Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230270

Classification of Clinical Notes from a Heart Failure Telehealth Network

Fabian WIESMÜLLER[a,b,c,1], Aaron LAUSCHENSKI[b], Martin BAUMGARTNER[b,c], Dieter HAYN[a,b], Karl KREINER[b], Bettina FETZ[d], Luca BRUNELLI[e], Gerhard PÖLZL[e], Bernhard PFEIFER[f,g], Sabrina NEURURER[f,g] and Günter SCHREIER[b,c]

[a]*Ludwig Boltzmann Institute for Digital Health and Prevention, Salzburg, Austria*
[b]*AIT Austrian Institute of Technology, Graz, Austria*
[c]*Institute of Neural Engineering, Graz University of Technology, Graz, Austria*
[d] *Landesinstitut für Integrierte Versorgung – LIV Tirol, Innsbruck, Austria*
[e]*Department of Internal Medicine III, Cardiology and Angiology, Medical University Innsbruck, Innsbruck, Austria*
[f]*Tyrolean Federal Institute for Integrated Care, Tirol Kliniken GmbH, Innsbruck, Austria*
[g]*Division for Digital Medicine and Telehealth, UMIT TIROL - Private University for Health Sciences and Health Technology, Hall (Tyrol), Austria*
ORCiD ID: Wiesmüller Fabian https://orcid.org/0000-0001-6567-7782

Abstract. Heart failure is a common chronic disease which is associated with high re-hospitalization and mortality rates. Within the telemedicine-assisted transitional care disease management program HerzMobil, monitoring data such as daily measured vital parameters and various other heart failure related data are collected in a structured way. Additionally, involved healthcare professionals communicate with one another via the system using free-text clinical notes. Since manual annotation of such notes is too time-consuming for routine care applications, an automated analysis process is needed. In the present study, we established a ground truth classification of 636 randomly selected clinical notes from HerzMobil based on annotations of 9 experts with different professional background (2 physicians, 4 nurses, and 3 engineers). We analyzed the influence of the professional background on the inter annotator reliability and compared the results with the accuracy of an automated classification algorithm. We found significant differences depending on the profession and on the category. These results indicate that different professional backgrounds should be considered when selecting annotators in such scenarios.

Keywords. Clinical notes, Annotation, Text classification, Natural Language Processing

1. Introduction

Due to the aging population the need for adequate treatment of chronic diseases becomes an ever more pressing issue. One of the most severe chronic illnesses is heart failure,

[1] Corresponding Author: Fabian Wiesmüller, AIT Austrian Institute of Technology, Graz, Austria, E-Mail: fabian.wiesmueller@ait.ac.at.

with a 12-month all-cause hospitalization rate of 44% for hospitalized and 32% for ambulatory heart failure patients [1,2]. To improve the outpatient care of heart failure patients and to reduce the hospitalization rates the telemedical disease management program HerzMobil was developed by the AIT Austrian Institute of Technology in cooperation with Landesinstitut für Integrierte Versorgung – LIV Tirol, the UMIT TIROL - Private University for Health Sciences and Health Technology, Hall (Tyrol), the telbiomed Medizintechnik und IT Service GmbH and the Tirol Kliniken [2]. During a three-month period after hospital discharge the patients track selected vital parameters like e.g., the blood pressure and upload them on a daily basis. Whilst most of the data are transmitted in a structured form, healthcare professionals (HCP) can additionally communicate via free text clinical notes. These notes can contain valuable information, such as reasons for medication changes, the absence of a patient or an HCP or a contact with a patient [2]. However, this information is currently only available in an unstructured format and thus impractical for analysis. Due to the rapidly increasing number of clinical free texts, various groups made efforts to work towards an automated analysis of such texts by analyzing and classifying clinical free texts by natural language processing (NLP), a subfield of machine learning [3-8]. In previous studies, automatic extraction of date and time references [9] and classification into predefined categories [10] were performed on notes of the HerzMobil program. These NLP solutions aim to improve the workflow of HerzMobil by providing additional information and reducing manual work (e.g., with filter functions).

To develop and train supervised machine learning models, a set of annotated data with an established ground truth is necessary. Therefore, a subset of the available HerzMobil notes had to be manually annotated and classified into categories.

This work focuses on the annotation process of clinical notes and evaluates the inter-observer agreement in between different professional groups (physicians, nurses, engineers). Additionally, the annotations are compared with the classifications of a regular expression-based machine learning model developed for this work.

2. Methods

2.1. Dataset

All the notes used for this work have been de-identified and split into their individual sentences before any further processing, based on a pre-existing algorithm [11]. This resulted in 636 individual text snippets for the annotation process. In the following the word note will refer to such a text snippet on a sentence level.

2.2. Note categories

In a multi-step, user-centered process with various HerzMobil stakeholders, eight different categories were identified to be of high interest for the involved personnel: *Absence, Home visitation, Contact HCP, Contact patient, Contact others, Education, Technical problems,* and *Therapeutic regime*. Each note could be assigned to zero, one or multiple categories.

2.2.1. Annotators and Guideline

Nine professionals from the project team, all familiar with the HerzMobil system, annotated the entire dataset: two physicians, four telehealth nurses, and three engineers. An annotation guideline was developed containing definition of the category. Throughout the user-centered process of developing the categories, the annotation guide was updated multiple times to reduce vagueness and result in clear category descriptions. The guide has been applied by the experts in three iteration steps: During the first iteration eight of the nine annotators annotated all 636 notes by either ticking or not each category for each note, which resulted in 0 to eight ticks per note. After this iteration, based on those notes and categories that featured inhomogeneous classification results, the annotation guide was adapted to clarify discrepancies and to reduce ambiguity. For the annotation of the notes, a pre-existing annotation tool was used [12].

With this updated guide, a second round of annotations was conducted by two engineers independently, one from the previous group of eight experts and one additional scientist (annotator number nine). During the second iteration, not only the notes themselves but also the ticks from the first iteration were considered by the annotators, while applying the new version of the guideline.

In a third iteration, those notes that were annotated differently by the two engineers were manually inspected, whereas the third engineer of the annotator team decided whether a category should be ticked or not to establish the final ground truth.

2.3. Statistical Analysis

The inter annotator reliability was calculated through the Cohen's Kappa (κ) by comparing the annotations of each annotator with every other annotator. The κ was used since it is one of the standard tools to measure reliability on binary classification tasks [13]. The κ was calculated per category as well as per annotator role, to evaluate whether specific categories were annotated with less ambiguity and if a difference in the annotator roles was noticeable. Subsequently, the agreement of the individual annotators with the ground truth was analyzed. Additionally, an automated classification algorithm, based on regular expressions, was compared to the annotators. The level of significance was determined with the Friedman test. $p < 0.05$ was considered statistically significant. The statistical analyses were performed with the Predictive Analytics Toolbox for Healthcare [14] based on MATLAB (The MathWorks, Nattick, MA). The studies were approved by the ethics committee of the Medical University Innsbruck (vote nr. 1035/2022).

3. Results

3.1. Comparing Roles

Table 1 shows the reliability between the three roles over all eight categories. A highly significant difference ($p < 0.0001$) between the professions has been identified.

Table 1. Comparison of the Cohen's Kappa between the different roles, combined over all categories.

	Engineers	Nurses	Physicians
Engineers	0.606	0.447	0.386
Nurses	0.447	0.371	0.352
Physicians	0.386	0.352	0.270

3.2. Comparing categories

The analysis of the annotations per category is shown in Table 2 and resulted in a highly significant ($p < 0.0001$) difference between the categories using the Friedman test.

Table 2. Comparison of the Cohen's Kappa between the different categories over all eight annotators.

Metrics	Absence	Home visitation	Contact HCP	Contact patient	Contact others	Training	Technical problems	Therapy regime
Median	0.288	0.557	0.321	0.433	0.282	0.335	0.553	0.527

3.3. Comparison with the ground truth

The comparison of a) each annotator and b) the regular-expression-based algorithm with the established ground truth is displayed in Table 3. Cohen's Kappa differed significantly between categories ($p < 0.01$), whilst no significant difference has been proven when aggregated per annotator ($p = 0.107$).

Table 3. Comparison of the individual annotators and the regular expressions with the ground truth, using the Cohen's Kappa (κ). Eng... Engineer, Phy... Physician, Nur... Nurse, ReEx... Regular Expressions

Category	Eng 1	Eng 2	Phy 1	Phy2	Nur 1	Nur 2	Nur 3	Nur 4	Total *	RegEx
Absence	0.858	0.860	0.000	0.484	0.573	0.556	0.405	0.085	0.520	0.720
Home visitation	0.872	0.969	0.381	0.708	0.773	0.357	0.844	0.656	0.741	0.912
Contact others	0.536	0.512	0.191	0.119	0.171	0.589	0.502	0.533	0.507	0.158
Contac HCP	0.714	0.400	0.329	0.629	0.485	0.690	0.131	0.667	0.557	0.147
Contact patient	0.277	0.182	0.106	0.167	0.323	0.241	0.128	0.425	0.212	0.050
Education	0.353	0.674	0.688	0.320	0.648	0.743	0.514	0.042	0.581	0.191
Technical problems	0.496	0.400	0.529	0.753	0.533	0.538	0.736	0.533	0.533	0.824
Therapy regime	0.511	0.640	0.413	0.359	0.617	0.659	0.491	0.659	0.564	0.048
Total**	0.524	0.576	0.355	0.422	0.553	0.572	0.496	0.533		0.174

* Median κ of all annotators per category ** Median κ of all categories per annotator

4. Discussion and Outlook

For this work free text clinical notes from a heart failure telehealth program have been annotated by nine experts using an annotation guide. These annotations have been analyzed regarding the agreement between the annotators among themselves and between the annotators and the ground truth.

Table 1 shows that the engineers had the highest agreement out of the three expert roles over all categories, whilst the least coherent annotations were done by the physicians with a difference in the median of κ of 0.336. It can also be seen that comparing the physicians and nurses with the engineers, results in a higher reliability than the respective groups achieved themselves. This result must be put into perspective, since even though the nurses and physicians contributed significantly to the annotation guide, it was mainly developed by the engineers, which ultimately means that the engineers invested more time in the annotation process.

Table 2 shows that the reliability of the annotators varies throughout the categories which was also indicated by a highly significant difference yielded by the Friedman test. Table 2 also shows that the reliability of commentators varied across categories, as

confirmed by a highly significant Friedman test. This confirms our assumption that some categories are easier to define and categorize than others. For example, *Home visitation*, *Technical problems* and *Therapeutic regime* were all classified with a high reliability ($\kappa > 0.525$) whilst *Absence* and *Contact with others* were annotated with less agreement ($\kappa < 0.290$). This can be explained by the nature of the categories, since *Technical problems* had a smaller margin of interpretation and could therefore be defined rather precisely compared to categories like *Education* which includes a broader range of topics.

The comparison with the ground truth shown in Table 3 revealed that the regular expressions were able to outperform the annotators drastically in the categories *Absence, Home visitation* and *Technical problems* with a respective difference in κ of 0.200, 0.171 and 0.291. However, the algorithm did not surpass the manual annotations in the remaining five categories. This prototypical regular expression algorithm will, however, be further developed and optimized on the HerzMobil notes in future work.

The knowledge gained in this work provides a helpful basis for future annotation tasks. For example, the established annotation guide can be used to annotate larger data sets that can be used as regular expressions in future projects for training and validating machine learning models. This would benefit not only the workflow in the HerzMobil program, but also other telehealth systems with a similar accumulation of clinical notes.

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

References

[1] Ponikowski P, Voors AA, Anker SD, et al. 2016 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J. 2016 Jul;37(27):2129-2200.
[2] Von der Heidt A, Ammenwerth E, Bauer K, et al. HerzMobil Tirol network: rationale for and design of a collaborative heart failure disease management program in Austria. Wien Klin Wochenschr. 2014 Nov;126(21-22):734-741
[3] Kong HJ. Managing Unstructured Big Data in Healthcare System. Healthc Inform Res. 2019 Jan;25(1):1-2.
[4] Mora S, Attene J, Gazzarata R, et al. A NLP Pipeline for the Automatic Extraction of Microorganisms Names from Microbiological Notes. Stud Health Technol Inform. 2021 Oct 27;285:153-158
[5] Yehia E, Boshnak H, AbdelGaber S, et al. Ontology-based clinical information extraction from physician’s free-text notes. J Biomed Inform. 2019;98:103276.
[6] Lee HJ, Zhang Y, Jiang M, et al. Identifying direct temporal relations between time and events from clinical notes. BMC Med Inform Decis Mak. 2018;18(Suppl 2):49.
[7] Fernandes MB, Valizadeh N, Alabsi HS, et al. Classification of neurologic outcomes from medical notes using natural language processing. Expert Syst Appl. 2023 Mar 15;214:119171
[8] George A, Johnson D, Carenini G, et al. Applications of Aspect-based Sentiment Analysis on Psychiatric Clinical Notes to Study Suicide in Youth. AMIA Jt Summits Transl Sci Proc. 2021 May 17;2021:229-237.
[9] Wiesmueller F, Eggerth A, Kreiner K, et al. Automated Extraction of Time References from Clinical Notes in a Heart Failure Telehealth Network. Computing in Cardiology Conference (CinC). 2020.
[10] Wiesmüller F, Kreiner K, Eggerth A, et al. Natural Language Processing for Free-Text Classification in Telehealth Services: Differences Between Diabetes and Heart Failure Applications. In dHealth 2021: From eHealth to dHealth 2021 (pp. 105-109). IOS Press.
[11] Baumgartner M, Schreier G, Hayn D, et al. Impact Analysis of De-Identification in Clinical Notes Classification. Studies in Health Technology and Informatics. 2022 May 16;293:189-196.
[12] Kreiner K, Hayn D, Schreier G. Twister: A Tool for Reducing Screening Time in Systematic Literature Reviews. In: EFMI-STC; 2018.
[13] Vach W. The dependence of Cohen's kappa on the prevalence does not matter. Journal of clinical epidemiology. 2005 Jul 1;58(7):655-61.
[14] Hayn D, Veeranki S, Kropf M, et al. Predictive analytics for data driven decision support in health and care. it - Information Technology. 2018;60(4):183-94.