dHealth 2023
B. Pfeifer et al. (Eds.)

doi:10.3233/SHTI230048

Impact of Professional Background on Inter-Annotator Variability and Accuracy During Annotation of Clinical Notes

Kristina WEISHÄUPL[a,1], Mario SCHUSTER[a], Christina MAYRL[a], Diana HUMBERGER[a], Antonia HOFFSTÄDTER[a], Theresa FISCHER[a], Bianca PUNZ-REIDLINGER[a], Fabian WIESMÜLLER[b,c], Aaron LAUSCHENSKY[c], Dieter HAYN[b,c], Karl KREINER[c], Bettina FETZ[d], Luca BRUNELLI[e], Gerhard POELZL[e], Bernhard PFEIFER[f,g], Gerald SLAMANIG[f,g], Sabrina Barbara NEURURER[f,g] and Günter SCHREIER[c,a]

[a] *University of Applied Sciences Wiener Neustadt, Wiener Neustadt, Austria*
[b] *Ludwig Boltzmann Institute for Digital Health and Prevention, Salzburg, Austria*
[c] *AIT Austrian Institute of Technology, Graz, Austria*
[d] *Landesinstitut für Integrierte Versorgung – LIV Tirol, Innsbruck, Austria*
[e] *Medical University Innsbruck, Innsbruck Austria*
[f] *Tirol Kliniken GmbH, Innsbruck, Austria*
[g] *UMIT Private University for Health Sciences & Health Technology, Hall, Austria*

Abstract. Background: The aging population's need for treatment of chronic diseases is exhibiting a marked increase in urgency, with heart failure being one of the most severe diseases in this regard. To improve outpatient care of these patients and reduce hospitalization rates, the telemedical disease management program HerzMobil was developed in the past. Objective: This work aims to analyze the inter-annotator variability among two professional groups (healthcare and engineering) involved in this program's annotation process of free-text clinical notes using categories. Methods: A dataset of 1,300 text snippets was annotated by 13 annotators with different backgrounds. Inter-annotator variability and accuracy were evaluated using the F1-score and analyzed for differences between categories, annotators, and their professional backgrounds. Results: The results show a significant difference between note categories concerning inter-annotator variability ($p<0.0001$) and accuracy ($p<0.0001$). However, there was no statistically significant difference between the two annotator groups, neither concerning inter-annotator variability ($p=0.15$) nor accuracy ($p=0.84$). Conclusion: Professional background had no significant impact on the annotation of free-text HerzMobil notes.

Keywords. Telemedicine, Heart Failure, Natural Language Processing, Electronic Health Records, Austria

[1] Corresponding Author: Kristina Weishäupl, University of Applied Sciences Wiener Neustadt, Wiener Neustadt, Austria, E-Mail: k.weishaeupl@outlook.com

1. Introduction

Heart failure (HF) has reached epidemic proportions and is now the most common cause of hospitalization [1, 2]. The growing need for adequate treatment of chronically ill patients is becoming increasingly pressing, with HF being one of the most severe issues. Telemonitoring of patients with HF is a concept aimed at early detection of impending acute decompensation, to prevent hospitalization improve patients' quality of life, and also reduce costs [2, 3]. Telemonitoring systems, such as the HerzMobil (HM) system, allow healthcare professionals (HCPs) to closely monitor patients post-discharge following a cardiac decompensation, maintain contact, interpret daily vital parameters, and adjust medication dosage over time to reach target doses [1, 4]. Most of the data collected in HM are collected in a structured way. However, HCPs also communicate and document via additional free-text clinical notes within the telemonitoring system, which are currently only available in an unstructured format.

To analyze the growing number of free-text clinical notes and to classify and extract information, natural language processing (NLP) is used. Previous studies have extracted date and time references and categorized notes from HM Tirol [5, 6]. These NLP solutions aim to improve the program's workflow by providing additional information and reducing manual work.

For supervised machine learning models to be developed and trained, a set of annotated data with a defined ground truth is necessary. Therefore, an annotation guideline was developed and a subset of HM Tirol notes was manually annotated by nine independent annotators with different professional backgrounds into eight categories in previous work [7]. Annotations differ significantly between annotators of different professions.

The goal of this scientific paper was to further analyze the inter-annotator variability among different professional groups (healthcare, engineering) involved in the annotation process. While in [7], all annotators were users of HM Tirol, i.e., they had different views on the related data based on their professional role, the present paper analyses the impact of the professional background only. All annotators in the present paper had the same information (annotation guide with definition of categories) available for annotating the notes, without further knowledge of HM Tirol and its patients. The results of this analysis will provide valuable insights into the reliability and validity of the ground truth data, which will aid in the development of NLP models for the efficient and effective management of clinical notes.

2. Methods

2.1. Dataset

The dataset used for this work consisted of clinical notes from the HM Tirol program. Before further processing, all notes have been de-identified and split into individual sentences, using a pre-existing algorithm [8]. The resulting 1,300 text snippets were further used for the annotation process and will be simply referred to as “notes” in the following.

2.2. Note Categories

Prior to this work, eight categories have been identified to be of high interest to the involved HM Tirol stakeholders. For this work annotators assigned notes to zero, one, or multiple of the following categories. These are listed in Table 1. Ground truth was established in a previous work [7].

Table 1. Definition of the eight categories.

Categories	Definition
Absence	Absence during the HM Tirol monitoring phase
Home visitation	Explicit contact for home visits from a HM Tirol partner
Contact HCP	Contact between HM Tirol partners and HM Tirol partners
Contact patient	Contact between HM Tirol partners and patients
Contact others	Contact between HM Tirol partners and other involved persons
Education	Explicit contact with patients for training and education
Technical problems	Problems with data transmission, functionality or errors of the system
Therapeutic regime	Medical content of diagnosis, therapy, medication, symptoms etc.

2.3. Annotators and Guideline

13 students from the extra-occupational master's program Health Care Informatics at the University of Applied Sciences Wiener Neustadt annotated the dataset. The annotators were divided into two groups, depending on their professional backgrounds: Healthcare (n=7) and Engineering (n=6). The notes were divided between the annotators so that every note was annotated by two annotators of each background resulting in approximately 400 notes per annotator and four annotations per note. A previously developed annotation guide was used to ensure a uniform definition of the categories [7].

2.4. Statistical Analysis

The inter-annotator variability was calculated as the mean F1 score of each annotator with every other annotator. Additionally, the accuracy was determined by the F1 score. The F1 score was used since the dataset was rather unbalanced and there were only a few positive cases in some categories [9]. For both measures, the two groups with different backgrounds were compared against each other. Each individual annotator's performance was compared to other annotators and differences between note categories were analyzed. P-value < 0.05 were considered statistically significant. The Predictive Analytics Toolbox for Healthcare in MATLAB (The MathWorks in Nattick, MA) was used for statistical analysis [10].

3. Results

3.1. Differences between groups

Figure 1 shows the inter-annotator variability over all categories grouped by annotator group. No statistically significant (p=0.15) difference between the two groups was identified using the two-tailed Wilcoxon test.

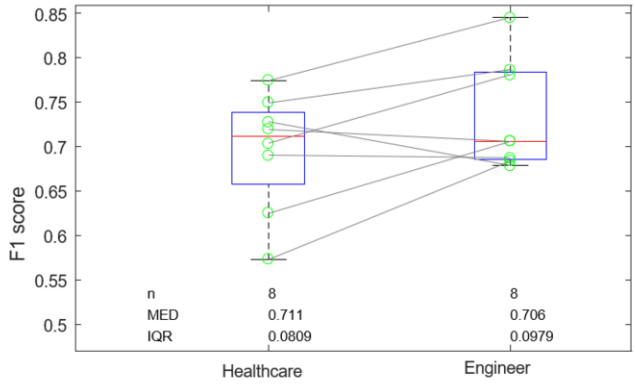

Figure 1. Boxplots of the F1 score for the inter-annotator variability over all categories. F1 scores per category are plotted as green circles for each group. Circles in the two groups representing the same category are connected by grey lines.

Figure 2 shows the accuracy grouped by the two annotator groups as compared to the ground truth. There was no statistically significant difference between the groups (p=0.84).

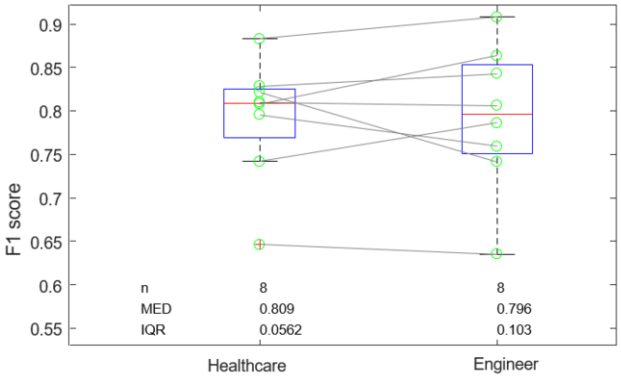

Figure 2. Boxplots of the F1 score for the accuracy (as compared to the ground truth) of each profession over all categories. F1 scores per category are plotted as green circles for each group. Circles in the two groups representing the same category are connected by grey lines.

Comparing the accuracy of each annotator as compared to the ground truth individually resulted in a statistically significant difference ($p<0.0002$) between the annotators. The individual F1 scores over all categories varied between 0.61 and 0.88.

3.2. Differences between note categories

The inter-annotator variability was calculated for all annotators grouped by the categories and for all categories grouped by the annotator group. Figure 3 displays the agreement in between the annotators per category, revealing a significant ($p<0.0001$, Friedman test) difference in between the categories. Annotators agreed most concerning the category "Home visitations", whereas the agreement was lowest for "Absence". Additionally, the F1 score of the accuracy differed significantly between the eight categories ($p<0.0001$, Friedman test).

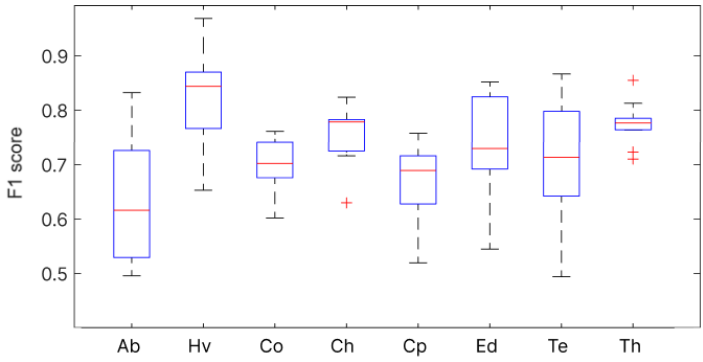

Figure 3. Inter-annotator variability between all annotators per category. *Absence (Ab), Home visitation (Hv), Contact others (Co), Contact HCP (Ch), Contact patient (Cp), Education (Ed), Technical problems (Te),* and *Therapeutic regime (Th)*

4. Discussion

For this paper, 1,300 clinical notes, originating from the telemedicine network HM Tirol, have been manually annotated and classified into eight categories. The annotations have been analysed concerning their agreement between the annotators themselves and a previously established ground truth.

Figure 3 shows the inter-annotator variability of the annotations. Similar to the previous work [7] the variability varied significantly throughout the categories. While in [7], all annotators were also active users of the HM Tirol system, leading to specific knowledge concerning notes and patients depending on their group, all annotators in the present paper were presented with the same amount of information. The category *Home visitation* had the best F1 score concerning the inter-annotator variability, followed by the category *Therapeutic regime,* whilst the category *Absence* had the lowest score. This can be explained by the different complexity and hence varying room for interpretation of the categories. Similar results were also found in [7], which could be an indication to precise the category. Moreover, it must be considered that there were only a few true positives in some categories, so deviations are higher weighted.

While [7] showed significant differences between the professions, the present study could not confirm these results (Figure 1). As shown in Figure 1, the inter-annotator variability between HCPs and engineers did not reach statistical significance. HCP had only a slightly higher agreement than engineers. We conclude that the differences identified in [7] rather stem from the different background knowledge of the HM Tirol

system and patients, not from the professional background. Moreover, the engineers in [7] invested more time into the annotation process as the final version of the annotation guide was mainly developed by them [7].

Compared to the Ground Truth the engineers had a slightly lower accuracy compared to the HCPs, however, this difference did not reach statistical significance (Figure 2).

5. Conclusion and Outlook

Despite using an annotation guide to reduce inter-annotator variability, there was a significant difference in agreement between the annotators and the ground truth in some of the categories. Nevertheless, there was no significant difference, neither comparing the two groups' inter-annotator variability nor their accuracy, which indicates that professional background may not be important for annotating when an annotation guide is used. Besides using an annotation guide the precise definition of the categories with little room for interpretation seems essential and may improve the accuracy of the annotation. However, due to the small number of annotators in this study, further investigation with a higher number of annotators is needed to confirm this hypothesis.

Acknowledgment

Parts of this work were supported by the Land Tirol, in the framework of the project “d4Health Tirol”.

References

[1] B. G. Celler and R. S. Sparks, Home telemonitoring of vital signs - Technical challenges and future directions, *IEEE Journal of Biomedical and Health Informatics* **19**(1) (2015).

[2] S. Kitsiou et al., Effects of home telemonitoring interventions on patients with chronic heart failure: An overview of systematic reviews, *Journal of Medical Internet Research*, **17**(3) (2015).

[3] I. C. Gyllensten et al., Early indication of decompensated heart failure in patients on home-telemonitoring: A comparison of prediction algorithms based on daily weight and noninvasive transthoracic bioimpedance, *JMIR Medical Informatics* **4**(1) (2016).

[4] M. Kropf et al., Telemonitoring in heart failure patients with clinical decision support to optimize medication doses based on guidelines, *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014* (2014).

[5] Wiesmueller, Fabian & Eggerth, Alphons & Kreiner, Karl & Hayn, Dieter & Hanke, Sten & Pölzl, Gerhard & Egelseer-Bründl, Tim & Schreier, Günter. (2020). Automated Extraction of Time References from Clinical Notes in a Heart Failure Telehealth Network. 10.22489/CinC.2020.186.

[6] Wiesmüller, Fabian, et al. "Natural Language Processing for Free-Text Classification in Telehealth Services: Differences Between Diabetes and Heart Failure Applications." *dHealth*. 2021.

[7] Wiesmueller et al., Classification of Clinical Notes From a Heart Failure Telehealth Network, *submitted to MIE (2023)*

[8] M. Baumgartner et al., Impact Analysis of De-Identification in Clinical Notes Classification, *Studies in Health Technology and Informatics* **293** (2022), 189–196.

[9] N. Chinchor, MUC-4 evaluation metrics, *4th Message Understanding Conference, MUC 1992 - Proceedings* (1992).

[10] D. Hayn et al., Predictive analytics for data driven decision support in health and care, *IT - Information Technology* **60**(4) (2021).