assigned to a clinical note. The fact that for this BV response, no relevant notes with ENR mentions were found, correctly represents no decreased Energy functioning. That being said, since we also cannot find any ENR qualifier level of 3 or 4, we have no way of representing medium and light fatigue in the clinical notes. Since the ENR qualifier levels only reach up to 2.5, we do not expect to find a strong correlation.

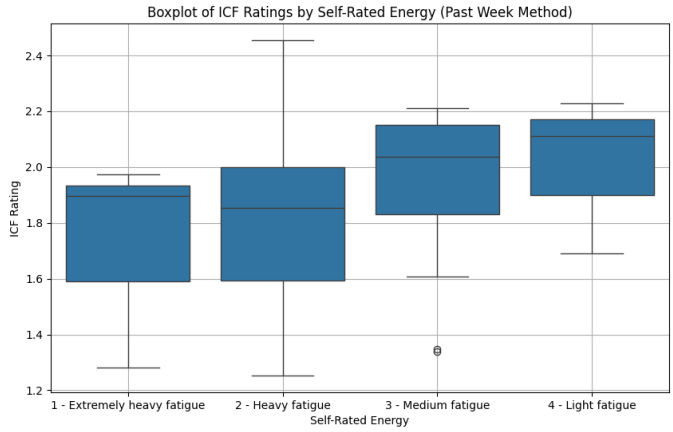


Figure 4.18: Box plot of ICF ENR ratings for each BV Energy question response

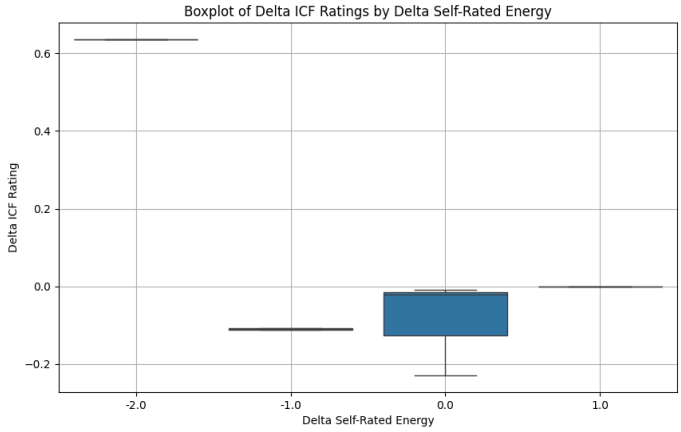


Figure 4.19: Box plot of Delta ENR ratings for each BV Energy question response

ICF Ratings / Self-Ratings	**1**	**2**	**3**	**4**	Total
ENR = 1	1	3	2	0	6
ENR = 2	2	14	16	7	39
Total	3	17	18	7	45

Table 4.12: Contingency Table of ENR ICF Ratings vs Self-Ratings with 1: Extremely heavy fatigue, 2: Heavy fatigue, 3: Medium fatigue, 4: Light fatigue

Based on figure 4.18, it is difficult to identify a relationship between the datasets. The plot shows a weak relationship, as the distribution shifts up when self-rated energy levels increase. However, we do not expect this relationship to be significant. A similar observation can be made from table 4.12.

Due to the infrequency of the ENR labels, computing the delta change for a 1-week period results in fewer data points, as only patients with two subsequent measurements are included. Therefore, we used the average of the past 2 weeks instead. This still did not result in a large dataset, but we could not increase the window further, as it would cause overlapping windows for different timepoints.

ICF Ratings / Self-Ratings	**-2**	**-1**	**0**	**1**	Total
-1	0	0	2	0	2
0	0	7	18	8	33
1	2	0	1	0	3
Total	2	7	21	8	38

Table 4.13: Contingency Table of ENR delta ICF Ratings changes vs delta Self-Ratings changes

Inspecting figure 4.19, we can deduce that there is likely no relationship between changes over time for the self-reported values and the extracted ICF ratings. This is likely because most delta BV scores are around 0, indicating no change between measurements. We can see the same effect in table 4.13. It can also be due to the infrequency of ENR labels. In section 4.1.1, we identified that a patient has 10 ENR labels on average. Additionally, in section 4.1.2, we reported that 7 out of 10 notes appear during hospitalization, with only 3 on average in the post-discharge period.

This is similar to Weight Maintenance (MBW), where also only 10 mentions were identified on average, but since that class referred to the last 1 month and 6 months respectively, it did not result in many empty observations. The BV question about fatigue and energy is repeated in each questionnaire, focusing mainly on the post-discharge period, except for the intake questionnaire (T0). As a result, we expect only 3 out of 4 post-discharge questionnaire periods to contain an ICF value on average. Since these plots visualize changes between measurements, we end up with only 1 or 2 subsequent measurements per patient on average.

	Pairwise	Delta change
Kendall's Tau (unrounded)	**Coef: 0.26, p = 0.03**	Coef: 0.05, p = 0.87
Kendall's Tau (rounded)	Coef: 0.21, p = 0.14	Coef: -0.59, p = 0.12

Table 4.14: Kendall's Tau Correlation Coeffecient (Coef) for exact ICF values and for rounded ICF values vs BV responses for the energy question for both the pairwise comparison as the delta change between timepoints

Table 4.14 shows that for the exact values, we did not find a statistically significant correlation using Kendall's Tau correlation. This is different than the result for Mobility (FAC), where the statistically significant correlation was only found for the rounded values. We suspect this is because the Mobility questions refer to the day itself, making categorical ICF ratings yield better results. In contrast, Mood and Energy questions refer to a period, requiring an aggregation method to find a single representative value. Rounding the ICF values before aggregation likely loses some nuances that the ICF classifier model picks up.

The delta change for ICF against BV ratings shows no significant relationship. Moreover, the delta change for the rounded values even shows a negative relationship, albeit not statistically, which is contrary to expectations. This is likely due to data sparsity for this class.

Figure 4.20 and 4.21 show the average ICF ratings versus the average BV ratings for all patients and the average change in values over time. Note that the change over time suffers from the same infrequency of labels as the individual delta change graphs.

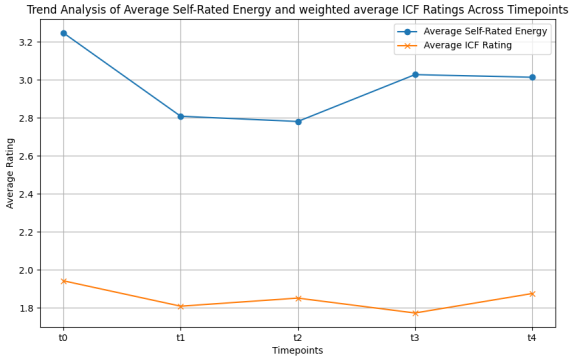


Figure 4.20: Overall Trend ENR ratings

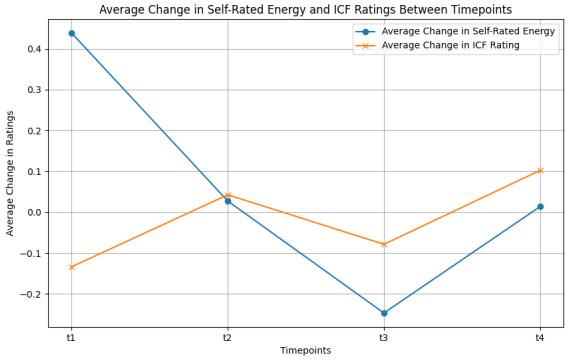


Figure 4.21: Overall Trend delta ENR ratings

These graphs also show no obvious relationship between the two datasets. We hypothesize that this is mainly due to infrequent labels rather than a mismatch in measured constructs. We can see the effect of the data sparsity very clearly by comparing both graphs, as it is clear that the delta change (4.21) over time is not correctly reflecting figure 4.20. Where the former shows a decrease in ratings between T0 and T1, the latter says that self-ratings increased from T0-T1. This is likely because of missing values, which only affects the average value on the same timepoint in figure 4.20, but affects the delta change to and from the timepoint containing the missing value.

Mood (STM)

The mood-related BV question was:

> "Please select one option for each of the groups in the list below that contains a sentence that best describes your health situation today. Mood:"

With the following possible responses and their encodings in below graphs and tables: "I am not anxious or feeling down" (3), "I am somewhat anxious or feeling down" (2), "I am very anxious or feeling down" (1).

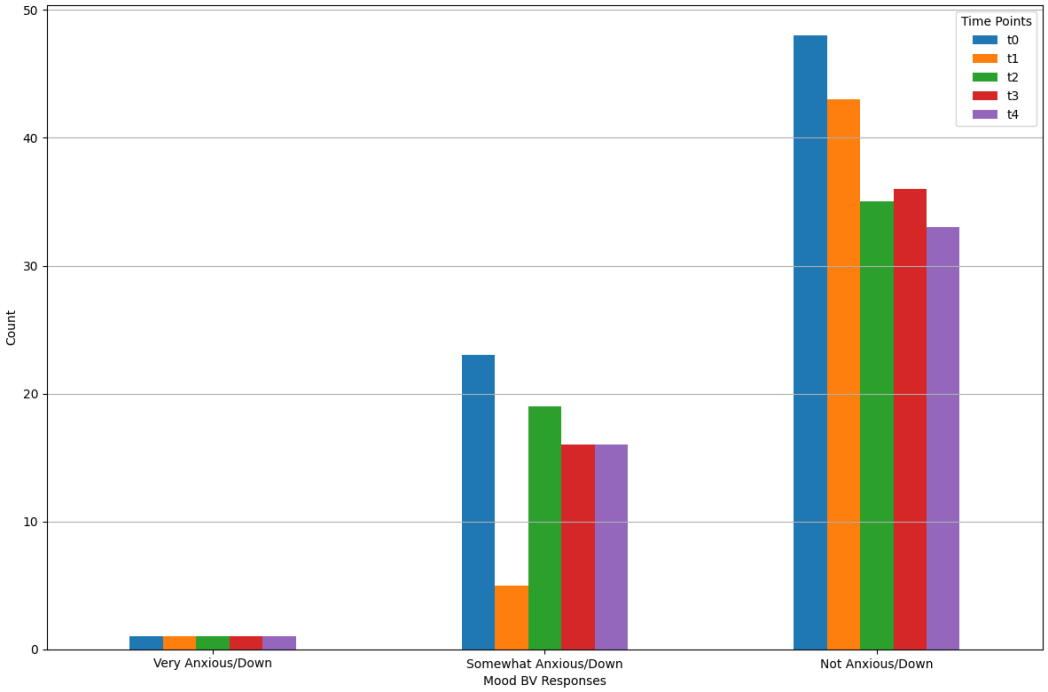


Figure 4.22: Histogram STM

Figure 4.22 shows that only a single patient rated themselves as very anxious or down for each timepoint. Most patients, however, responded with value 3, indicating no issues with their mood. Similar to the other ICF categories, we can see that participants were more likely to respond to questionnaire at timepoint T0 than for the subsequent questionnaires.

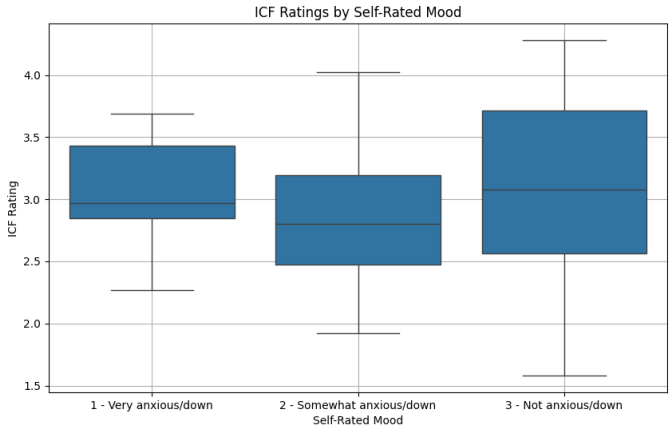


Figure 4.23: Box plot of ICF STM ratings for each BV Mood question response

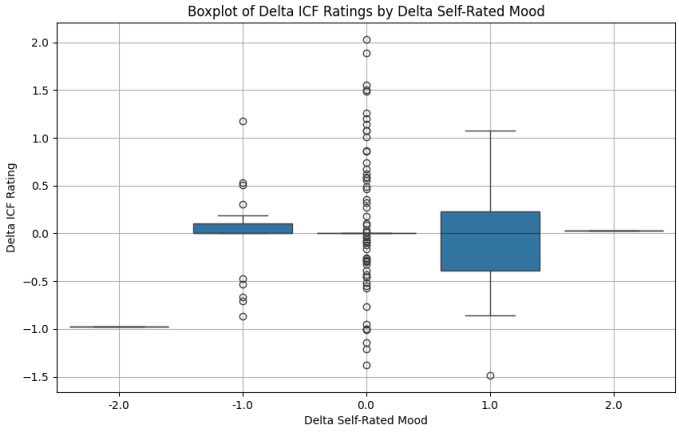


Figure 4.24: Box plot of Delta STM ratings for each BV Mood question response

ICF Ratings / Self-Ratings	**1**	**2**	**3**	Total
STM = 2	1	21	43	65
STM = 3	3	46	83	132
STM = 4	1	12	59	72
Total	5	79	185	269

Table 4.15: Contingency Table of STM ICF Ratings vs Self-Ratings with 1: Very anxious/down, 2: Somewhat anxious/down, 3: Not anxious/down

In both figure 4.23 and table 4.15, most data points for the BV responses are at level 3. The relevant ICF ratings for these levels are scattered across 2, 3 and 4, centering around 3 and with slightly more (59) at ICF level 4 than level 3 (43), indicating good mood-related functioning. For

a BV response of 2, the distribution of ICF ratings shifts to slightly lower values (21 at level 2 versus 12 at level 4), suggesting a weak to medium correlation between both datasets.

When inspecting figure 4.24 containing the delta change over time, the relationship between the datasets becomes less clear. For a delta change of -1 up to +2 for self-rated Mood levels, the median of delta changes in ICF ratings stays around 0. This means that regardless of a change in self-rated Mood levels, the ICF ratings tend to stay the same. This can also be observed when inspecting the contingency table 4.16 with delta ratings for the rounded ICF values. Values centered around a diagonal in this table would indicate a strong relationship, but we can see the values being centered across the middle column and row instead, indicating that for most changes in the values between subsequent timepoints, the other variable tends to stay static.

ICF Ratings / Self-Ratings	**-2**	**-1**	**0**	**1**	**2**	Total
-2	0	0	1	1	0	2
-1	1	4	13	4	0	22
0	0	21	93	15	1	130
1	0	4	14	5	0	23
2	0	0	4	0	0	4
Total	1	29	125	25	1	181

Table 4.16: Contingency Table of STM delta ICF Ratings vs delta Self-Ratings

Both figure 4.24 and table 4.16 do not clearly show a relationship between the change of ratings between timepoints. The correlation metrics for the delta change confirm this, as can be seen in the table 4.17.

Mood (STM)	Pairwise	Delta change
Kendall's Tau (unrounded)	**Coef: 0.12, p = 0.02**	Coef: -0.02, p = 0.76
Kendall's Tau (rounded)	**Coef: 0.12, p = 0.04**	Coef: 0.01, p = 0.77

Table 4.17: Kendall's Tau Correlation Coeffecient (Coef) for exact ICF values and for rounded ICF values vs BV responses for the mood question for both the pairwise comparison as the delta change between timepoints

Table 4.17 contains the Kendall's Tau correlation coefficients for the pairwise comparison and the delta change over time. We can see that for both the unrounded and the rounded ICF values, a significant correlation was found, however, when inspecting the delta change over time, we do not find such a relationship. Similar to the other analyses in case 3, we also plot the average rating per timepoint and the average delta change in ratings between each timepoint to inspect the general functioning curve described by both datasets. These can be found in figures 4.25 and 4.26. An interesting observation is that the self-rated Mood tends to be higher rated than the ratings as assigned by the ICF classifier models. The self-ratings from the BV questionnaire - on average - stays between 4 and 5, while the average ICF rating at each timepoint does not exceed 3.25. However, we can still observe a peak at timepoint T1, which is the first questionnaire post-discharge, followed by a decrease afterwards.

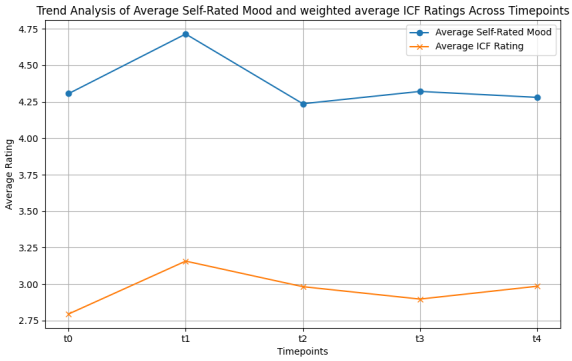


Figure 4.25: Overall Trend STM ratings

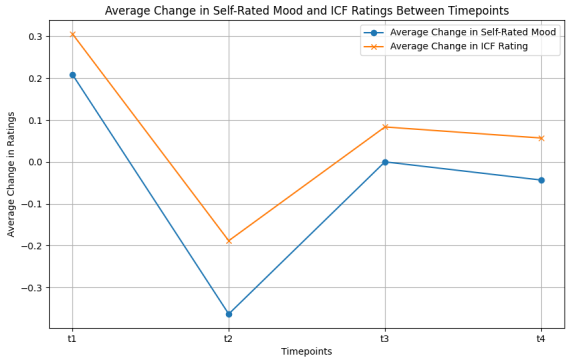


Figure 4.26: Overall Trend delta STM ratings

4.3 Experiment 1: Error Analysis

In chapter 3, we argued that the implications of the ICF classifier models were not sufficiently explored and quantified. Experiment 1 serves as an external validation of the ICF classifier output. For each of the relevant ICF classes, we compared the constructs measured by the questionnaire data with the extracted ICF qualifier levels. For each class, we found that at least one aggregation method of ICF ratings showed significant correlations with the self-reported values.

While this is valuable information for verifying the usefulness of the ICF classifiers, we need to perform an error analysis to fully understand these relationships. In this section, we will extract examples for each ICF class where the functioning data from both datasets overlaps and, perhaps more importantly, where the datasets disagree.

Weight Maintenance (MBW)

For Weight Management, we are particularly interested in identifying patients who gave themselves a rating of 2, indicating no unintentional weight loss, but were still assigned an ICF rating of 2.5 or lower, indicating some unintentional weight loss. We de-identified the notes (indicated with [square brackets]) and removed any irrelevant sentences for this analysis.

We identified a patient who responded to the question about losing 3kg unintentionally in the last month with a rating of 2 for T0 and 1 for T4, indicating they had lost weight in the month prior to hospitalization, but not in the month before their last questionnaire.

The lowest assigned ICF MBW rating in the same period is from a phone consultation 20 days prior to their intake. This EHR entry was assigned an MBW qualifier level of 1.83, indicating unintentional weight loss of 3kg in the last month. However, the patient indicated they had not lost 3kg. Upon inspection of the clinical note, we found the following description:

> "Height: [170+] cm Weight: 73 kg Body Mass Index is [24+] kg/m2. Weight 1 month ago (% change compared to current weight): 75 kg (-2.67%) Weight 6 months ago (% change compared to current weight): 78 kg (-6.41%) Usual weight (% change compared to current weight): 78 kg (-6.41%)"
>
> (Original: "Lengte: [170] cm Gewicht: 73 kg Body Mass index is [24+] kg/m$\hat{2}$. Gewicht 1 maand geleden (% verandering tov huidig gewicht): 75 kg (-2.67%) Gewicht 6 maanden geleden (% verandering tov huidig gewicht): 78 kg (-6.41%) Gebruikelijk gewicht (% verandering tov huidig gewicht): 78 kg (-6.41%)")

This example describes a (supposedly unintentional) weight loss, but it amounts to 2kg in the last month. Therefore, the patient correctly responded to the BV question that they had not lost 3kg, while the ICF classifier model still assigned a low qualifier level. This is not unexpected behavior for the model, as it uses word embeddings to represent words and numbers, likely losing the arithmetic information present in the numbers. In other words, a language model does not inherently understand the rational quality of numbers; it knows 5 is bigger than 4 but does not necessarily understand the exact difference between 5 and 4 compared to 4 and 3. The model likely learned from the annotations that any weight loss results in a lower MBW ICF score, regardless of the size of the loss.

This patient answered the same question for T4 with 1, indicating they had lost 3kg weight unintentionally. Upon inspection, we found another phone consult that was assigned MBW level 2.04.

> "Weight [date of note]: 71 kg Admission weight: 74.8 kg Usual weight: 78 kg."
>
> (Original: "Gewicht [datum van notitie]: 71 kg Opnamegewicht 74,8 kg Gebruikelijk gewicht: 78 kg.")

This description corresponds with the patient's self-reported weight loss of over 3kg in the last month. The ICF label was correctly assigned as 2. It is interesting to note that this ICF rating is higher, unrounded, than the previous rating, even though the weight loss exceeded 3kg only the second time. The first example has more separate mentions of weight loss, but without a sentence boundary, all mentions are part of the same sentence in the model pipeline. We expect that the higher density of weight loss mentions in the first example caused the model to assign a lower level, even when the weight loss was smaller.

Mobility (FAC)

In the analysis of Walking or Mobility-related functioning, we noticed that out of 3 patients who said they were bedridden at some point, 2 received a relatively high FAC rating in the nearest clinical note. Upon inspection of the note in question, however, we did not find a specific walking related mention. The note seems to be a progress report focusing on gastrointestinal issues. However, there is a single sentence present that we expect might have affected the ICF output:

> Translation: "During the day, frequently gags and burps up mucus."
>
> Original: "Loopt overdag veel te kokhalzen en opboeren van slijm."
>
> Transliteration: "Walking during the day a lot gagging and burping of mucus."

The translation of this sentence does not make it clear why the model might have misidentified this mention. However, the transliteration shows how the original Dutch sentence starts with words that could be translated as "Walking during the day a lot." We suspect the model learned to identify mentions like these as mobility-related. The fact that it did, suggests the model did not generalize well during training.

Another patient, who also rated their mobility as 1 (indicating they were bedridden), received an ICF rating of 4. However, after inspecting the relevant note, it does not seem the model made a similar mistake this time.

"Around the surgery, spent two months in the hospital, of which two weeks in the ICU. Four weeks after discharge, things were going very well, could walk 2 km without a walker, all aids were removed from the house. Now 2 months at home, things have been getting progressively worse, especially in the last 3 weeks. Not short of breath at rest. But with very mild exertion already breathless. Panting, high breathing, really has to make an effort to regain composure. It hinders him in everything."

(Original: "Rondom operatie twee maanden in ziekenhuis gelegen, waarvan twee weken op IC. Vier weken na ontslag ging het heel goed, kon 2km lopen zonder rollator, alle hulpmiddelen uit het huis weggehaald. Nu 2 maanden thuis, het gaat sindsdien steeds slechter, met name de laatste 3 weken. Niet kortademig in rust. Maar bij hele milde inspanning al benauwd. Hijgen, hoge ademhaling, moet echt zijn best doen weer tot zinnen te komen. Het belemmert hem in alles.")

From this excerpt, we can conclude that the model mistakenly identified a historic mention of good mobility as a current event, even though the rest of the excerpt explains otherwise. This excerpt contains multiple sentences, indicated by full stops. Since the ICF classifier models analyze each sentence individually, and only the positive sentence specifically mentions the word "walk," only this sentence was identified as a mobility mention. The other sentences, which describe worsened functionality, do not contain the word "walk" and are likely identified as "exercise tolerance" mentions instead. The same note was assigned an Exercise Tolerance qualifier of 2.5.

This example highlights the importance of temporal disambiguation. While annotators marked any annotation specifically when referring to past or future events, this was not included in the training set. Since all historic and future mentions were filtered out, the model has never seen such a sample during training and was never evaluated on how it handles these cases. If these disregarded notes had been left in the training set and labeled as 'no mention,' the model would have used this loss to generalize and improve its performance, which would have been reflected in the evaluation. We recommend improving the current model setup by retraining and evaluating it with temporal disambiguation.

Energy (ENR)

For the Energy ICF classes, we noted that the only existing identified levels by the ICF models were 1 and 2. Therefore, we are mainly interested in cases where the self-rated energy level reflects no issues with energy and fatigue. The following excerpt is from a note that was assigned an ICF ENR value of 1.90, which corresponds to the following description from the annotation guidelines:

Moderate fatigue; the patient gets easily tired from light activities or needs longer time to recover after an activity.

"[Date 2 weeks prior to current date]: [...] Less energy and power."

(Original: "[Datum 2 weken voor huidige datum]: [...] Minder energie en kracht.")

Similar to the case we discussed about Mobility, this note contains a history of previous phone contact and describes worsened energy levels 2 weeks before the date of the note. As such, the ICF classifier assigns labels to this date, even though it does not reflect a description of functioning on that date.

Another patient rated their own energy levels as being good upon hospital admission, and while the note seems to confirm that, the ENR rating was still 2.2, indicating some energy-related issues. The description, however, is very positive:

> "History: Triple therapy went well, few side effects, and now in good condition: tennis, exercise bike, skiing. Not fatigued, no dyspnea or cough. [...] General: Vital man, appearance conforming to calendar age, does not look ill, no pallor, no jaundice."
>
> (Original: "Anamnese Tripletherapie goed gegaan, weinig bijwerkingen en nu goede conditie: tennis, hometrainer, gietsen. Niet vermoeid, geen dyspnoe of hoest. [...] Algemeen: Vitale man, uiterlijk conform kalenderleeftijd, oogt niet ziek, bleek -, icterus.")

It is unknown why the ICF classifier model misidentified these sentences. The MAE score for the IAA computation was 0.39, indicating the average disagreement in annotated qualifier levels. Additionally, the MAE score for the sentence-level evaluation is 0.48. This means we can expect a maximum mismatch between the assigned ICF level and the real functioning level of, on average, 0.87. Even assuming this standard error, the maximum ENR rating divergence would be 3.1, which still does not completely reflect this description.

In chapter 3, we argued that reporting the distribution of annotated data is essential to interpreting these results. Since the ICF classifier output on the notes used during experiment 1 was always around 1 or 2, we expect the annotated data had similar distributions. Therefore, the model might not have learned that any ICF label over 2 was possible. Without knowing this for certain, we cannot fully explain these results. However, if the distribution of annotated ICF ENR labels is similar to the distribution in this dataset, we can assume that the originally selected notes for annotation were not representative of the full variety of clinical notes.

Mood (STM)

For the mood class, we identified that most BV responses were label 3. We only noticed a small shift of distributions when comparing the ICF distribution at BV label 3 and BV label 2, which indicated a small effect size, but a linear relationship. The correlation metrics showed a similar effect, where the effect size was small, but the relationship significant. In this error analysis, we will take a look at any case where there was a big mismatch between the self-reported mood and the ICF-based mood score.

> "Psycho-Social: has had trouble a week after the treatment, nausea, malaise, no appetite. Was happy that the blood count had improved."
>
> (Original: "Psycho-Sociaal: heeft een week na de kuur last, misselijk- malaise- geen eetlust. Was blij dat het bloedbeeld was verbeterd.")

We expect that this misidentification of the mood level stems from the description "was happy that the blood count had improved". This is identified as mood being good, while in reality only one day later the patient responded to the question in the BV questionnaire with "I am somewhat anxious or feeling down".

The next example is the other way around, the extracted Mood level is higher than the patient's own Mood rating. They responded with rating 1: I am very anxious or feeling down.

> "Spent a long time talking with mr in the role of case manager after discharge. Physically, it is going better, mr is tapering off the pain medication. [...] Mr feels mentally not well. Feels like he has fallen into a black hole now that he is home and the surgery is behind him. Has also had two panic attacks. Is very worried about the future/prognosis. Feels lonely. Has already been to the GP and will receive supportive conversations. Mr was happy with our conversation."
>
> (Original: "Lange tijd met dhr gesproken in de rol van casemanager na ontslag. Fysiek gaat het eigenlijk wel, dhr is de pijnstilling aan het afbouwen. [...] Dhr voelt zich mentaal echter nite goed. Heeft het idee dat hij in een zwart gat is gevallen nu hij thuis is en de operatie achter de rug is. Heeft ook twee keer een paniek aanval gehad. Maakt zich erg zorgen om de toekomst/prognose. Voelt zich eenzaam. Is al bij de HA geweest en zal ondersteunende gesprekken krijgen. Dhr was blij met ons gesprek.")

This is an interesting example because almost the entire note is negative about the patient's mood, clearly stating that he has had panic attacks and feels lonely. However, the last sentence says the patient was happy with the conversation. We expect this unevenly affected the final rating, as this sentence was likely assigned a high level, pushing the average score up.

This also shows how sentence splitting could affect the final rating. If the note author was inconsistent with sentence full stops, as seen in previous examples, the number of sentences rated low could vary. This affects the weight the final good score has on the total average. If we delete the full stops between all sentences describing the patient's mood as negative, but keep the final sentence separate and run the classifier again, the final rating becomes 3.53 instead of 2.73, without changing the note's content.

Therefore, we know the model is sensitive to sentence boundaries, even though the meaning or content of the note is not significantly affected by them. We recommend conducting a broader sensitivity analysis to determine how other text alterations that do not affect the meaning impact model output in future research.

In the this section, we reported the results of the external validation of the ICF model output using questionnaire data. We found that 5 out of the 9 ICF classes that the ICF classifier models were trained on were sufficiently described and evaluated for the original publication. For each of these classes, there is some correlation between the extracted functioning level and the self-reported functioning classes.

4.4 Experiment 2: Predicting post-discharge functioning using in-hospital functioning data

The purpose of Experiment 2 is to quantify the predictive value of in-hospital functioning data on post-discharge functional recovery. In Experiment 1, we described the utility of the ICF classifier models output by comparing them with a patient's self-reported functioning experience. Now that we have established the correlation between functioning data as assigned by the ICF classifier model and self-reported functioning data, we will use the ICF classifier output as a proxy of a patient's functioning.

In section 3.4, we described the methodology for training the prediction model. We experimented with different combinations of input representation techniques, such as rounding the ICF qualifier levels to fit the original ordinal and discrete distribution instead of the continuous lev-

els. We also experimented with the addition of structured clinical data to see its effect on model performance. For all these experiments, we used an LSTM architecture.

Next to testing different input representations, we also test different ways of aggregating the post-discharge data into a target label. In the previous section, we showed that there are multiple ways of aggregating multiple ICF ratings into a single rating, used for correlation analysis. Two methods were taking a weighted average of multiple ratings in a specified window and taking two of the closest measurements on both sides of the target date and weighing them based on the distance to the target date. For the target labels for the prediction model training, we will use both aggregation methods as well.

The LSTM model uses the Mean Absolute Error (MAE) as the loss function, similarly to how the qualifier level prediction loss was calculated. We evaluate the performance on the test set using the MAE.

We train a prediction model for each of the target ICF categories. For each of these, we rebalance the overrepresented samples with negative target labels (missing values represented as 5) to contain as many samples as positive labels. Figure 4.27 shows the distribution of the target labels for the ADM class if the dataset is not rebalanced:

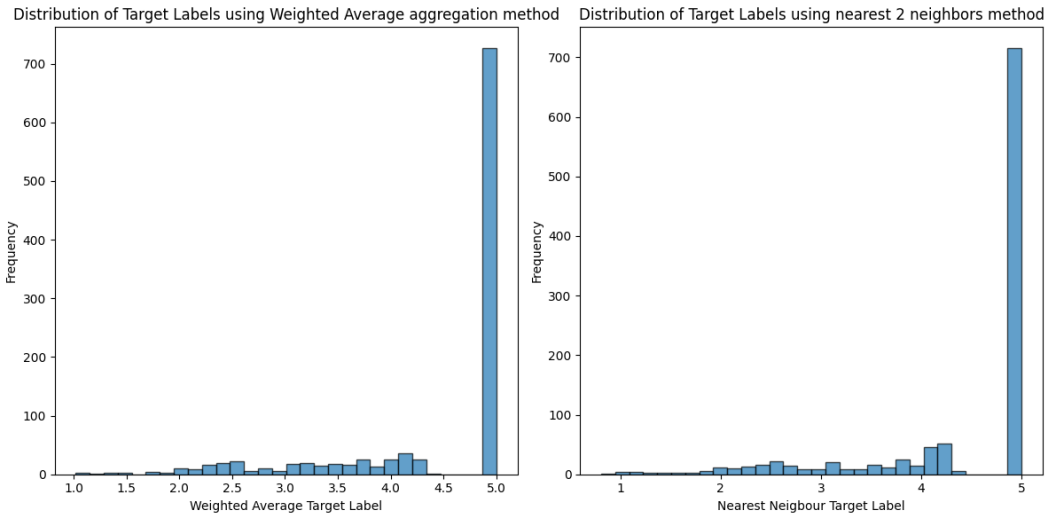


Figure 4.27: Distribution of Respiratory Functioning target labels

This shows that the aggregation method of selecting the two nearest neighbors provides us with a distribution with more variation. Therefore, we expect slightly worse performance metrics for this aggregation method.

We trained the LSTM models using the same setup. The optimizer is Adam, the loss function MAE and the activation function in the last layer linear and for the hidden layers it's tanh and the recurrent activation function in sigmoid. We use a batch size of 8. The LSTM models contain an input layer, output layer and two hidden layers. The input layer has no specified dimensions to allow for variable input sizes, the hidden layers exist of 64 and 32 cells respectively and the output layer is a dense layer of size 1. We train for 20 epochs with the amount of steps per epoch specified as the length of the training set divided by the batch size. This model setup is also detailed in 3.13.

Tables 4.18 to 4.22 report the MAE on the test set for each of the 5 ICF classes: Repiratory Functioning (ADM), Mobility (FAC), Mood (STM), Weight Maintenance (MBW) and Energy (ENR).

Respiratory Functioning (ADM)	Unrounded		Rounded	
	Weighted Average Target	Two nearest neighbors Target	Weighted Average Target	Two nearest neighbors Target
Excluding Structured Data	0.48	1.67	0.49	1.62
Including Structured Data	0.49	1.77	0.50	1.76

Table 4.18: Comparison of MAE score on test set for Respiratory Functioning for each combination of ICF representation (Unrounded/Rounded), target label aggregation method (weighted average/two nearest neighbors) and inclusion of structured data (including/excluding)

Mobility (FAC)	Unrounded		Rounded	
	Weighted Average Target	Two nearest neighbors Target	Weighted Average Target	Two nearest neighbors Target
Excluding Structured Data	0.52	1.81	0.52	1.73
Including Structured Data	0.53	1.74	0.53	1.94

Table 4.19: Comparison of MAE score on test set for Mobility for each combination of ICF representation (Unrounded/Rounded), target label aggregation method (weighted average/two nearest neighbors) and inclusion of structured data (including/excluding)

Mood (STM)	Unrounded		Rounded	
	Weighted Average Target	Two nearest neighbors Target	Weighted Average Target	Two nearest neighbors Target
Excluding Structured Data	0.41	1.61	0.41	1.62
Including Structured Data	0.41	1.49	0.39	1.55

Table 4.20: Comparison of MAE score on test set for Mood for each combination of ICF representation (Unrounded/Rounded), target label aggregation method (weighted average/two nearest neighbors) and inclusion of structured data (including/excluding)

Weight Maintenance (MBW)	Unrounded		Rounded	
	Weighted Average Target	Two nearest neighbors Target	Weighted Average Target	Two nearest neighbors Target
Excluding Structured Data	0.47	1.09	0.47	1.10
Including Structured Data	0.49	1.09	0.49	1.11

Table 4.21: Comparison of MAE score on test set for Weight Maintenance for each combination of ICF representation (Unrounded/Rounded), target label aggregation method (weighted average/two nearest neighbors) and inclusion of structured data (including/excluding)

Energy (ENR)	Unrounded		Rounded	
	Weighted Average Target	Two nearest neighbors Target	Weighted Average Target	Two nearest neighbors Target
Excluding Structured Data	0.27	1.00	0.26	1.04
Including Structured Data	0.26	0.94	0.26	0.94

Table 4.22: Comparison of MAE score on test set for Energy for each combination of ICF representation (Unrounded/Rounded), target label aggregation method (weighted average/two nearest neighbors) and inclusion of structured data (including/excluding)

The first thing we notice when reviewing the results, is the minimal change in MAE scores between rounding and not rounding the ICF values. We can also see that the weighted average aggregation method yields better results. The inclusion of structured data does not seem to affect the model performance much.

Upon first inspection of these results, it looks like ENR is performing the best. This, however, is likely due to the lower variation in the ENR labels. This illustrates the importance of reporting on the label distribution, because a score like MAE is dependent on the range of possible labels. However, we have no reason to expect that the distribution of ENR labels differs in our datasets from the annotated dataset used to train those models, so we can still use this score to compare with the MAE scores as reported about the ICF classifier models.

In conclusion, most prediction models predict the physical functioning of patients after discharge with a similar error margin as the ICF classifier models have for their functioning level predictions. We expect that if the ICF model's performance were to increase, the predictive value of those labels on post-discharge functioning would increase as well.

4.4.1 Baseline

We compare these results with baseline models for each ICF class, consisting of an LSTM prediction model trained on solely the structured notes, with the exact same setup as the prediction models from Experiment 2, to be able to reason about the added predictive power of the ICF ratings. This model was trained using only the structured data from the EHR, excluding the assigned ICF levels during hospitalization. The MAE scores on the test set, including a comparison with the best-performing prediction model, can be found in table 4.23.

Baseline	Weighted Average Target (MAE)	Two nearest neighbors Target (MAE)
ADM	MAE: 0.48 (prediction model MAE: 0.48)	MAE: 1.81 (prediction model MAE: 1.62)
FAC	MAE: 0.75 (prediction model MAE: 0.52)	MAE: 1.79 (prediction model MAE: 1.73)
STM	MAE: 0.93 (prediction model MAE: 0.41)	MAE: 1.62 (prediction model MAE: 1.49)
MBW	MAE: 0.67 (prediction model MAE: 0.47)	MAE: 1.30 (prediction model MAE: 1.09)
ENR	MAE: 0.25 (prediction model MAE: 0.26)	MAE: 1.20 (prediction model MAE: 0.94)

Table 4.23: MAE score on test set for the baseline model, for both the Weighted Average Target and the Two nearest neighbors target. In brackets the MAE scores for the best performing prediction models

This comparison against a baseline shows us that even just using the structured data has some predictive value on post-discharge functioning. It would be very interesting to see which of these

variables affected the output most. We can also see how the MAE score for the ENR and ADM predictions are almost the same for the baseline model as for the experimental models. For ENR, this is likely the lowest score due to the lower variation in target labels. ENR was the ICF class that we identified as showing the weakest correlation with self-reported functioning in Experiment 1, which can explain why this is the class where the baseline model and experimental models do not differ as much as some other classes.

The case for ADM is more difficult to explain. It appears that the structured data alone can identify the post-discharge ADM score to the same degree as the in-hospital functioning data. However, this might be because the ICF classifier models were originally trained for a COVID-19 project, where respiratory functioning issues are more likely to appear. As a result, the dataset used for this thesis likely differs most in terms of the frequency of labels from the dataset used to train the ICF models for the ADM class. Closer inspection of these results is required to understand this better.

4.4.2 Error Analysis

For each of the 5 ICF classes, we explore the errors and correctness of the prediction models by analysing individual predictions. Since the target label for this modeling task is an average of the ICF ratings around the target date of the 6th week post-discharge, this inspection is of qualitative nature. It entails assessing whether the average rating corresponds with the described functioning (i.e. was the ICF classifier model correct to begin with) and reasoning about what the error margin of the predictions mean in a hospital context. Ideally, in the error analysis, we pay attention to all features the model used for prediction, which includes looking at whether the assigned ICF labels to the in-hospital notes were correct. However, a qualitative inspection of all notes during hospitalization and all notes after discharge is too time consuming. Therefore we focus mostly on the prediction and the potential error while not speculating on which input features might have caused the error.

Respiratory Functioning (ADM)

Correct Prediction:

We identified a patient with the target ADM rating of 3.121. The predicted ADM rating is 3.183. This patient received a two identical ADM ratings in the discharge period, both 3.121, resulting in the same value for the average. While both notes are different EHR entries, they both contain the identical sentence, which is why both received the same ADM rating.

> "Accompanied by mild dyspnea symptoms, with the sensation of occasionally not getting enough air or skipping breaths."
>
> Dutch: "Hierbij lichte dyspnoe klachten, idee af en toe geen lucht te krijgen of ademhaling over te slaan."

While the note describes shortness of breath, it does not describe whether this happens only during exercise (ADM: 3) or in rest (ADM: 2). Therefore, it is difficult to assess whether this rating is correct. In any case, we know that the prediction model correctly predicted the same rating the ICF classifier models gave to this description.

Incorrect Prediction:

This paragraph shows an example of a wrong prediction by the model. This patient received the average rating of ADM: 1.445. Again, this average was computed from three identical ratings, all assigned to different EHR entries, two for progress reports and one for a telephone consultation. We identified only a single word referring to respiratory functioning in the first note:

> "dyspnea"
>
> Dutch: "dyspnoe"

In the second and third note, we found the following description:

> "Bothered by a dry, tickling cough for which codeine is used as needed."
>
> Dutch: "Last van droge kriebelhoest waarvoor zo nodig codeïne."

First of all, it interesting to see that all three notes received the exact same ICF ADM rating, up to 8 decimals precision, which is not expected from a regression model trained on such a large dataset given the difference between both descriptions.

All three notes describe readmission of the patient after discharge. While readmissions often indicate decreased functioning, none of these descriptions describe a respiratory functioning level of 1. The example of of a correct prediction above is much more specifically describing a decrease in respiratory functioning, yet they received a higher ADM rating.

The ADM prediction model from Experiment 2 predicted an ADM functioning level of 3.971. While this is also not reflective of the described respiratory functioning, it is a more realistic rating given these descriptions. The rating of 1.455 which was assigned to all three notes would correspond to the patient needing additional oxygen, which does not correspond to these EHR entries. This is an example of why we expect the prediction models to improve in performance when the ICF classifier models are able to more precisely identify functioning description in clinical notes.

Mobility (FAC)

Correct Prediction:

The patient we selected for this review, is one who received a single FAC = 3.883 rating for a progress report near the 6th week post-discharge target. The prediction model predicted it would be 4.435. Both reflect the same ICF qualifier level, with the following description in the annotation guidelines: "Patient requires verbal supervision for walking, without physical contact. It was difficult to find the corresponding part in the note, because it does not seem to describe decreased mobility specifically."

We identified these phrase that might have affected the assigned ICF rating:

> "Medical history: Progressing slowly. Feels pressure in the lower abdomen and sensitivity there after standing and walking for a long time."
>
> Dutch: "Anamnese: Gaat langzaam vooruit. Houdt druk op onderbuik en gevoeligheid daar na lang staan en lopen."

This does not describe a decreased mobility. However, the first part "Slowly improving", is literally written in Dutch (transliteration): "is slowly moving forward." This might be incorrectly identified as decreased mobility. The ICF classifier model's input were word embeddings, trained on a large medical dataset. We are not certain how this sentence would be represented, whether the embedding model was able to differentiate this "moving forward" from literal mobility-related "moving forward". It would be very interesting to experiment with the embedding model to determine if it is able to represent the different meanings these words can have.

The second phrase does talk about mobility (walking), but only in relation with pain in the abdomen. If anything, the only real mobility-related statement is that patient is able to walk for a longer period, which would get an ICF label of 4.

Incorrect Prediction:

An example of a wrong prediction, is the following. This is the telephone consult that was taken with the patient's husband. It received a Mobility (FAC) rating of: 3.354. The prediction model predicted the following post-discharge FAC functioning: 4.849. This rating would reflect a perfect functioning or the absence of any ratings post-discharge.

""Yesterday walked 3000 steps. [...] In contact with Erasmus."

Dutch: "Gisteren 3000 stappen gelopen. [...] Contacten met Erasmus lopen."

The rating of 3.354 corresponds with the same qualifier level description as the previous example. While this is not specifically described in the note, the context does show some difficulties with walking. The fact that the 3000 steps are mentioned at all, means that mobility related issues were probably expected. Additionally, since they mention the mobility department of another hospital, we know that this functioning level is not yet perfect. Therefore, while it is difficult to assess whether the ADM rating of 3 is correct based on this limited information, we do know that the predicted level of 5 is incorrect. We asked to one of the original annotators to assign a label to this sentence as they would have during the annotation task, and they identified this is a level 4. Therefore, is seems as if the prediction model is as far off as the extraction model, but we have to keep in mind that the prediction model was trained on the extraction model's assessment of the target period.

Mood (STM)

Correct Prediction:

For the correct prediction, we identified a patient with 4 mood-related ICF ratings assigned to post-discharge notes: 2.970 (progress report), 3.596 (Plan of Care), 3.352 (Plan of Care) and 2.586 (Plan of Care). These round to 3 and 4, which have the following descriptions in the annotation guidelines respectively: "Slight problem with emitional functioning: irritable, gloomy, etc." and "No problem with emotional functioning, emotions are appropriate, well regulated, etc."

We extracted the relevant phrases from each note.

STM = 2.970: "Eating is excellent, very satisfied with the care at [care location]."

Dutch: "Eten gaat voortreffelijk, zeer tevreden over verzorging in de [zorglocatie]."

STM = 3.596: "Pain is bearable. Mr. seems somewhat gloomy."

Dutch: "Pijn is draaglijk Dhr komt wat somber over"

STM = 3.352: "Mr. seems gloomy, which he knows."

Dutch: "Dhr komt somber over, is hiermee bekend."

STM = 2.586: "Mr. looks very gloomy, found everything very difficult this morning, and is tired."

Dutch: "Dhr oogt erg somber, en vond het vanmorgen allemaal erg moeilijk en is moe."

Based on these phrases, it does not seem as if the assigned mood levels are correct. First of all, out of the four examples, the first describes the most positive mood, but was assigned a similar rating as the others. The other descriptions are all very similar, they are describing sadness. The last example, however, is much more specific and therefore received a slightly worse score.

The prediction model from experiment 2 also predicted this patient would have a decreased mood.

Incorrect Prediction:

Another patient was assigned a mood level of 2 in their post-discharge period, while the prediction model predicted a slightly higher functioning level: 3. The EHR entry to which this level was assigned, did not have a note type assigned in our database.

"This is experienced by the patient as very distressing, and she panics and initially pulls out the scope. [...] Given the persistent agitation/panic in the patient and the drop in oxygen saturation, it was decided to stop the procedure and reschedule it under propofol."

Dutch: "Dit wordt door patiënte als zeer vervelend ervaren en zij raakt in paniek en trekt aanvankelijk de scoop eruit. [...] Gezien de persisterende onrust/ paniek bij patiënte en saturatiedaling wordt besloten de procedure te staken en opnieuw in te plannen onder propofol."

The main difference between a mood level of 2 and 3, according to the annotation guidelines, is that for 2, a severe problem with emotional functioning had to be identified and for 3, it is only regarding a moderate problem. While the example above seems to describe a more severe emotional functioning than the examples we gave for the correct predictions, it is difficult to asses whether the ICF model is more correct than what our prediction model predicted.

Weight Maintenance (MBW)

Correct Prediction:

For the correct prediction, we identified a patient with MBW rating 2.728 for a progress report and a predicted MBW rating of 2.613.

"Medical history: Before surgery 91 kg. After surgery 94 kg with fluid, still during hospitalization 89 kg. Currently 87.5 kg, approximately 4 kg lost. Has lost muscle mass."

Dutch: “Anamnese Voor de ok 91 kg. Na ok 94 kg met vocht, nog tijdens opname 89 kg. G nu 87,5 kg, Ca. 4 kg kwijt Is spiermassa kwijt.”

The patient lost 4 kg weight, which is more than 3 kg (MBW level 2), but not over 6 kg (MBW level 1). Therefore, both the extracted ICF level as well as the predicted level are correct.

Incorrect Prediction:

Another patient received a single MBW ICF rating for a Plan of Care. It was assigned a level of 2, while the prediction model predicted it to be 3.

“Ms. has gained 2 kg compared to yesterday, with no additional complaints.”

Dutch: “Mw os 2kg aangekomen t.o.v. gister, geen klachten erbij.”

An MBW rating of 2 corresponds with weight gain or loss of over 3kg. MBW 3 corresponds with some unintentional weight gain or loss, but without a specific amount. Therefore, we can see that the predicted value is more correct than the ICF rating as extracted by the ICF classifier models.

Energy (ENR)

Correct Prediction:

In both experiment 1 and 2 we identified that the Energy (ENR) class is the worst performing of all ICF classes. Even though the MAE scores for this class are the lowest, which normally indicates better performance, we know that ENR ratings range mostly from 1 to 2. We identified a patient whose post-discharge Plan of Care note was assigned an ENR level of 1.545. The prediction model predicted the level to be 1.614. This corresponds to the description: “Moderate fatigue; the patient gets easily tired from light activities or needs a long time to recover after an activity.”

“Medical history: Mr. indicated that he doesn’t feel well and is tired. He did shower independently. Walked around the ward.”

Dutch: “Anamnese: Dhr gaf aan zich niet lekker te voelen en moe te zijn. Heeft wel zelfstandig gedoucht. Liep rondjes over de afdeling.”

This description does fit the ENR level of 2. Both the ICF classifier models and the prediction model correctly identified it as such.

Incorrect Prediction:

Another patient received a rating of 1.712 for a progress report. However, the prediction model predicted it would be 3.976.

“Recovery is disappointing. Still tired. Physiotherapy twice a week.”

Dutch: “Herstel valt tegen. Nog moe. 2 maal per week fysio.”

The only part of this example directly refering energy and fatigue, is “still tired”. It does not talk about any limitations due to this fatigue. While it is clear that the predicted level of 4 is incorrect, we also do not believe the label of 2 is correct either. This includes getting tired easily from light activities, and no such thing was described in the note. Similarly to all previous “Incorrect Predictions” for the other classes, except for FAC, the incorrectness of the prediction stems partly from the incorrectness of the target label.

In this error analysis we showed the kind of mistakes the prediction models make. Interestingly, for the most of the incorrect predictions, we found that it was the ICF classifier model that misidentified the qualifier level and therefore the prediction models from experiment 2 were evaluated against this misrepresentation. Only for Mobility (FAC) it seems that the identified incorrect prediction was evaluated against a correctly assigned ICF rating. It is important to note that none of these examples were specifically selected, we evaluated a set of random patients from the test set. As a result of the inconsistencies among the target labels, the MAE scores we reported as performance metrics might have gotten inflated. We know we are evaluating certain predictions as being wrong that upon inspection proved to be right. We expect that, if the ICF models are retrained and make less mistakes, that the prediction models will benefit from this increase in target correctness.

Chapter 5

Discussion

The findings of this research provide significant insights into the relationship between in-hospital International Classification of Functioning, Disability, and Health (ICF) ratings and post-discharge functioning, validated against patient self-reports. This section discusses the usability of the ICF classifier model, its predictive capabilities, and the correlation with patient-reported outcomes. We compare our findings with existing literature and we provide recommendations for future research.

Usability of ICF Classifier Model

This thesis describes the usability of the previously developed ICF classifier models in two ways. Once by comparing the model's output with self-reported functioning data (Experiment 1) and once by exploring the predictive relationship by the model's output on in-hospital clinical notes and on the model's output on post-discharge clinical notes (Experiment 2). In these experiments, we identified a few limitations of the ICF classifier models. This section discusses these aspects of the model.

The ICF classifier model demonstrated reasonable accuracy in extracting functioning levels from clinical notes. However, certain limitations were identified. Some annotated ICF classes had a low Inter-Annotator Agreement score, in particular Exercise Tolerance (INS). The classifier pipeline consists of two models in sequence, but both models were only evaluated on gold data, with as a result that there is no quantification of the interdependency of errors. It was insufficiently explored how sensitive the models were to changes in the text that did not affect its meaning. Additionally, it is challenging to reason about the transferability and usability of the model due to the lack of distributional data on the datasets.

It is also essential to quantify what portion of the model's increased performance, after aggregating data on a note level and evaluating against the aggregated annotations in the same note, is due to the redistribution of data and downsizing of the total sample size, and what portion is due to an actual increase in model performance. We therefore recommend controlling for dataset imbalance in future studies.

In section 1.3 we explained that the professional's opinion and the self-reported functioning is not expected to be the same, even if both measurements are without mistakes. Ideally, we would have prefaced experiment 1 with a similar comparison, but using a professional's assessment of the clinical notes from the BV project participants prior to using the ICF model output. It would have been useful to assess how professional opinions (i.e., as annotated by the annotators) relate to self-reported functioning levels prior to comparing the model outputs with self-reported functioning data. That would show us to what extent correlations can be explained by the relationship between

ICF classes and self-reported functioning data and what proportion can be explained by wrongfully predicted labels by the model.

Nevertheless, in this thesis we showed several examples of the ICF classifier output and the corresponding note text. Regardless of all limitations we identified earlier, the trend of the extracted ICF levels over time seems to correspond with what the notes are reporting.

Correlation with Patient-Reported Outcomes

The first experiment focused on validating the ICF classifier model by correlating its outputs with patient-reported outcomes from the Beter Voorbereid (BV) project. Significant correlations were observed for all classes except Weight Maintenance (MBW) for at least one of the input representations. Moreover, we found that for questions referring to a single day, a correlation is only found when taking the rounded value for each ICF rating, while for questions about a longer period of time, the unrounded exact ICF ratings proved more useful. We suspect this is because - while aggregating multiple ratings over a period of time - the unrounded ICF ratings carry more nuanced information. Conversely, when a question refers to only a single day and we select the closest ICF rating, rounding the rating to reflect the original ordinal scale of the ratings, gave us a stronger correlation.

We showed that there is some commonality between what both measurements measure about a patient's functioning. This overlap between both datasets is the basis for experiment 2, since we know that the average rating in a window around a target date reflects a patient's own experience of their functioning.

While we found a significant correlation for 4 out of 5 relevant ICF classes, the correlation coefficients were relatively small, indicating a weak to medium correlation. We also found that the correlation between the delta change of ICF values and self-reported BV responses was smaller or non-existent than for the pair-wise comparison. We hypothesise that this is mainly because smaller sample sizes, since any missing value affects the presence of a delta change value to and from the time point of the missing value.

We identify a few limitations to our approach. First and foremost, only 74 patients from the Beter Voorbereid study were considered. This is a relatively small sample size and we expect that using a bigger dataset would provide better insights in the relationship between the functioning data as extracted by the ICF clasifier models and the self-reported functioning data. Additionally, only 5 out of the 9 ICF classes were considered for this project due to data scarcity, insufficient performance reported by previous researchers and non-overlapping constructs between the questionnaires and the classifier models.

Predictive Value of In-Hospital ICF Data

The second experiment attempted to quantify the predictive value of in-hospital ICF data on post-discharge functioning. We found that the in-hospital ICF classes are able to confidently predict post-discharge functioning with a similar error rate as was reported for the original models (Kim et al., 2022). We expect that if the original ICF classifier models increase in performance, the predictive value of these labels increases as well.

In Experiment 2, we also identified limitations to our approach. We employed the Mean Absolute Error evaluation metric, which considers all data to be on a rational scale. We used this evaluation metric because it allows us to directly compare the predictive value of the models with the average error rate on the ICF ratings as assigned by the ICF classifier models.

For Respiratory funcioning (ADM), the best performing prediction model had an MAE score of 0.48. This means that - on average - the model is 0.48 off. Table 3.6 reported on the MAE scores for the ICF model output itself, and ADM scored 0.37 here. This means that the prediction model is 0.48 off of predicting a score that itself has an error margin of 0.37. So while using another evaluation metric that better reflects the ordinal nature of the ICF labels would be more useful in terms of interpreting the result in a hospital context, it complicates a direct comparison between the ICF models performance and the prediction models performance in Experiment 2.

While it is a good sign that the prediction models make mistakes within the same error margin as the ICF classifier models, this does not necessarily mean that these models can be directly used within a hospital context. For example, for the ADM class, where we have an average absolute error of 0.48, on top of a potential error in the target label of - on average - 0.38, means that the predicted ADM outcome of patients is also, on average, off by 0.4 to 0.9. Realistically, this means that we can predict a level 4, which means "no problems", while the target is ADM = 3.5 and while the real respiratory functioning was 3: "shortness of breath during exercise". Similarly for the other cases: while an error of 1 does not look like a lot, it can be the difference between predicting a patient will suffer moderate weight loss (MBW 3) while in reality the patient did experience some weight loss, but also regained it afterwards (MBW 4).

We showed that the addition of structured data to the training samples does not necessarily improve the model performance. However, we also showed that the baseline model, which contains only the structured data, performs slightly worse than when only the ICF data is used, meaning that the structured data alone does have some predictive value on post-discharge functioning. We believe that experimentation with different ways of combining the ICF data and the structured EHR data would result in better performance of the models overall.

Additionally, due to time restrictions and the size of the current analyses, we were not able to optimize the LSTM prediction models themselves by finding optimal (hyper)parameter settings.

Comparison with Existing Literature

While it is challenging to compare the outcomes of this thesis with existing literature, due to the unique approach based on a previously developed ICF classifier model, this is an important aspect of placing this research into the scientific field. Even without a direct comparison, we can still link this research to others by evaluating each aspect separately.

The first comparison we would like to make, is for Experiment 1, where we compare the automatically extracted ICF functioning data with the same patient's self-reported experience. Velikova et al. (2001) conducted a research into the relationship between Quality of Life information in medical records as annotated by professional's and the self-reported Quality of Life as reported in Quality of Life Questionnaire (QLQ)-C30 from the European Organization for the Research and Treatment of Cancer (EORTC). They identified for each of the Quality of Life categories that an overall higher proportion of the patients reported symptoms and functional problems in the questionnaire than was found in the medical records. The categories for which the lowest rate of problem recording in medical records was found as opposed to the self-rated rate of recording were: pain, fatigue, emotional distress and nausea. While our project did not focus on pain and nausea, we did also find a relative low reporting rate for fatigue (Energy) and emotional distress (Mood).

Secondly, we want to compare the outcomes of Experiment 2 with existing literature. Tankumpuan et al. (2015) wrote an article about the predictors of physical functioning in postoperative brain tumor patients. They found that mood and fatigue were the best predictors of post-discharge physical functioning. However, we also found that fatigue and mood are least frequently reported

in clinical notes, and according to Velikova et al. (2001) this is not because those Quality of Life categories are not an issue for patients, but rather because they are underrepresented in medical records. These two findings together explain why, even though the prediction models in Experiment 2 perform similarly in terms of the MAE scores as the ICF classifier models do, the outcome is not satisfactory to be usable in a healthcare setting.

Finally, we discussed the challenges of using an NLP model for downstream modeling tasks, like we did in Experiment 2. We explained how a research in isolation can provide good results, but when used in practice, the proposed approach performs worse. This is because of methodological choices made that complicate downstream tasks, such as filtering out samples are expected to decrease the model's performance (i.e. filtering out all historic and future mentions), but also because any errors made by the first model will automatically be transferred to the downstream model. Therefore, we are essentially building a model that uses incomplete information. Velupillai et al. (2018) wrote an article on the use of NLP models for health outcomes research and they describe the same challenges. They describe that an NLP model can never be 100% accurate, cannot be easily transferred from one domain to another, are rarely developed with clinical experts in mind and that they should be externally validated with the clinical production environment in mind. This is what we attempted to do with Experiment 1.

Recommendations for Future Research

Based on the findings, several recommendations can be made to enhance the model's performance and usability. We focus on both recommendations to improve the quality and therefore utility of the ICF classifier models and on recommendations to improve the analyses introduced in this thesis.

Recommendations for ICF Classifier models

Reannotating data and calculating an Inter-Annotator Agreement Score on a bigger subset would likely be very beneficial for both model performance and interpretability. However, we recognize that annotating data is often very time-consuming and expensive. Therefore, we also identify steps to evaluate the consistency in annotated data without needing a new annotation round. The first recommendation is to stratify the test set on annotator. This would yield some insight in which annotator diverged most from the generalized patterns the model extracted. Additionally, a sensitivity analysis could be very useful. This would include, based on examples of where the annotators disagreed on a label of qualifier level, qualitatively assessing why that might be the case and construct a custom dataset, possibly with synthetic data, to do a sensitivity analysis on those variations.

Finally, we recommend the following adaptations to the methodology to further increase the usability of those models. Future studies should focus on rebalancing the dataset to ensure a more even distribution of ICF ratings. This could involve oversampling underrepresented classes or using advanced data augmentation techniques. Additionally, in order to better describe the quality of annotations, calculating an Intra-Annotator Agreement could also be useful to see how consistent annotators are with themselves.

Recommendations for approach in this thesis

While these recommendations focus on the previous research, which we expect would affect the

performance of downstream tasks as well, there were also a few limitations to the methodology used in this thesis. Primarily due to time limitations, we were not able to employ cross-validation to better represent the model's performance. We expect that this would benefit the model, given that we used a relatively small dataset.

We also did not conduct a grid search to find the optimal hyperparameter settings, as this would require training multiple models and our experiment 2 already contained 40 trained models. Our focus was on comparing different methods of representing the input data rather than finding the optimal modeling setup.

In our experiments, we measured the relationship between self-reported functioning values and automatically extracted ICF functioning levels. We explained that these constructs are not measuring the same aspects of a patient's functioning, which is part of the reason why the correlation coefficients were relatively small. We expect that by taking different datasets with a clearer connection would increase the value of the output of this experiment. We recommend redoing a similar experiment with questionnaires containing questions that directly correlate with the ICF functioning levels.

For experiment 2, we created prediction models that predicts the average post-discharge functioning level at the 6-week mark. While we know this average rating relates to a patient's experience, it is a vaguely defined target label. In order to solidify the predictive value of in-hospital functioning level on post-discharge functioning, we recommend extending this experiment with prediction tasks of other post-discharge functioning data. For example, predicting readmission chances of patients post-surgery would be a good measure of a patient's functioning, while being a clearly defined target that can easily be extracted from the Electronic Health Records.

Additionally, we used the Mean Absolute Error (MAE) score to evaluate the model to allow direct comparison with the ICF models' performance. However, this metric does not correctly reflect the ordinal categorical nature of the ICF qualifier levels. Using evaluation metrics specifically designed for classification tasks would likely yield better interpretable descriptions of the model's performance.

Code and Data Availability

Since this project included privacy-sensitive patient data, we are not allowed to share the dataset. The same holds for all code that preprocesses and selects the data, since this includes hard-coded privacy sensitive data as well. The rest of the code will be published at a later date.

Chapter 6

Conclusion

In this thesis, we made an attempt at showing how unstructured clinical data during a patient's hospitalization can be used to gain insights in their functioning status after they are discharged from the hospital. We have shown how this extracted data relates to a patient's own experiences and how this can subsequently be used to predict post-discharge functioning. While the results are promising, we realize additional evaluation steps on both this research and the previous research are needed to quantify the usability of these prediction models in a healthcare setting.

In the introduction of this thesis, we introduced two research questions:

1. How does a professional's judgment of a patient's functioning, as extracted from unstructured clinical data, correlate with the patient's own perception of their functioning?

2. How effectively can NLP-extracted ICF functioning levels from in-hospital patient notes predict post-discharge functioning as documented in follow-up notes?

In Experiment 1, we answered the first question and showed that there is a significant correlation between a professional's judgment of a patient's functioning and the patient's own perception for 4 out of 5 ICF classes. For the ICF class Weight Maintenance we found no such significant correlation, which we suspect might be due to a smaller dataset. For any mismatches between the datasets, we could not assess whether these were due to a poor relationship between the professional's and the patient's judgment or due to the ICF classifier model making incorrect predictions.

In Experiment 2, we answered the second research question. We showed that the ICF functioning levels can predict post-discharge functioning with similar accuracy to the ICF models' extraction of those functioning levels. For most classes, except for MBW and ADM, this is a better performance than the baseline, consisting of a prediction model trained on only the structured clinical data. We expect that if the quality of the ICF models increases, a similar increase in performance will be observed when retraining the prediction models using the same setup.

In this thesis, we showed how Natural Language Processing can be used in healthcare to extract useful information from unstructured clinical notes that was previously not available in the Electronic Health Records. We made initial steps towards showing how this information can then be used for downstream modeling tasks, such as predictions tasks, which could help target the correct intervention to patients. While there is a lot of room for improvement, we believe that NLP can benefit healthcare immensely.

Bibliography

Brajcich, M., Gao, Y., Livingston, H., and McDonald, E. G. (2021). Barriers to post-discharge monitoring and patient-clinician communication: a qualitative study. *BMC Health Services Research*, 21:111.

Brandberg, C., Ali, L., Wallengren, C., Jakobsson, U., Schüssler, P., and Frykberg, G. (2021). Self-management challenges following hospital discharge for patients with multimorbidity: a longitudinal qualitative study of a motivational interviewing intervention. *BMC Health Services Research*, 21:176.

Bull, M. J., Hansen, H. E., and Gross, C. R. (2000). Predictors of elder and family caregiver satisfaction with discharge planning. *Journal of Cardiovascular Nursing*, 14(3):76–87.

Burke, R. E., Guo, R., Prochazka, A. V., Misky, G. J., Coleman, E. A., and Ginde, A. A. (2016). Four-year results of a national study of home and community based services outcomes. *Journal of the American Geriatrics Society*, 64(10):2024–2028.

Cella, D., Yount, S., Rothrock, N., et al. (2007). The patient-reported outcomes measurement information system (promis): progress of an nih roadmap cooperative group during its first two years. *Medical Care*, 45:S3–S11.

Chang, R., Penoyer, D. A., Borckardt, J. J., Gathright, M. M., and Kussin, P. S. (2012). Re-engineering the post-discharge appointment process for general medicine patients. *Journal of General Internal Medicine*, 27(6):727–732.

Cole, M. G. (2001). The impact of geriatric post-discharge services on mental state. *International Journal of Geriatric Psychiatry*, 16(12):1171–1177.

Coppa, K., Chastain, K., Kipp, R., and Wiggins, R. (2021). Examination of post-discharge follow-up appointment status and 30-day readmission. *Journal of Nursing Care Quality*, 36(2):161–167.

De Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., and Nissim, M. (2019). Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Feuering, R., Vered, E., Kushnir, T., Jette, A., and Melzer, I. (2014). Differences between self-reported and observed physical functioning in independent older adults. *Disability and Rehabilitation*, 36:1395–1401.

Forster, A. J. and van Walraven, C. (2007). Interactive voice response systems for improving delivery of ambulatory care. *The Cochrane Database of Systematic Reviews*, 4.

Haynes, A. B., Weiser, T. G., Berry, W. R., Lipsitz, S. R., Breizat, A., Dellinger, E. P., Herbosa, T., Joseph, S., Kibatala, P. L., Lapitan, M. C., Merry, A. F., Reznick, R. K., Taylor, B., and Gawande, A. A. (2009). A surgical safety checklist to reduce morbidity and mortality in a global population. *New England Journal of Medicine*, 360(5):491–499.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Huddar, S. R., Tiwari, S., and Jain, M. (2016). Using nlp-extracted features from clinical notes to predict complications for patients in critical care. *Journal of Biomedical Informatics*, 60:295–304.

Hyde, C. J., Robert, F., and Sinclair, A. J. (2000). The effects of supporting discharge from hospital to home in older people. *Age and Ageing*, 29(3):255–259.

Inouye, M. (2001). Predicting outcomes of patients in japan after first acute stroke using a simple model. *American Journal of Physical Medicine & Rehabilitation*, 80(9):645–649.

Kabaev, S., Muller, H., and Ahmed, M. (2023). Advancements in nlp for healthcare: Sentiment analysis and named entity recognition. *IEEE Journal of Biomedical and Health Informatics*, 27(3):529–536.

Khamis, H. J. (2008). Measures of association: How to choose? *Journal of Diagnostic Medical Sonography*, 24(3):155–162.

Kim, J., Verkijk, S., Geleijn, E., Leeden, M. v. d., Meskers, C., Meskers, C., Veen, S. v. d., Vossen, P., and Widdershoven, G. (2022). Modeling dutch medical texts for detecting functional categories and levels of covid-19 patients. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4577–4585.

Lee, J., Kim, S.-Y., Lee, S. H., Cho, H. K., Lee, C. W., Lee, K.-Y., Jeong, W. J., Rhee, C.-S., Min, J.-H., and Kim, H.-Y. (2023). Feasibility and utility of a smartphone application-based longitudinal cough monitoring in chronic cough patients in a real-world setting. *Respiratory Research*, 24:84.

Leppin, A. L., Gionfriddo, M. R., Kessler, M., Brito, J. P., Mair, F. S., Gallacher, K., Wang, Z., Erwin, P. J., Sylvester, T., Boehmer, K., Ting, H. H., and Montori, V. M. (2014). Preventing 30-day hospital readmissions: a systematic review and meta-analysis of randomized trials. *JAMA Internal Medicine*, 174(7):1095–1107.

Lineback, C., Thompson, C., and Brown, J. (2021). Nlp techniques for predicting unplanned readmission within 30 days. *Journal of Medical Systems*, 45(6):78.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

McIlvennan, C. K., Eapen, Z. J., and Allen, L. A. (2015). Readmissions after heart failure: the impact of hospital volume on rates and outcomes. *Annals of Internal Medicine*, 163(3):233–234.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Mistiaen, P., Francke, A. L., and Poot, E. (1999). Predictive validity of the brass index in screening patients with post-discharge problems. *Journal of Advanced Nursing*, 30(5):1050–1056.

Mistiaen, P. and Poot, E. (2006). Telephone follow-up, initiated by a hospital-based health professional, for postdischarge problems in patients discharged from hospital to home. *The Cochrane Database of Systematic Reviews*, 4.

Mousa, A. Y., Broce, M., Bozzay, J., Davis, E., Lucas, L., McDermott, S., and Slim, A. (2019). Results of telehealth electronic monitoring for post discharge complications and surgical site infections following arterial revascularization with groin incision. *Journal of Vascular Surgery*, 69(1):122–131.

Murff, H., FitzHenry, F., Matheny, M. E., Gentry, N., Kotter, K. L., Crimin, K., and Dittus, R. S. (2011). Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA*, 306(8):848–855.

Parker, S. G., Peet, S. M., McPherson, A., Cannaby, A.-M., Abrams, K., Baker, R., Wilson, A., Lindesay, J., Parker, G., and Jones, D. (2002). A systematic review of discharge arrangements for older people. *Health Technology Assessment*, 6(4):1–183.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q. V., Litsch, K., Mossin, A., Tansuwan, J., Wang, D., Wexler, J., Wilson, J., Ludwig, D., Volchenboum, S. L., Chou, K., Pearson, M., Madabushi, S., Shah, N. H., Butte, A. J., Howell, M. D., Cui, C., Corrado, G. S., and Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):18.

Sabbag, S. A., Twamley, E., Vella, L., Heaton, R., and D., H. P. (2012). Predictors of the accuracy of self assessment of everyday functioning in people with schizophrenia. *Schizophrenia Research*, 137:190–195.

Shepperd, S., Parkes, J., McClaran, J., and Phillips, C. (2004). Comprehensive discharge planning with postdischarge support for older patients with congestive heart failure: a meta-analysis. *JAMA*, 291(11):1358–1367.

Shirkavand, R., Zhang, F., and Huang, H.-C. (2023). Prediction of post-operative renal and pulmonary complication using transformers. *ArXiv*, abs/2306.00698.

Sousa, T., Oliveira, H. G., and Alves, A. (2020). Exploring different methods for solving analogies with portuguese word embeddings. In *Proceedings of the 9th Symposium on Languages, Applications and Technologies (SLATE 2020)*, pages 9:1–9:14.

Tankumpuan, T., Utriyaprasit, K., Chayaput, P., and Itthimathin, P. (2015). Predictors of physical functioning in postoperative brain tumor patients. *Journal of Neuroscience Nursing*, 47(1):E11–E21.

Tevis, S. E. and Kennedy, G. D. (2014). Postoperative complications and implications on patient-centered outcomes. *Journal of Surgical Research*, 181(1):106–113.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, , and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Velde, M. v. d., Valkenet, K., Geleijn, E., Kruisselbrink, M., Marsman, M., Janssen, L. M., Ruurda, J. P., van der Peet, D. L., Aarden, J. J., Veenhof, C., and van der Leeden, M. (2021). Usability and preliminary effectiveness of a preoperative mhealth app for people undergoing major surgery: Pilot randomized controlled trial. *JMIR Mhealth Uhealth*, 9(1):e23402.

Velikova, G., Wright, P., Smith, A., Stark, D., Perren, T., Brown, J., and Selby, P. (2001). Self-reported quality of life of individual cancer patients: Concordance of results with disease course and medical records. *Journal of Clinical Oncology*, 19:2064–2073.

Velupillai, S., Suominen, H., Liakata, M., Roberts, A., Shah, A. D., Morley, K., Osborn, D., Hayes, J., Stewart, R., Downs, J., et al. (2018). Using clinical natural language processing for health outcomes research: overview and actionable suggestions for future advances. *Journal of biomedical informatics*, 88:11–19.

Verkijk, S. and Vossen, P. (2021). Medroberta.nl: A language model for dutch electronic health records. In *Computational Linguistics in the Netherlands Journal*, page 141–159.

Walraven, C. v., Seth, R., Austin, P. C., and Laupacis, A. (2002). Continuity of care and patient outcomes after hospital discharge. *Journal of General Internal Medicine*, 17(11):792–798.

Wang, T., Li, J., and Zhang, X. (2022). Evaluating machine learning techniques for readmission risk prediction. *PLOS ONE*, 17(4):e0267554.

Whiteneck, G., Gassaway, J., Dijkers, M., Heinemann, A., and Kreider, S. (2012). Relationship of patient characteristics and rehabilitation services to outcomes following spinal cord injury: The scirehab project. *The Journal of Spinal Cord Medicine*, 35(6):484–502.

World Alliance for Patient Safety (2009). Patient complications post-surgery.

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

World Health Organization (2013). *How to use the ICF: A practical manual for using the International Classification of Functioning, Disability and Health (ICF)*. World Health Organization, Geneva.

Wu, Y., Wang, H., Li, D., Wang, Y., and Liu, H. (2015). Developing a clinical decision support system using named entity recognition and sentiment analysis. *Journal of Biomedical Informatics*, 56:319–326.

.1 Appendix

Level	Generic Qualifier	MBW	STM	FAC	ENR
4	No problem	Healthy weight, no unintentional weight loss or gain, SNAQ 0 or 1.	No problem with emotional functioning: emotions are appropriate, well regulated, etc.	Patient can walk independently anywhere: level surface, uneven surface, slopes, stairs.	No problem with the energy level.
3	Mild problem	Some unintentional weight loss or gain, or lost a lot of weight but gained some of it back afterwards.	Slight problem with emotional functioning: irritable, gloomy, etc.	Patient can walk independently on level surface but requires help on stairs, inclines, uneven surface; or, patient can walk independently, but the walking is not fully normal.	Slight fatigue that causes mild limitations.
2	Moderate problem	Moderate unintentional weight loss or gain (more than 3 kg in the last month), SNAQ 2.	Moderate problem with emotional functioning: negative emotions, such as fear, anger, sadness, etc.	Patient requires verbal supervision for walking, without physical contact.	Moderate fatigue; the patient gets easily tired from light activities or needs a long time to recover after an activity.
1	Severe problem	Severe unintentional weight loss or gain (more than 6 kg in the last 6 months), SNAQ ≥ 3.	Severe problem with emotional functioning: intense negative emotions, such as fear, anger, sadness, etc.	Patient needs continuous or intermittent support of one person to help with balance and coordination.	Severe fatigue; the patient is capable of very little.
0	Complete problem	Severe unintentional weight loss or gain (more than 6 kg in the last 6 months) and admitted to ICU.	Flat affect, apathy, unstable, inappropriate emotions.	Patient cannot walk or needs help from two or more people; or, patient walks on a treadmill.	Very severe fatigue; unable to do anything and mostly lays in bed.

Table 1: Annotation Guidelines for ICF classifier training data: Weight Maintenance (MBW), Mood (STM), Mobility (FAC), Energy (ENR)

Level	ADM
4	No problem with respiration, and/or respiratory rate is normal (EWS: 9-20).
3	Shortness of breath in exercise (saturation $\geq$90), and/or respiratory rate is slightly increased (EWS: 21-30).
2	Shortness of breath in rest (saturation $\geq$90), and/or respiratory rate is fairly increased (EWS: 31-35).
1	Needs oxygen at rest or during exercise (saturation $<$90), and/or respiratory rate $>$35.
0	Mechanical ventilation is needed.

Table 2: Annotation Guidelines for ICF classifier training data: Respiratory Functioning (ADM)