Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2317–2326
Marseille, 20-25 June 2022

An Annotated Corpus of Textual Explanations for Clinical Decision Support

Roland Roller[1], Aljoscha Burchardt[1], Nils Feldhus[1], Laura Seiffe[1], Klemens Budde[2], Simon Ronicke[2*], Bilgin Osmanodja[2*]

[1]German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
[2]Charité – Universitätsmedizin Berlin, Germany
firstname.secondname@{dfki/charite}.de

Abstract

In recent years, machine learning for clinical decision support has gained more and more attention. In order to introduce such applications into clinical practice, a good performance might be essential, however, the aspect of trust should not be underestimated. For the treating physician using such a system and being (legally) responsible for the decision made, it is particularly important to understand the system's recommendation. To provide insights into a model's decision, various techniques from the field of explainability (XAI) have been proposed whose output is often enough not targeted to the domain experts that want to use the model. To close this gap, in this work, we explore how explanations could possibly look like in future. To this end, this work presents a dataset of textual explanations in context of decision support. Within a reader study, human physicians estimated the likelihood of possible negative patient outcomes in the near future and justified each decision with a few sentences. Using those sentences, we created a novel corpus, annotated with different semantic layers. Moreover, we provide an analysis of how those explanations are constructed, and how they change depending on physician, on the estimated risk and also in comparison to an automatic clinical decision support system with feature importance.

Keywords: Natural Language Explanations, XAI, Clinical Decision Support

1. Introduction

Machine-learning-based recommendation or decision support systems have become a commodity in online shopping, entertainment platforms and other consumer apps and services. In many professional domains, the uptake has been comparably slower, and especially in "high-risk" domains such as medical services, a number of requirements and concerns still need to be addressed before applications from research will find broad usage, e.g., in hospitals and clinical care.

In the context of automatic clinical decision support, research has shown that machine learning can outperform physicians on very particular, narrow tasks, or can help physicians to work more efficiently (see, e.g., Gulshan et al. (2016) or Rajpurkar et al. (2017)). Still, good performance is only one building block towards the final goal of trustworthy AI. Among the many issues that need to still be figured out in these complex socio-technical systems, explainability is a prominent one (Markus et al., 2021).

Trying to 'explain' a decision made by a machine learning model (or the model itself) is currently a research topic with an increasing popularity - not only in the medical context, but overall in the machine learning community. In recent years, a large variety of novel techniques were presented in the context of making machine learning models – in particular neural networks – and their decisions more transparent, e.g., by presenting the most relevant input features in the form of saliency maps (Feldhus et al., 2021), generating counterfactuals (Wu et al., 2021), explaining them in natural language (Wiegreffe and Marasović, 2021), or finding influential instances in the training data (K and Søgaard, 2022).

Most of these explainable AI (XAI) approaches including our own previous work were technology-driven in the sense that they served for model debugging purposes in the first place. Often, the results have not been evaluated by domain experts, i.e., potential users of the system, at all. And if so, are the system generated explanations useful to understand a decision made? Conversely, a relevant question would be, how would the potential users like to get the explanation?

In this work, we approach the question of how an explanation for clinical decision support could ideally look like, from a different angle. Instead of focusing on the machine learning component, we targeted the physician and asked, how would physicians provide an explanation? Within previous work, we carried out a study comparing the performance of a machine learning model and different physicians to estimate the risk of some negative patient outcomes (Roller et al., 2022). In addition to the actual study, we asked the participating physicians to justify their estimation. This collection of textual justifications in the context of clinical decision support shows the perspective of physicians, how they would describe and explain their decision - possibly to a peer. This dataset has been annotated on different levels and analysed in detail to help and endorse the development of trustful clinical decision support systems in the future. The dataset is made publicly available here[1]. Although the sentences of our dataset

* Shared last authorship.

[1]https://github.com/DFKI-NLP/Ex4CDS

are in the German language, we try to generalize certain explanation patterns, in order to make a general contribution for the research community.

2. A Dataset of Textual Explanations

In the following, we describe how the data was collected and in which way we annotated it, to provide a more detailed analysis.

2.1. Context and Data Collection

Our dataset of textual explanations has been collected within a previous study[2] in which we tested the performance of a machine learning (ML) model against physicians predicting some future outcomes (Roller et al., 2022). In the context of kidney disease, the task was to predict a score (likelihood) from 0-100, if the given patient would suffer a 1) rejection, 2) death-censored graft loss, and 3) infection, within the next 90 days. We refer to those three risks, as *endpoint*. The experiment has been carried out as a reader study using retrospective data, with overall 120 different patients at a certain point of their patient life. Using this given point in time, the physician (and the decision support system) could consider all data of that patient until this point in time, and made an estimation.

The experiment has been conducted in two phases: A first phase without decision support and a second phase in which the physician first received the estimation of the machine learning model, along with a dashboard presenting the risk score together with a 'traffic light', and the most influential features, responsible for the model decision. Most influential features were divided into *global* and *local* features.

Overall, eight physicians participated in this study, four junior and four senior physicians. Each physician received 15 patients in each phase - both times different patients. Each physician had up to 30 minutes time to analyse the history of each patient, in order to make the estimations for each endpoint. Along with each risk estimation, the physician provided an explanation to support the decisions. This collection of human explanations describes the foundation of our dataset. In addition to our annotated dataset, we also publish the explanations (relevant features) of the machine learning component.

2.2. Corpus

The original reader study included 120 patients at a particular time in their life. For each of those patients, each physician (4 senior, 4 junior) had to make three estimations, one for each endpoint. Moreover, the experiment has been carried out in two rounds, once without automatic decision support and a second round with decision support. Each time, the participating physician analysed 15 patients, and in each round a different set of patients. Therefore the dataset results in 720 different notes (120 patients, times 3 endpoints, and two rounds).

[2]The main study is not the focus of this work.

2.3. Annotation Layers

According to a first manual analysis, explanations can be constructed from the following parts: events of the past (which might be still valid), a description of the current situation, and an outlook and conclusion. Those explanation 'blocks' can be ordered in different ways, and can even be mixed with each other. Most explanations contain a description about the current situation. Each of those parts mentions different factors which increase or decrease the overall risk that the given endpoint occurs. In most cases those factors are diseases, symptoms, but also particular negative/positive developments of lab values or the intake of particular medications. Similarly as in normal clinical text, explanations include a large number of negations and other factors which change the level of truth of entities.

The main goal of our annotations is to find a structured way to analyse the human explanations. Mainly we want to find out how justifications/explanations are structured and which content they provide. Moreover, we are particularly interested in aspects which are responsible for increasing or decreasing the risk that one of the endpoints occurs. For this reason, the annotation has been carried out on different levels, as described below:

Temporal Aspects As mentioned above, explanations include information about the past, the present, but also about the future of the patient. In order to cover and examine this aspect in more detail, we assign different labels to the corresponding phrases of the explanation, as presented in Table 1.

Label	Description
past	Event occurred in the past and is over.
past present	Event occ. in the past, but is still valid in present time.
present	Event which is present/relevant for the present time.
future	Event which might occur in future.

Table 1: Annotated Temporal Aspects

Entities and Relations In order to cover the most frequent and most relevant information in the explanations, we define a set of entities and relations to be annotated, as presented in in Table 2 and 3. The schema itself was built upon the work of Roller et al. (2020).

While the upper part of Table 2 describes the core entities, such as the positive or negative health status or laboratory values which are present or absent, the lower part of the table relates to some essential risk factors: Age of the transplanted patients, age of the transplant and also time since the last transplantation (there can be multiple ones in the life of a patient) can have a significant influence on the risk of certain endpoints. *Age*, *Donor_Age* and *Tx_Time* are all expanded by the

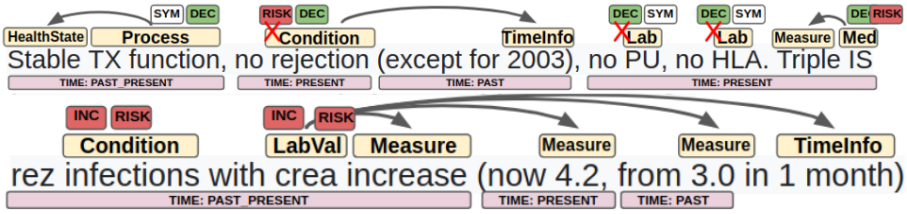

Figure 1: Two annotated explanations translated from German to English, including negations (red X), increasing (INC) and decreasing (DEC) factors. The upper explanation is partially extracted from a positive (endpoint=*rejection*, score=*0*) and the lower one partially from a negative (endpoint=*graft loss*, score=*78*) explanation.

attributes *high*, *middle*, *low*, which describe a high/low age or a long/short time since transplantation.

Label	Description
Condition	A pathological medical condition of a patient, can describe for instance a symptom or a disease.
DiagLab	Particular diagnostic procedures which have been carried out.
Lab Values	Mentions of lab values.
HealthState	A positive condition of the patient.
Measure	Mostly numeric values, often in context of medications or lab values, but can also be a description if a value changes, e.g. *raises*.
Medication	A medication.
Process	Describes particular process, such as *blood pressure*, or *heart rate*, often related to vital parameters.
TimeInfo	Describes temporal information, such as *2 weeks ago* or *January*.
Other	Additional relevant information which influence the health condition, and the risk
Age	Describes the age of the patient
Donor_Age	Describes the age of the donor
Tx_Time	Time since the transplantation

Table 2: Annotated Entities

Label	Description
has_Measure	Connects *Measurements* to mainly *Medications* and *Lab Values*.
has_State	Connects *Condition* and *HealthState* to other entities, such as *Process* or *Lab Values*.
has_TimeInfo	Connects *TimeInfo* entities to other entities, such as *Condition*.

Table 3: Annotated Relations

Factuality Typically clinical notes contain a large number of negations and vague expressions, as it makes a difference if something is currently not present (e.g. a symptom), or cannot be completely verified at this point. This phenomena can be also observed in our textual explanations, and can express something positive or negative. Related work in clinical context often targets factuality regarding symptoms and diseases and include *negations* and *speculations (hedges)*. As the clinical world tends to be more complex than negations and speculations, we extend the standard schema with the attributes: a) *positive*, b) *negated*, c) *speculated*, d) *unlikely*, e) *minor*, and f) *possible future*.

The last three items describe an extension of the original schema used for NegEx (Chapman et al., 2001). *unlikely* defines a kind of speculation, but expresses a tendency towards negation. *minor* expresses that something is present, but to a lower extent or in a lower amount. Finally, *possible future* expresses that something is not there, but might occur in the future. The attribute *positive* is not annotated explicitly.

Progression To analyse the explanations in more detail, entities are extended with some additional information, as presented in Table 4. Firstly we label, if an entity increases or decreases the risk that the endpoint occurs, and if the entity is a risk or a symptom of the endpoint itself.

Label	Description
risk factor	A state/process that causes/prevents the respective endpoint (upstream in a causal chain). Increases/decreases risk causally and probabilistically.
symptom	A state/process whose occurrence/absence is a consequence of the respective endpoint (downstream in a causal chain). Increases/decreases risk probabilistically, but not causally.
increase	increases the risk that endpoint occurs
decrease	decreases the risk that endpoint occurs
conclusion	physician makes a concluding statement

Table 4: Annotation of the progression

2.4. Annotation Process

The annotation process has been carried out in different steps, using the brat rapid annotation tool (Stenetorp et al., 2012). First we ran an automatic annotation. This included the partial automatic labelling of the semantic layer using mEx (Roller et al., 2020) for named entity recognition and relation extraction. Although the schema of mEx partially differs, it includes various of

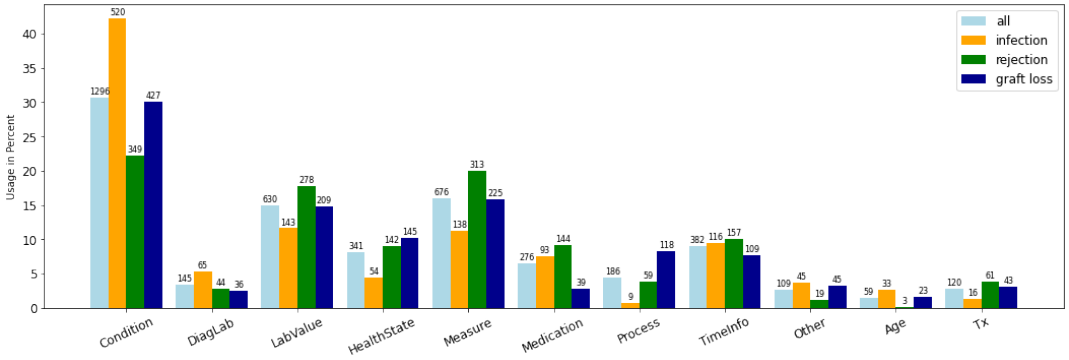

Figure 2: Overview about the normalized frequency of the different entities across all endpoints.

our entities and is a good start to speed up the annotation. Then we applied the German adaptation of NegEx (Cotik et al., 2016) (*negation*, *speculation*) to the *Condition* labels of the automatic generated labels.

In the next step, two physicians corrected the automatically generated labels and included the additional annotation layers *temporal aspects* and *progression*. Both physicians annotated the complete dataset and resulted in an inter-annotator agreement (IAA) of 0.825 mean F1 for the entities on token level[3]. Finally the disagreements of the physicians have been analysed by the first author to make a final decision. Two example annotations are provided in Figure 1, more examples can be found on our github repository.

3. First Quantitative Data Analysis

In the following we provide an analysis of the textual explanations on different levels.

3.1. Document Length

First, explanations were split into single tokens. On average, each document consists of 18.62 tokens with a standard deviation of 11.45. Analysing the documents on endpoint level we see slight variations: Explanations about graft loss tend to be slightly longer than rejection, with 20.60 (12.40) avg tokens per document, in comparison to 19.30 (11.73) avg tokens. Infection has the lowest number of avg tokens with 15.96 (9.60). Moreover, the mean content word usage per explanation (*Lexical Density*) is 79.62%, which shows a very high information density of each explanation.

Figure 3 presents the avg number of tokens per physician. The table shows how the average length of the formulated explanations differs between the physicians. While one physician in particular (S8) uses very short explanations (avg. 9 tokens), many others write much longer texts, even with up to 28 tokens (S5) on average per document.

3.2. Annotated Information

Figure 2 shows the frequency of each entity in the final dataset, overall and also according to each endpoint. The bars in the figure are normalized (divided by the overall number of annotated entities); the true frequency is displayed above each bar. In the case of *Condition*, for instance, the entity occurs 1296 times in the *overall* dataset, and 520 times in the *infection* subset. The figure shows that human explanations for the explanation subset *infection* include the concept *Condition* more frequently (in percentage) compared to *rejection* or *graft loss*. Similarly, it seems that *LabValue*, *Process*, *HealthState* and *Measure* are more frequently used for the explanations of *rejection* and *graft loss*. Overall, *Condition* and *LabValue* are the most frequent entities in the dataset. Therefore, we infer that those are particularly important to define an explanation.

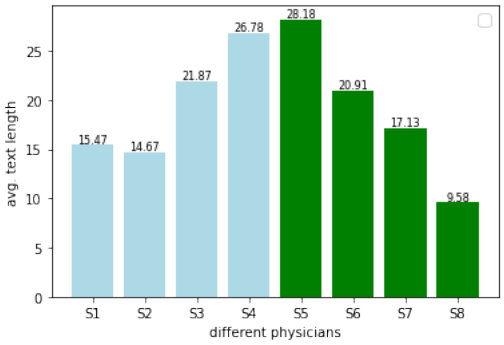

Figure 3: Avg. length of the different physician explanations (junior in orange and senior in green).

Token	Description	Entity	#
Krea	creatinine	LabValue	176
IS	immunosuppression	Medication	73
CRP	c-reactive protein	LabValue	65
stabile	stable	HealthState	64
PU	proteinuria	LabValue	61
aktuell	currently	TimeInfo	60
DSA	donor-specific antibody	LabValue	54
Tx Funktion	transplant function	Process	54
Rejektion	rejection	Condition	52
stabil	stable	HealthState	42

Table 5: Top-10 Frequency of Annotated Tokens

The Top-10 most frequent annotated strings in the dataset are presented in Table 5. Since the most crucial information in the explanations was annotated, the list presents words of high relevance to our explanations,

[3]https://github.com/kldtz/bratiaa

such as *creatinine* (Krea) or CRP. Both laboratory values are of great interest for the given tasks. Creatinine gives information about kidney function, which can be impaired in the case of rejection, graft loss or infection. CRP is a marker of inflammation and is increased during infection. Depending on the endpoint, the list of most frequent annotations change.

Overall, 692 *hasMeasure*, 397 *hasState* and 343 *hasTimeInfo* relations have been annotated. *hasMeasure* is mostly a connection between a *LabValue*, a *Medication* or a *Condition*, together with a *Measure* entity, *hasState* describes in most cases a connection between a *Process* or a *LabValue* together with a *HealthState*, and finally *hasTimeInfo* connects in most cases *Condition* or *LabValue* with *TimeInfo*.

Factuality: Regarding factuality, the dataset includes a majority of negations in comparison to the other elements, as listed in Table 6. In most cases factuality attributes are connected to *Condition* (74%) and *LabValues* (16%) entities.

negation	specul.	pos._future	minor	unlikely
#318	#111	#89	#32	#20
55.8%	19.5%	15.6%	5.6%	3.5%

Table 6: Annotated factuality attributes: Overall frequency (upper line), and percentage in comparison to other factuality attributes (lower line).

3.3. Influence of Risk Score

Now we try to find out if the explanations differ depending on the expected risk score of the physician. For this purpose, the explanations with their associated risk score have been sorted into four different bins. As shown in Table 7, the largest number of explanations are sorted in the bin with a low risk score. Although estimations of the physicians are not necessarily correct, this roughly corresponds to the true distribution - in most cases the endpoint did not occur. Moreover, the table shows that explanations which include a higher risk score, tend to be longer, than explanations with a lower score.

	Risk Score Bins			
	[0-25]	[26-50]	[51-75]	[76-100]
#	217	94	47	75
length	17.85	18.73	19.6	20.11

Table 7: Overview about the number of documents assigned to each risk score bin (#, upper part), and the average length of each explanation in each bin (length, lower part).

Figure 4 depicts the relative usage of the different (selected) entity types within the different risk score bins. As seen in the figure, the usage of the entity *Condition* is much lower, in case of a low risk in comparison to the other risk groups. At the same time the frequency of *HealthState* is much higher in the explanation, the lower the risk of the estimation. While this phenomena appears to be obvious, we can also observe other effects, such as a stronger usage of *Measure* in case of a high risk, or a lower mentioning of *Medication*.

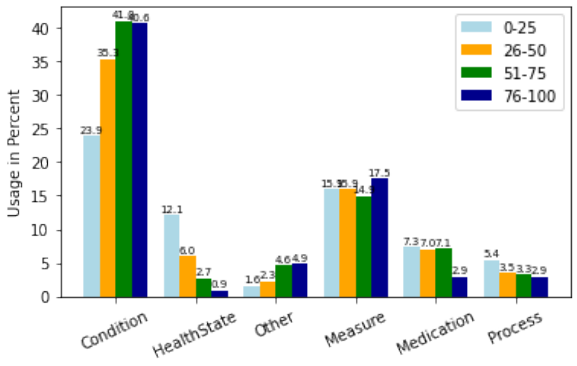

Figure 4: Overview usage of different entity types according to the different risk score bins [0-25], [26-50], [51-75] and [76-100].

Table 8 presents the average usage of progression in each explanation according to the different risk score bins. We can see that explanations with a lower risk score tend to include more decreasing (and less increasing) factors in comparison to explanations with a higher risk score. Similarly in case of explanations with a higher risk score, *finding* is annotated more frequently.

risk bin	increase	decrease	risk factor	finding
0-25	1.04	1.91	1.86	1.37
26-50	1.97	1.04	1.88	1.42
51-75	2.96	0.36	2.09	1.50
76-100	3.58	0.10	1.83	2.10

Table 8: Average occurrence of given progression attributes per explanation - according to risk score bin.

3.4. General Analysis

As seen in Figure 2, *Condition* and *LabValue* are used very frequently in our explanations. However, other entities also play an important role describing a high, medium or low risk in the context of a particular endpoint. The entities *TimeInfo* and *Measure* themselves might contribute only marginally to the explanation, but get higher values by connecting them to other entities. While *TimeInfo* defines when something happened, *Measure* adds information about certain measurements of lab values, tendencies or medication dosages, etc.. Together with the connected entity, a lab value might then transform into something positive (decrease risk) or something negative (increase risk). While the interpretation of such an explanation requires expert knowledge, in various cases the core

entity can be extended by some *HealthStatus* (e.g. 'stable', 'good') or *Condition*. Similar scenarios can be observed also for *Processes* and other entities. Overall, explanations contain much implicit information.
Some explanations go even further and provide (partially) high-level justifications such as '*Risk results from anamnesis (case history).*' (risk score of 90). For this explanation, the author just refers to the given parameters in the database without further explanation and assumes that the situation and the risk factors are obvious. Other explanations include positive mentions (*HealthState*) such as 'good general condition' which appears to be more of a summarization of different factors, or 'good transplant function'.

4. Thoughts about developing explainable decision support

In this section, we first analyse the feature-based explanations of our machine learning system to illustrate on which basis current systems arrive at their predictions. We then present different explanation types which have been found across the human dataset and discuss the possibility to integrate them into an textual explanation of an automatic clinical decision support system.

4.1. Feature-based model explanations

Although the machine learning model and its explanations are rather simple, the analysis might be still of interest for the development of future decision support systems.
The features of the machine learning model mainly take structured data into account. This includes, for instance, demographics, lab values, vital parameters, medication (changes) and diagnoses. As patients tend to have several visits per year, the models included the last two occurring values in many cases. In addition to that, some features were explicitly modelled, such as for instance the temporal distance to particular events (e.g. transplantation, or last rejection), number of measured values, length of stay in a hospital, or the increase/decrease of values. However, in comparison to the physicians, the machine learning model could not access detailed information about hospitalizations, and did not consider textual information (e.g. clinical text).
Table 9 presents the overall ("global") most relevant features for each model and for each endpoint. In comparison to the most frequently annotated information in the explanations of the physicians (see Table 5), we can see a certain overlap. Generally, the explanations do consider the lab values *creatinine* and *CRP*, and also talk about the time since transplantation, or recent infections or rejections. However, it is interesting to note in particular that time distances or frequencies appear to be of valuable information rather than a particular value, such as the number of infections in the last 360 days. Most notable are the features *#lab values in the last 60 days* and *days since last lab value*. Both features appear to be very different from explanations a human would provide. On the other hand, the features might make sense and carry valuable high-level information, such as the 'patient underwent a lot of different examinations' implying the treating physician thinks that the patient might be in a serious condition.

	feature	**value**	**import.**
Rejection	last creatinine value	float	12.78
	months since transplantation	int	7.66
	had rejection in last 180 days	binary	6.69
	days since transplantation	int	6.69
	#lab values in the last 60 days	int	3.18
Graft Loss	months since transplantation	int	35.12
	#transplantations	int	23.76
	last creatinine value	float	8.99
	days since transplantation	int	3.04
	gfrhp	float	2.26
Infection	#infections in the last 180 days	binary	8.75
	#days since last lab value	int	7.67
	crphp outside norm	binary	6.42
	#infections in the last 360 days	int	6.41
	rdweb (last value)	float	6.28

Table 9: Most relevant (global) model features of decision support system according to the different endpoints, including their importance.

feature	frequency
body size	134
blood pressure (diastolic)	118
last creatinine value	94
# hospitalizations in last year	67
mean CRP value	59
current weight	57
last hsthp value	55
age	54
mean creatinine value	53
body temperature	48

Table 10: Most frequent (local) features presented by system. Features have been extracted from the Top-5 list of all 720 predictions.

Table 10 presents the most frequent (local) features for the different endpoint predictions. Besides some obvious lab values and age, the table shows that the ML system takes additional features into account, which have not been mentioned by the physicians. Firstly, it considers some mean lab values to be relevant. While general norm values for e.g. particular age groups exist, patients might also have a personalized score which could in theory be naturally higher/lower compared to other patients. Moreover, features such as body temperature (possibly an indication for infection) or weight are taken into account. In general, an increase in weight within a short period of time can be a negative symptom, but the weight itself does not seem to be a useful feature. Finally, some features, such as *body size*,

appear to be not relevant at all. However, besides that *body size* tends to occur with low risk predictions, this feature might not be necessarily the reason for the model prediction - instead it could be connected to a sub-optimal ('local') feature extraction.

4.2. Aspects of Human Explanations

From a technical point of view, a textual explanation for an automatic clinical decision support system can be implemented - at least in parts. In the following we present different aspects in human explanations we observed, and discuss how those could be technically implemented by an XAI model which has access to the relevant local features for each decision.

Explicit Description is the most obvious explanation pattern of the physicians, which mentions a given feature and its value, for instance a systolic blood pressure of 125. This can be easily implemented from a technical point of view, using the current value of the relevant local feature, and then inserted into a template.

Tendencies and fluctuations of values are frequently used in the explanations. However, they might be more difficult to be identified by XAI methods, as they describe a correlation between subsequent values. If those aspects play an important role for the given prediction task, it is not enough to report just actual measurement of the value. Instead, it might be helpful to summarize the information into natural language statements, e.g. 'the value rises (sharply)'. In this way, information could more easily be perceived by a human reader. However, it is not clear how much sequential information needs to be taken into account, or how long the target period should be. This might depend on the context and the given task.

Factuality partially also plays an important role, e.g. the explicit absence of symptoms, lab values or medication can describe something positive or negative. The explicit absence of information might be more difficult to take into consideration to an XAI learning model. One way to address this issue could be an ablation study which explores how the decision of a model changes if something would be present. As the number of missing values can be arbitrarily large, it would be important to identify the really relevant ones, otherwise the explanation would be filled with irrelevant missing facts. Aspects such as *speculations* or *possible future* might be much more difficult to introduce.

Values in-/outside the norm are frequently reported in the explanations and represent a high level description of values, connected with an assessment - something is rather positive if inside the norm. This explanation might make it easier for a reader, as the actual value might not be that important, but rather the aspect that something is a borderline case, etc. From a technical perspective, this can be implemented, but requires additional world knowledge and could be included in the form of rules. Note, if a value is within the norm often depends on multiple factors, such as person itself, age and gender.

Interpretation of given facts are used by the physicians and has similarities to the aforementioned explanation aspect. The physicians often use interpretations in such a way that values are either *good*, *bad*, or *stable*, rather than on a continuous scale. This, however, might depend on the physician's experience and existing medical guidelines. While the experience of a physician could be difficult to capture within an XAI (and is possibly not wanted), aspects related to medical guidelines can be implemented similarly as in the case above. However, to go from e.g. *being inside the norm* to something *being good* is even one step further and moves from pure facts to interpretation. Therefore, it can be implemented, but might be more difficult and opens additional susceptibility to errors.

High-level interpretations such as '*good/bad general condition*' or '*fit patient*' are an easy way to provide an easily understandable and fast overview about the patient's health condition. From XAI perspective, those explanations might be difficult to generate. Technically, a set of such conditions ('fit', 'good condition') could be defined, identified within a classification (regression) task, and then mapped to an explanation. However, this would require sufficient training data and adds new sources of error (misinterpretation). Moreover, it is not clear if those kinds of high-level interpretations would actually be wanted by users of an automatic decision support system. We presume that if the XAI provides less information, and mainly high-level interpretations, this could not necessarily increase its trustworthiness.

5. Related Work

Explainable artificial intelligence (XAI) in the medical domain State-of-the-art machine learning approaches, usually based on neural networks, reveal a black-box nature, so automated predictions and decisions are barely comprehensible to humans. This harms trust which is crucial in high-stakes settings such as the medical domain (Rudin, 2019; Bruckert et al., 2020). Since computer-aided diagnosis can have a direct influence on the well-being of patients, transparent and justifiable explanations are of utmost importance for real-world clinical practice (Lucieri et al., 2020).

According to Holzinger (2020), there is a need for causality in XAI for medicine: "In the same way that usability encompasses measurements for the quality of use, causability encompasses measurements for the quality of explanations produced by XAI." XAI systems should "allow a domain expert to ask questions to understand why an AI came up with a result, [...] to gain insight into the underlying independent explanatory factors of a result". We argue that dataset contributions like ours advance the field in this direction: We follow Holzinger (2020) in that we contribute explicit knowledge in the form of natural language explanations made by domain experts. It can be used to build interfaces which generate more plausible explanations.

Natural language explanations Wiegreffe and Marasović (2021) provided a survey of datasets containing human-annotated natural language explanations (NLEs, also referred to as rationales). They identified three distinct categories: Highlights, free-text, and structured explanations. Our work concerns itself with free-text explanations which are not constrained to the words or modality of the explanandum.

There exist very few corpora containing free-text NLEs in the medical domain as NLEs themselves are traditionally underrepresented in the medical XAI canon (Tjoa and Guan, 2021). Firstly, DeYoung et al. (2020) introduced the EVIDENCE INFERENCE corpus consisting of biomedical articles that describe randomized control trials and associated prompts. These can be used for an evidence extraction task, i.e. the task of inferring the relationship between a treatment and a comparator with respect to an outcome. Annotators were asked to provide rationales for their answers (what type of relationship: increase, decrease, no difference) in the form of text highlights. Secondly, Kotonya and Toni (2020) presented the PUBHEALTH corpus for explainable fact-checking of claims in the public health setting. It contains the full text of the fact-checking article discussing the veracity of the claim and a justification as explanation for the veracity label. The work of Kotonya and Toni (2020) and DeYoung et al. (2020) both focus on explanations in the context of biomedical articles. Our work instead, targets explanations made by physicians, for clinical decision support with certainty scores as the primary application.

Regarding certainty scores in NLP, Chen et al. (2020) proposed a refinement of natural language inference, enhancing datasets with additional human judgments about the likelihood of a categorical label on a probabilistic scale. Our dataset also contains likelihood estimates and allows for subtle distinctions.

Lastly, Taylor et al. (2021) aimed to predict the likelihood of patients being re-admitted to a hospital after a prior ICU stay and generate an NLE alongside it. However, the adoption of the MIMIC-III dataset was rather unsuccessful and the authors reported the need for clinically-derived ground truths.

Diagnostic captioning We also draw a connection to the task of diagnostic captioning (or biomedical image captioning; radiology report generation). While the task has an inherent multimodal nature building upon the task of image-to-text generation, the textual reports written by doctors are detailed diagnoses that are self-explanatory to domain experts (Lucieri et al., 2020).

The 'Impression' ("a short summary of the most immediately relevant findings") and 'Findings' ("a natural language description of the important aspects in the image") metadata fields of MIMIC-CXR (Johnson et al., 2019) instances can be interpreted as a decision-explanation pair. Nevertheless, explicit justifications for classification labels or annotations are not present. Spinks and Moens (2019) pointed out several issues with the annotations of such datasets. This means that for a proper evaluation of an NLE, the data has to be drastically limited to specific subsets and labels.

Many models trained on these datasets have been examined with different local explanation methods such as saliency maps (Messina et al., 2022). A drawback of their interpretability, however, is that their outputs provide no information about why particular pixels are important for the outcome or to what class they belong (Spinks and Moens, 2019). Moreover, they have yet to be evaluated for plausibility to explainees and faithfulness to the underlying model and – as previously stated – ground truth explanations covering both image and text dimension are hard to come by. Our dataset circumnavigates this issue, since it strictly contains textual data and rich annotations allowing for thorough causal analyses.

6. Conclusion

In this work we presented a dataset of human justifications/explanations in the context of predicting possible outcomes in medicine. Due to the annotations of different layers (semantic, factuality, temporal information) our data will certainly be a useful resource in the context of clinical text processing. Moreover, although it might be difficult to use this data to directly setup an XAI system, this work provides insights into how humans would provide an explanation on the given task. We draw a comparison to a feature based XAI system, present different explanation patterns of human physicians, and discuss how they could possibly be technically implemented.

We think that this corpus potentially constitutes a valuable resource for different scholars interested in ethical, legal and social implications (ELSI) of human-machine interaction (to be precise: human-human interaction enhanced by machine output in our scenario). From an ethical standpoint, one might, e.g., be interested in how responsibility is shared, which may lead over to legal questions of liability and probably several other related topics.

What we envision as a particularly interesting future work would be to cluster and abstract the explanations into certain explanation types (e.g., the formalised checklist type vs. the global estimation type). From our first studies of the corpus, we are convinced that it will not be possible to come up with a one-size-fits-all kind of blueprint for the "ideal" explanation. We think it will rather be a question of personal taste that should be studied in future experiments, e.g., using Wizard-of-Oz systems that are configurable regarding the explanation types and that assess the effectiveness and efficiency of the explanations for different test subjects.

7. Acknowledgements

This research was supported by the German Federal Ministry of Education and Research (BMBF) through the project vALID (01GP1903A).

8. Bibliographical References

Bruckert, S., Finzel, B., and Schmid, U. (2020). The Next Generation of Medical Decision Support: A Roadmap Toward Transparent Expert Companions. *Frontiers in Artificial Intelligence*, 3:75.

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., and Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310.

Chen, T., Jiang, Z., Poliak, A., Sakaguchi, K., and Van Durme, B. (2020). Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online, July. Association for Computational Linguistics.

Cotik, V., Roller, R., Xu, F., Uszkoreit, H., Budde, K., and Schmidt, D. (2016). Negation detection in clinical reports written in German. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*, pages 115–124.

DeYoung, J., Lehman, E., Nye, B., Marshall, I., and Wallace, B. C. (2020). Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132, Online, July. Association for Computational Linguistics.

Feldhus, N., Schwarzenberg, R., and Möller, S. (2021). Thermostat: A large collection of NLP model explanations and analysis tools. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 87–95, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410.

Holzinger, A. (2020). Explainable AI and Multi-Modal Causability in Medicine. *i-com*, 19(3):171–179.

Johnson, A., Pollard, T., Berkowitz, S., Greenbaum, N., Lungren, M., Deng, C., Mark, R., and Horng, S. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8.

K, K. and Søgaard, A. (2022). Revisiting methods for finding influential examples. *AAAI*.

Kotonya, N. and Toni, F. (2020). Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online, November. Association for Computational Linguistics.

Lucieri, A., Bajwa, M. N., Dengel, A. R., and Ahmed, S. (2020). Achievements and challenges in explaining deep learning based computer-aided diagnosis systems. *ArXiv*, abs/2011.13169.

Markus, A. F., Kors, J. A., and Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113:103655.

Messina, P., Pino, P., Parra, D., Soto, A., Besa, C., Uribe, S., Andía, M., Tejos, C., Prieto, C., and Capurro, D. (2022). A Survey on Deep Learning and Explainability for Automatic Report Generation from Medical Images. *ACM Computing Surveys (CSUR)*.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.

Roller, R., Seiffe, L., Ayach, A., Möller, S., Marten, O., Mikhailov, M., Alt, C., Schmidt, D., Halleck, F., Naik, M., et al. (2020). Information Extraction Models for German Clinical Text. In *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–2. IEEE.

Roller, R., Budde, K., Burchardt, A., Dabrock, P., Möller, S., Osmanodja, B., Ronicke, S., Samhammer, D., and Schmeier, S. (2022). When Performance is not Enough – A Multidisciplinary View on Clinical Decision Support. *ArXiv*, abs/2204.12810.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215.

Spinks, G. and Moens, M.-F. (2019). Justifying diagnosis decisions by deep neural networks. *Journal of biomedical informatics*, page 103248.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France, April. Association for Computational Linguistics.

Taylor, N., Sha, L., Joyce, D. W., Lukasiewicz, T., Nevado-Holgado, A. J., and Kormilitzin, A. (2021). Rationale production to support clinical decision-making. *ArXiv*, abs/2111.07611.

Tjoa, E. and Guan, C. (2021). A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32:4793–4813.

Wiegreffe, S. and Marasović, A. (2021). Teach Me to Explain: A Review of Datasets for Explainable NLP. In *Proceedings of NeurIPS*.

Wu, T., Ribeiro, M. T., Heer, J., and Weld, D. (2021). Polyjuice: Generating counterfactuals for explain-

ing, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online, August. Association for Computational Linguistics.