Delft University of Technology

Human-centered XAI

Developing design patterns for explanations of clinical decision support systems

Schoonderwoerd, Tjeerd A.J.; Jorritsma, Wiard ; Neerincx, Mark A.; Van Den Bosch, Karel

DOI
10.1016/j.ijhcs.2021.102684

Publication date
2021

Document Version
Final published version

Published in
International Journal of Human Computer Studies

Citation (APA)
Schoonderwoerd, T. A. J., Jorritsma, W., Neerincx, M. A., & Van Den Bosch, K. (2021). Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. *International Journal of Human Computer Studies*, *154*, 1-25. Article 102684. https://doi.org/10.1016/j.ijhcs.2021.102684

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Int. J. Human–Computer Studies 154 (2021) 102684

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Human-centered XAI: Developing design patterns for explanations of clinical decision support systems

Tjeerd A.J. Schoonderwoerd [a], Wiard Jorritsma [a], Mark A. Neerincx [*,a,b], Karel van den Bosch [a]

[a] TNO, Soesterberg, Kampweg 55, the Netherlands
[b] Technical University of Delft, Delft, Mekelweg 5, the Netherlands

ARTICLE INFO

Keywords:
Explainable AI
Explainability
Causability
Human-centered design
Clinical decision making
Decision-support system
User study
Design patterns

ABSTRACT

Much of the research on eXplainable Artificial Intelligence (XAI) has centered on providing transparency of machine learning models. More recently, the focus on human-centered approaches to XAI has increased. Yet, there is a lack of practical methods and examples on the integration of human factors into the development processes of AI-generated explanations that humans prove to uptake for better performance. This paper presents a case study of an application of a human-centered design approach for AI-generated explanations. The approach consists of three components: Domain analysis to define the concept & context of explanations, Requirements elicitation & assessment to derive the use cases & explanation requirements, and the consequential Multi-modal interaction design & evaluation to create a library of design patterns for explanations. In a case study, we adopt the DoReMi-approach to design explanations for a Clinical Decision Support System (CDSS) for child health. In the requirements elicitation & assessment, a user study with experienced paediatricians uncovered what explanations the CDSS should provide. In the interaction design & evaluation, a second user study tested the consequential interaction design patterns. This case study provided a first set of user requirements and design patterns for an explainable decision support system in medical diagnosis, showing how to involve expert end users in the development process and how to develop, more or less, generic solutions for general design problems in XAI.

1. Introduction

The research community of eXplainable AI (XAI) has a large technological focus on gaining insight in black-box machine learning models with output explanations (e.g., Adadi and Berrada, 2018; Guidotti et al., 2018b). For example, Ribeiro et al. (2016) extract relevant input features of a machine learning model, along with their contribution to the output, which can be used in explanations to a human. Such methods are valuable because they enable extraction of relevant information from complex machine learning models, potentially increasing the transparency of such models. However, they do not elucidate whether such explanations are fit for purpose for the humans in the concerning operational context (e.g., Anjomshoae et al., 2019; Hoffman et al., 2018; Miller, 2018).

There is a clear need for methods and models to pervasively integrate the human factor into the research and development of XAI (e.g., Kirsch, 2017; Miller, 2018; Neerincx et al., 2018; Ras et al., 2018; Schneider and Handali, 2019; Thellman et al., 2017; see Anjomshoae et al., 2019 for a comprehensive literature review). Such methods and models should capture the notion that an explanation always is a response to a particular, implicit or explicit, request by an explainee in a specific context (e.g., Caro-Martinez et al., 2018; Hoffman et al., 2018; Kirsch, 2017; Lombrozo, 2006; Miller, 2018; Ribera and Lapedriza, 2019). So, explanations need to be grounded in an understanding of the primary purpose of the AI-system, its users, and its intended use in order to learn if, why, what and when explanation is required. Note, for example, that different types of users (e.g., developers, domain experts, and lay users) require different kinds of explanations (Burnett, 2020; Ribera and Lapedriza, 2019).

Take the example of a Clinical Decision Support System (CDSS) for diagnosis, typically making use of machine learning (e.g., multi-class classification) algorithms in order to predict the likelihood of a diagnosis based on numerical data in a case (e.g., demographic data, clinical history and answers on questionnaires). The predictions have to be explained in order to have clinical relevance and to facilitate long-term use by clinicians (Guida et al., 1997; Ye and Johnson, 1995). In order to

* Corresponding author at: TNO, Soesterberg, Kampweg 55, the Netherlands.
E-mail address: M.A.Neerincx@tudelft.nl (M.A. Neerincx).

https://doi.org/10.1016/j.ijhcs.2021.102684
Received 25 March 2021; Received in revised form 14 June 2021; Accepted 16 June 2021
Available online 18 June 2021
1071-5819/© 2021 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

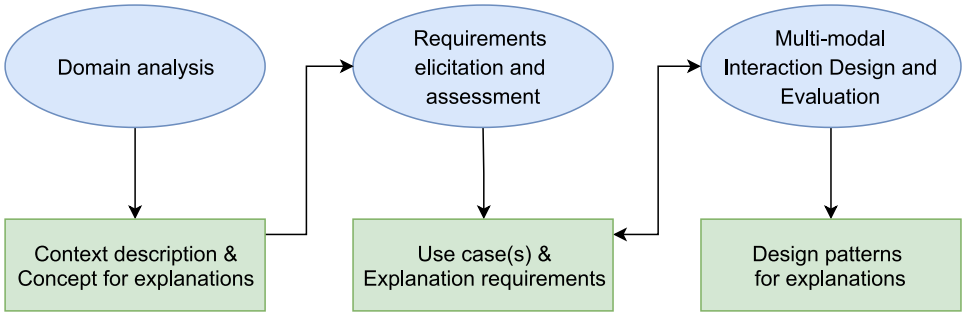


Fig. 1. Flow diagram of the DoReMi-practice for human-centered XAI design. It consists of three components:*Domain analysis, Requirements elicitation & assessment*, and *Multi-modal Interaction design & evaluation*. Each component produces outcomes (indicated by rectangles) that serve as input for another component in the process.

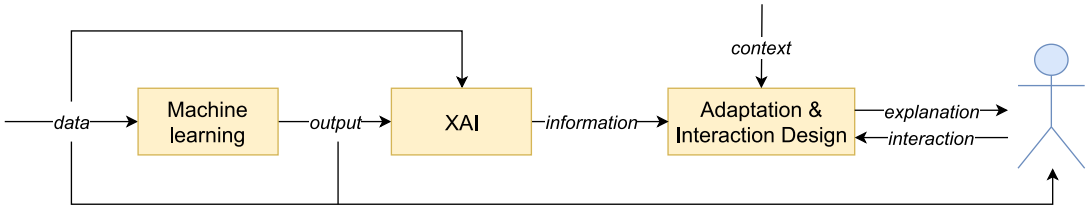


Fig. 2. Framework for explanation generation and communication to a user.

Table 1
Information elements that can be used in clinical decision-making, and the explanation-category they belong to.

#	Information element	Category
1	Patient information that is used to make the diagnosis.	General
2	The reason why this information is indicative of this diagnosis.	General
3	A description of the condition.	General
4	The prevalence of the condition.	General
5	The information that supports this diagnosis.	Evidence
6	The information that contradicts this diagnosis.	Evidence
7	Other diagnoses that are conceivable based on the case information.	Contrastive
8	The reason why it is this diagnosis, and not another one.	Contrastive
9	The likely diagnosis if feature X had not been A but B.	Counterfactual
10	From what value of feature X the diagnosis would have been different.	Counterfactual
11	Similar cases that received the same diagnosis.	Case-based
12	How this case relates to another, similar case.	Case-based
13	A typical case of someone with this diagnosis.	Case-based
14	How the current case relates to the typical case of this diagnosis.	Case-based
15	How certain this diagnosis is.	Certainty
16	The information that would increase the certainty of the diagnosis.	Certainty
17	How it was determined that feature X was or was not present.	Input data
18	The information that is relevant in making this type of diagnosis.	Input data
19	Cases that are different, yet received the same diagnosis.	Case-based
20	The performance of the system for other, similar cases.	System performance

design fit-for-purpose explanations for clinicians, XAI developers require an in-depth understanding of the medical decision-making process, the use of explanations in medical diagnosis, and the explanation needs of clinicians. Moreover, they need insight into the context in which explanations are provided: what explanation is effective in one context may be different in another (e.g., explaining to a colleague-clinician or to a patient). Moreover, personal preferences of the user might play a role (e.g., clinicians might focus on different information in a case). Explanations from a system may also have to be sensitive to the degree of agreement between the hypothesized diagnosis by the clinician and the system. For example, when a clinician does not agree with the advised diagnosis, it might be especially relevant to show information that helps to determine the cause of the discrepancy. Additionally, it might be relevant to take into account the amount of experience that a clinician has with the system.

Table 2
Social conditions in which we investigated explanation requirements.

Situation ID	Situational description
Situation 1a	Providing explanation to an impartial colleague
Situation 1b	Providing explanation to an agreeing colleague
Situation 1c	Providing explanation to a disagreeing colleague
Situation 2a	Providing explanation to an agreeing parent of the patient
Situation 2b	Providing explanation to a disagreeing parent of the patient
Situation 3a	Receiving congruent explanation from a CDSS
Situation 3b	Receiving incongruent explanation from a CDSS

More recently, researchers in the field of XAI have increased the attention to the fact that the development of explanations from an AI-system requires a tailor-made approach. The focus within XAI development has started to shift from a mainly technical approach to a more integrated sociotechnical one in which human-centered design (HCD) is paramount (e.g., Lim et al., 2019; Mittelstadt et al., 2019; Neerincx et al., 2019; Madumal, Miller, Sonenberg, Vetere, see Arrieta et al., 2020 for an overview of recent papers on HCD for XAI). In order for this approach to succeed, it is required that the XAI research community establishes a common understanding of terms such as explainability and transparency. In this paper, we adopt the terminology as described by Arrieta et al. (2020):

- **Explanation**: an interface between human and system that accurately approximates the model of the system and is comprehensible to the human (Guidotti et al., 2018b).
- **Explainability**: the ability to deliver explanations. The model that is used by the system needs to be interpretable to be able to provide an explanation (Guidotti et al., 2018b).

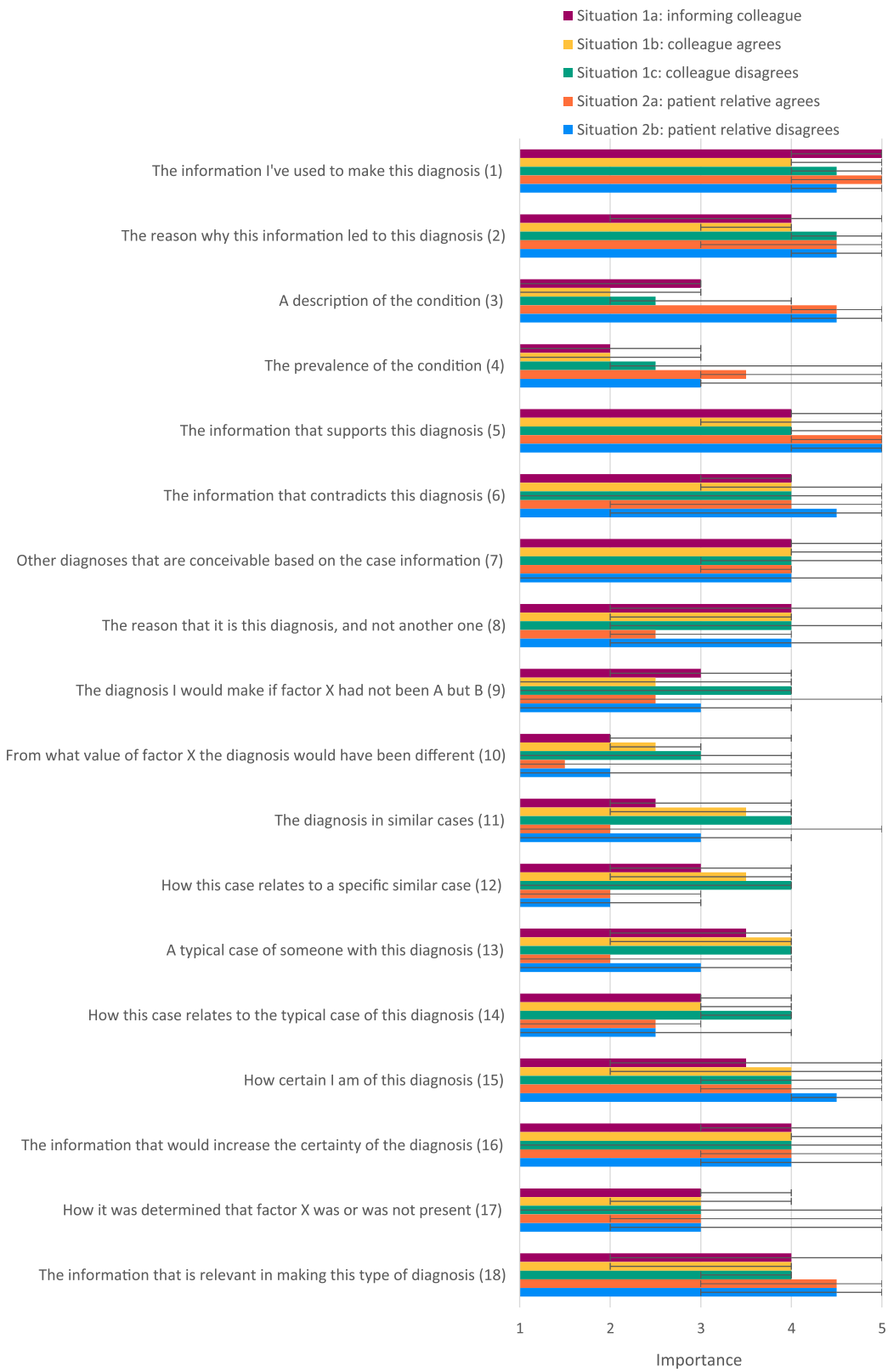


Fig. 3. Median importance of information elements, as rated by clinicians for each situation. Error bars indicate minimum and maximum values.

- **Causability**: the ability the enable a user to achieve causal understanding with effectiveness, efficiency, and satisfaction in a specified context of use (Holzinger et al., 2019).
- **Interpretability**: the ability to provide meaning to a human in understandable terms (Guidotti et al., 2018b).
- **Transparency**: a model is transparent if it is understandable by itself (Adadi and Berrada, 2018).
- **Comprehensibility**: the ability of a model to represent its knowledge in an understandable fashion (Adadi and Berrada, 2018).

Fig. 4. Median importance of information elements, as rated by clinicians for each situation. Error bars indicate minimum and maximum values.

- **Understandability**: the ability to make a human understand the model's function without the need to explain its internal structure or the algorithms that are used (Montavon et al., 2018).

Thus, an *understandable explanation* from a system provides a human with information that is extracted from and/or based on its internal model and makes a human understand (part of) the functioning of the model, in order to understand an output of this model. Importantly, explainability deals with extracting explanations from a system's model, which are not inherently human-understandable. Therefore, an explainable model and explanation interface is required to create explanations that can be understood by humans (Holzinger et al., 2019). Thus, while XAI is concerned with developing methods to make machine models transparent and traceable, causability is about measuring the quality of such explanations to increase causal understanding of a user (Holzinger et al., 2020). The human-centered design methodology provides methods to determine exactly what information is understandable and useful to humans and thus should be used in explanations from the system. With some exceptions (e.g., Ehsan and Riedl, 2020; Eiband et al., 2018), there are still very few examples that show how to apply human-centered design in XAI development. In this paper, we present a case-study of human-centered design for post-hoc, local explanations of diagnoses that are made by a clinical decision-support system. We describe a best-practice approach for human-centered XAI design that we call DoReMi, consisting of three components: *D*omain analysis, *Re*quirements elicitation, and *M*ulti-model *i*nteraction design. This approach instantiates human-centered design, specifically aiming at generic design solutions for explanations from AI-systems that are grounded in the requirements and needs of the users and operational context.

In the remainder of the paper, we describe and apply DoReMi in order to answer our main research question: *How to design post-hoc local explanations of artificial intelligence that are processed and understood by humans in such a way that the overall human-AI performance is effective, efficient and satisfactory?* The research context in which we investigate this question is decision-support for child health diagnosis. Clinicians are involved in all three components of the human-centered DoReMi practice. In the latter two components of DoReMi, experienced paediatricians take part in two user studies. In the first user study we discover what explanations the CDSS should be able to provide. For these explanations, interaction design patterns are specified, instantiated in a mock-up, and, subsequently, tested in the second user study.

2. Human-centered design for XAI

The need for the human factor in XAI design and development has

Table 3
UI design patterns (DPs) and the information needs they address.

ID	Problem description (the user needs to know...)	Information elements from Table 1
DP 1: Class information	A description of the class & The prevalence of the class	3, 4
DP 2: Available/ relevant information	The information that is used to make the classification & The information that is relevant in making this type of classification	1, 18
DP 3: Certainty	How certain the system is of this classification	15
DP 4: Supporting/ contradicting information	The information that supports this classification & The information that contradicts this classification	5, 6
DP 5: Feature value origin	How it was determined that feature X was or was not present	17
DP 6: Alternative classifications	Other classifications that are conceivable based on the case information	7
DP 7: Contrastive explanation and thresholds	The reason why it is this classification, and not another one & From what value of feature X the classification would have been different	8, 10
DP 8: Counterfactuals w.r.t. classification	The likely classification if feature X had not been A but B	9
DP 9: Counterfactuals w.r.t. certainty	The information that would increase the certainty of the classification	16
DP 10: Comparison to other cases	The classification in similar cases & How this case relates to a specific similar case & The most different cases with the same classification	11, 12, 19
DP 11: Comparison to typical cases	A typical case with this classification & How this case relates to the typical case of this and other classifications	13, 14
DP 12: Performance on similar cases	The performance of the system for other, similar cases	20

been addressed in numerous recent papers (e.g., Amershi et al., 2019; Eiband et al., 2018; Hall et al., 2019; Holzinger et al., 2019; Liao et al., 2020; Neerincx et al., 2019; Ribera and Lapedriza, 2019; Ehsan, Riedl; Markus, Kors, Rijnbeek). These papers provide more or less specific design guidelines for researchers and practitioners to support development of effective explanations that meet human needs. In addition, some also describe how to systematically apply HCD methods to the XAI design and development process in order to better understand the social and technical requirements of human-AI interaction (e.g., Eiband et al., 2018; Hall et al., 2019; Neerincx et al., 2019; Ehsan, Riedl). In general, the human-centered approach to XAI focuses at uncovering what, when, and how to explain to human end users, by iteratively involving the users in the development process (e.g., through interviews, hypothetical scenarios, focus groups, and questionnaires). For example, Eiband et al. (2018) describe an approach to increase transparency of existing intelligent systems by involving users in a stage-based design process to develop explanations. Moreover, Wolf (2019) shows how to use scenarios early in the system development process to identify the user needs for explanations, which can then serve as basis for further development of explanations.

These cooperative design methodologies help researchers and developers to identify and understand the values of users and the social and operational context of the human-AI interaction. The needs and requirements that are uncovered can be further evaluated by assimilating them into explanations and by presenting them to end users in a (simplified) work context. DoReMi applies these typical phases of HCD (i.e., understand, define, design, and evaluate) to XAI research and development, and contributes to existing theory and methods by facilitating the production of *reusable* knowledge for XAI design in the form of generic design patterns. DoReMi explicitly links requirements to design solutions through a design rational, thereby providing best practices for explanations from AI systems.

2.1. Process

Fig. 1 shows a process flow diagram for human-centered XAI design. We distinguish three components that are important in the design process: domain analysis, requirements analysis, and interaction design. Each component produces outcomes that serve as input for another component. For each component, we will explain its goal in the design process, list methods that are appropriate to achieve this goal, and describe its outcomes.

2.1.1. Domain analysis

A starting point for all human-centered design approaches is to gain understanding about the context of use (e.g., Ehsan and Riedl, 2020; Eiband et al., 2018; Hall et al., 2019; Tomsett et al., 2018). In DoReMi, this is called the domain analysis, in which the focus is on gaining understanding about the context in which the system will be introduced in order to develop a first concept for explanations. The goal here is to determine if and why explanations are required, and what information can be considered relevant in the context that is studied. Important questions that the XAI developer should try to answer are: *Who is going to use the system?, What are typical tasks that these users perform?, What is the expected benefit of using the system?, What is the primary function of explanations within this context?*, and *What type of explanations potentially improve human-system interaction in this context?*. This domain analysis facilitates value-sensitive design (Friedman et al., 2008), in which developers explicitly account for human values in the design process by assessing the role of the system and the relevance and function of explanations.

Methods to carry out a domain analysis are for example: consulting available literature, performing interviews with domain-experts, presenting hypothetical scenarios, and observing experts while they perform tasks. For our medical diagnosis case, we consulted the vast amount of literature that is already available on clinical decision-support systems (e.g., Berner and La Lande, 2007; Friedman et al., 1999; Kawamoto et al., 2005; Ozaydin et al., 2016).

The outcome of the domain analysis is a description of the context in which the system will provide explanations, and a first concept for the explanations based on the information that is relevant to end users.

2.1.2. Requirements elicitation and assessment

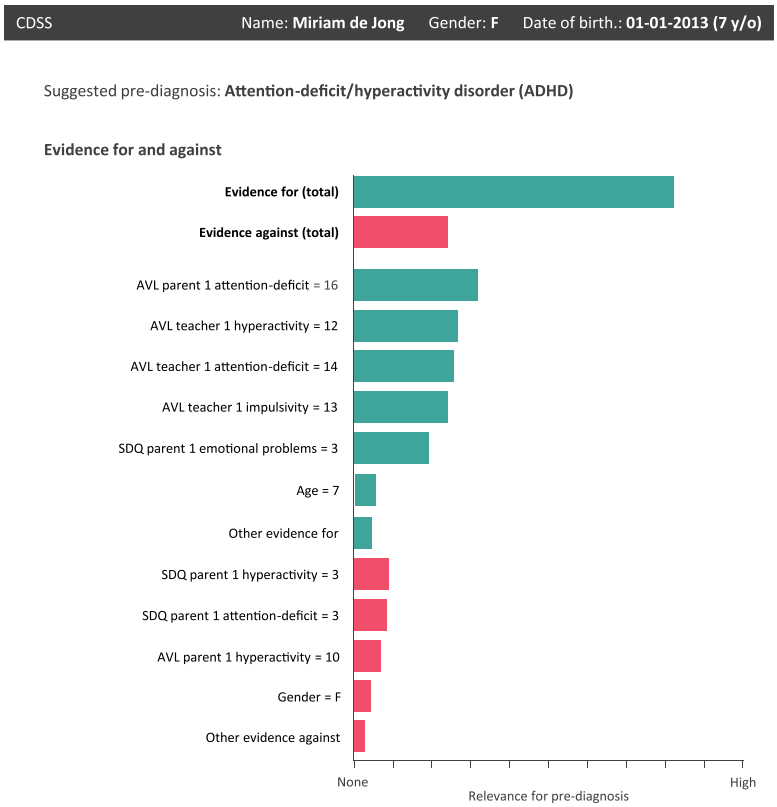
Another part of the DoReMi-method consists of identifying the requirements that users pose for explanations that are provided by the system. The goal in this part is to determine what kind of explanations the system should be able to provide, and to identify potential contextual dependencies. Relevant questions in this part of the process are: *What are the requirements that users pose for explanations from the system?*, and *(How) should explanations be adapted to the specific context in which they are provided?*.

There are multiple methods that can be used to elicit such requirements, all of which actively involve end users (see Paetsch et al., 2003 for an overview). For example: think-aloud protocols, questionnaires, discussions with end users, or requirements prioritization. Key in all of these methods is to provide users with a sufficiently rich context (i.e., a use case or scenario) from which their requirements can emerge (Maguire and Bevan, 2002). For example, Wolf (2019) proposes to present users with potential scenarios of use in which explanations are likely to be relevant (i.e., explainability scenarios). When using DoReMi, such use cases can be based upon the context description that is obtained in the domain analysis. By asking users to provide explanations themselves based upon the use cases, we can obtain preliminary insight into the kind of explanations that users expect to receive from the system.

Table 4
DP 4: Supporting/contradicting information.

Problem description	The user needs to know: - The information that supports this classification - The information that contradicts this classification
UI design example	

Problem description	The user needs to know: - The information that supports this classification - The information that contradicts this classification
UI design example	



Rationale	The design was based on common feature importance visualizations from the XAI literature (e.g., [58, 3, 59]). Inspired by Poulin and colleagues [58], we placed the total supporting and contradicting information at the top to allow for easy visual comparison of the difference between the two. The individual features comprising these total bars are presented underneath. We chose to visualize positively and negatively contributing features in the same direction and distinguish them by color (as in e.g., [3]), as opposed to showing them in opposite directions (as in e.g., [59, 43], because that makes it easier to compare the sizes of the positive and negative feature bars.

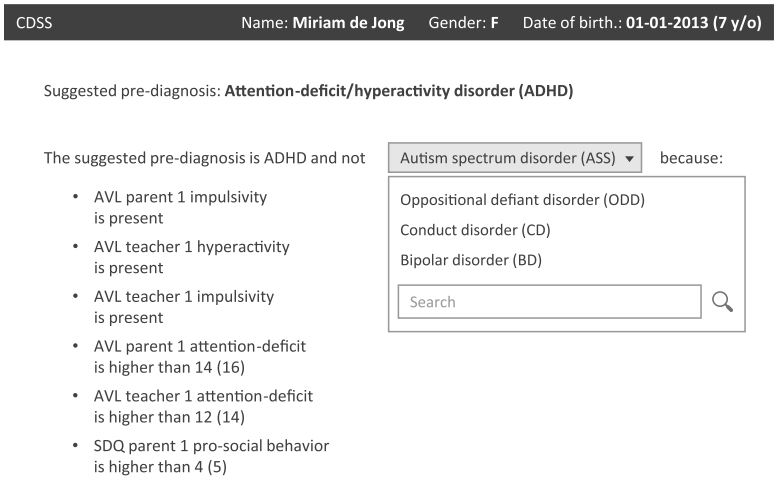
Rationale	The design was based on common feature importance visualizations from the XAI literature (e.g., Poulin et al., 2006; Ribeiro et al., 2016; Strumbelj and Kononenko, 2010). Inspired by Poulin and colleagues (Poulin et al., 2006), we placed the total supporting and contradicting information at the top to allow for easy visual comparison of the difference between the two. The individual features comprising these total bars are presented underneath. We chose to visualize positively and negatively contributing features in the same direction and distinguish them by color (as in e.g., Ribeiro et al., 2016), as opposed to showing them in opposite directions (as in e.g., Strumbelj and Kononenko, 2010; Wang et al., 2019, because that makes it easier to compare the sizes of the positive and negative feature bars.

Moreover, the concept for explanations that was developed in the domain analysis can be presented in order to identify additional requirements for system explanations.

2.1.3. *Multi-modal interaction design and evaluation*

The requirements analysis provides insight into the kind of information that users want to receive in explanations from the system. However, it does not illuminate how this information can effectively be presented in an explanation. The goal of this part of the DoReMi-

Table 5
DP 7: Contrastive explanation and thresholds.

Problem description	The user needs to know: - The reason why this classification is correct, and not another one - From what value of feature X the classification would have been different
UI design example	
	Problem description: The user needs to know: - The reason why this classification is correct, and not another one - From what value of feature X the classification would have been different
	UI design example: 
	Rationale: The design was based on [60]. It allows the user to select an alternative diagnosis to compare the hypothesized diagnosis to, and gives a rule-based explanation of why the system suggested the one and not the other. Similar rule-based explanations, but without the comparison to a specific class, can be found for example in [61] and [43]. For rules that contain a threshold, we added the feature value in parentheses so that users can assess by how much the threshold was exceeded.
Rationale	The design was based on (Waa et al., 2018). It allows the user to select an alternative diagnosis to compare the hypothesized diagnosis to, and gives a rule-based explanation of why the system suggested the one and not the other. Similar rule-based explanations, but without the comparison to a specific class, can be found for example in Guidotti et al. (2018a) and Wang et al. (2019). For rules that contain a threshold, we added the feature value in parentheses so that users can assess by how much the threshold was exceeded.

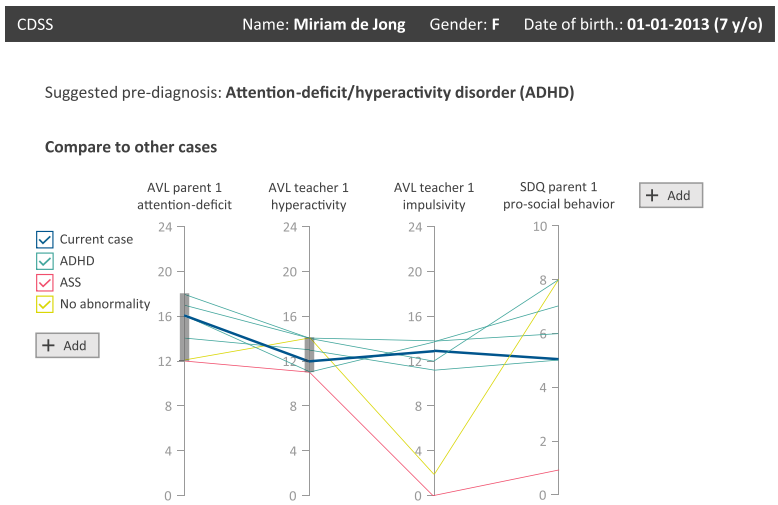
approach is to discover how the information generated by an XAI system can be effectively communicated to the user. This involves choosing appropriate modalities to present the information (typically a multimodal combination of visual and textual content (Holzinger et al., 2021), and creating mock-up interfaces. Although there are multiple examples of interface prototypes for explanations (e.g., Berner and La Lande, 2007; Cai et al., 2019; Pu and Chen, 2007; Wang et al., 2019, there are no general XAI practices concerning the visual or textual presentation of explanations. Therefore, DoReMi stimulates the development of *generalizable* design solutions for explanations by explicitly linking the obtained explanation requirements to *interaction design patterns* for XAI. Relevant questions in this part are: *What existing XAI methods are suitable to obtain the required information for the explanations from the system?, What existing interface design methods are suitable to present the information?, What generic design patterns can be derived from the explanation design problem (i.e., the requirements) and the proposed solution?*, and *Are the proposed design patterns able to increase appropriate use of the system by facilitating understanding and trust in the users?*.

Design patterns are developed for the generation, sharing, use and evolution of design knowledge. They describe the core of a solution to a generic or recurring design problem, which can be reused for the concerning type of design problems (Alexander, 1977; Van Welie et al., 2001). In other words, a design pattern is a structured description of an invariant solution to a recurrent problem within a context. When considering the UI design problem of communicating information generated by an XAI system to a user, a UI design pattern is then a solution to the problem 'the user needs to know X', where X can be any type of information generated by an XAI system (e.g., the information that supports the classification, or how certain the system is of the classification). An XAI design pattern description should include: (1) a problem description (i.e., the user needs to know X); (2) a UI design example that illustrates the solution; and (3) a rationale for why this is a good solution to the problem. These design patterns serve as building blocks that anyone designing a UI for an XAI system can adapt and reuse. Because design patterns are solutions to fine-grained UI design problems, designers can pick and choose the patterns that are most relevant for their context of use.

Design patterns can be generated by a pattern engineering process (Neerincx et al., 2016). First, key design problems are being identified for which, subsequently, existing patterns are searched for. If no pattern can be found, a new Proto Pattern description is created, instantiated and tested. The UI design examples that are constructed should be

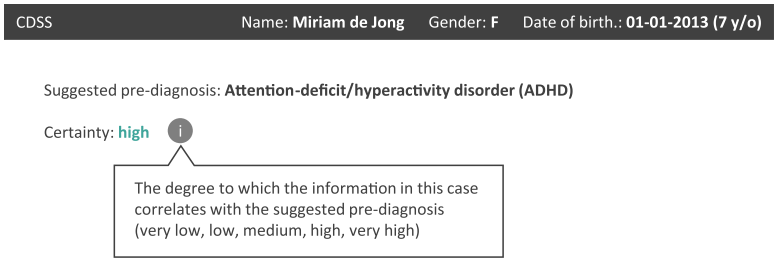
Table 6
DP 10: Comparison to other cases.

Problem description	The user needs to know: - The classification in similar cases - How this case relates to a specific similar case - The most different cases with the same classification
UI design example	

Problem description	The user needs to know: - The classification in similar cases - How this case relates to a specific similar case - The most different cases with the same classification
UI design example	
Rationale	This design uses the parallel coordinates technique [62], which is a common way to visualize high-dimensional data, to help users determine the similarity and differences of the current case with that of other relevant cases. It shows the feature values of the current case (thick blue line) and of a number of similar cases and their diagnoses (other lines) across four features. Initially, only the most relevant features are shown based on the feature importances determined by the system, but users can add more features or rearrange them if they want. Users can also deselect or add certain diagnoses. The grey rectangle on an axis indicates the range of a feature. Cases that fall outside the range of one or more features are not shown. Initially, the system determines what are similar cases by setting the ranges for all features. The user can then choose to widen or narrow this definition of similarity by adjusting the ranges on the feature axes. If an axis does not have a grey rectangle, it means that its full range is used. Users can draw a rectangle on the axis to adjust the range. This design is very similar to existing parallel coordinates implementations. For an interactive example see the Plotly website [63]. Although parallel coordinates are a common data visualization technique, to the best of our knowledge they have not been used as part of explanations in the XAI literature. In [64], a different visualization technique was developed for a similar purpose, but we found it less intuitive than parallel coordinates. Furthermore, it does not allow users to adjust the systems definition of similarity.

Rationale	This design uses the parallel coordinates technique (Inselberg, 1997), which is a common way to visualize high-dimensional data, to help users determine the similarity and differences of the current case with that of other relevant cases. It shows the feature values of the current case (thick blue line) and of a number of similar cases and their diagnoses (other lines) across four features. Initially, only the most relevant features are shown based on the feature importances determined by the system, but users can add more features or rearrange them if they want. Users can also deselect or add certain diagnoses. The grey rectangle on an axis indicates the range of a feature. Cases that fall outside the range of one or more features are not shown. Initially, the system determines what are 'similar cases' by setting the ranges for all features. The user can then choose to widen or narrow this definition of similarity by adjusting the ranges on the feature axes. If an axis does not have a grey rectangle, it means that its full range is used. Users can draw a rectangle on the axis to adjust the range. This design is very similar to existing parallel coordinates implementations. For an interactive example see the Plotly website (Parmer et al., 2020). Although parallel coordinates are a common data visualization technique, to the best of our knowledge they have not been used as part of explanations in the XAI literature. In (Lamy et al., 2019), a different visualization technique was developed for a similar purpose, but we found it less intuitive than parallel coordinates. Furthermore, it does not allow users to adjust the system's definition of similarity.

Table 7
DP 3: Certainty.

Problem description	The user needs to know: - How certain the system is of this classification
UI design example	
	Problem description: The user needs to know: - How certain the system is of this classification
	UI design example: 
	Rationale: We initially based our design on the Intuitive Confidence Measure [65], and defined certainty as the probability that a classification is correct according to the system. However, we found that his definition can be confusing, especially when the certainty is not high. For example, users could mistake a 50% certainty as just a random guess, and erroneously assume that a certainty for a diagnosis that is below 50% means that another diagnosis is more likely. This can cause confusion when interpreting the suggested diagnosis of the system. We decided to adopt a correlation-based definition instead, and to limit the certainty levels to a five-point scale using natural language. This makes it more analogous to a human saying, for example, 'My best guess is ADHD, but I'm not quite sure'.
Rationale	We initially based our design on the Intuitive Confidence Measure (van der Waa et al., 2018), and defined certainty as the probability that a classification is correct according to the system. However, we found that his definition can be confusing, especially when the certainty is not high. For example, users could mistake a 50% certainty as just a random guess, and erroneously assume that a certainty for a diagnosis that is below 50% means that another diagnosis is more likely. This can cause confusion when interpreting the suggested diagnosis of the system. We decided to adopt a correlation-based definition instead, and to limit the certainty levels to a five-point scale using natural language. This makes it more analogous to a human saying, for example, 'My best guess is ADHD, but I'm not quite sure'.

evaluated with end users in terms of understandability and usefulness, for example by having them perform a (simplified) typical task while using the system's explanations (e.g., Doshi-Velez and Kim, 2017; Wolf, 2019). The quality of the explanations and the explanation interface can for example be measured by using the System Causability Scale (Holzinger et al., 2020), which contains items about information completeness, level of detail, understandability, and causality. It might also be useful to consult general XAI design guidelines (e.g. Amershi et al., 2019; Eiband et al., 2018) in order to identify the strengths and potential weaknesses of the patterns. The outcomes of the evaluation further refine the explanation requirements that were obtained in the requirements assessment. The new design patterns are then added to the concerning library.

2.2. Explanation framework

Fig. 2 presents a high-level framework for explanation generation and communication by a system that uses a machine learning model. The system is trained to make predictions (e.g., diagnoses) based on a particular kind of input data (e.g., medical data). When presented with a new case, the system calculates the predicted output. The user has access to the same input data, and is presented with the prediction of the machine learning process. An XAI method makes use of the inputs and outputs in order to explain the prediction to a user. Such methods are able to extract relevant information (e.g., features and feature importances) that is used by the system to calculate an output (see Adadi and Berrada, 2018 for an overview of recent XAI methods for machine learning algorithms). While this information could provide a user with some insight into the reasoning process of the system, presenting it directly typically does not enable a user to understand *why* a particular output was obtained. Therefore, outputs will have to be explained in order to enable users to understand the system's rationale. Providing a rationale enables a user to obtain a sense of the trustworthiness of a particular advice from the system. In addition, the experience that a user gains with explanations of outputs for many types of patient cases over time enables a user to develop a level of trust that is adequately calibrated, meaning that the user's perception of the trustworthiness of the system is aligned with its actual trustworthiness (de Visser et al., 2019). In turn, this allows for adequate use of the system (Hoffman et al., 2018). Thus, the information that is extracted by the XAI method should be presented in a way that enables users to better understand the system and develop calibrated trust. The DoReMi-approach can be used to find out how the content (i.e., the information) and the form (i.e., the presentation) of the explanation must be adapted to the context of use, in such a way that the resulting explanation is understandable and useful to the user. This also includes the design of interactive elements in the explanation that enables users to obtain answers to additional questions they might have (e.g., to find out why the system's output is A *instead of B*).

3. Domain analysis of clinical decision-making

The main goal of the domain analysis is to determine if and why explanations are required within the context of use, and what

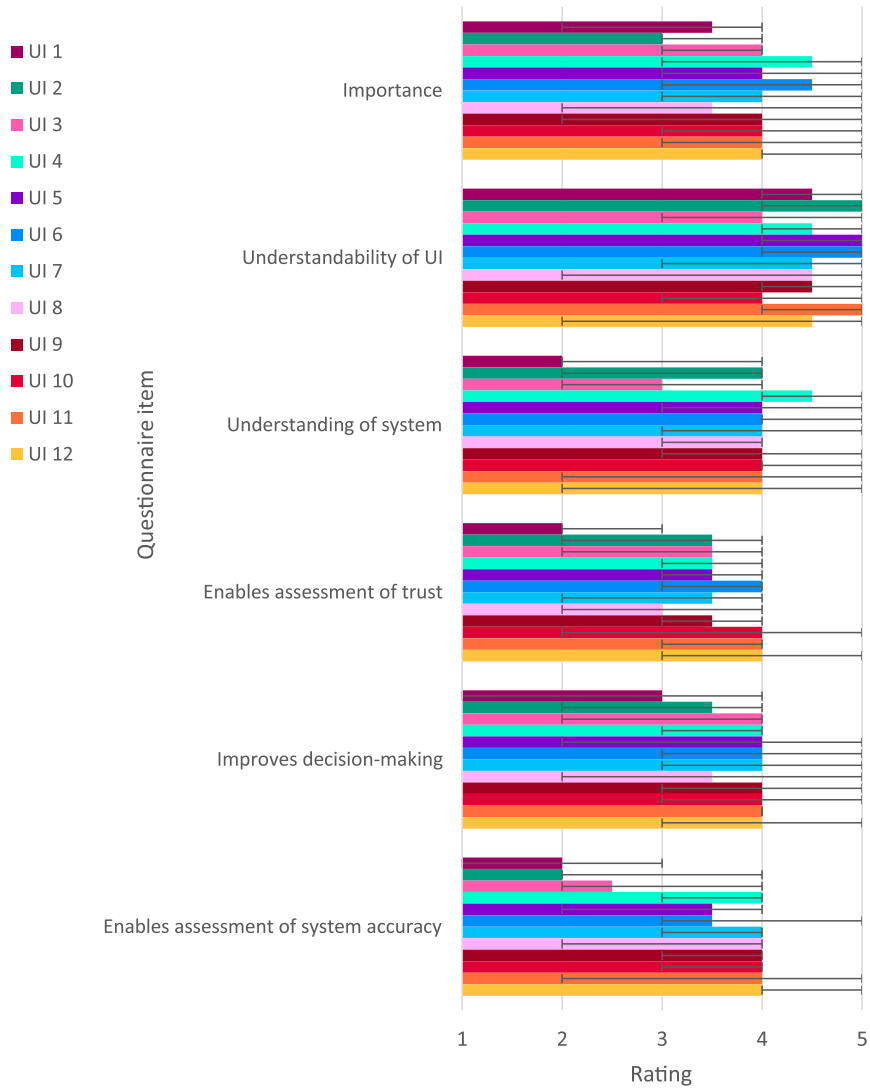


Fig. 5. Median Likert-scale values on the six questionnaire items for each user interface. Error bars represent minimum and maximum values.

information is potentially relevant for explanations of the system. We describe the context of clinical decision-making and decision-support systems within the domain of healthcare, which we investigated by performing a literature study on these topics. Moreover, we provide a list of information elements that we identified in the literature and could be relevant for explanations of diagnoses.

3.1. Context description

The purpose of decision-support systems is to enable users to make better informed decisions by collecting, analyzing, structuring, and presenting information relevant to the decision-making process. By facilitating clinical decision-making of a clinician at the point in time that these decisions are made, clinical decision-support systems should reduce medical errors and improve patient care (Berner and La Lande, 2007; Kawamoto et al., 2005). By using machine learning to find statistical patterns in large amounts of patient data, such systems can also provide accurate advice regarding the diagnosis of individual patients (Friedman et al., 1999; Hunt et al., 1998; Ozaydin et al., 2016). A key requirement for the successful adoption of CDSSs into clinical practice is that such systems can explain their advised diagnoses to clinicians, as this allows clinicians to understand the system, and to build confidence in the performance of the system (e.g., Guida et al., 1997; Ye and Johnson, 1995). Moreover, the system could support clinicians in their communication to a patient by producing explanations that are tailored to the patient. So far, clinicians have generally shown low acceptance of CDSSs, which is mainly attributed to the inability of such systems to provide understandable and meaningful explanations (Berner and La Lande, 2007; Holst et al., 2000; Kawamoto et al., 2005). As a result, clinicians do not develop an adequate level of trust in the system (i.e., calibrated trust (de Visser et al., 2019)).

3.2. Concept for explanations

We consulted the literature on explanations within the medical domain in order to identify the typical information that is mentioned in explanations of medical diagnoses (e.g., Bussone et al., 2015; Kononenko, 2001; Wang et al., 2019; Xie et al., 2019. It is important to realize that the diagnostic process is context-dependent and involves analysis of large amounts of data. Clinicians often cross-validate relevant patient information, in order to maintain high sensitivity (i.e., correctly diagnosing patients with their disease) and specificity (i.e., correctly diagnosing healthy patients as having no disease) in their decision-making. Therefore, clinicians can be supported by information that helps them to remain prudent in their decision-making, such as supportive- and counter-evidence, a comparison of the current case with similar cases, or evidence for potential differential diagnoses. Table 1 shows the 20 information elements that we abstracted from the literature, and which

Table A8
DP 1: Class information.

Problem description	The user needs to know: - A description of the class - The prevalence of the class
UI design example	

Problem description	The user needs to know: - A description of the class - The prevalence of the class
UI design example	

CDSS Name: **Miriam de Jong** Gender: **F** Date of birth.: **01-01-2013 (7 y/o)**

Suggested pre-diagnosis: **Attention-deficit/hyperactivity disorder (ADHD)**

Disorder
Attention-deficit/hyperactivity disorder (ADHD) is a disorder marked by an ongoing pattern of inattention and/or hyperactivity-impulsivity that interferes with functioning or development.

Prevalence
The prevalence of ADHD in children (below the age of 18) in the Netherlands is estimated at 2.9%. 75% of children with ADHD are male and 25% are female. The prevalence of ADHD in adults is estimated at 2.1%.

Table A9
DP 2: Available/relevant information.

Problem description	The user needs to know: - The information that is used to make the classification - The information that is relevant in making this type of classification
UI design example	

Problem description	The user needs to know: - The information that is used to make the classification - The information that is relevant in making this type of classification
UI design example	

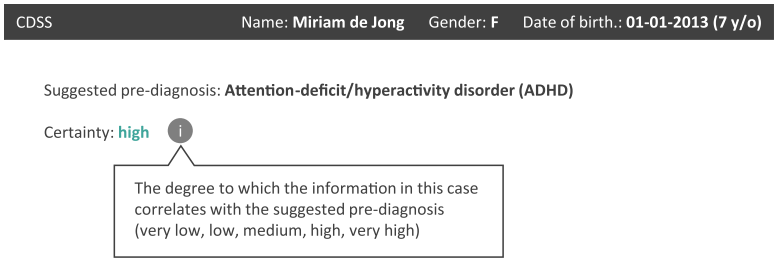
CDSS Name: **Miriam de Jong** Gender: **F** Date of birth.: **01-01-2013 (7 y/o)**

Suggested pre-diagnosis: **Attention-deficit/hyperactivity disorder (ADHD)**

Available information in this case

Information	Available
Patient demographics	
Gender	✓
Age	✓
AVL	
AVL parent 1	✓
AVL parent 2	✗
AVL teacher	✓
AVL other	✗
SDQ	
SDQ parent 1	✓
SDQ parent 2	✗
SDQ teacher	✗
SDQ other	✗

Table A10
DP 3: Certainty.

Problem description	The user needs to know: - How certain the system is of this classification
UI design example	
	Problem description: The user needs to know: - How certain the system is of this classification
	UI design example: 
	Rationale: We initially based our design on the Intuitive Confidence Measure [65], and defined certainty as the probability that a classification is correct according to the system. However, we found that his definition can be confusing, especially when the certainty is not high. For example, users could mistake a 50% certainty as just a random guess, and erroneously assume that a certainty for a diagnosis that is below 50% means that another diagnosis is more likely. This can cause confusion when interpreting the suggested diagnosis of the system. We decided to adopt a correlation-based definition instead, and to limit the certainty levels to a five-point scale using natural language. This makes it more analogous to a human saying, for example, 'My best guess is ADHD, but I'm not quite sure'.
Rationale	We initially based our design on the Intuitive Confidence Measure (van der Waa et al., 2018), and defined certainty as the probability that a classification is correct according to the system. However, we found that his definition can be confusing, especially when the certainty is not high. For example, users could mistake a 50% certainty as just a random guess, and erroneously assume that a certainty for a diagnosis that is below 50% means that another diagnosis is more likely. This can cause confusion when interpreting the suggested diagnosis of the system. We decided to adopt a correlation-based definition instead, and to limit the certainty levels to a five-point scale using natural language. This makes it more analogous to a human saying, for example, 'My best guess is ADHD, but I'm not quite sure'.

could potentially support clinicians in evaluating hypotheses concerning clinical diagnoses. We categorized these information elements based on characteristics of explanations that have been identified in the social sciences (Hilton, 1990; Miller, 2018): information that is contrastive (element 7 and 8), counterfactual (9, 10), example/case-based (11–14, 19), or involves certainty (15 and 16). Moreover, element 5 and 6 refer to supportive and counter-evidence respectively, element 17 and 18 refer to the input data that are used by the system, and element 1–4 consist of descriptive information about the diagnosis. Element 20 concerns the performance history of the system. The choice of categories was validated with a number of XAI-experts (colleagues), who confirmed that the categories corresponded to those typically used in XAI, and that the information elements were categorized sensibly.

4. Requirements elicitation and assessment for explanations from a CDSS

The goal in this part of the DoReMi-approach is to determine what kind of explanations the CDSS should be able to provide to clinicians, and to identify potential contextual dependencies for explanations by CDSSs. To achieve this goal, we conducted a user study among paediatricians that consists of two parts. We first developed a questionnaire that contained the information elements that we identified in the domain analysis. In collaboration with an experienced clinician, we also constructed a realistic use case of a child patient. In the first part of the user study, we investigated the contextual dependencies of explanations, by presenting clinicians with the use case, and asking them to indicate the importance of each information element *in their own explanations* within multiple hypothetical social situations. In the second part of the study, we investigated requirements for the explanations from the system by sequentially presenting clinicians with two hypothetical system outputs based on the use case (i.e., a positive and negative diagnosis), and by asking them to indicate what questions they want to ask the system.

4.1. Methods

4.1.1. Participants

A total of six paediatricians (all native Dutch speakers) were contacted via an e-mail in which we generally introduced our research context and requested their participation in our study. All clinicians were experienced in diagnostic decision-making (M = 18.6 years working experience as clinician, SD = 9.8 years). The paediatricians were working at different health centres in the same region in the Netherlands. Overall, participants indicated that they were slightly familiar with artificial intelligence in general, and not at all familiar with clinical decision support systems. Clinicians received a small monetary reward as token of appreciation to fill in a questionnaire, which on average took 50 minutes to complete.

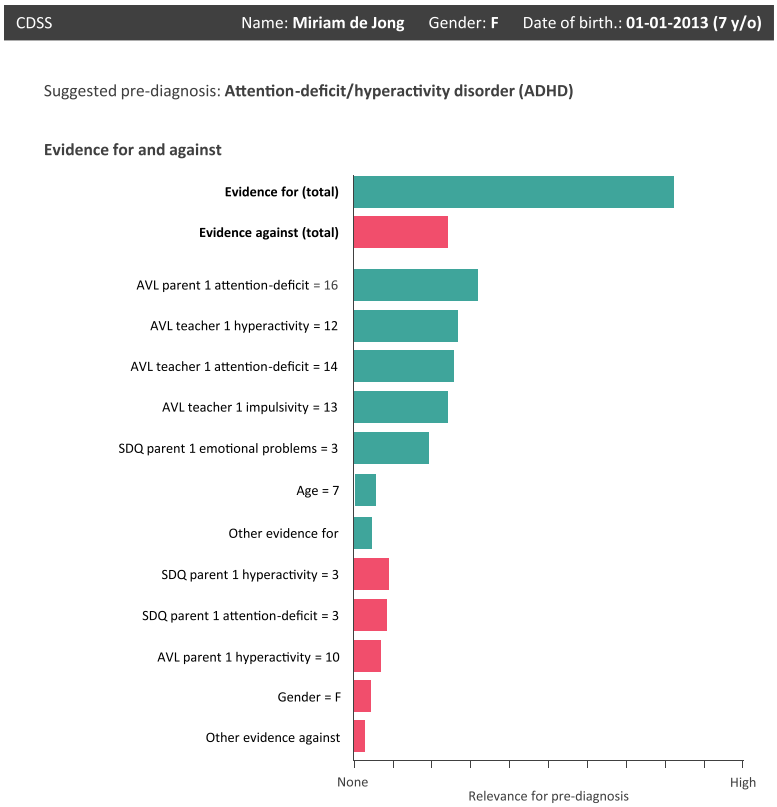
4.1.2. Materials

Use case User requirements are best discovered when presenting users with a particular context for which they can express their needs (Doshi-Velez and Kim, 2017; Maguire and Bevan, 2002). Therefore, we

Table A11
DP 4: Supporting/contradicting information.

Problem description	The user needs to know: - The information that supports this classification - The information that contradicts this classification
UI design example	

Problem description	The user needs to know: - The information that supports this classification - The information that contradicts this classification
UI design example	



Rationale	The design was based on common feature importance visualizations from the XAI literature (e.g., [58, 3, 59]). Inspired by Poulin and colleagues [58], we placed the total supporting and contradicting information at the top to allow for easy visual comparison of the difference between the two. The individual features comprising these total bars are presented underneath. We chose to visualize positively and negatively contributing features in the same direction and distinguish them by color (as in e.g., [3]), as opposed to showing them in opposite directions (as in e.g., [59, 43], because that makes it easier to compare the sizes of the positive and negative feature bars.

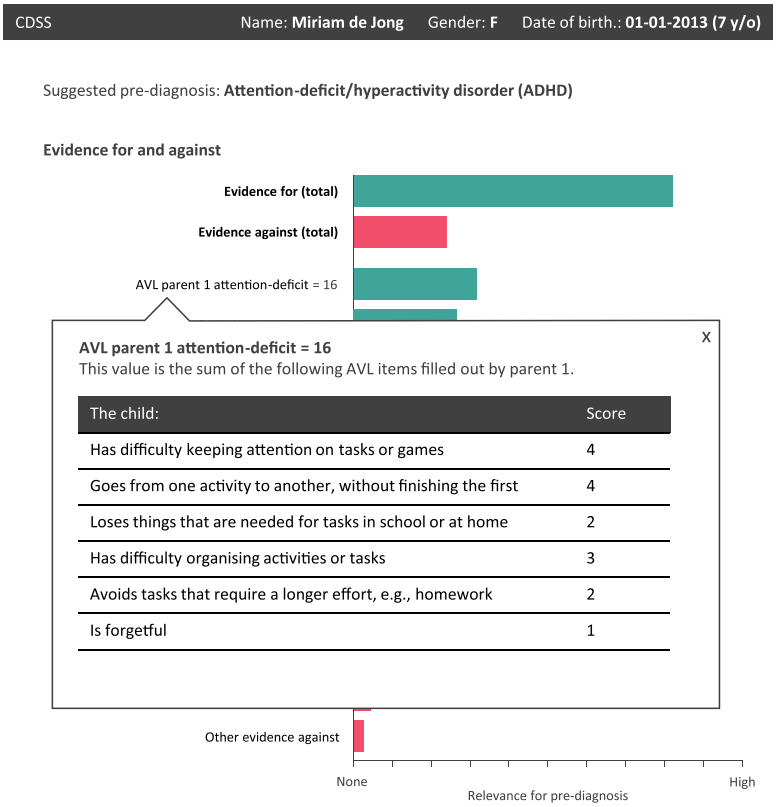
Rationale	The design was based on common feature importance visualizations from the XAI literature (e.g., Poulin et al., 2006; Ribeiro et al., 2016; Strumbelj and Kononenko, 2010). Inspired by Poulin and colleagues (Poulin et al., 2006), we placed the total supporting and contradicting information at the top to allow for easy visual comparison of the difference between the two. The individual features comprising these total bars are presented underneath. We chose to visualize positively and negatively contributing features in the same direction and distinguish them by color (as in e.g., Ribeiro et al., 2016), as opposed to showing them in opposite directions (as in e.g., Strumbelj and Kononenko, 2010; Wang et al., 2019, because that makes it easier to compare the sizes of the positive and negative feature bars.

collaborated with an experienced paediatrician in order to develop a fictitious use case that contains sufficient detail that allows making and explaining a diagnosis. The fictitious patient in the use case is a 7-year old girl called Miriam, who shows symptoms of Attention Deficit/-Hyperactivity Disorder (ADHD). Although paediatricians do not make a definitive diagnosis, they do engage in a diagnostic process in which they identify potential problems and indicate the medical diagnosis that is likely (i.e., a pre-diagnosis). Based on this pre-diagnosis, they can refer a child to a specialist such as a mental healthcare professional. The use case was carefully constructed to ensure a sufficient level of realism and

Table A12
DP 5: Feature value origin.

Problem description	The user needs to know how it was determined that feature X was or was not present.
UI design example	

Problem description	The user needs to know how it was determined that feature X was or was not present.
UI design example	



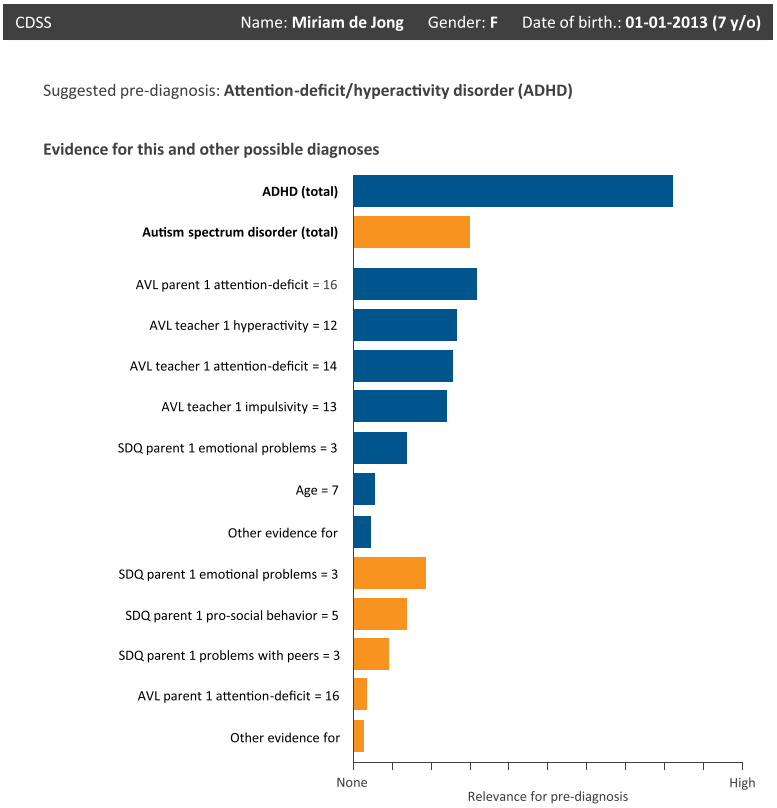
adequate alignment with current work processes. Although the case points to ADHD as primary diagnosis, it also leaves some room for discussion. The use case consists of four datasources: one short textual report, and three completed Likert-scale questionnaires. The report consist of a description of a fictitious conversation with Miriam and her mother in which relevant observations are mentioned (e.g., 'Miriam never finishes her work'). In addition to this textual report, we also presented filled-in versions of two types of questionnaires that are typically administered when a behavior disorder is suspected in a child: The Strengths and Difficulties Questionnaire (SDQ), which is a brief survey aimed at capturing emotional and behavioral problems in children, and the ADHD Questionnaire (AVL), which is used to identify behavioral symptoms of ADHD. The SDQ contains 25 items which are scored on a scale of 1 (not true) to 3 (true), for example: 'The child takes into account the feelings of others'. The AVL contains 18 items that are scored on a scale of 1 (never) to 5 (very often), for example: 'The child is easily distracted'. Our use case contains one SDQ, filled in by the mother of Miriam, and two AVLs that are filled in by the mother and the teacher.

Social contexts Table 2 lists the seven social conditions in which we investigated information requirements for explanations. We came up with these social situations by considering the outcomes of the domain analysis. We verified the validity of our identified situations with the same expert that helped us to develop the use case. In situation 1a-1c, clinicians explained their diagnosis to a colleague who is either impartial, agreeing (i.e., also diagnosing ADHD), or disagreeing (i.e., no abnormality) with the diagnosis of the clinician. In situation 2a and 2b, the mother of the child-patient requests an explanation of the diagnosis with which she either agrees or disagrees. In our use case, we chose to include the parent of the patient, because diagnoses of child-patients are often communicated to parents or caretakers, and not directly to the patient itself. In situation 3a and 3b, the clinician assumed the role of explainee and indicated what information they find important to receive from a CDSS in case it provides a diagnosis that is either congruent (3a, ADHD) or incongruent (3b, no abnormality) with the diagnosis of the clinician.

Questionnaire and interview We chose to use a questionnaire followed by a short, unstructured interview in order to investigate what information clinicians would like to receive from a decision support system, and how the content of explanations depends upon the social context in which they are provided. To mitigate the risk of overlooking relevant information, we adopted both open and closed questions to enable clinicians to indicate their information preferences. While there are numerous methods to elicit requirements (e.g., interviews or focus groups), we chose to use a questionnaire as the main method. The main reason for this choice is that a questionnaire enables structured collection and comparison of answers. Based on the literature, we already developed a fairly extensive overview of concrete information elements that are potentially relevant within the context of clinical decision-

Table A13
DP 6: Alternative classifications.

Problem description	The user needs to know other classifications that are conceivable based on the case information.
UI design example	
	Problem description: The user needs to know other classifications that are conceivable based on the case information.
	UI design example:

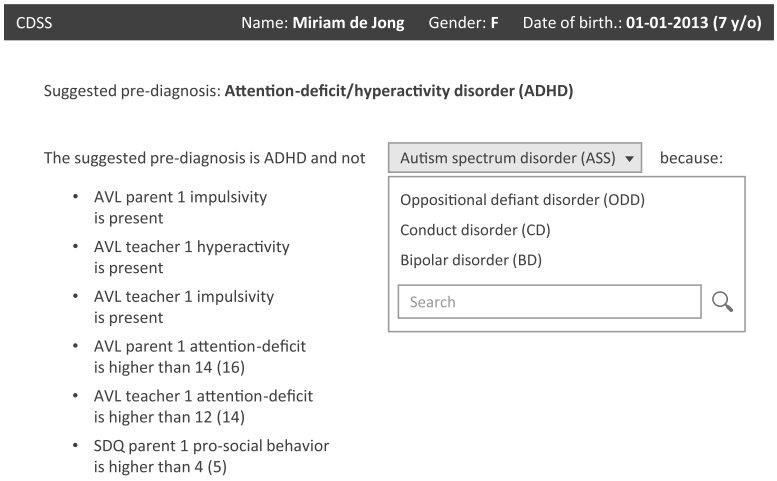


support. The goal in this stage of DoReMi is to verify and validate whether these elements are indeed relevant for clinicians, and to understand why. A questionnaire is a structured way of quantifying the importance of each element for different kind of contexts. Moreover, this approach allows comparison of answers across domains. In order to improve our understanding of why certain information is relevant, we also had a short unstructured interview session with the group of clinicians, in which we asked about their attitude towards AI-support of medical diagnosis, whether they wanted to elaborate on some of their answers, and whether they had any additional remarks.

The questionnaire consisted of two parts. In the first part, we asked clinicians to make a diagnosis based on the use case, and to indicate what information they have used in their decision-making. Then, we sequentially introduced participants to the social situations, for which they were asked to indicate the importance of the information elements from Table 1 that they could use in their *own* explanation. We created brief introductions for each social situation. Situations 1a-2b each started with a description of the explainee (e.g., whether this is a colleague or the mother of the patient, and whether they agree or disagree with the diagnosis). We also included a short explanation of the diagnosis by the colleague, in which information was mentioned that was taken into account in the decision-making by the colleague (e.g., 'Miriam is easily distracted and quickly loses the main thread during a conversation... Although she shows no signs of over-activity, I think there are sufficient indicators of ADHD.'). The importance of each information element in each situation was rated on a five-point Likert-scale with the following levels: (1) not at all important, (2) not important, (3) somewhat important, (4) important, and (5) very important. We also included an open question in each situation, which asks clinicians to indicate any additional information elements that they can think of.

In the second part, we introduced clinicians with a fictitious clinical decision support system. We wrote an introduction including a general description of its purpose and operation (i.e., finding likely diagnoses for a particular case, by running the case through a model that was constructed by learning from large amounts of patient data), its potential supportive functionalities, its limitations caused by its dependence on registered data, and its performance (i.e., the system works well in many, but not all cases). Situation 3a and 3b consisted of the system suggesting the diagnosis ADHD or no abnormality respectively, for which no explanation is provided. In both situations, we presented the list with information elements, including the two system-specific elements (19 and 20). We rewrote each element into a question (e.g., 'How certain is the system of this diagnosis?'), in order to enable clinicians to indicate how important they think it is to be able to ask each question to the system within each situation. Additionally, we included an open question about whether clinicians suspect changes in the importance ratings when imagining that they have been working with the system for

Table A14
DP 7: Contrastive explanation and thresholds.

Problem description	The user needs to know: - The reason why this classification is correct, and not another one - From what value of feature X the classification would have been different
UI design example	
	Problem description: The user needs to know: - The reason why this classification is correct, and not another one - From what value of feature X the classification would have been different
	UI design example: 
	Rationale: The design was based on [60]. It allows the user to select an alternative diagnosis to compare the hypothesized diagnosis to, and gives a rule-based explanation of why the system suggested the one and not the other. Similar rule-based explanations, but without the comparison to a specific class, can be found for example in [61] and [43]. For rules that contain a threshold, we added the feature value in parentheses so that users can assess by how much the threshold was exceeded.
Rationale	The design was based on (Waa et al., 2018). It allows the user to select an alternative diagnosis to compare the hypothesized diagnosis to, and gives a rule-based explanation of why the system suggested the one and not the other. Similar rule-based explanations, but without the comparison to a specific class, can be found for example in Guidotti et al. (2018a) and (Wang et al., 2019). For rules that contain a threshold, we added the feature value in parentheses so that users can assess by how much the threshold was exceeded.

some time and have learned that most of its advice is correct.

4.1.3. Procedure

We invited the paediatricians to a central health centre, where we booked a conference room to conduct the study. The first author was present to provide a short verbal introduction to the research, and to answer any questions during the completion of the questionnaire. Although participants were present in the same room, no communication was allowed between them. Prior to filling out the questionnaire, participants were presented with a general introduction of the study, in which it was made clear how their data would be used, and that the study was evaluated on ethics and quality by an internal review committee at our research institute. After signing the informed consent, we presented the questionnaire, followed by the short interview. Afterwards, paediatricians were thanked for their participation and received their reward.

4.1.4. Data analysis

Because of the exploratory nature of our study, we did not perform inferential statistical analysis on the data. Instead, we investigated the quantitative and qualitative data from the clinicians in order to find interesting trends. In addition, we determined the median importance rating for each information element, and the minima and maxima in order to obtain an indication of between-subjects variability. Unfortunately, it appeared that one participant did not fill out six questionnaire items, and another participant did not fill out four items, leading to ten information elements with four ratings instead of five. The median importance ratings and their variability were compared between the information elements and the situations. The first two authors analyzed the qualitative data to aid the interpretation of these comparisons. We used an open coding scheme in which we independently evaluated the answers and wrote down any interesting remarks. We then discussed our findings.

4.2. Results

4.2.1. Part 1: providing explanations in different social contexts

All six clinicians diagnosed Miriam with ADHD, indicating that the information in the use case provides sufficient reason to designate this diagnosis as being most likely. In their explanations, all participants mention similar factors that are indicative of this diagnosis (e.g., low attention span, restless, learning difficulties). Some also mention more

Table A15
DP 8: Counterfactuals w.r.t. classification.

Problem description	The user needs to know the classification that is likely if feature X had not been A but B.
UI design example	

Problem description	The user needs to know the classification that is likely if feature X had not been A but B.
UI design example	

general characteristics that they took into account, such as the perceived impact of the behavior of Miriam at school and at home, and the stable relationship with her mother.

Fig. 3 shows the median, minimum, and maximum importance ratings of the 18 information elements in all five social situations. In general, median importance ratings were high for all information elements in all social situations. However, some scores also varied heavily across participants, as depicted by the error bars in Fig. 3. There are some elements that clinicians unanimously rated as (very) important (scores of 4 or 5) in at least one situation, which are elements 1 (in all situations), 2 (only in both disagreeing situations), 3 (only in explanations to the parent of the patient), 5 (in all situations, except in case of the disagreeing colleague), 7 (only when informing a colleague, or in case of an agreeing colleague), and 15 (only in case of a disagreeing parent of the patient). There are also elements that unanimously received relatively lower values (maximum of 3) for some situations: element 3 and 4 (only when informing a colleague, or when the colleague agrees), 10 (only in case a colleague agrees), 12 (only when explaining to a parent of the patient), and 14 (only in case the parent agrees). Note that for some elements, the majority of clinicians provided low ratings (median values lower than 3), while only one clinician provided a higher rating (as indicated by the maximum value). Inspection of the data revealed that it was not a single clinician who consequently gave either high or low ratings, which indicated that participants are rather pronounced about the importance of items.

For situation 1c (in which a colleague-clinician disagrees with the diagnosis that is made by the participant), two clinicians also indicated (in the free-text entry box) high importance of mentioning the negative consequences of *not* diagnosing Miriam, given her behavioral problems.

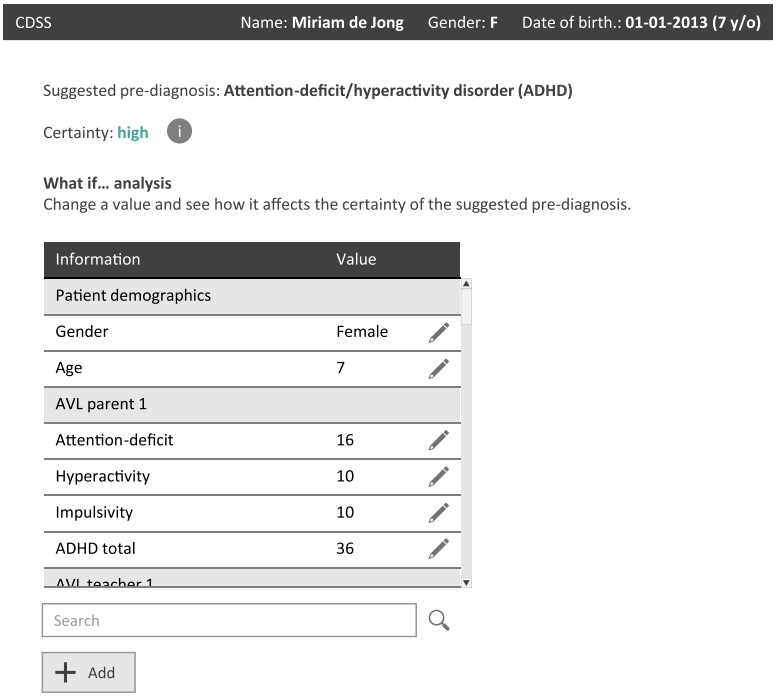
4.2.2. Part 2: receiving explanations from a CDSS

Fig. 4 shows the median, minimum, and maximum importance ratings of the 20 information elements, given the context in which the system provides a diagnosis that is either equal to that of the clinician (i.e., situation 3a), or differs (i.e., situation 3b). Overall, median importance ratings are high, with the exception of elements 3, 4, 9, 13, and 17, which have median values of 3 (somewhat important) or lower in situation 3a and/or 3b. However, every element received a rating of 5 (highly important) from at least one clinician. There are seven elements that most clinicians rated with a 5 and all other clinicians with a 4 (important) in *both* situations. These are elements 1, 5, 6, 7, 15, 16, and 20, all of which have median values of 5. There are two elements that received lower ratings from all clinicians (i.e., maximum value of 3 or lower), which are element 4 (but only in case the clinician does *not* agree with the diagnosis of the system), and element 13 (but only when the clinician *does* agree with the system). For other elements, the minimum and maximum ratings are more varied between individuals.

Apart from the importance ratings, clinicians also indicated whether they still would want to receive the same information in an explanation of the CDSS, when imagining that they have been working for some time with the system. Unanimously, clinicians predict this information to be different (i.e., they all responded with 'No'). They all provided the reason that they are likely to get familiar with the decision-making of the system over time, thereby requiring less extensive explanations. Moreover, they stated that some information is likely to remain important over time (supporting- and counter evidence, and certainty of the diagnosis), while other information is likely to become less important (what information is used to make this diagnosis, and the diagnosis in similar cases).

Table A16
DP 9: Counterfactuals w.r.t. certainty.

Problem description	The user needs to know the information that would increase the certainty of the classification.
UI design example	

Problem description	The user needs to know the information that would increase the certainty of the classification.
UI design example	

4.2.3. Interview with clinicians

All six clinicians shared a very positive attitude towards the use of a CDSS to aid their decision-making, saying that they could 'certainly see how such a system could be beneficial in diagnosing' and that they valued the information that was provided by the system. In particular, they valued the opportunity to be able to explore the case in more detail, by being able to compare supporting and contradicting information for multiple diagnoses. Multiple clinicians also explicitly mentioned that the explanations help them to better evaluate the different diagnoses that are conceivable given the patient's complaints.

4.3. Discussion

The goal of our first study was to obtain insight into the kind of information that clinicians find important to be part of an explanation of a diagnosis. We explored whether the information needs differ depending on various social contexts. Furthermore, we explored what explanations clinicians want to receive from a system that supports diagnostic decision-making. We first discuss the results on explanations as provided by clinicians (part 1), and then on explanations from a CDSS (part 2).

4.3.1. Explanations from clinicians

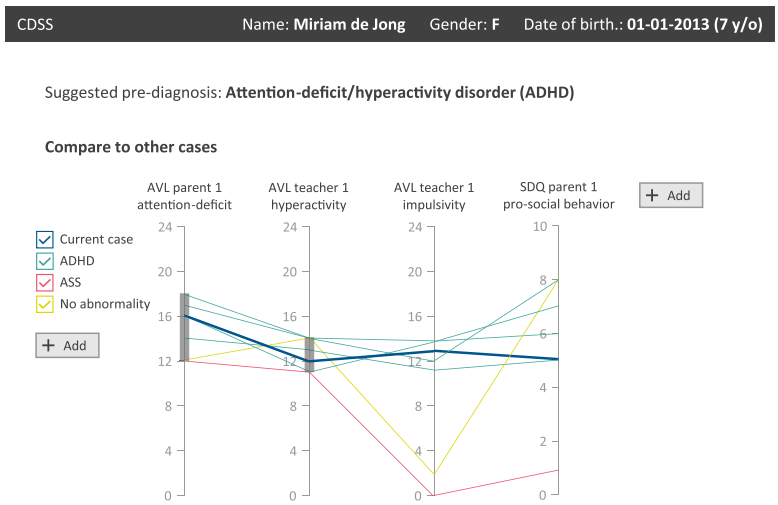
Considering the results as a whole, there are three outcomes that immediately stand out: (1) importance ratings for most information elements are high, (2) there is a relatively large variation in individual scores, while (3) scores for most information elements only show subtle differences between the five situations. Given that we carefully constructed the list of information that is often used in clinical decision-making, it is not surprising that overall, clinicians rate this information as being important or highly important (i.e., 78% of all median values are above 3). The exhaustiveness of the list is also evident by the fact that nearly none of the clinicians mentioned additional information. Across all social situations, clinicians consider it highly important to mention the information that is used to make a diagnosis (for example, data from questionnaires and physiological measurements), and the information that supports the diagnosis (for example, the prosocial-, attention-, and behavioral scale scores on the intake questionnaires) in their explanations. This information can be considered to be the primary evidence for their decision. Interestingly, there is no information that is unanimously evaluated by clinicians as being less important.

While the overall median importance ratings for the information elements are high, there are large differences in some scores between clinicians. This indicates that they have personal preferences concerning the information that they provide in their diagnosis. For example, while one clinician highly values counterfactual explanations (e.g., I think it is ADHD and not ASD, because Miriam is described as social, but is also very easily distracted) across all situations (median of scores is 4 across elements 9 and 10), another clinician attributes much less importance to this information (median of 2). While these differences might be somewhat strengthened by potential contrast effects (Sherif et al., 1958), it still shows that there are individual preferences for particular information elements, which is an interesting result.

We observed only small variations in scores when comparing the importance of information elements across different social situations.

Table A17
DP 10: Comparison to other cases.

Problem description	The user needs to know: - The classification in similar cases - How this case relates to a specific similar case - The most different cases with the same classification
UI design example	

Problem description	The user needs to know: - The classification in similar cases - How this case relates to a specific similar case - The most different cases with the same classification
UI design example	
Rationale	This design uses the parallel coordinates technique [62], which is a common way to visualize high-dimensional data, to help users determine the similarity and differences of the current case with that of other relevant cases. It shows the feature values of the current case (thick blue line) and of a number of similar cases and their diagnoses (other lines) across four features. Initially, only the most relevant features are shown based on the feature importances determined by the system, but users can add more features or rearrange them if they want. Users can also deselect or add certain diagnoses. The grey rectangle on an axis indicates the range of a feature. Cases that fall outside the range of one or more features are not shown. Initially, the system determines what are similar cases by setting the ranges for all features. The user can then choose to widen or narrow this definition of similarity by adjusting the ranges on the feature axes. If an axis does not have a grey rectangle, it means that its full range is used. Users can draw a rectangle on the axis to adjust the range. This design is very similar to existing parallel coordinates implementations. For an interactive example see the Plotly website [63]. Although parallel coordinates are a common data visualization technique, to the best of our knowledge they have not been used as part of explanations in the XAI literature. In [64], a different visualization was developed technique for a similar purpose, but we found it less intuitive than parallel coordinates. Furthermore, it does not allow users to adjust the systems definition of similarity.

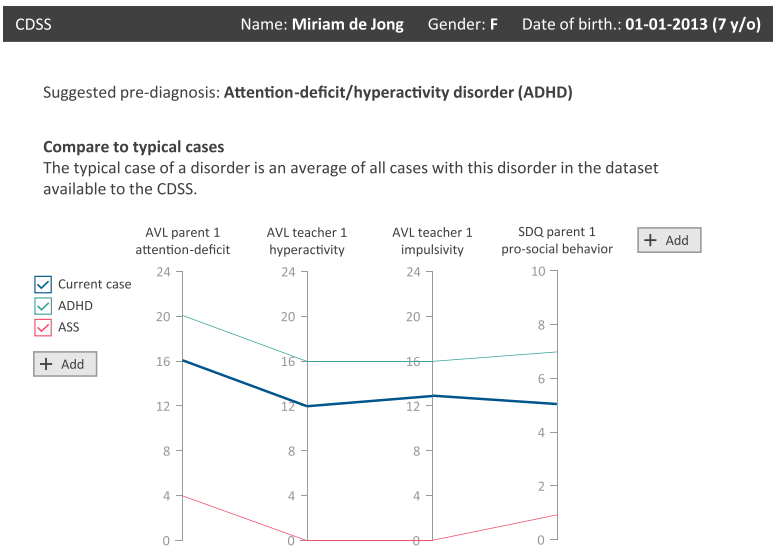
Rationale	This design uses the parallel coordinates technique (Inselberg, 1997), which is a common way to visualize high-dimensional data, to help users determine the similarity and differences of the current case with that of other relevant cases. It shows the feature values of the current case (thick blue line) and of a number of similar cases and their diagnoses (other lines) across four features. Initially, only the most relevant features are shown based on the feature importances determined by the system, but users can add more features or rearrange them if they want. Users can also deselect or add certain diagnoses. The grey rectangle on an axis indicates the range of a feature. Cases that fall outside the range of one or more features are not shown. Initially, the system determines what are 'similar cases' by setting the ranges for all features. The user can then choose to widen or narrow this definition of similarity by adjusting the ranges on the feature axes. If an axis does not have a grey rectangle, it means that its full range is used. Users can draw a rectangle on the axis to adjust the range. This design is very similar to existing parallel coordinates implementations. For an interactive example see the Plotly website (Parmer et al., 2020). Although parallel coordinates are a common data visualization technique, to the best of our knowledge they have not been used as part of explanations in the XAI literature. In (Lamy et al., 2019), a different visualization was developed technique for a similar purpose, but we found it less intuitive than parallel coordinates. Furthermore, it does not allow users to adjust the system's definition of similarity.

Table A18
DP 11: Comparison to typical cases.

Problem description	The user needs to know: - A typical case with this classification - How this case relates to the typical case of this and other classifications
UI design example	

Problem description	The user needs to know: - A typical case with this classification - How this case relates to the typical case of this and other classifications

UI design example



This suggests that, overall, clinicians only subtly adapt their explanations to the social situation in which they need to be used. When we consider the results in more detail, it appears that clinicians find it more important to include general information, such as a description and the prevalence of the condition, in an explanation to a parent of the patient instead of a colleague, while explaining how the patient's case relates to other, similar cases is less important. Furthermore, mentioning alternative diagnoses for the current case is most important when explaining to a colleague.

Mentioning how certain the clinician is of the diagnosis is most important when explaining to a parent of the patient who disagrees with the diagnosis, while comparing its case with that of a typical case is least important when explaining to a parent who agrees. These results are in line with the patient-centered approach of communication within healthcare (Baker, 2001), which is aimed at facilitating understanding of the diagnosis in patients and their relatives, and individualism (i.e., treating patients as individuals instead of one of many that received the same diagnosis).

Lastly, clinicians also find it highly important to mention the relevance of the information for their diagnosis when the explainee (either a parent of the patient *or* a colleague) believes there is no abnormality.

This result indicates the importance of the sensitivity/specificity trade-off in clinical diagnosing (Berner and La Lande, 2007; Kawamoto et al., 2005). That is, while clinicians try to mitigate the amount of false positive diagnoses, they are also careful to avoid false negatives. This is confirmed by the free-text answers of two clinicians, who explicitly state the high importance of mentioning the negative consequences of *not* diagnosing Miriam, given her behavioral problems.

4.3.2. Explanations from a CDSS

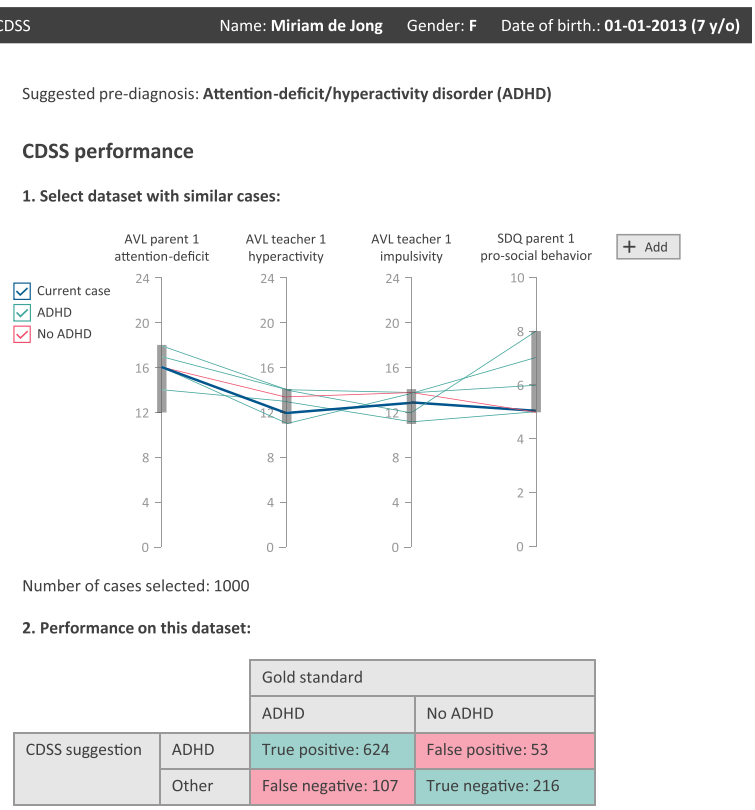
First of all, clinicians unanimously indicated that they find it highly important to know about the information that the system used to make the diagnosis (element 1). This indicates that clinicians do indeed require an explanation that accompanies the diagnosis that is suggested by a CDSS. As in the results of part 1, the overall importance ratings of information elements are high (i.e., 80% of median values are above 3, and all elements were rated as highly important by at least one participant). Moreover, once again there are relatively large variations in individual scores. This suggests that not all clinicians want to receive the same explanations from a CDSS, which argues for personalization based on their information preferences. However, the fact that many information elements are important according to clinicians does not mean that it is wise to present all this information in an explanation at once. That is, explanations that are too detailed may lead to over-reliance on the system, (Bussone et al., 2015; de Visser et al., 2019), causing clinicians to adopt decisions of the system without careful consideration (i.e., inadequate use).

Another finding is that the information preferences of clinicians do not seem to be dependent upon whether they agree with the system's advice. Regardless of whether they agree with the system, clinicians unanimously find it (highly) important to receive an explanation that contains: (1) the information that the system used to make the diagnosis, (2) supporting and contradicting information (e.g., ADHD is suggested, and the patient scores high on questionnaire items measuring attention deficiencies, but low on items measuring hyperactivity), (3) how certain the system is of the diagnosis, (4) the information that would increase the certainty (e.g., items that measure impulsivity are important predictors for ADHD, but are missing in the current case), (5) other diagnoses that are conceivable, and (6) the performance of the system for other, similar cases. These results are highly similar to those of the first part in which clinicians assessed the importance of information in explanations that they provide. In both cases, clinicians find it important to

Table A19
DP 12: Performance on similar cases.

Problem description	The user needs to know the performance of the system for other, similar cases.
UI design example	

Problem description	The user needs to know the performance of the system for other, similar cases.
UI design example	



indicate the information that is used in the decision-making process, the supporting information, other likely diagnoses, and the certainty with which they make the diagnosis. Interestingly, information that contradicts the diagnosis, and information that would increase the certainty of the diagnosis that is made, is considered more important by clinicians in explanations from CDSSs. That is, while not all clinicians find it important to mention contradicting information in explanations to others, all clinicians want to receive this information as part of the explanation from the system. These results reveal a particular attention of clinicians to information that enables them to obtain a reliable, accurate, and substantiated picture of all possible diagnoses given the information in the case. This is confirmed by their additional remarks during the interview, in which they indicate that they value the system mostly because it enables them to be more critical, which is also found by Wang and colleagues (Wang et al., 2019). That is, regardless of whether they agree with the system, clinicians mostly want to use the system to better interpret the information in the patient's case, by evaluating the evidence for and against the hypothesized diagnosis, other likely diagnoses (i.e., contrastive explanations), and the system's certainty and performance.

Lastly, clinicians unanimously expected that their information preferences for explanations of the system will change over time, as they build up experience with its advice and explanations. That is, they expect their understanding of the system to increase over time, causing them to be needing less extensive explanations. More specifically, clinicians stated that supporting- and counter evidence, and the certainty of the system will likely remain important in explanations, while information that is used to make the diagnosis, and the diagnosis in similar cases is likely to become less important over time. It might be interesting to further investigate this result in a longitudinal study in which users work with the system and its explanations for longer periods of time.

5. Multi-modal interaction design and evaluation of explanations from CDSS

In this part of the DoReMi-method, the goal is to discover how the information that is generated by an XAI system can be effectively communicated to the user. To investigate this, we first created UI design patterns for explanations on the basis of the results that were obtained in the first user study. Because we did not find clear contextual dependencies for explanations, and all clinicians found most information elements important or highly important, we decided to develop UI design patterns for all elements. The steps in the design process for each pattern were: (1) specify the explanation design problem (based on the user study), (2) specify the information elements that are relevant for the design solution, (3) consult XAI literature and AI-experts to identify techniques that can obtain the information from the system, (4) consult design experts to specify how this information can be presented in an

understandable manner, (5) create a concept-interface based on the pattern, (6) refine the concept based on internal discussions with AI and interaction design experts.

After creating the patterns and user interfaces that contained explanations of a diagnosis based on the use case from the first user study, we conducted a second user study in which clinicians evaluated the designs. In the following sections, we describe how we created the UI design patterns, and how the user study was set up.

5.1. UI design patterns for explanations by a CDSS

We developed design patterns for all information elements from Table 1, which we contextualized by using the information in the use case from the first study. Some information elements were combined into a single pattern, because it would be illogical or impossible to separate them. Table 3 shows the distribution of information elements over the twelve UI design patterns that we created. Initial designs were made by author 2 based on his personal interaction design knowledge and experience and examples from the XAI literature. The designs were then iteratively refined based on reviews with the other authors, and AI experts and interaction design experts from our organisation.

Tables 4 – 7 show the most interesting design patterns (i.e., the ones containing non-trivial design choices). All other patterns can be found in Appendix A. Each pattern includes a problem description (i.e., the user needs to know *X*), a UI design example that illustrates the solution, and a rationale for why this is a good solution to the problem (only included for non-trivial design choices). In the rationale, we also discuss the XAI methods that can be used to extract the information for the UI designs.

5.2. Methods

5.2.1. Participants

The same six paediatricians from the first study also participated in this second user study. Participants were once again recruited via an e-mail, in which we explained the general purpose of the second study and requested their participation. Clinicians received a box of chocolate bonbons as token of appreciation.

5.2.2. Materials

Questionnaire We based the questionnaire items for the second study on questions that have been identified by Hoffman and colleagues (Hoffman et al., 2018) as capable of measuring subjective accounts of understanding and trust in users of XAI systems. More specifically, we used three questions from the scale that measures explanation satisfaction in users (see Hoffman et al., 2018 for the full survey). Although this scale contains a total of eight items, we only chose questions that directly relate to understanding and trust, as our goal was to let clinicians evaluate the interfaces in terms of these concepts. We added three other questions in order to measure the importance of each explanation component (i.e., the information element(s) presented in one UI), understandability of the UI, and the extent to which clinicians think that the UI improves the decision-making process. This brought the total number of items in the questionnaire to six, all of which were presented for each explanation-component:

- *This explanation-component is important.*
- *This explanation-component is understandable.*
- *From the explanation-component, I understand how the system works.* (taken from Hoffman et al., 2018)
- *This explanation-component lets me judge when I should trust and not trust the system.* (taken from Hoffman et al., 2018)
- *This explanation-component improves my decision-making process.*
- *This explanation-component shows me how accurate the system is.* (taken from Hoffman et al., 2018)

We measured importance in order to find out whether the ratings are congruent with those obtained in the first study. Low consistency can mean that the communicated information by the interface differs from that which the interface is designed to communicate, in which case it is essential to determine what causes this difference. For example, it could be that the information in the interface is difficult to comprehend for users, which impedes being able to assess its importance. Therefore, we also included understandability of the interface as questionnaire item.

Another desirable characteristic of explanations by a system is that they enable users to improve their understanding of the system (Hoffman et al., 2018; Miller, 2018). That is, the information should provide an accurate representation of the capabilities of the system and should be carefully designed in order to reduce the risk of developing misunderstanding (especially in high-risk domains such as healthcare), in order to facilitate adequate use of the system. To the same end, explanation interfaces should also enable users to develop a level of trust in the system that accurately reflects the system's performance (de Visser et al., 2019), for example by including information that enables assessment of the system's accuracy.

The main goal of a decision-support system is to support decision-making. Therefore, we also included the question about whether clinicians feel like the explanations in the interface are able to improve their decision-making process. Of course, ultimately we want to obtain objective results on all six topics that were included in the questionnaire, but this requires users to work with the system and gain experience with its explanations for a longer period of time. However, in this part of the design process of explanations, it is valuable to obtain subjective measurements to have an indication about the effectivity of the explanation interfaces, and to learn how users value the presentation of information (i.e., the face validity of the design patterns).

The full questionnaire consisted of the 12 interface designs, which were subsequently presented on the left page of a brochure-style booklet. On each right page, the six questionnaire items to evaluate the interface on the left were presented. This method of presenting enabled participants to see the interface while providing answers to the questions. All items were rated on a five-point Likert-scale. The first item was rated on the same levels as in study 1: (1) not at all important, (2) not important, (3) somewhat important, (4) important and (5) very important. All other items were rated on the following levels: (1) strongly disagree, (2) disagree, (3) neutral, (4) agree, (5) strongly agree. In addition to the closed questions, we included an open question at the end of the questionnaire in which we asked whether there is any additional information that participants would like to receive.

5.2.3. Procedure

The study was once more conducted in a conference room at a local health centre. The first author was present to introduce the study and to answer any questions of the paediatricians. Prior to filling out the questionnaire, participants read about the introduction of the study, and were once again presented with the use case of the 7-year old called Miriam. Then, we subsequently presented the 12 user interfaces and the six questions.

5.2.4. Data analysis

For each UI, we determined the median, minimum, and maximum of ratings on all questionnaire items separately. We evaluated the 12 user interfaces on these measures. The consistency of individual importance ratings between study 1 and 2 was determined by calculating the mean difference between importance ratings of each interface and the average importance of the corresponding information elements (see Table 3) in situation 3a (system agrees) from the first study. Moreover, the first two authors analyzed the qualitative data from the last question from the questionnaire in order to evaluate whether participants had any additions to the information in the interfaces.

5.3. Results

Fig. 5 shows the results on all questionnaire items for each user interface. Scores on the items are quite consistent between participants, as most scores differ by 1 or 2 points (on a 5-point Likert-scale). Overall, scores on all six items are high (82% of medians per interface are above 3). Moreover, as in study 1, there are many relatively high minimum values (i.e., 4 or 5) and no relatively low maximum values (i.e., 1 or 2), which indicates that, overall, participants gave high ratings. First of all, importance ratings are high for all interfaces, especially for UI 12, which clinicians unanimously rated with either 4 (important) or 5 (highly important). Interestingly, there is relatively strong consistency between individual importance ratings in study 1 and study 2 (mean difference between scores for each participant = 0.85, standard deviation = 0.64). Moreover, importance ratings for information pertaining to the certainty (UI 3) and performance (UI 12) of the system were rated with 4 or 5 by almost all clinicians in both studies.

The median score on understandability was 4 or higher for all interfaces, although there were two clinicians who did not find interface 8 and 12 understandable, respectively. For interfaces 4, 6, and 10, all clinicians (strongly) agreed with the statement concerning the ability of the interface to facilitate system understanding. On average, ratings on the ability of the interfaces to enable clinicians to judge the level of trust in the system were slightly lower, with interface 1 receiving the lowest ratings (median of 2, maximum of 3), and interface 10 and 12 receiving the highest ratings (median of 4, maximum of 5). With respect to the ability of the interface to improve decision-making, interface 1 received lowest ratings, while only interface 10 received unanimously high ratings of 4. Lastly, interface 12 was rated by all clinicians to best enable them to assess the system's accuracy, while interface 1 and 2 received the lowest scores.

Apart from the ratings on the 5-point Likert-scale, participants unanimously indicated that there was no additional information that they would like to receive on top of the information that is presented in the user interfaces.

5.4. Discussion

In this study, clinicians evaluated interface designs for separate components of explanations of diagnoses from a CDSS. More specifically, they indicated the importance of the information that the interfaces present, and their level of agreement with statements regarding the understandability of the interfaces, their ability to enable clinicians to understand how the system works, to judge when they should (not) trust the system, to improve the decision-making, and to learn how accurate the system is. Interestingly, the within-subjects consistency in importance ratings between study 1 and 2 for each information element is high, which implies that the interface designs are able to communicate the information that they are intended to communicate.

Moreover, there is likely to be an interaction between importance and understandability of the interface, as understanding the information that is presented is a prerequisite for being able to assess its importance. In this study, nearly all clinicians agreed or agreed strongly with the statement about understandability of each interface. However, there was one clinician who did not find UI 8, which enables making and evaluating counterfactual statements about patient information, understandable. Another clinician did not find UI 12, presenting a confusion-matrix that indicates the system's performance, understandable. Although most clinicians indicated to understand all interfaces, the fact that some did not underlines the importance of proper training with the system. That is, all end users should have a similar level of understanding about the system's functionalities and their potential usefulness, in order to enable similarly adequate use of the system. User interfaces should thus be designed in such a way that the information that they present is understandable, while also supporting learning about the system. It is therefore also interesting to know what interfaces can increase system understanding. For most of the interfaces that we designed, clinicians felt that it increases their understanding of the system, especially the information concerning supporting and contradicting factors (UI 4), differential diagnoses (UI 6), and the partial dependency plots that enable case-based reasoning (UI 10). Lowest ratings were provided for interface 1, which contained a description of the diagnosis, and 3, in which the certainty of the system was expressed in natural language, which can be expected as this kind of information tells little about how the system works.

Next is the ability of the interfaces to enable users to obtain an appropriate level of trust in the system. The results of our study show that merely providing the system output along with a description of the diagnosis was judged as being least informative about the trustworthiness of the system's advice. This is understandable, as this information does not enable users to learn about the reasoning behind the advice, which is why explanations are required. The interface that shows the system's performance was rated as being most helpful to assess the trustworthiness of the system. This is line with previous research in which performance was found to be the most powerful predictor of human trust in a system (see Hancock et al., 2011 for a meta-analysis). Overall, the statement regarding trust in the system received the most neutral scores, which indicates that clinicians are most indifferent about this item. This could indicate that clinicians have difficulty evaluating this characteristic of the interface, which emphasizes the importance of an objective evaluation of trust (e.g., by measuring reliance). This finding is also congruent with the notion that trust in AI systems is not solely based on explanations of outputs, but also on the performance and attributes (e.g., adaptability and personality) of the system (Hancock et al., 2011).

Overall, clinicians were positive about the contribution of the information that is presented in the interfaces to their decision making, except for general information about the diagnosis, for which most clinicians were neutral (and one even strongly disagreed). Moreover, participants unanimously agreed that case-based information containing the typical case of a patient with the diagnosis that is suggested, and its relation to the current case would improve their decision-making.

Lastly, most UI design patterns are judged by clinicians as enabling them to assess the accuracy of the system, which is important to be able to determine how reliable the output of the system is. On average, information about the diagnosis, the information that is used in the decision-making, and the certainty of the system inform clinicians the least about the system's accuracy. Not surprisingly, interface 12, which contains the confusion-matrix that shows the sensitivity and specificity of the system for all or a selected amount of cases that received the same diagnosis was judged as providing the most information about the system's accuracy.

6. General discussion

We described the DoReMi-approach for the development of XAI, in which end users are actively and repeatedly involved in order to develop fit-for-purpose explanations. While the user is central to the approach, the design of explanations that are adequate for use in a particular context requires collaboration of a multi-disciplinary research and development team from the start. That is, the method includes activities that cover multiple areas of expertise: research and development of XAI methods, interactions with domain experts, design and execution of human factors research, and design of interaction between user and XAI. Thus, successful development of XAI requires a team that consists of end users, and experts on AI, human factors, and (human-computer) interaction design.

By collaborating in a multi-disciplinary fashion, we applied the DoReMi-approach to investigate explanations for clinical decision support in the domain of child health. This provided us with the desired design specifications (see Fig. 1): a description of the context of use, a suitable use case to investigate and evaluate explanations, explanation

requirements from clinicians, and ultimately a first set of UI design patterns for explanations. The DoReMi-practice enabled us to efficiently obtain these first design specifications by involving a relatively low number of end users and by using reusable questionnaires. Note that the results of DoReMi should be refined and evaluated further through their use by other XAI researchers and developers.

The DoReMi-approach is intended to be used right from the start of XAI development, in order to find out what content and what form is suitable for explanations within the context(s) in which the system will operate. The main purpose of explanations is ultimately to enable users to learn about the system (Lombrozo, 2006; Williams et al., 2013), so that they understand how the system works and are able to predict when and to what extent its output can be trusted, which facilitates adequate use of the system (Hoffman et al., 2018). These effects of explanations can be estimated, but not determined with subjective evaluations. Instead, they require objective measurements within the context of use. Therefore, the first activity after obtaining all outcomes from the DoReMi-method is to build a prototype of the XAI system based on these findings, and to have users interact with the system in the context of use (which may be simulated for high-risk domains) in order to measure their behavior. By combining subjective (e.g., Hoffman et al., 2018) with objective (e.g., Samek et al., 2017) measurements, it can be established what the effects of explanations are on users' understanding of, trust in, and reliance on the system. This can also provide more insight into how and when the explanations are used by the user. For example, it could reveal context-dependencies that were not discovered during domain analysis or requirements elicitation, and it also provides opportunity to investigate the requirements and effects of explanations over time. Additionally, potential (undesirable) side-effects can also be revealed, such as the effect of the amount of detail in explanations on the degree of reliance on the system (Bussone et al., 2015).

7. Conclusion

This paper presented the application of a human-centered design approach for eXplanations of Artificial Intelligence (XAI) as a practice for user requirements analysis and deriving corresponding *reusable* design solutions. It distinguishes three components: Domain analysis to define the concept & context of explanations, Requirements elicitation & assessment to derive the use cases & explanation requirements, and the consequential Multi-modal interaction design & evaluation to create a library of design patterns for explanations. This DoReMi-approach provided the first set of user requirements and UI design patterns for an explainable decision support system in child health, showing how to involve expert end users in the development process, and how to derive, more or less generic solutions for general design problems in XAI from the domain & requirements analysis and current interaction design knowledge. Whereas current XAI-studies mainly focus on single purpose explanations (most often transparency) and natural-language explanation formats (Nunes and Jannach, 2017), the DoReMi-approach provided a richer XAI design space with (1) a (first) set of user requirements for explanations that service the different user goals, and (2) a (first) set of reusable design patterns for multi-modal explanatory user interfaces.

The evaluations with clinicians showed that they really need explanations of the AI-output, particularly to help mitigate false positive diagnoses, while avoiding false negatives. Such explanations support the required critical attitude, reducing the risk of over-reliance. All clinicians who participated in the study indicated that the explanations of the different design patterns are important and understandable, supporting trust-calibration and decision-making. The set of UI design patterns seem to cover all relevant information elements to be included in the different explanations. There were some individual differences in the rating of the explanations that point to the need for personalization.

Currently, the explanations and design patterns are being integrated in a CDSS prototype, and will be further tested on their understandability, trust development and decision-making performance. This study will entail different sessions to study learning, trust development and performance change over time. To extend the scope and test the genericity of the explanation design patterns, the development and evaluation will include different types of use cases (e.g., child mental health, adult diabetes management) and users (e.g. clinicians and patients).

CRediT authorship contribution statement

Tjeerd A.J. Schoonderwoerd: Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization. **Wiard Jorritsma:** Writing - original draft, Data curation, Formal analysis, Visualization. **Mark A. Neerincx:** Writing - review & editing, Validation. **Karel van den Bosch:** Writing - review & editing, Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research has been funded by the Netherlands Organisation for applied scientific research (TNO), under the Early Research Programme in Applied Artificial Intelligence.

Appendix A. UI design patterns for XAI

This appendix includes the full set of UI design patterns that were created for the explanation requirements and, subsequently, used in the second user study (Tables A8 – A19).

References

Adadi, A., Berrada, M., 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138–52160.
Alexander, C., 1977. A Pattern Language: Towns, Buildings, Construction. Oxford University Press.
Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P.N., Inkpen, K., et al., 2019. Guidelines for human-ai interaction. Proceedings of the 2019 chi Conference on Human Factors in Computing Systems, pp. 1–13.
Anjomshoae, S., Najjar, A., Calvaresi, D., Främling, K., 2019. Explainable agents and robots: results from a systematic literature review. Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. International Foundation for Autonomous Agents and Multiagent Systems, pp. 1078–1088.
Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al., 2020. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible ai. Inf. Fusion 58, 82–115.
Baker, A., 2001. Crossing the quality chasm: a new health system for the 21st century.
Berner, E.S., La Lande, T.J., 2007. Overview of clinical decision support systems. Clinical Decision Support Systems. Springer, pp. 3–22.
Burnett, M., 2020. Explaining ai: fairly? Well? Proceedings of the 25th International Conference on Intelligent User Interfaces, pp. 1–2.
Bussone, A., Stumpf, S., O'Sullivan, D., 2015. The role of explanations on trust and reliance in clinical decision support systems. 2015 International Conference on Healthcare Informatics. IEEE, pp. 160–169.
Cai, C.J., Jongejan, J., Holbrook, J., 2019. The effects of example-based explanations in a machine learning interface. Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 258–262.
Caro-Martinez, M., Jimenez-Diaz, G., Recio-Garcia, J.A., 2018. A theoretical model of explanations in recommender systems. ICCBR, p. 52.
Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
Ehsan, U., Riedl, M. O., 2020. Human-centered explainable ai: towards a reflective sociotechnical approach. arXiv preprint arXiv:2002.01092.
de Visser, E.J., Peeters, M.M., Jung, M.F., Kohn, S., Shaw, T.H., Pak, R., Neerincx, M.A., 2019. Towards a theory of longitudinal trust calibration in human–robot teams. Int. J. Soc. Robot. 12, 459–478.

Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., Hussmann, H., 2018. Bringing transparency design into practice. 23rd International Conference on Intelligent User Interfaces, pp. 211–223.
Friedman, B., Kahn, P.H., Borning, A., 2008. Value sensitive design and information systems. The Handbook of Information and Computer etzhics, pp. 69–101.
Friedman, C.P., Elstein, A.S., Wolf, F.M., Murphy, G.C., Franz, T.M., Heckerling, P.S., Fine, P.L., Miller, T.M., Abraham, V., 1999. Enhancement of clinicians' diagnostic reasoning by computer-based consultation: a multisite study of 2 systems. JAMA 282 (19), 1851–1856.
Guida, G., Mussio, P., Zanella, M., 1997. User interaction in decision support systems: the role of justification. 1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation, 4. IEEE, pp. 3215–3220.
Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., Giannotti, F., 2018a. Local rule-based explanations of black box decision systems. arXiv preprint arXiv:1805.10820.
Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D., 2018. A survey of methods for explaining black box models. ACM Comput. Surv. (CSUR) 51 (5), 93.
Hall, M., Harborne, D., Tomsett, R., Galetic, V., Quintana-Amate, S., Nottle, A., Preece, A., 2019. A systematic method to understand requirements for explainable ai (XAI) systems. Proceedings of the IJCAI Workshop on eXplainable Artificial Intelligence (XAI 2019), Macau, China.
Hancock, P.A., Billings, D.R., Schaefer, K.E., Chen, J.Y., De Visser, E.J., Parasuraman, R., 2011. A meta-analysis of factors affecting trust in human-robot interaction. Hum. Fact. 53 (5), 517–527.
Hilton, D.J., 1990. Conversational processes and causal explanation. Psychol. Bull. 107 (1), 65.
Hoffman, R. R., Mueller, S. T., Klein, G., Litman, J., 2018. Metrics for explainable ai: challenges and prospects. arXiv preprint:1812.04608.
Holst, H., Åström, K., Järund, A., Palmer, J., Heyden, A., Kahl, F., Trägil, K., Evander, E., Sparr, G., Edenbrandt, L., 2000. Automated interpretation of ventilation-perfusion lung scintigrams for the diagnosis of pulmonary embolism using artificial neural networks. Eur. J. Nucl. Med. 27 (4), 400–406.
Holzinger, A., Carrington, A., Müller, H., 2020. Measuring the quality of explanations: the system causability scale (SCS). KI-Künstliche Intell. 6 (34), 193–198.
Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H., 2019. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip. Rev. 9 (4), 1–13.
Holzinger, A., Malle, B., Saranti, A., Pfeifer, B., 2021. Towards multi-modal causability with graph neural networks enabling information fusion for explainable ai. Inf. Fusion 71, 28–37.
Hunt, D.L., Haynes, R.B., Hanna, S.E., Smith, K., 1998. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. JAMA 280 (15), 1339–1346.
Inselberg, A., 1997. Multidimensional detective. Proceedings of VIZ'97: Visualization Conference, Information Visualization Symposium and Parallel Rendering Symposium. IEEE, pp. 100–107.
Kawamoto, K., Houlihan, C.A., Balas, E.A., Lobach, D.F., 2005. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ 330 (7494), 765.
Kirsch, A., 2017. Explain to whom? Putting the user in the center of explainable AI. Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 co-located with 16th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017). Bari, Italy, pp. 1–5.
Kononenko, I., 2001. Machine learning for medical diagnosis: history, state of the art and perspective. Artif. Intell. Med. 23 (1), 89–109.
Lamy, J.-B., Sekar, B., Guezennec, G., Bouaud, J., Séroussi, B., 2019. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. Artif. Intell. Med. 94, 42–53.
Liao, Q.V., Gruen, D., Miller, S., 2020. Questioning the ai: informing design practices for explainable ai user experiences. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–15.
Lim, B.Y., Yang, Q., Abdul, A.M., Wang, D., 2019. Why these explanations? Selecting intelligibility types for explanation goals. IUI Workshops.
Lombrozo, T., 2006. The structure and function of explanations. Trends Cogn. Sci. 10 (10), 464–470.
Madumal, P., Miller, T., Sonenberg, L., Vetere, F., 2019. A grounded interaction protocol for explainable artificial intelligence. arXiv preprint arXiv:1903.02409.
Maguire, M., Bevan, N., 2002. User requirements analysis. IFIP World Computer Congress, TC 13. Springer, pp. 133–148.
Markus, A. F., Kors, J. A., Rijnbeek, P. R., 2020. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. arXiv preprint arXiv:2007.15911.
Miller, T., 2018. Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. 267, 1–38.
Mittelstadt, B., Russell, C., Wachter, S., 2019. Explaining explanations in ai. Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 279–288.
Montavon, G., Samek, W., Müller, K.-R., 2018. Methods for interpreting and understanding deep neural networks. Digit. Signal Process. 73, 1–15.
Neerincx, M.A., van Diggelen, J., van Breda, L., 2016. Interaction design patterns for adaptive human-agent-robot teamwork in high-risk domains. International Conference on Engineering Psychology and Cognitive Ergonomics. Springer, pp. 211–220.
Neerincx, M.A., van der Waa, J., Kaptein, F., van Diggelen, J., 2018. Using perceptual and cognitive explanations for enhanced human-agent team performance. International Conference on Engineering Psychology and Cognitive Ergonomics. Springer, pp. 204–214.
Neerincx, M.A., van Vught, W., Henkemans, O.B., Oleari, E., Broekens, J., Peters, R., Kaptein, F., Demiris, Y., Kiefer, B., Fumagalli, D., et al., 2019. Socio-cognitive engineering of a robotic partner for child's diabetes self-management. Front. Robot. AI 6, 1–16.
Nunes, I., Jannach, D., 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. User Model. User-Adapted Interact. 27 (3-5), 393–444.
Ozaydin, B., Hardin, J.M., Chhieng, D.C., 2016. Data mining and clinical decision support systems. Clinical Decision Support Systems. Springer, pp. 45–68.
Paetsch, F., Eberlein, A., Maurer, F., 2003. Requirements engineering and agile software development. WET ICE 2003. Proceedings. Twelfth IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, 2003. IEEE, pp. 308–313.
Parmer, J., Parmer, C., Johnson, A., 2020. Plotly website. http://www.plotly.com.
Poulin, B., Eisner, R., Szafron, D., Lu, P., Greiner, R., Wishart, D.S., Fyshe, A., Pearcy, B., MacDonell, C., Anvik, J., 2006. Visual explanation of evidence with additive classifiers. Proceedings of the National Conference on Artificial Intelligence, 21. AAAI Press; MIT Press; 1999, Menlo Park, CA; Cambridge, MA; London, p. 1822.
Pu, P., Chen, L., 2007. Trust-inspiring explanation interfaces for recommender systems. Knowledge-Based Syst. 20 (6), 542–556.
Ras, G., van Gerven, M., Haselager, P., 2018. Explanation methods in deep learning: Users, values, concerns and challenges. Explainable and Interpretable Models in Computer Vision and Machine Learning. Springer, pp. 19–36.
Ribeiro, M.T., Singh, S., Guestrin, C., 2016. Why should i trust you?: explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 1135–1144.
Ribera, M., Lapedriza, À., 2019. Can we do better explanations? A proposal of user-centered explainable AI. IUI Workshops, pp. 1–7.
Samek, W., Wiegand, T., Müller, K.-R., 2017. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296.
Schneider, J., Handali, J., 2019. Personalized explanation in machine learning. CoRR abs/1901.00770.
Sherif, M., Taub, D., Hovland, C.I., 1958. Assimilation and contrast effects of anchoring stimuli on judgments. J. Exp. Psychol. 55 (2), 150.
Strumbelj, E., Kononenko, I., 2010. An efficient explanation of individual classifications using game theory. J. Mach. Learn. Res. 11 (Jan), 1–18.
Thellman, S., Silvervarg, A., Ziemke, T., 2017. Folk-psychological interpretation of human vs. humanoid robot behavior: exploring the intentional stance toward robots. Front. Psychol. 8, 1962.
Tomsett, R., Braines, D., Harborne, D., Preece, A., Chakraborty, S., 2018. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. arXiv preprint arXiv:1806.07552.
Van Welie, M., Van Der Veer, G.C., Eliëns, A., 2001. Patterns as tools for user interface design. Tools for Working with Guidelines. Springer, pp. 313–324.
van der Waa, J., van Diggelen, J., Neerincx, M.A., Raaijmakers, S., 2018. ICM: an intuitive model independent and accurate certainty measure for machine learning. ICAART, pp. 314–321.
van der Waa, J., Robeer, M., van Diggelen, J., Brinkhuis, M., Neerincx, M., 2018. Contrastive explanations with local foil trees. 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018).
Wang, D., Yang, Q., Abdul, A., Lim, B.Y., 2019. Designing theory-driven user-centric explainable ai. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, p. 601.
Williams, J.J., Lombrozo, T., Rehder, B., 2013. The hazards of explanation: Overgeneralization in the face of exceptions. J. Exp. Psychol. 142 (4), 1006.
Wolf, C.T., 2019. Explainability scenarios: towards scenario-based XAI design. Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 252–257.
Xie, Y., Gao, G., Chen, X., 2019. Outlining the design space of explainable intelligent systems for medical diagnosis. arXiv preprint: 1902.06019.
Ye, L.R., Johnson, P.E., 1995. The impact of explanation facilities on user acceptance of expert systems advice. MIS Q. 157–172.