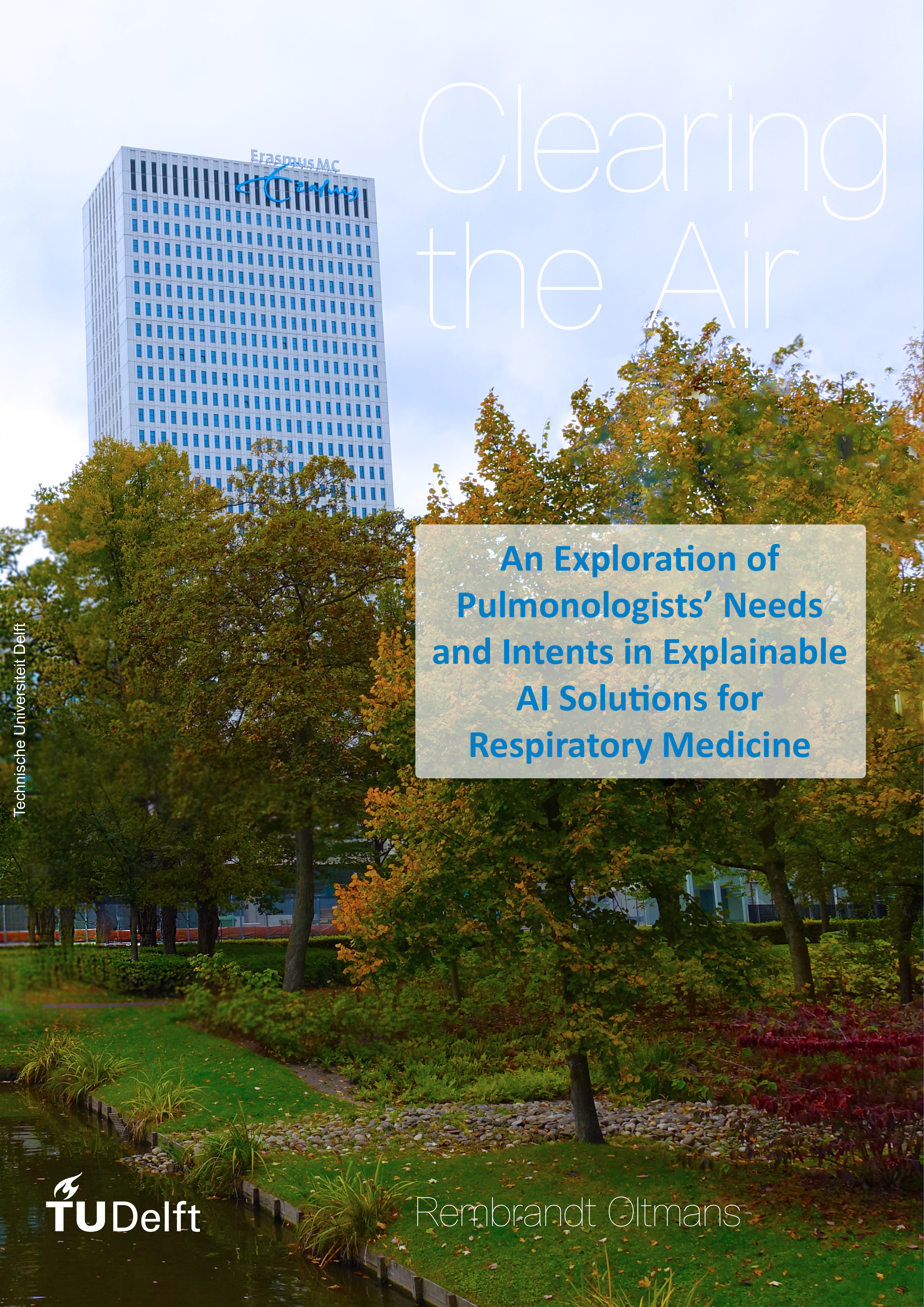


Clearing the Air

An Exploration of Pulmonologists' Needs and Intents in Explainable AI Solutions for Respiratory Medicine

Rembrandt Oltmans

Technische Universiteit Delft

TU Delft

Clearing the Air

An Exploration of Pulmonologists' Needs and Intents in Explainable AI Solutions for Respiratory Medicine

by

R.F.A. Oltmans

to obtain the degree of Master of Science

at the Delft University of Technology,

to be defended publicly on Thursday August 31, 2023 at 09:00 AM.

Student number:	4534840	
Project duration:	December 1, 2022 – August 31, 2023	
Thesis committee:	dr. C. Lofi,	TU Delft, Thesis advisor
	dr. J. Yang,	TU Delft, Daily supervisor
	dr. J. Jung,	TU Delft, Erasmus MC, External supervisor
	dr. Z. Yue,	TU Delft, External committee member

This thesis is confidential and cannot be made public until January 31, 2024.

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.//

Cover image: Museumpark @ Rotterdam by Guilhem Vellut, Flickr

TUDelft

Abstract

Despite the low adoption rates of artificial intelligence (AI) in respiratory medicine, its potential to improve patient outcomes is substantial. To facilitate the integration of AI systems into the clinical setting, it is essential to prioritise the development of explainable AI (XAI) solutions that improve the understanding of the AI predictions. These XAI solutions empower clinicians to collaborate effectively with AI systems, thereby enhancing the overall outcomes for patients in respiratory medicine. Unfortunately, the lack of user-centric studies in this domain has made it challenging to identify the specific aspects of explainability that are most effective in improving the adoption of AI in the real-world environment. To address this gap, we conducted a mixed-methods study of clinicians in respiratory medicine to identify the most relevant and crucial aspects of XAI solutions. Our study focused on understanding how XAI can be effectively translated into clinical practice by leveraging the expertise of doctors in the field. Because of the lack of knowledge about XAI concepts among pulmonologists a different approach is taken to regular user-centric XAI research and no direct examples of state-of-the-art XAI solutions are used. Rather the expertise of doctors is used to make them implicitly identify their needs and intents. Our findings reveal that the successful adoption of XAI solutions in respiratory medicine requires tailored solutions that address communication barriers, promote patient-centric care, and overcome AI adoption challenges. The study highlights the significance of task-specific visualisations, comprehensive explanations, preferred granularity, and the ability to mimic human judgement in successful XAI solutions. Trust and collaboration between clinicians and AI systems are essential for effective adoption, wherein AI is perceived as a colleague rather than a replacement. This ensures that clinicians can easily understand and work with the model predictions, ultimately leading to improved patient outcomes. By aligning XAI design with the needs and intents of pulmonologists, we established the importance of Co-designing solutions with domain experts and embedding XAI within clinical workflows emerged as key strategies. Our research underscores the imperative of transparency, extended validation, and continuous alignment of AI technologies with medical values. By following these principles, XAI solutions can be developed to enhance the diagnosis and treatment of respiratory illnesses, ultimately improving patient outcomes in respiratory medicine.

Executive Summary

This report delves into the integration of explainable artificial intelligence (XAI) solutions in healthcare, with a particular focus on enhancing patient care in the field of pulmonology. The study recognises the pivotal role of XAI in bridging the gap between complex AI outputs and human comprehension for informed clinical decisions. The report unfolds in several chapters, each illuminating critical aspects of XAI's integration into medical practices.

Introduction and XAI Overview The introduction sets the stage by emphasising the potential collaboration between doctors and AI in healthcare to overcome AI's lack of transparency and enhance patient outcomes. The need for explainable AI becomes evident, given the challenges posed by the opacity of AI systems in medical decision-making. The report highlights the importance of interpretable explanations for AI decisions to enable effective collaboration between doctors and AI systems.

Exploring XAI Algorithms and Medical Applications The subsequent chapter delves into various XAI algorithms such as LIME and Anchors, emphasising their role in providing understandable explanations for complex AI models. The report underscores the need for user studies, particularly in the medical domain, to tailor XAI algorithms to clinicians' unique needs. It also stresses the importance of adapting XAI methods for diverse medical data and involving medical experts for successful implementation.

Research Focus: XAI in Pulmonology The research centers on enhancing collaboration between doctors and AI in the context of idiopathic pulmonary fibrosis (IPF) care. IPF, a chronic lung disease, poses diagnostic and treatment challenges. The study underscores the potential of XAI in improving patient outcomes in respiratory medicine, particularly in IPF care. The report explores XAI algorithms' suitability for doctors and their impact on enhancing medical decisions in IPF cases.

Human-Centered XAI Frameworks and Design The report emphasises human-centered frameworks for designing XAI solutions in healthcare. It highlights the importance of collaboration, transparency, and usability testing to meet doctors' needs and expectations. The chapter also outlines the significance of co-design approaches that involve end-users in the design process, creating solutions that align with their requirements.

Contribution and Future Work The study's contributions lie in identifying opportunities for XAI application in pulmonology, understanding medical staff's needs, and designing human-centered XAI experiences. However, the report acknowledges certain limitations, including contextual specificity, sample size, and prototype dynamics. The recommendations section outlines future research directions, including broader medical speciality engagement, longitudinal studies, ethical considerations, patient involvement, and collaborative frameworks.

Conclusion and Call for Action In conclusion, the report underscores the significance of XAI's integration in healthcare, particularly pulmonology, to enhance medical decision-making and patient care. It asserts that XAI solutions should align with doctors' needs, provide tailored explanations, and foster collaboration and trust. The study's insights and recommendations call for ongoing research, collaboration, and the responsible integration of XAI to revolutionise healthcare practices and improve patient outcomes.

This report paints a comprehensive picture of XAI's potential in healthcare, offering a roadmap for integrating AI technologies responsibly and effectively into medical practices.

Executive Summary Dutch

Dit rapport gaat in op de integratie van verklaarbare kunstmatige intelligentie (XAI)-oplossingen in de gezondheidszorg, met speciale aandacht voor het verbeteren van de patiëntenzorg op het gebied van longziekten. Het onderzoek erkent de cruciale rol van XAI bij het overbruggen van de kloof tussen complexe AI-resultaten en menselijk begrip voor geïnformeerde klinische beslissingen. Het rapport bestaat uit verschillende hoofdstukken, die elk kritieke aspecten van de integratie van XAI in de medische praktijk belichten.

Inleiding en overzicht van XAI De inleiding zet de toon door de nadruk te leggen op de potentiële samenwerking tussen artsen en AI in de gezondheidszorg om het gebrek aan transparantie van AI te overwinnen en de resultaten voor patiënten te verbeteren. De behoefte aan verklaarbare AI (XAI) wordt duidelijk, gezien de uitdagingen die de ondoorzichtigheid van AI-systemen in de medische besluitvorming met zich meebrengt. Het rapport benadrukt het belang van uitlegbare verklaringen voor AI-beslissingen om effectieve samenwerking tussen artsen en AI-systemen mogelijk te maken.

XAI-algoritmen en medische toepassingen verkennen Het volgende hoofdstuk gaat in op verschillende XAI-algoritmen zoals LIME en Anchors, en benadrukt hun rol in het geven van begrijpelijke verklaringen voor complexe AI-modellen. Het rapport onderstreept de noodzaak van gebruikersstudies, met name in het medische domein, om XAI-algoritmen af te stemmen op de unieke behoeften van clinici. Het benadrukt ook het belang van het aanpassen van XAI methoden voor diverse medische gegevens en het betrekken van medische experts voor een succesvolle implementatie.

Focus onderzoek: XAI in longziekten Het proefschrift richt zich op het verbeteren van de samenwerking tussen artsen en AI in de context van de zorg voor idiopathische longfibrose (IPF). IPF, een chronische longziekte, stelt ons voor uitdagingen op het gebied van diagnose en behandeling. Het onderzoek onderstreept het potentieel van XAI in het verbeteren van patiëntresultaten in de respiratoire geneeskunde, met name in de IPF-zorg. Het rapport onderzoekt de geschiktheid van XAI-algoritmen voor artsen en hun impact op het verbeteren van medische beslissingen in IPF-zaken.

Mensgericht XAI-raamwerken en -ontwerp Het rapport legt de nadruk op mensgerichte raamwerken voor het ontwerpen van XAI-oplossingen in de gezondheidszorg. Het benadrukt het belang van samenwerking, transparantie en bruikbaarheidstesten om tegemoet te komen aan de behoeften en verwachtingen van artsen. Het hoofdstuk schetst ook het belang van co-design dat eindgebruikers betrekt bij het ontwerpproces en oplossingen creëert die aansluiten bij hun behoeften.

Bijdrage en toekomstig werk De bijdragen van het onderzoek liggen in het identificeren van mogelijkheden voor XAI-toepassingen in de longziekten, het begrijpen van de behoeften van medisch personeel en het ontwerpen van mensgerichte XAI-ervaringen. Het rapport erkent echter bepaalde beperkingen, waaronder contextuele specificiteit, steekproefgrootte en prototypedynamiek. In het hoofdstuk met aanbevelingen worden toekomstige onderzoeksrichtingen beschreven, waaronder een bredere betrokkenheid van medische specialismen, longitudinale studies, ethische overwegingen, betrokkenheid van patiënten en samenwerkingsverbanden.

Conclusie en oproep tot actie Concluderend onderstreept het rapport het belang van de integratie van XAI in de gezondheidszorg, met name pulmonologie, om de medische besluitvorming en patiëntenzorg te verbeteren. Het rapport stelt dat XAI-oplossingen moeten worden afgestemd op de behoeften van artsen, uitleg op maat moeten geven en samenwerking en vertrouwen moeten bevorderen. De inzichten en aanbevelingen van het onderzoek vragen om doorlopend onderzoek, samenwerking en een verantwoorde integratie van XAI om een revolutie teweeg te brengen in de gezondheidszorg en de resultaten voor patiënten te verbeteren.

Dit rapport schetst een uitgebreid beeld van het potentieel van XAI in de gezondheidszorg en biedt een routekaart voor een verantwoorde en effectieve integratie van AI-technologieën in medische praktijken.

Contents

List of Figures

List of Tables

Glossary

Artificial Intelligence (AI) Branch of computer science that aims to create intelligent machines capable of performing tasks that typically require human intelligence, such as learning, problem-solving, decision-making, and language understanding.

Bidirectional Encoder Representations from Transformers (BERT) Is a pre-trained language model that utilises a bidirectional approach, where it considers the context from both left and right sides of a word, leading to improved performance in various natural language processing tasks over other language models.

Clinicians Doctor with direct contact with patients rather than being involved in studies

Co-design Co-design is a collaborative approach to designing products, services, or systems that involves active participation from end-users or stakeholders throughout the design process to create solutions that better meet their needs and preferences.

Explainable Artificial Intelligence (XAI) This field focuses on creating interpretable representations of AI systems and models, enabling them to provide clear explanations of their decision-making process to humans. These explanations can be generated post-hoc or embedded into the models themselves for interpretability.

Idiopathic Pulmonary fibrosis (IPF) A progressive lung disease of unknown cause, characterised by scarring and thickening of the lung tissue, which can lead to breathing difficulties and decreased lung function. In many cases leading to a premature death.

In Silico Experimental techniques performed by computers; not on real patients.

Multidisciplinary Consultation (MDO) Common in Dutch hospitals, consists of a structured meeting where different medical professionals collaborate to provide optimal patient care by discussing complex cases and developing joint treatment plans to advise the primary care doctor.

Pulmonary Fibrosis (PF) Umbrella term encompassing various causes of lung fibrosis, including but not limited to IPF. Other forms of pulmonary fibrosis can result from occupational and environmental exposures, connective tissue disorders, drug toxicity, radiation therapy, and certain infections.

Pulmonologist Doctor who diagnoses and treats diseases of the respiratory system.

1

Introduction

Imagine a future where doctors and AI work hand in hand, seamlessly combining their expertise to produce better results, reduce time constraints, alleviate working pressure, and deliver individualised care of unparalleled precision. Such a future is within reach, but it requires bridging the current barrier of explainability in AI.

Presently, when faced with complex machine learning models, doctors are confronted with a challenging choice: blindly trust the AI system's diagnosis or invest countless hours unravelling the intricate pathways leading to its conclusions. Current practise shows that often this first option is chosen, having already real influence on patient their life's. Yet, as the prevalence of AI in medicine grows, this dichotomy becomes increasingly untenable. The lack of transparency and interpretability in AI poses a substantial obstacle to its widespread adoption in healthcare [47]. Imagine doctors gaining access to intuitive explanations, shedding light on how the AI system arrives at its diagnoses and treatment recommendations. Such insights enable medical professionals to make informed decisions, enhancing patient care and outcomes. By providing interpretable explanations for the decisions made by these models it empowers doctors to work effectively with the work from intelligent machines. This empowerment, grounded in transparency and human-understandable insights, can preemptively identify and resolve potential issues before it harms patients.

But while artificial intelligence faces the fundamental challenge of lacking transparency, the concept of transparency itself may be even more enigmatic than AI [67]. Within the realm of XAI, a tapestry of algorithms unfolds, each weaving a unique path toward interpretability. LIME, Anchors, and a plethora of other model-agnostic and application-agnostic algorithms take center stage, capable of shedding light on any black-box model. It is crucial to emphasise that explainable artificial intelligence techniques rarely provide a universal solution that fits every scenario. The nature of human-machine interaction will vary based on the objectives of the individuals involved [32]. In medicine the objective and characteristics from the end-users, doctors, has not yet been researched extensively. Therefore, a critical gap remains in the field of XAI research: the lack of user studies exploring the unique needs and perspectives of clinicians.

In this thesis, we delve into the realm of doctors working efficiently from AI, leveraging XAI to overcome the current limitations and unlock the full potential of AI in healthcare. Not working collaboratively with AI, as that involves mutual goal understanding, preemptive task co-management and shared progress tracking [151], but rather enabling doctors to work from the expertise delivered by AI. We explore what is needed to implement XAI in the clinical workflow, focusing on the specific domain of respiratory medicine, with a particular emphasis on the treatment of patients suffering from pulmonary fibrosis. This progressive and debilitating lung disease demands accurate and timely diagnosis, making it an ideal use case for investigating XAI in highly complex and high stakes areas of medicine.

1.1. Explainable AI

In the past problems have occurred with the implementation of these complex models into practise, even though the theoretical performance of the models was high. Over the past years, the adoption of radiology AI in the Netherlands for example has shown growth. However, signs of stagnation are starting to emerge, with some implementations being discontinued due to budget constraints or disappointing experiences [83]. A leading real life example of failing AI systems is Watson for Oncology [144] which was IBM's flag ship medical AI being implemented for cancer diagnosis recommendation in hospitals world wide. However, lack of robustness meant that the performance in some places such as in South Korea was far below usable standards. At the hospital the system was implemented its top recommendations for 656 colon cancer patients matched those of the experts in less then half of time, potentially causing real patients harm. Caruana et al. [28] demonstrated the potential harm caused by a medical AI system that improperly handled biases in its training dataset in the context of pneumonia risk prediction. The system mistakenly associated certain lung diseases with lower risk of severe pneumonia effects due to the high level of care received by those patients, highlighting the importance of addressing bias properly to avoid jeopardising the well-being of broader patient populations.

XAI provides a guidance to address these challenges of robustness and bias by providing interpretable explanations for the decisions made by these models. This not only enhances the trust of medical professionals in the AI system but also facilitates collaboration between human experts and machines. Enabling developers and doctors to find these problems with the AI systems before they cause real harm to patients. Recent research finds that collaboration between explainable artificial intelligence and pulmonologists improves the accuracy of pulmonary function test interpretation significantly [35].

When using no explainability for complex machine learning models you currently have two options: blindly trust the AI system's diagnosis or spend countless hours trying to understand how the AI arrived at its conclusion. With the increasing use of AI in the medical domain, the above scenario is becoming more common. While AI has the potential to revolutionise healthcare, its lack of transparency and interpretability poses a significant challenge to its widespread adoption [47]. This is where Explainable AI comes into play. XAI is a subfield of AI that aims to develop algorithms and techniques that can explain the decisions made by AI systems in a human-understandable way.

What are these Explainable AI Algorithms?

Many types of novel XAI algorithms have been proposed in the field, such as LIME, Anchors, and many others [60]. Most of these algorithms are model-agnostic and application-agnostic meaning that they can be applied in any context and on any black box model. However, these general algorithms are not always adequate for healthcare data, which can sometimes be of a sequential nature. Different XAI techniques have been developed to accommodate such sequential data types [115]. Researchers have also developed benchmarks to compare these algorithms for specific use cases [15], and some studies have compared different XAI techniques to determine which is most effective in certain domains [111, 81]. While the absence of standardised evaluation methods in the field remains a challenge [98].

These explanations are designed to be understandable to humans, providing insight into how the model reached its conclusions. Each algorithm uses different techniques and methods to achieve this. The current constraints in utilising these explanation algorithms involve insufficient consideration of end users when providing explanations, rather these algorithms are made as if they would be used by the developers themself [44]. Even though the type of explanations are dependent on the end-users who need to use them [73]. Doctors and healthcare professionals in general have a limited knowledge of AI principles [158, 1] but high amounts of knowledge about the underlying factors reaching certain healthcare related decisions. The suitability of explanations therefore requires an understanding of what the specific user group, in this case doctors, wants to know from the specific AI application [86] and how that works together with their own mental models for reaching those predictions.

Towards Implementing XAI in the Clinical Workflow

AI in medicine has yet to propagate to the clinical setting [65]. There are currently limited examples of such techniques being successfully deployed into clinical practice. Respiratory medicine is no exception [53]. It is therefore important to consider how the use of AI fits into clinical workflow [107]. In the medical domain, XAI is particularly critical, as physicians need to understand how an AI system arrived at its diagnosis or treatment recommendation to make informed decisions. In this thesis, we will explore the field of Explainable AI and its relevance in the medical domain. Notably, in an Italian hospital, experimental systems that incorporate explainability, including explainability maps that visualize network activity in lung areas, are already being deployed in two Diagnostic Radiology Units [140]. This serves as evidence that the adoption of AI and explainability in clinical settings is more imminent than commonly perceived.

There is overall distinct lack of application of XAI in the context of healthcare AI systems and, in particular, a lack of user studies exploring the needs of clinicians [9]. Therefore a different approach will be taken, which takes into account the low level of knowledge about AI and explainability that doctors posses. We will complete a user study of clinicians in the field of pulmonology to take away one of the major barriers of implementing AI in clinical practice and trying to pave the way for achieving a higher adoption rate of AI in the clinical setting by including medical doctors consider that they must participate in the design process [100]. The main contributions this work makes are related to finding opportunities in the respiratory medicine domain and about finding the parameters that are related to creating a Human-centred XAI experience for the practitioners in this field. With a primary aim of promoting future research and advancement in the seamless integration of AI applications in the medical domain by minimising implementation barriers.

The focus of this thesis is on trying to improve the explainability of AI in a specific use case, which is common with papers researching XAI in the medical domain. We consider our scope the department of respiratory medicine, with a specific focus on the treatment of patients suffering from pulmonary fibrosis. The use case is interesting to research as pulmonary fibrosis is a progressive and debilitating lung disease with no cure, and accurate and timely diagnosis and treatment plan is critical for effective care. The care for IPF is complex as it is hard to diagnose, involving a multidisciplinary team [14]. Furthermore, the disease has a large overlap with other fields such as pathology and radiology. Making the treatment of IPF patients an excellent use case for investigating XAI in highly complex areas and offers opportunity to improve the current care by the adoption of AI.

Employing Practical Scenarios

Research into explainable artificial intelligence solutions for doctors has shown promising results, particularly when specific use cases are employed. However, many studies in this area fail to provide a detailed analysis of the specific use case, instead opting for general questions about XAI concepts to medical staff in that use case. We employ this use case focusing on the specific disease and its treatment and utilising it to extract the XAI needs. Specifically, we will employ the use case of idiopathic pulmonary fibrosis to explore the potential of XAI solutions for improving patient outcomes.

By employing the IPF use case as a foundation for our study, we aim to provide a more in-depth exploration of XAI solutions for doctors. Specifically, we will examine the role of XAI in Pulmonary Fibrosis Care at the Erasmus MC hospital and their pulmonology department. This approach will allow us to gain a better understanding of how XAI can be utilised to improve patient outcomes in a specific context, rather than just exploring abstract concepts in a general sense. This chapter will start with a general overview of the disease and then is followed by an introduction of the care provided at Erasmus MC to IPF patients.

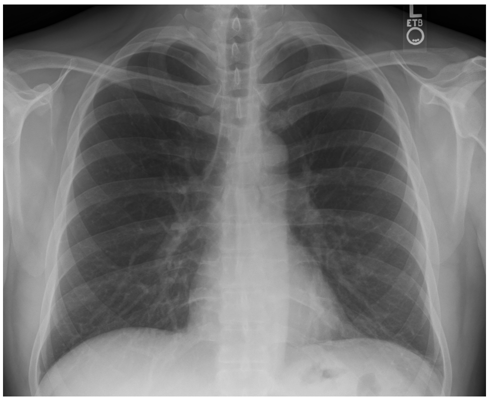


Figure 1.1: Chest X-ray of healthy lungs.

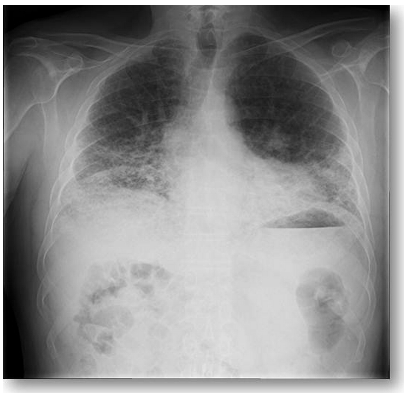


Figure 1.2: Chest X-ray of a person suffering from IPF, from Spagnolo et al. [143].

1.2. The Disease

IPF is a chronic and progressive lung disease characterised by the formation of scar tissue within the lungs, leading to shortness of breath and a persistent cough. While there is no cure for IPF, various treatment options are available that can help to slow its progression and manage symptoms. The Erasmus MC hospital, located in Rotterdam, the Netherlands, has a dedicated pulmonology department that provides specialised care for patients with IPF. This section will give a brief overview of the disease and its treatment.

Idiopathic Pulmonary Fibrosis

The most common form of idiopathic interstitial pneumonia is idiopathic pulmonary fibrosis. It is a chronic and progressive lung disease that primarily affects older adults. The disease causes lower lobe scaring within the lungs which can be seen on chest X-rays. Figure 1.1 shows a person with healthy lungs, the difference with Figure 1.2 is obvious as abnormalities and a lower total lung volume can be seen [143]. IPF has an unknown cause but is commonly identifiable through characteristic imaging and histologic appearances. Misdiagnosis and inappropriate treatment with immunosuppressive therapy is a common issue with IPF. However, treatments that can slow the progression of the disease are since recently available [78]. Because the disease is heterogeneous and difficult to diagnose the estimate for the number of people suffering from it are imprecise. However, an estimation is that between seventy thousand and one million people suffer from this disease world wide [97]. The average lifespan of untreated patients with IPF is only 3 to 4 years [95].

Symptoms, Diagnosis and Effects on Life

Patients commonly present with unexplained exertional dyspnea, chronic dry cough, or Velcro-like crackles on examination. In practice, patients with interstitial lung disease often initially receive a diagnosis of heart failure or chronic obstructive pulmonary disease. As the disease has a very low occurrence and knowledge outside of specialised care centres about the disease is low. In some cases it can take years before the correct diagnosis is found [78]. Most of the time diagnosis consists on high-resolution CT (HRCT) imaging patterns, taking a detailed medical history, serological testing for connective tissue disease and in some cases surgical lung biopsy [130]. As diagnosing IPF is very difficult other diagnostic steps might be taken to rule out other diseases.

Chronic cough can have a considerable impact on the quality of life of patients with IPF by causing interruptions in sleep, limitations in speech, significant desaturation, musculoskeletal pain, and urinary incontinence. It is not surprising that coughing may restrict social interactions for these patients [61]. Another common effect of IPF dyspnea can have significant consequences on a patient’s quality of life, leading to difficulties in performing daily activities and affecting their psychosocial and economic well-being. Dyspnea has been independently connected to depression [79].

Treatment

In the context of treating IPF, two antifibrotic therapies, nintedanib and pirfenidone, have been approved to slow the decline in lung function and decrease the risk of acute respiratory deterioration. While clinical trials have not shown a reduction in mortality, pooled data and observational studies suggest that antifibrotic therapies can improve life expectancy of patients suffering from IPF [95]. Even though the side effects of both medication can be severe, most individuals can tolerate antifibrotic therapy, and dose adjustment can reduce side effects without compromising efficacy. A holistic approach to care that includes symptom management and supportive care tailored to the individual's needs is also beneficial for patients with IPF. The only ”cure” for IPF is a single or double lungs transplantation with healthy lungs. Because of several reasons such as a maximum age limit, limited life expectancy after transplantation and a shortage of donor lungs not available or chosen as treatment for all patients.

The unpredictable nature of the progression of IPF and the risk of sudden acceleration resulting in life-threatening exacerbation often leads to IPF patients dying in hospitals, receiving life-prolonging interventions. On the other hand, referring IPF patients to palliative care too early when the disease is still mild may cause short-term declines in their quality of life. Despite this, there is evidence to suggest that palliative care may help alleviate respiratory symptoms and improve the quality of life of patients with advanced IPF [61].

1.3. Research Gap

Looking at all the related work we can identify that in the up-and-coming field of XAI there are still major unsolved problems which severely limit the applicability of AI in high-stake domains such as healthcare. We will begin by identifying what the problem is and how this research gap will be closed. After which we will state which contributions will be made in this thesis and how we will address such a gap.

The Role of AI in Healthcare

Doctors are limited in their ability to recognise patterns, aggregate, and distil large amounts of disparate data into a clear, evidence-based treatment plan [107]. The large amount of data that is available now can be used in combination with machine learning algorithms to create the simple tools needed [124] that help diagnose or judge the severity of different respiratory conditions.

Indeed, several works have studied the intersection between AI research and the healthcare domain [6]. But despite promising applications of AI in respiratory medicine [53] and in specific overlapping fields such as radiography which were quick to adopt AI [6], AI in medicine has yet to propagate to the clinical setting [65]. There are currently limited examples of such techniques being successfully deployed into clinical practice. Respiratory medicine is no exception [53]. It is therefore important to consider how the use of AI fits into clinical workflow [107] and to look at the factors which prevent these systems from being adopted. Which include those intrinsic to the science of machine learning, logistical difficulties in implementation, and consideration of other barriers to adoption as well as of the necessary sociocultural or pathway changes [65]. Furthermore some parts of the respiratory medicine field are still very underrepresented.

There remains an enormous potential for AI to embrace domains outside of imaging such as pulmonary function tests and physiological biosignals [53]. But also time consuming administrational tasks such as writing responses to patient questions could be semi-automated by AI saving doctors valuable time, while they are considered to be more emphatic towards patients [11]. Overall, AI is likely to play a key role in aiding clinicians in the diagnosis [6] and differentiation of respiratory diseases in the future, and it will be exciting to see the benefits that arise for the patients and doctors from its use in everyday clinical practice [107].

Current Problems with AI in Healthcare

The current generation of AI algorithms in the medical domain suffer from problems of reliability caused by the lack of robustness [16] and transparency [128]. In many cases, the most powerful machine learning techniques purchase diagnostic or predictive accuracy at the expense of our ability to access [91]. The features that are used to distinguish between data categories are not readily translated into verbal or visual 'rules' that a human can understand [53]. Which the European Commission Joint Research Centre sees as causing negative consequences for EU citizens and organisations as AI is starting to play a crucial part in systems for decision-making and other processes [129]. A number of real-life high-profile cases have shown these concerns to be valid as machine learning can learn dangerous rules from the training data [9].

XAI as Solution for Problems with AI in Healthcare

Among other requirements, scholars and medical practitioners have highlighted the need to provide explanations for AI-based decision systems [72]. To develop methods for improved interpretability of machine learning predictions. If these goals can be achieved, the benefits for patients are likely to be transformational [65].

At the same time, while research is being carried out in the field of XAI, this is still mostly algorithmic-centred and does not yet meet the bar for deployment in the real-world [86, 88]. Many of the algorithms used are also not used by end users in practise [18]. Explainability is a social process [109] — it relies on the perceptions of the reader — but the recipients of explanations and their needs are seldom take into account when designing and proposing XAI solutions. As different users require different forms of explanation in different contexts [101]. Research has shown the limitations of existing XAI approaches both technically and from a human perspective looking at concepts such as understandability and actionability [87]. And this shows as clinicians' views sometimes differ from existing notions of explainability [146].

How to best design explainable AI systems is a non-trivial problem [98]. End-users carrying out a task have a goal or purpose behind trying to achieve it. To do that, they may need or desire to have some kind of additional information to base their decision on. Different properties of explainability can be more or less important depending on the reason to demand explainability [98]. In healthcare these reasons can sometimes be very consequential for a patients health. On top of that, the way the information is presented and how the presentation supports end-users in their activities plays an important role specific to the medical domain [120]. In this sense, the emerging field of human-centered XAI, which focuses on answering the question ”who” the explanations are for, plays a crucial role in helping to answer these questions. Human centered XAI recommends to broaden the view on XAI to incorporate methods and frameworks from social science [109] and design with the goal of shifting the current techno-centric paradigm of XAI toward creating user experiences for XAI. An experience that ensure humans' right to understand and contest AI decisions [45]. The suitability of explanations is question dependent and requires an understanding of user questions for a specific AI application [86] and promotes a question-driven framework to embody these needs [88].

Doctors and health care professionals in general have a limited knowledge of AI principles [158, 1]. Furthermore, most doctors are unaware of the advantages and most common challenges to artificial intelligence applications in the health sector [1] and are less aware or concerned about higher level issues in AI use such as fairness, bias, and health inequalities [100]. Many papers suggest schooling the doctors in principles of AI, however the busy clinical schedule does not always allow for this [58]. Most explainability algorithms are in practise only used by machine learning developers [18] and even machine learning practitioners with different levels of experience with computer vision can have a hard time envisioning uses of certain explanations [12].

Addressing the Research Gap

There is overall distinct lack of application of XAI in the context of healthcare AI systems and, in particular, a lack of user studies exploring the needs of clinicians [9]. Therefore a different approach will be taken, which takes into account the low level of knowledge about AI and explainability that doctors posses. But does include medical doctors, consider that they must participate in the design process [100]. We will complete a user study of clinicians in the field of pulmonology to take away one of the major barriers of implementing XAI in clinical practice and attempt to pave the way for future researchers to achieve a higher adoption rate of AI in the clinical setting.

1.4. Research Questions and Contributions

Research Questions

For this reasons, we seek to contribute to the existing body of work around human-centred XAI by focusing on the information pulmonologists desire or need to see in explanations during the care they provide for patients. This is achieved by taking into consideration the patient journey. The patient journey refers to the entire process that a patient undergoes when seeking healthcare, from the initial symptoms or concerns that prompt them to seek medical attention, through diagnosis, treatment, and follow-up care. It encompasses various touchpoints such as medical consultations, diagnostic tests, treatment plans, medication administration, and monitoring progress. By considering the patient journey, we aim to gain a comprehensive understanding of the specific needs and preferences of pulmonologists at different stages of patient care, ensuring that the explanations provided by XAI systems align with their requirements throughout this continuum of care. Specifically, in this study, we interview pulmonologists and medical PhD students to investigate the following main research question.

Main RQ — *How do pulmonologists' needs and intents shape the design of XAI solutions?*

Because of the complexity of the field and lack of common consensus about the exact definition of explainability the main research question is subdivided into three smaller research questions which are easier to investigate. The split is taken from Xu et al. [156] who split their research question into three simple sub-questions: when, what and how. The following questions can be read along the same line of reasoning:

RQ1 — *Where in the patient journey are moments for clinicians where XAI can be helpful?*

RQ2 — *What information do doctors seek in explanations?*

RQ3 — *Under which conditions do doctors engage with explanations?*

These sub-research questions will be separately analysed and then combined in the discussion to answer the main research question. Where RQ1 looks at when they need explainability, RQ2 looks at what they want from the explainability and RQ3 looks at how they engage with it, if at all.

Contributions

The main contributions this work makes are related to finding opportunities in the respiratory medicine domain and about finding the parameters that are related to creating a Human-centred XAI experience for the practitioners in this field. With the key ambition of trying to work towards enabling future research into lowering the barrier for the implementation of AI applications in the medical domain. The exact contributions will be classified as follows:

Contribution 1 — *Opportunity finding: where and how AI and explanations could be applied in the pulmonology domain.*

Contribution 2 — *Identifying: Needs, wants, and goals of medical staff about explanations.*

1.5. Methodology

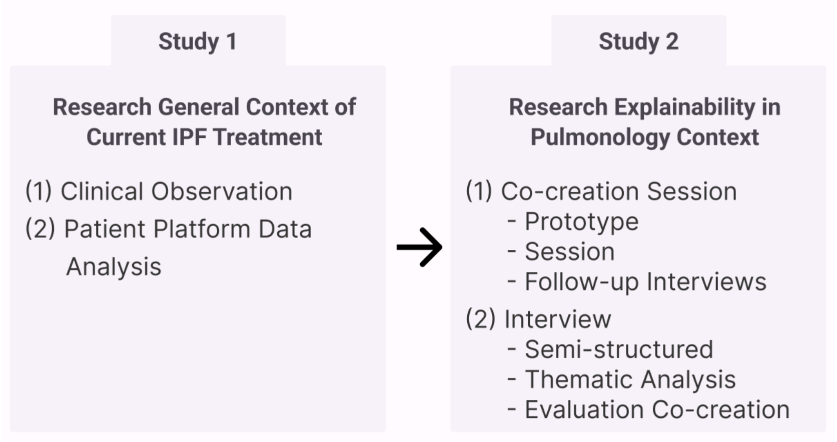


Figure 1.3: Flow of the research into the explainability context.

The following chapter aims to provide a detailed account of the research approach and a general view of what each method employed is about, including the data collection methods, analysis techniques, and validation processes employed throughout the studies. By following this systematic research flow, we aim to contribute valuable insights into the explainability needs. The exact details about each method employed can be found in the relevant sections.

The research is structured into two primary studies, each contributing to the overall understanding of XAI in pulmonology. Study 1 focuses on examining the general context of current IPF treatment. It involves clinical observations of pulmonologists in their practice to gain firsthand insights into patient care and identify the real challenges faced by medical professionals. Additionally, a thorough analysis of platform data is conducted to identify the specific needs and gather valuable insights into patients with pulmonary fibrosis.

Study 2 delves into the explainability context within pulmonology. This part of the research encompasses various components. First, a prototype is developed based on the knowledge and insights gathered from Study 1, serving as a foundation for subsequent stages. The prototype is then utilised in a multidisciplinary co-creation session, where pulmonologists and experts collaborate to gather initial feedback on explainability usage and explore the specific needs and requirements of pulmonologists regarding XAI. This session aims to generate valuable insights and provide a platform for interdisciplinary discussions. To further deepen the understanding of pulmonologists' perspectives, semi-structured interviews are conducted. These interviews serve as a means to obtain additional input on the utilisation of explainability and uncover any additional needs. Thematic analysis methods are applied to analyse the collected interview data systematically, extracting meaningful insights. The results obtained from the thematic analysis are then validated through a co-creation session involving diverse experts who review and provide feedback on the findings.

The remainder of the thesis is structured following the structure of the research performed. The research flow can be seen in Figure 1.3, and is outlined as follows:

Study 1: Research into the General Context of the Current IPF Treatment

The first part is mainly connected to the first and second chapters: First the use case is provided in section 3.1 which will be used to perform the research. The medical background and context will be presented in chapter 3 to give the reader an overview of the research performed to generate relevant medical background information for the second study.

1. **Clinical Observation of Subjects**: The researchers observe pulmonologists in their clinical practice to gain insights into the care provided and identify the real problems faced by these professionals. This observation and its results can be found in section 3.1.

2. **Platform Data Analysis**: Big data research is performed on platform data to identify the actual needs of patients with pulmonary fibrosis and extract valuable insights section 3.2.

The main goal of this part is to gain an understanding of the current care processes in pulmonology find the problems that clinical staff deals with now, understanding the patient journey, and generating data and artefacts that will be utilised in the later stages of the research in study 2.

Study 2: Research into the Explainability Context in the Pulmonology Context

The second part is mainly focused on the actual collection of the scientific data and connected to the following sections: The methodology of the research together with in-deep discussion of the research method can be found in section 4.3. The outcomes of this study are presented in section 4.4.

1. **Co-creation**

   Prototype Creation: A prototype is developed based on the gathered knowledge and insights from the previous study. This prototype will be utilised in subsequent stages of the research subsection 4.1.2.

   Co-Creation Session: The prototype is utilised in a multidisciplinary co-creation session, where pulmonologists and other experts collaborate. This session aims to gather initial feedback on how explainability is used and explore the needs and requirements of pulmonologists regarding XAI subsection 4.1.3.

   Follow-Up Interviews: After the co-creation session, follow-up interviews are conducted to further elaborate on the needs identified during the session and gather more in-depth insights from the participating pulmonologists subsection 4.1.4.

2. **Interview**

   Semi-Structured Interviews: Additional semi-structured interviews are carried out with a number of pulmonologists to obtain further input on their utilisation of explainability and identify any additional needs. The methodology for this can be found in section 4.2.

   Thematic Analysis: The interviews conducted in the previous section are subjected to a systemic analysis. Thematic analysis methods are employed to analyse the combined output of the interviews and extract meaningful insights. The methodology for this can be found in subsection 4.3.1 and the results in subsection 4.3.3.

The main goal of this part is to use the artefacts during the co-creation session and subsequent interviews to understand how pulmonologists use explainability and determine their specific needs. The feedback gathered will be used to improve the prototype and refine the journey map, ensuring more relevant and valid tools for the research questions. The other parts are to validate the research findings and opportunities for explainability within the medical domain of PF care. It aims to answer the research questions regarding pulmonologists' needs and intents in shaping the design of XAI solutions. Additionally, the research contributes to identifying opportunities for applying AI and explanations in the field of pulmonology and understanding the needs, wants, and goals of medical staff regarding explanations. By following this research flow, the thesis aims to provide valuable insights into the explainability needs in pulmonology and contribute to the development of XAI solutions that align with the requirements of pulmonologists.

To increase the validity of the research and because this is often a limiting factor in the usefulness of previous research all relevant materials relating to the co-creation session and interviews can be found in the Appendix. With Appendix A for the co-creation session , Appendix B containing material for the follow-up interviews, Appendix C consisting of the material for the semi-structured interviews and Appendix D containing the final checklist helping the future development of XAI for healthcare.

1.6. Connection with Erasmus MC

This research is closely linked to the expertise and experience of the Erasmus University Medical Center in Rotterdam. All of the participants were linked to this hospital and all events, such as the observations, took place at their premises. By establishing a connection with Erasmus MC and their expertise in healthcare, this research aims to bridge the gap between explainable artificial intelligence and the specific needs and challenges faced by pulmonologists in their own clinical setting. By collaborating with medical professionals at Erasmus MC, valuable insights can be gained into the practical implementation of explainability methods, in the example context of IPF care, ensuring that in the future developed solutions align with the requirements and goals of pulmonologists and could ultimately contribute to improving patient outcomes.

The respiratory department

The Erasmus University Medical Center based in Rotterdam, the Netherlands, is affiliated with Erasmus University and home to its faculty of medicine. It is a government owned scientific University Medical Center and performs research in numerous health related directions. The respiratory medicine research from the medical centre is highly ranked within western Europe [1].

In the Netherlands there are around 3200 known people suffering from lung fibrosis[2]. Because of the multidisciplinary and complex care for the patients suffering from IPF most of these patients are under treatment in one of the three national expertise centres. These are large university hospitals The Erasmus Medical Centre is one of the 3 expertise centres in the Netherlands for this type of lung diseases. Even though the disease does not really fluctuate between countries, the availability of treatment and diagnosis methods vary even in Europe [69].

Regular care for IPF patients

The regular care path for patients in the Erasmus MC with IPF is described in a document, branched from the regular interstitial lung disease care path, containing information about the registration, diagnosis, first consult, medication, multidisciplinary meetings and follow-up treatment regarding new patients with (suspected) IPF. The document describes the standardised care that a patient should receive. However, because of the complexity and heterogeneity of the disease together with the fact that the disease does not only effect the patient at the hospital but also their life at home, the actual events a patient experiences can be different. From one of the doctors perspective the regular care path at the Erasmus MC and comparably at other centres in the Netherlands goes approximately as follows:

”The patient comes in, you see him then you do a history of the patient, physical examination depending on what has already been done. Following that you discuss the patient in an MDO where you see if medication should be started. You discuss that with the patient and talk about the different side effects of the medication and the personal wishes of the patient. You start the patient on the medication plan, then you follow up with that patient concerning the side effects he suffers, perform lab checks, lung function tests. The patient comes back once in a while, in the Erasmus MC that time is around 3 months, every time you evaluate with how the patient is doing. During these visits you talk about it being a fatal progressive disease and what that patient wants at the end of his life. During this you change the medication based on the individual patient, not only in the medication to treat the disease but also to fight the symptoms, that is not really predictable in advance that is different with every patient.”

[1] Erasmus MC ranking of respiratory medicine in western Europe www.scimagoir.com

[2] Number of people suffering from Pulmonary Fibrosis Netherlands - Numbers taken from the Dutch patient organisation Longfonds www.longfonds.nl/longziekten/longfibrose

1.7. Outline

The remainder of the thesis will be structured as follows. In chapter 2, we will review the related work, providing an overview of existing research and literature on the topic. Moving on to chapter 3, we will delve into Study 1, which focuses on the research conducted in the general context. Within this study, we will first discuss the clinical observation phase, including the observation sessions and the results obtained. This will be followed by the patient platform data analysis, where we will outline the process of scraping patient experiences, clustering the data using BERTopic, and transforming the clusters into a patient journey map. Finally, we will present the results derived from the patient platform data analysis, concluding the section on the research in the general context. Chapter 4 will introduce Study 2, which delves into the research on explainability in the pulmonology context. This study begins with a preliminary study that involves developing a functioning explainability prototype based on the knowledge and insights gathered from the previous study. We will then describe the co-creation session, where pulmonologists and other experts collaborate to provide initial feedback on the use of explainability and explore the needs and requirements of pulmonologists regarding explainable artificial intelligence. Follow-up interviews will be conducted to further elaborate on the identified needs, followed by a discussion on the methodology employed for the semi-structured interviews. Thematic analysis will be performed on the interview data, and the results will be presented. The validation of the findings will be carried out through a co-creation session involving diverse experts. We will conclude this section by summarising the research on explainability in the pulmonology context. The discussion and conclusion will be covered in chapter 5 and chapter 6. Here, we will engage in a comprehensive discussion of the findings from both studies and their implications. We will also provide a concise conclusion, highlighting the key findings, contributions of the research, and any identified limitations with corresponding recommendations for future studies. In the conclusion we will also present a framework for adopting XAI in pulmonology, which aims to provide guidance on incorporating explainable AI solutions into the field. Appendices will be included, providing additional information such as the questions used for co-creation and interviews, design prompts, and details about the co-creation validation session.

2

Related Work

Large amount of research has already been performed into the field of explainability. In this section the relevant directions within this research field will be described together with relevant literature from each of those directions. First the algorithms that make up the basis of the XAI field will be described together with how to decide which one to use and how they are used in practise. Then a quick overview will be given on how these methods are used in the medical domain and human-centered frameworks will be described that use their focus on the end-users to create systems that try to prevent the barriers of adoption that other implementations pose. Finally, relevant literature will be provided that uses different methods for finding opportunities for XAI in the medical domain. The relevant literature and basic overview of the different directions within the XAI field provide a basis for the research gap that will be presented in the next section.

Explainable AI algorithms in general

XAI algorithms aim to provide human-understandable explanations for decisions of complex machine learning models. Many types of novel algorithms have been proposed in the field, such as: LIME, Anchors, Integrated Gradients, Causal Models and many more [60]. The XAI algorithms that exist are very diverse in the characteristics they possess. Some examples of the characteristics they diverge on are ante-hoc or post-hoc, global or local, vision or text [38]. Other categorisations are for example hierarchical systems and also look more into the input and output format [149]. However, no specific commonly accepted categorisation of the methods exists. Each algorithm approaches the objective using different techniques and methods and has different limitations, strengths and weaknesses. But in all of them the objective is the same, provide a better understanding and control of the complex model they try to explain.

The number of XAI algorithms has grown so rapidly in the last few years that some researchers looked into creating benchmarks making it easier to pick which of the algorithms to use in specific use cases [15]. However, in practise researchers compare the different methods themselves for their specific use cases. Examples of this are Moscato et al. who looked into comparing state-of-the-art methods for explaining the features used for credit risk assessment [111] while Lee et al. compared different XAI visualisation techniques for explaining defect classification to domain experts [81]. McDermid et al. connect different explainability methods to different types of stakeholders depending on their needs and ethics [101].

The methods used to compare the explanations generating algorithms differs in each of these papers and in Lee et al. is performed by having a simplification of the real end users. This is primarily caused by the absence of standardised evaluation methods [98]. Other research sees explanations as playing a vital role in improving human-agent interactions and performance in decision support systems and claim that the metric should measure the increased performance of collaborating with the XAI [135]. Instead of picking one, the output of multiple XAI algorithms can be used simultaneously. Krishna et al. [71] looks at what to do when these XAI algorithms disagree with each other. Bhatt et al. [18] look into how the algorithms are actually used in practise. They find that explanations almost exclusively used by internal stakeholders rather then external ones. Motivated by this conclusion they try to find what the limitations are from current explainability techniques that prevent them from being used by external end users. Colin et al. [33] conducted a study to evaluate the usefulness of explainability methods in real-world scenarios. The results demonstrated that explainability methods have shown progress in assisting human users in bias detection and identifying new strategies. However, these methods struggled to provide meaningful insights in understanding failure cases.

Open problems for explainable AI algorithms include: a remaining lack of agreement on what explainability means, no clear guidance how to choose amongst explainable AI methods, and as stated previously the absence of standardised evaluation methods [98] and importantly with a limited research emphasis being placed on supporting alternate domains [120]. The effects of this are apparent when looking at the real world use of explainability methods, were there is no focus on the end users but explanations are mostly used by machine learning developers [18]. Furthermore, research has suggested that users often over-rely on system suggestions - even if the suggestions are wrong. Providing explanations could potentially mitigate misplaced trust in the system and over-reliance. [25]. Other research indicates that data scientists over-trust and misuse interpretability tools. Furthermore, few of our participants were able to accurately describe the visualisations output by these tools [64].

XAI in the medical domain

Deep learning models have made significant advancements in healthcare, especially in diagnostics and surgery, outperforming doctors in some diagnosis tasks. However, the lack of transparency in these models poses challenges in explaining their results and integrating them effectively into clinical settings [17]. A survey looking at healthcare based XAI systems which were trained on electronic health record data [121] found that researchers primarily employ ”if-then” rules 28% to improve the interpretability of complex machine learning methods. Another prevalent approach is to enhance the performance of less complex ML methods while preserving their interpretability through optimisation techniques 21%. Additionally, dimensionality reduction techniques 19% are also frequently utilised in this context. They found however that many of the methods were not evaluated, which was likely caused by the absence of a consensus on the definition of interpretability contributing to the lack of a standardised approach for assessing the outcomes of XAI methods. From these unevaluated methods they still draw recommendations for future XAI methods in healthcare.

To investigate the potential of XAI in the medical domain, most researchers use specific use cases, to explore the use of XAI algorithms and models in that specific domain. These use cases provide a practical and focused approach to researching XAI in healthcare, allowing researchers to identify the benefits, limitations, and ethical implications of using XAI in real-world settings. From this more focused research a broader conclusion about XAI in the medical domain can be taken. This section will describe some of the relevant use cases that are used in the past to research XAI in the medical domain.

The general XAI algorithms are not always adequate to use as healthcare data can sometimes be sequential or ontology-linked data. Therefore Panigutti et al. [115] came up with a local model-agnostic explanations XAI technique applicable to this specific type of sequential ontology-linked data. But also for non-sequential and ontology-linked data not all explainability methods are relevant to the applications in the medical domain. Therefore distinctions have to be made in the methods which ones are relevant. One example of this is Poceviciute et al. [126] where they look at the field of digital pathology, the digitisation of microscopy images, which is a medical imaging sub-discipline where gigapixel images are analysed in order to find out many different properties of the underlying sample. They present the case for the use of XAI in that specific field and list all state-of-the-art methods that they consider relevant for the image domain and give their advantages and drawbacks. They, however, don’t state which or how to pick which method to use in what relevant case or review any method with real pathologists.

The use of XAI in the medical domain is researched in different ways. Signoroni et al. [140] developed a deep learning architecture for predicting a semi-quantitative score which conveys the degree of lung compromise. The model was created together with Italian hospitals which lead to the inclusion of highly resolved explainability maps to help visualise the network activity on the lung areas. The model is currently undergoing an experimental deployment in the two Diagnostic Radiology Units in an Italian hospital.

Lundberg et al. [93] look at explainable machine learning predictions to help anaesthesiologists prevent hypoxemia during surgery with real time explainable AI. Where the explainability is in the form of showing features that modify the chance that a patient will have hypoxemia. However they explainability method they used was not described as being thought of together with the medical staff and they did not test if the explainability actually helps the anaesthesiologists improve their ability to prevent hypoxemia. This is a common pitfall in the explainability literature where it is quickly assumed that the presence of explainability data means that end users will understand the AI method and that the benefits of fully explainable AI methods are automatically reached. This can also be seen in Davagdorj et al. [36] who proposes a framework to enhance the interpretability of a deep neural network model for predicting non-communicable diseases under patients. The framework utilises the Shapley values approach to provide explanations from both a population-based global perspective and a human-centered local explainability perspective. The importance scores of features in the DNN model construction are used to determine the most significant factors for non-communicable diseases prediction. Peng et al. [122] use the same type of explainability and evaluation of explainability method for deterioration risk prediction of hepatitis patients. Du et al. [40] again employed Shapley values and utilised the same explanation methods validation to explain the automatic prediction of gestational diabetes mellitus. However, to improve clinical usability, they conducted feature selection in collaboration with clinical experts. Their model included a small set of features, 4 or 5, to minimise data entry and make the system faster and easier to use. Education was removed as feature from the models based on expert consultation to ensure clarity and practicality. In all these cases the explainability method is chosen without considering other types of explainability methods and evaluated based on if the important decision features found match with the existing literature. Relying solely on theoretical benchmarks to evaluate attribution methods is risky as they are disconnected from human involvement [33]. Instead it is important to measure the performance of doctors working together with the AI system.

Another use case that is used for researching XAI in the medical domain includes Metta et al. [108] which looks at exemplars and counterexemplars explanations for Image Classifiers with the task of skin lesion labelling. This method, just like Lundberg et al., does not review the actual performance increase or any other factor from the use of the explainability with the relevant end users.

Human-centered explainable AI research frameworks

Most popular methods in the XAI field are model-agnostic and application-agnostic which means that the methods are not tailoring their explanations to the end-users presented with the explanations. Despite the importance of human-centeredness, and the differences the end-users make in the characteristics of explainability, these methods are seldom explored in the context of a specific type of end-users. The failure to thoroughly explore the application of these methods with targeted end-users results in ineffective implementations of XAI techniques in scenarios where the limitations are neither well-understood nor adequately considered. Human-centered explainable AI research frameworks are methodologies or approaches for designing and developing AI systems that prioritise the needs and experiences of a specific sets of users and try to prevent these pit-falls by taking the end-users into account.

Kim et al. [66] used a mixed method study where they conducted interviews with users of a bird identification app and tried to find out about their XAI needs, uses, and perceptions. They found that users want practically useful information that can improve their collaboration with the AI, instead of technical system details. These include calibrating trust, improving their task skills and changing their behaviour to supply better inputs to the AI. As the study was almost entirely focused on the bird identification app users, the findings might not transfer to the medical domain.

In healthcare human-centeredness is done by involving the doctors, nurses or patients in the design procedure to ensure that the explainability is aligned with the users' goals and values, makes the AI model transparent in its decision-making, and can provide understandable and relevant explanations to users. It also tackles the problem in the XAI field that explainability methods are often not reviewed properly on their performance being used by the end users, as in these frameworks the end users are continuously involved in the design processes and evaluations are commonly performed.

Panigutti et al. [116] create human-centered explainable AI for doctors by using the collaborative approach of co-design. Co-design involves active participation from end-users throughout the design process to create solutions that better meet their needs and preferences. They use an iterative design approach of prototyping, testing and then redesigning their explainable AI system. The testing with real medical users showed that naive implementation of state-of-the-art methods leads to negative factors which could prevent end-users from using the system such as information overload and ill-suited explanations for the target group.

Finding opportunities for XAI in the medical domain

Exploratory surveys are an essential tool for researchers to identify opportunities for implementing explainable artificial intelligence in the medical domain. By conducting surveys, researchers can gather valuable insights and feedback from end-users, stakeholders, and experts, and understand the specific challenges, needs, and expectations.

Tonekaboni et al. [146] interviewed clinicians in two acute care settings – Intensive Care Unit and Emergency Department – to develop notions of explainability and identify their needs towards building reliable ML systems for their respective clinical practice. They used hypothetical Scenarios and asked more general questions to find aspects of explainability that catalyse building trust with the underlying AI models and they identify classes of explanations that clinicians identified as most relevant.

Cai et al. [27] look into using a prototype of a deep neural network predicting prostate cancer diagnosis and survey before, during and after to learn the types of information that pathologists desire about the AI assistant. They found that clinicians desired upfront information about basic, global properties of the model, such as its known strengths and limitations, its subjective point-of-view, and its overall design objective. Where participants compared these information needs to the collaborative mental models they develop of their medical colleagues when seeking a second opinion. Evans et al. [125] use a mixed-method study of surveying and interviewing pathologist showing them samples of state-of-the-art AI explainability techniques to find their XAI needs. The authors agree that most research into explainability has been to algorithm centric instead of user-centric, however then go on to present these algorithm centric methods to the medical staff. We take a different approach where we take care to prevent limiting doctors to existing algorithm centric explainability methods. Lakkaraju et al. use a half hour semi-structured interview to ask questions about explainability use and needs to 16 doctors who already use explainability in their workflow to diagnose machine learning models [75]. They conclude that these doctors think feature importance is important and would want to use a interactive explainability interface. The difference with our research however is that we want use explainability working together with clinical doctors, some who have no experience using explainability or have knowledge about AI. The fact that the doctors in [75] have experience with machine learning means that a large part of the interest into the working of models comes not from their medical side, but rather their AI developer roll.

The role of XAI in pulmonology

In the field of pulmonology AI is not yet implemented to its fullest potential. Previous research has shown that it still deals with several problems, such as opaque decision making and the need for clinical validation. Using XAI in pulmonology would have several advantages in the working together between pulmonologist and XAI.

Problems

The overall benefits of explainability still need to be proven in practice in the field of healthcare [98]. Opaque decisions are more common in medicine than critics realise [91]. While this debate of whether it is acceptable to use nontransparent algorithms for patient care is unsettled, it is notable that many aspects of the practice of medicine are unexplained, such as prescription of a drug without a known mechanism of action [147]. However, there cannot be exceptionalism for AI in medicine it requires rigorous studies, publication of the results in peer-reviewed journals,

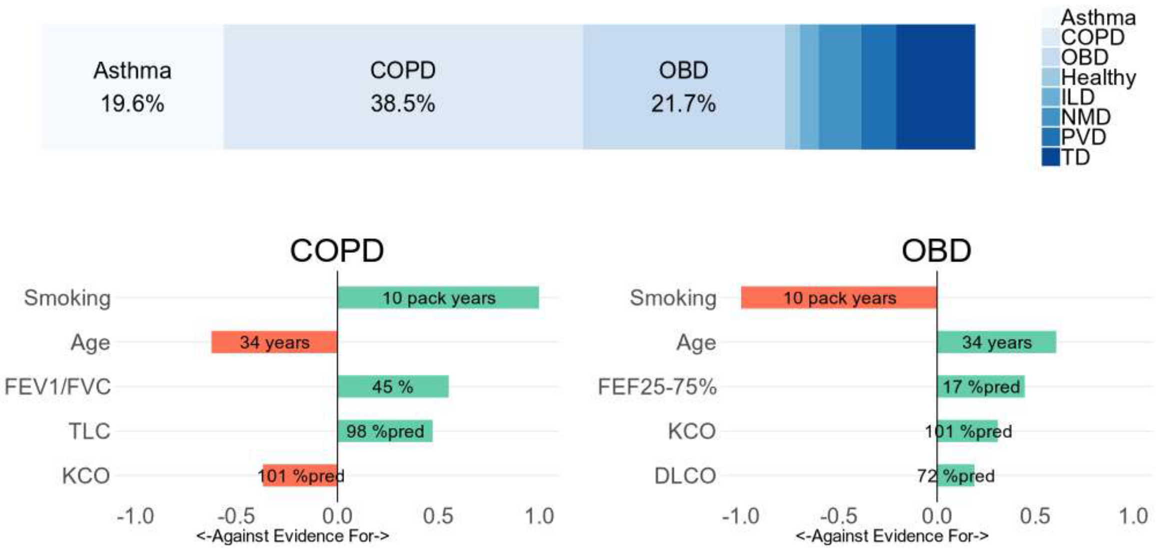


Figure 2.1: Explanation of diagnostic prediction of different lung diseases using Shapley values from Das et al. [35]

and clinical validation in a real-world environment, before roll-out and implementation in patient care [147]. Similarly to other medication of healthcare related devices.

Poursabzi-Sangdeh et al. found that using a clear model hampered participants' abilities to detect when the model had made a sizeable mistake, seemingly due to information overload caused by the amount of detail in front of them [127]. But that problem disappeared when giving the user more interactive information warning them. Results show that visual inspection of explanations alone can favour methods that may provide compelling pictures, but lack sensitivity to the model and the data generating process when looking at saliency maps [4] commonly used in the medical field.

Advantages

The advantages of using XAI in pulmonology have been underlined by Das et al. who used Shapley values as can be seen in Figure 2.1 to explain the diagnostic prediction of different lung diseases by an AI system [35]. They conclude that the doctors together with the explanations perform better then either system alone. Thus showing the positive effects that collaboration between doctors and AI systems have in cases where the AI is not a black box but rather can explain their decision making process. Their study also finds that collaboration between explainable artificial intelligence and pulmonologists improves the accuracy of pulmonary function test interpretation significantly. The explainability is performed by providing Shapley Values to the pulmonologist, which help to explain the decision-making process of a black-box model by assigning a numerical value to each feature based on its importance in the model's output. This research shows that the collaboration between XAI and pulmonologist can lead to better care for patients. This is the driving point behind our attempt to improve this collaboration.

3

Study 1: Research General Context

In this chapter, we describe our approach to researching the medical background and context of IPF. This chapter includes sections on observations from doctor-patient sessions, platform data analysis and patient journey mapping. In particular, we describe our use of clinical observations and the BERTopic framework for clustering patient experiences and the challenges we faced in turning those into topics that could be used in the patient journey map. By combining insights from medical literature, doctor-patient sessions, and patient experiences, we hope to gain a holistic understanding of IPF care that can inform us in the extraction of explainable AI needs from the pulmonology medical staff.

Our research approach to understanding the medical background and context of Idiopathic Pulmonary Fibrosis involves two key methods: clinical observations and patient platform analysis. These methods allow us to gain insights into the workflow of medical staff and the patient journey, respectively, with the aim of identifying the problems that doctors deal with and finding the real-life patient journey identifying possible opportunities for AI systems.

Clinical observation, a qualitative research method, involves recording the behaviours of medical staff in a real-world setting. By observing doctor-patient interactions and deviations from standard treatment protocols available from the hospital, we can identify underlying problems. These observations lay the foundation for creating interview questions, and further analysis in the following chapter. The non-reactive nature of the observation sessions ensures minimal interference with actual medical procedures, maintaining the natural behaviour of doctors and patients.

In addition to clinical observations, we employ patient platform analysis to gain further insights into patients' experiences and perspectives. By analysing online patient stories from community platforms, we will try to understand patients' coping strategies, concerns, barriers to health behaviour changes, and disease-related symptoms. Although online patient stories may suffer from selection bias and lack contextual cues, they provide valuable qualitative data that can supplement traditional sources. It's important to consider the strengths and limitations of online patient stories and use them in conjunction with other studies to gain a deeper understanding of patient behaviour.

By combining insights from clinical observations and patient platform analysis our research aims to gain a holistic understanding of IPF care. This understanding will inform the extraction of explainable AI needs from the pulmonology medical staff, described in the following chapter with the use of the insights and artefacts extracted. Ultimately, leading to the development of XAI solutions that address the identified problems and improve the working together between AI and medical staff. The knowledge extracted here will also help with holding the semi-structured interviews later as for holding a good interview knowledge of the subject is required by the interviewer [3].

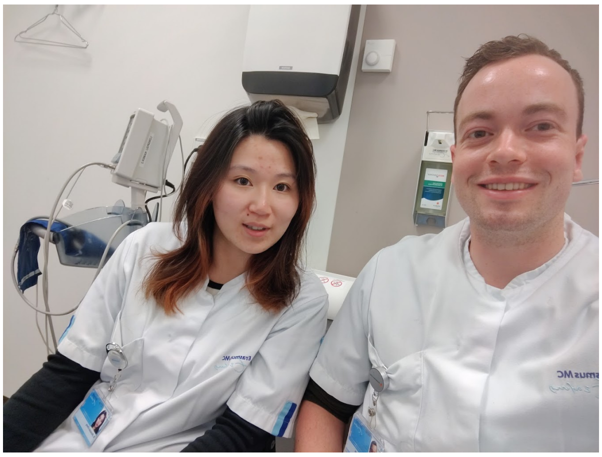


Figure 3.1: Image of one of the clinical observation sessions.

3.1. Clinical Observation

Naturalistic observation is a qualitative research method where the behaviours of the subjects, in this case medical staff, is recorded in a real world setting [8]. Observations can be used particularly well in a mixed-research study to get a basic understanding of the research field. They lay at the basis for creating surveys or interview questions [8]. Field notes are taken to record the actions and events that take place during the observation. As normal treatment protocols are known and doctors are thought those during their education, deviation of these protocols during the observation gives a clear indication that underlying problems exists. The type of deviations also give a indication into the direction that doctors see a solution. This type of problem finding is not unique and can be used for example in implicit requirement analysis [23].

3.1.1. Observation session

Two observation sessions were conducted to gain insights into the interactions between a pulmonologist and IPF patients during check-up consultations. The first observation session took place on December 14, 2022, and lasted for an entire day, while the second session occurred on May 15, 2023, lasting only half a day. The location of the observation sessions was the third floor of the RG building from Erasmus MC housing the pulmonology consultation and breathing test rooms. The number of patients seen on average during the observations was around 6-8, although the schedule for the doctor included seeing around 15.

The study was non-reactive where it is attempted to not influence the actions observed. This enabled us to study the problems that doctors suffer normally during their work. To not be too overt during the observation and to avoid disturbing actual medical procedures: doctor coats were worn by the observers and they sat unobtrusively in the back of the room during the observation. The presence of the observers was announced by the medical staff in advance and permission was asked to be present during their check-ups or lung function tests from the patients and their care-givers. In all cases they agreed to let the observers be present and they did not seem to be distracted by the presence of them, as they showed emotions and asked questions to the doctor that seemed to denote expected behaviour for patients.

Types of Patient Interactions Observed

During these observation sessions, various types of consultations were observed, including in-person consultations between patients and pulmonologists, interactions between patients and specialised nurses, phone consultations between pulmonologists and patients, and extended breathing tests conducted by specialised nurses. The patient population encountered during these sessions varied significantly, with some patients being in the advanced stages of the disease while others were still in relatively good health.

The interactions between the healthcare professionals and the patients, particularly those with the specialised nurse, showed how they provided detailed information and explanations to address specific patient questions, such as the use of home oxygen. The check-up consultations with the pulmonologist showed how they make quick assessments every three months to evaluate stability, make adjustments to medication, and determine if any changes to the treatment plan are required. Additionally, the extended breathing tests provided us with valuable insights into the care provided at the hospital and helped to identify the real problems faced by these professionals.

Types of Check-ups

The most significant moments during contact with out-patient patients occur during the regular check-ups, which typically take place approximately every 3 months, depending on the patients' and primary doctors' schedules. There are two main types of regular check-ups: the in-person and the phone consultations. Patients might decide on doing their consultation by phone because of their health condition, distance to the hospital, fear of becoming sick when coming to the hospital and other reasons such as negative experiences with the hospital. The actual check-ups can differ depending on the number of patients that day, the patient themselves, doctors and their experience, and type of the patient's disease and progression of it. The description of the two types of check-ups can therefore more be seen as a general view of how they are carried out, on average.

In-person consultation

1. The doctor prepares for the patient by reading their test results and medical file on their computer. This can also involve calling other doctors, that might manage a different aspect of the disease, to discuss.
2. The doctor checks in the system if the patient is already waiting in the waiting room.
3. The doctor walks through the hallway to the waiting room and collects the patient and in most cases their caregiver is also present.
4. The doctor asks to the patient how it is going while walking back to the consultation room. In this time they already have a quick overview of the state of the patient based on what they say and how out of breath they are from the walking.
5. In some cases the doctor hooks the patient up to the blood oxygen saturation monitor and checks their oxygen level.
6. When the patient performed any tests before the consultation, which most of the time is the case, the doctor talks about those results with the patient and caregiver or other person who joined them during the check-up.
7. The doctor asks about the symptoms of the disease and for side-effects of the medication that they might suffer from.
8. Based on this the doctor might change the medications taken, dose the patient is on or make any other adjustments to their treatment plan. This can also involve thinking about new test to perform on the patient or starting palliative care.
9. The doctor asks the patient if they have any questions, some patients bring small lists so they can remember their questions.
10. The doctor tries to answer the patients questions in a full and emphatic way, while providing them accurate information about their current health status.
11. The doctor renews and adapts medication prescriptions for the patient to collect at a pharmacy of choice.
12. The doctor plans new check-ups and other appointments with the hospital with the patient and informs them about other more practical concerns hindering their visits to the hospital such as imminent road works on their route.
13. The doctor might inform them and ask them about participating in a study.
14. The doctor stands-up and lets the patients out of the consultation room.

15. The doctor immediately starts preparing on their computer for the next patient.

Phone consultation

The phone check-ups are very similar to the in-person check-ups as the doctor has the same objectives during both types of consultations. With the obvious difference that step 2, 3 and 14 are skipped, 4 and 5 happen over the phone instead and the fact that in some cases the doctor talks with only the caregiver instead of the patient.

These phone consultations seem to take on average shorter then the in-person ones and also have less interaction with the asking and answering of questions, a observation that is in line with the literature [90] which states that telemedicine consultations usually take shorter because of dissatisfaction of doctors. The reason that the consultations were not in-person might also influence the length. One patient during the observation of their consultation chose for a online consultation, likely because they were hesitant to receive further treatment from the specific doctor because of a conflict, also likely leading to a shorter consultation.

3.1.2. Results

From the several different patient-doctor events observed we can summarise common problems that were encountered by the health staff in during these events. Because of the low number of observations there can be no statistical relevance tied to these observation results, however they can be indicative of larger problems that medical staff deal with that could be possibly solved by future AI solutions. Additionally they are helpful in the research into the context that the explainability solutions have to work in.

Significant time pressure Throughout the first observation session, it became evident that pulmonologists faced significant time pressure, which sometimes hindered their ability to fully prepare for patient check-ups. This time constraint also led to difficulties in addressing all patient questions in a comprehensive and unhurried manner. This point was commonly mentioned by the medical staff and the time pressure could be observed in their actions, or sometimes in the lack of certain actions.

Examples of this include not having time to have any breaks between patients during the consultations even for drinking tea. The death, before the appointment, of a patient caused some time to be freed up during one observation, however that time was quickly filled up by other tasks such as a newer pulmonologist asking for a second opinion on their own patient. Other effects of time pressure is that during the emotional breakdown of a patient about the fear for his impending death, the doctor expertly calmed the patient by explaining what would happen in his last moments, however after the doctor mentioned they felt hurried sometimes to calm the patient to continue with the consultation, instead of being able to provide them with the time they wanted to help their emotions.

Communication between medical staff Medical staff needs to communicate all information extracted from the patient, all information told to the patient and decisions made, to other medical staff. This leads to a large amount of communication directly and indirectly with current and future relevant persons involved in the patients care. While the medical staff understands the importance of the communication, they do think it takes an excessive amount of time. As time is a resource that they have a lack of, the large time need of communication leads to the use of shortcuts and to frustration.

Examples of communication between medical staff are:

- Noting down of the consult in the patients file.
- Writing letters to stakeholders such as the patients general practitioner about subjects such as the life expectancy that was told to the patient so it matches what the patients got told in the hospital.
- Calling other doctors involved in the patients treatment when they need to discuss adjustments to the treatment plan.

The lack of time and the need to have extensive communication leads to doctors copy-pasting a lot of information from their own templates or previous consults. Furthermore, the excessive use of abbreviations could lead to problems. During the observation of the lung function test only halfway through the hour long testing the nurse specialist noticed that, after some small talk with the patient, that the patient was suffering not from the chronic condition she thought. Rather the patient was suffering from a completely other disease that has the same commonly used 3 letter abbreviation. Although in this case it did not lead to complications, in other situations it could cause unnecessary testing or other negative consequences. The need for highly skilled specialised doctors to perform manual tasks, such as sending basic information emails to GP's, led to some frustration as they did not see it as something that should be part of their function, which it was not in the past. However, since recent cost cutting measures it was decided their medical secretaries were not necessary, thus the task became theirs.

Problems planning tests Patients with PF often suffer from other comorbidities [80] causing the need for scheduling other consultations and tests which can have long waiting times. This can make the treatment of the patients more difficult as the quick deterioration of patients leads to time pressure in treating symptoms. Coping methods for this used by the medical staff include scheduling appointments when they assume they might be needed in the future, cancelling them when they are at that moment in the future not necessary.

Patients not remembering their own needs and questions Another main problem observed was the fact that patients had trouble remembering the questions they wanted to ask to the doctor. A common tip given by the doctors and patient organisations [1] is to prepare a list of questions in advance so they can ask the questions they want during the consultation. But even in instances were patients brought a list of their questions, they still in some cases forgot to ask them or knew they wanted to ask more but forgot what exactly. However, because of the direct contact they can have with the nurse specialist this is not a large problem as they can still ask their questions later by email or phone call.

Not being in their home environment Patients often want to lead the doctor to believe the state of their life and health is very positive, something that the measurements, such as blood oxygen, or advanced questioning about their life and health contradicted. The patients have, in most cases, a large respect for the doctor and, according to the medical staff they sometimes want to present themself doing better then they actually are doing. This might cause problems for doctors in accurately knowing all problems the patients might deal with, preventing them from giving helpful advise and treatment plan adaptions to deal with problems.

Conflict with patients The emotional nature of chronic disease can easily cause conflict between patients, caregivers and doctors. This can cause major obstructions for the doctors in providing medical care to that patient. In one instance, the doctor finding another severe health condition within a patient caused that patient and caregiver to become hostile against them and the hospital, not wanting to come into in-person consultations and being hesitant for receiving additional hospital care.

The observations performed shed light on the challenges faced by pulmonologists in terms of time management, patient communication, administrative tasks, and getting an accurate image of the state of their patients and their needs. These results will help shape the co-creation session that will be described in the following chapter. Therefore contributing to the identification of specific areas where XAI systems could potentially offer support and enhance the efficiency and effectiveness of pulmonologists' work in delivering quality care to IPF patients.

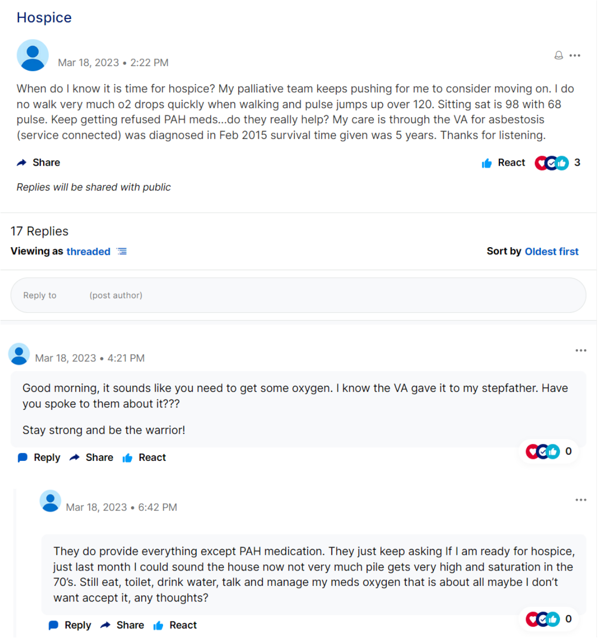
[1] Advise preparing a list of questions for consultations by PF patient organisation www.longfibrose.nl/zorg/gesprek-met-uw-arts/

3.2. Patient Platform Data Analysis

Through analysis of online patient stories from community platforms, researchers can gain insight into patients' coping strategies, concerns, barriers to health behaviour changes, and disease-related symptoms. These platforms provide a large-scale database with tens of thousands of first-hand patient stories about their care paths and illness experiences, offering an opportunity to supplement traditional data sources with qualitative data [63].

However, it's important to note that online patient stories may not represent the entire patient population and can suffer from selection bias. Patients who choose to share their experiences online may not be representative of the broader population and their stories may be influenced by previous content. Additionally, online stories may lack contextual cues, such as socio-economic background or emotional/situational states, unless users reveal such information online.

Despite these limitations, online patient stories can provide valuable insight into the patient experience and should be used in conjunction with other studies that employ more systematic and stratified sampling methods to reduce bias. By carefully considering the strengths and limitations of online patient stories, researchers can gain a deeper understanding of patient behaviour and inform the development of patient-centric healthcare services.



Hospice

Mar 18, 2023 • 2:22 PM

When do I know it is time for hospice? My palliative team keeps pushing for me to consider moving on. I do no walk very much o2 drops quickly when walking and pulse jumps up over 120. Sitting sat is 98 with 68 pulse. Keep getting refused PAH meds...do they really help? My care is through the VA for asbestosis (service connected) was diagnosed in Feb 2015 survival time given was 5 years. Thanks for listening.

Share React 3

Replies will be shared with public

17 Replies

Viewing as threaded Sort by Oldest first

Reply to (post author)

Mar 18, 2023 • 4:21 PM

Good morning, it sounds like you need to get some oxygen. I know the VA gave it to my stepfather. Have you spoke to them about it???

Stay strong and be the warrior!

Reply Share React 0

Mar 18, 2023 • 6:42 PM

They do provide everything except PAH medication. They just keep asking If I am ready for hospice, just last month I could sound the house now not very much pile gets very high and saturation in the 70's. Still eat, toilet, drink water, talk and manage my meds oxygen that is about all maybe I don't want accept it, any thoughts?

Reply Share React 0

Figure 3.2: Example of a patient post on the Inspire platform, with two other forum users answering the original question.

3.2.1. Scraping the Patient Experiences

The Patient Platform - Inspire

A large number of patients indicate that they use online patient discussion platforms to search for information about their disease [96]. One of the larger open patient discussion platforms that exists is Inspire [2]. With around 35.000 community forum participants, of which most are patients, posting their experiences and questions about their own disease trajectory. The forum is well moderated to prevent harmful information or spam messages to appear within the posts. Patients have been posting for more than 10 years on the platform, resulting in around 140.000 posts. The platform is set up together with the American Lung Association, all posts are written in English and a large percentage of the posts is about the disease trajectory of patients that life in the United States. An example of a patient post, together with other forum users answering them can be seen in Figure 3.2. The question the patient asks in the post is regarding if the advise their healthcare provider is wise to follow, which is a concern we more often encountered in post of patients on the platform.

[2] Inspire website https://www.inspire.com/groups/living-with-pulmonary-fibrosis/

The Inspire platform was chosen to be used for the data analysis as it is the largest open source platform and it is well moderated. Another advantage is that the platform has different sections for patients and caregivers. Therefore it is easier to only look at the journey of the patients when there are less posts from the perspective of the caregiver between, which could be hard to separate.

Scraping and Cleaning Data

As no API was available to us it was chosen to scrape the platform data. In order to use all information from the patients on the platform both top level posts and comments to those post where scraped. For ethical reasons it was chosen to remove all personal information such as email addresses, names, geographical names and specific hospitals. This information was automatically replaced by generic tags such as *<Location>* this was carried out in combination with regular expressions and BERT based named entity recognition. The use of these tags could influence the performance in two ways. It could increase the performance of the embedding model, because for our application different locations and people should be considered identical, or it could confuse the model and make it embed the documents further away than wanted. No easy method for testing this performance for all possible use cases is present, and therefore we will consider this as a necessity which we can't change, taking any performance hit as unavoidable.

To standardise the data further monetary units, date & times, medicine dosages where normalised, special characters and smileys where removed and medical abbreviations were expanded. Then everything was converted to lowercase. Because the forum has strictly controlled community guidelines all post where in English and no spam or large quantities of misinformation could be identified. Because BERT embeddings will be used in further steps no further preprocessing is necessary as the performance difference of preprocessing such as stemming is insignificant when using this type of embedding.

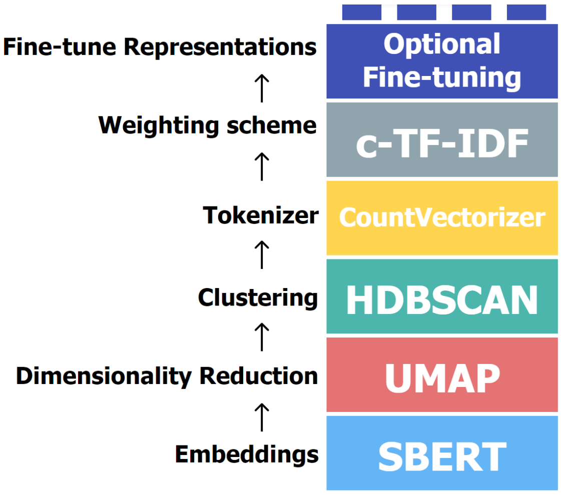


Figure 3.3: Internal modularity and workings of the BERTopic framework - Image made by M. Grootendorst.

3.2.2. Clustering Using BERTopic

Document embeddings provide a viable alternative to traditional topic modeling such as LDA at lower complexity and runtime [139]. The utilisation of BERTopic [54] for analysing short and unstructured text is the most promising approach for embedding-based topic modeling to find interesting topics [43]. BERTopic is a modular approach to document clustering based on Bidirectional Encoder Representations from Transformers embeddings. The simple description of BERTopic is a three-step process for generating topic representations. First, each document is transformed into its embedding representation using a pre-trained language model. Next, the dimensionality of the resulting embeddings is reduced to enhance the clustering process. Finally, topic representations are extracted from the clusters of documents using a custom class-based variation of TF-IDF [54]. The modularity of the system can be seen in Figure 3.3 where each component can be swapped for a different method or implementation.

Choosing Embedding Model

Different types of embedding models and methods exist, to find the best one fitting the use case the following main decisions, summed up below, were made. They consist of finding out which underlying structure the embedding model has to have, how it was trained and what should be embedded to get the best possible topics for finding out the patients their experience suffering from IPF. The following sections describe why it was chosen to use the configuration of embedding model that was used.

Type of internal architecture In recent years, pre-training language models have greatly enhanced the accuracy of Natural Language Processing tasks. Song et al. [142] introduces a novel pre-training method called Masked and Permuted Language Modeling (MPNet), aiming to overcome limitations present in existing approaches. While masked language modeling in models like BERT effectively captures bidirectional context, it overlooks the dependency among masked tokens. XLNet addresses this issue through permuted language modeling, but lacks complete position information during autoregressive pre-training. To address these limitations and leverage the strengths of both masked language modeling and permuted language modeling, MPNet splits tokens in a sequence into non-predicted and predicted parts. By considering the dependency among predicted tokens through permuted language modeling, MPNet resolves the discrepancy between pre-training and fine-tuning stages. The average performance of MPNet in the task of sentence embeddings is higher then most other BERT and Roberta based models [3] and therefore the most recent MPNet based model *all-mpnet-base-v2* is used.

General or medically trained The problem with picking medically trained models that are fined tuned or trained on medical data is that IPF is a very small and specific disease. The model therefore embeds all the post mentioning IPF or its medication and direct side effects in a very small space. Furthermore, the posts on the patient platform often talk about the home situation and other factors that are not commonly talked about in medical papers, thus not working well with the pre-trained medical models. Therefore, after taking these points into account a more general model was picked. While using the non-medical model we found that it had no trouble with identifying similar post with medical terms as related, even when different brand names of a specific IPF medicine were used in each of those post.

Embedding sentence, paragraph or post Another decision that had to be made was what to embed, the post, paragraph or entire document. The problem with top level posts on the platform is that they often don't contain a specific question or answer but might describe the recent or full experiences that a patient dealt with. This means that multiple topics could be present in one document, which can reduce the effectiveness of the clustering as it is unclear which cluster those documents belong and including them makes the cluster less specific. The reverse is also true, with sentence embeddings there might not be a topic of interest discussed in each sentence. Rather, clusters might be formed depending on too general and trivial topics that are not relevant for doctors e.g. saying hello. Both types of sentence and posts embeddings were tested. Sentence embeddings deemed to give a too large number of topics that were not all relevant for doctors to have information about. Some examples of topic clusters generated for sentence embeddings were ”Wishing people merry Christmas”, ”Greetings”, ”Looking good” which are not interesting topics on the platform but are very common sentences in posts to appear near the beginning or ending. Paragraph embeddings could have solved some of these problems. However, because of the lack of appropriate use of paragraphs on the forum used, this would involve first separating separate topics of interest within each post. Even though this is possible to perform, as it would complicate the platform analysis further it was chosen not to dive deeper into this direction. Rather this is seen as a limitation of the research performed and post embeddings where chosen.

[3] Performance of different types of embedding models www.sbert.net/docs/pretrained_models.html

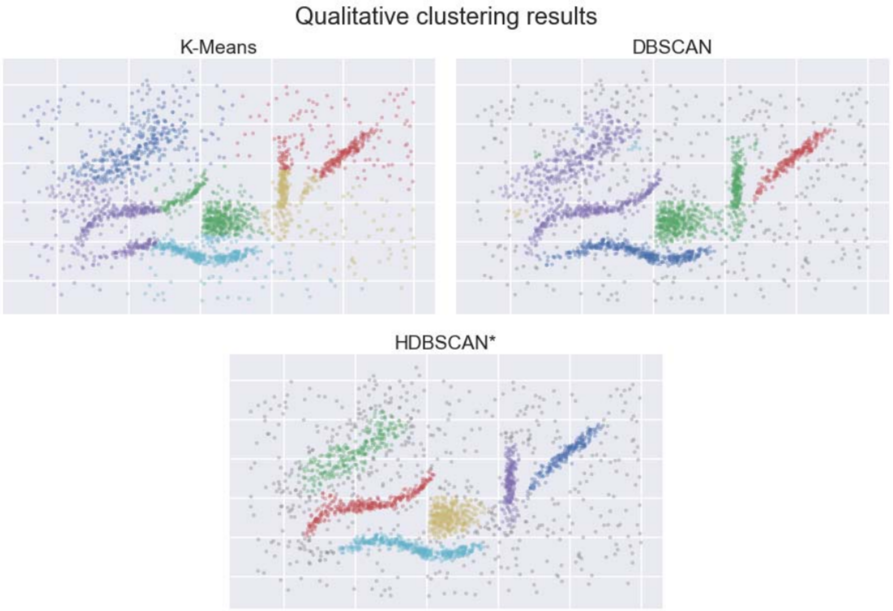


Figure 3.4: Colours denote cluster membership of nodes, the gray color is the noise cluster. Shows the difference between K-means, and 2 density-based clustering methods DBSCAN and HDBSCAN. Taken from [103]. The graph is a two dimensional projection of a higher dimensions space taken with arbitrary axis.

Clustering Algorithm

After the embedding of the posts into the multidimensional space the documents are clustered into a number of representative topics. It is important to differentiate between the terms ”partitioning” and ”clustering” as they have specific definitions. Clustering involves identifying subsets of data that exhibit natural grouping patterns, without necessarily assigning a cluster to every individual point. On the other hand, partitioning requires that each data point is assigned to a particular cluster. The partitioning approach can encounter difficulties in the presence of noise or when distinct clusters are not clearly discernible, resulting in suboptimal outcomes. As posts on this platform could be concerning an near infinite number of subjects, it was chosen to look at finding naturally occurring topics in the posts on the patient platform and therefore perform the task of clustering.

Different approaches exist to cluster the embedded posts. Density-based clustering [70] defines clusters as being able to be described as collections of data objects that exhibit high object density within contiguous regions of the data space. Notably, these clusters are distinguished from one another by contiguous regions of low object density. One of the advantages of density-based clustering methods is that they do not necessitate prior knowledge of the number of clusters as input parameters. Moreover, they do not rely on assumptions regarding the underlying density or variance within the clusters. This flexibility makes density-based clustering techniques highly adaptable to diverse datasets and enables them to effectively handle varying densities and irregularly shaped clusters. In figure 3.4 you can see the difference between K-means a partitioning algorithm and HDBSCAN a density-based clustering algorithm, where the K-means clusters contain a lot of outliers, which are ignored in the HDBSCAN algorithm [103].

As we are dealing with a unknown number of clusters and data distribution and we want to only identify the main topics that patients are dealing with without too many outliers we choose to use a density-based clustering algorithm. For this use case we make use of the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [104] algorithm as it is an optimised version of DBSCAN. HDBSCAN employs an iterative process by applying DBSCAN with different epsilon values and then combines the results to identify a clustering solution that exhibits optimal stability across the range of epsilon values. Where stability is the sum of iterations of epsilon values the points in the cluster stay in that cluster. This approach enables HDBSCAN to effectively discover clusters with varying densities, distinguishing it from DBSCAN. Additionally, this iterative integration of results enhances the robustness of HDBSCAN by mitigating challenges associated with parameter selection.

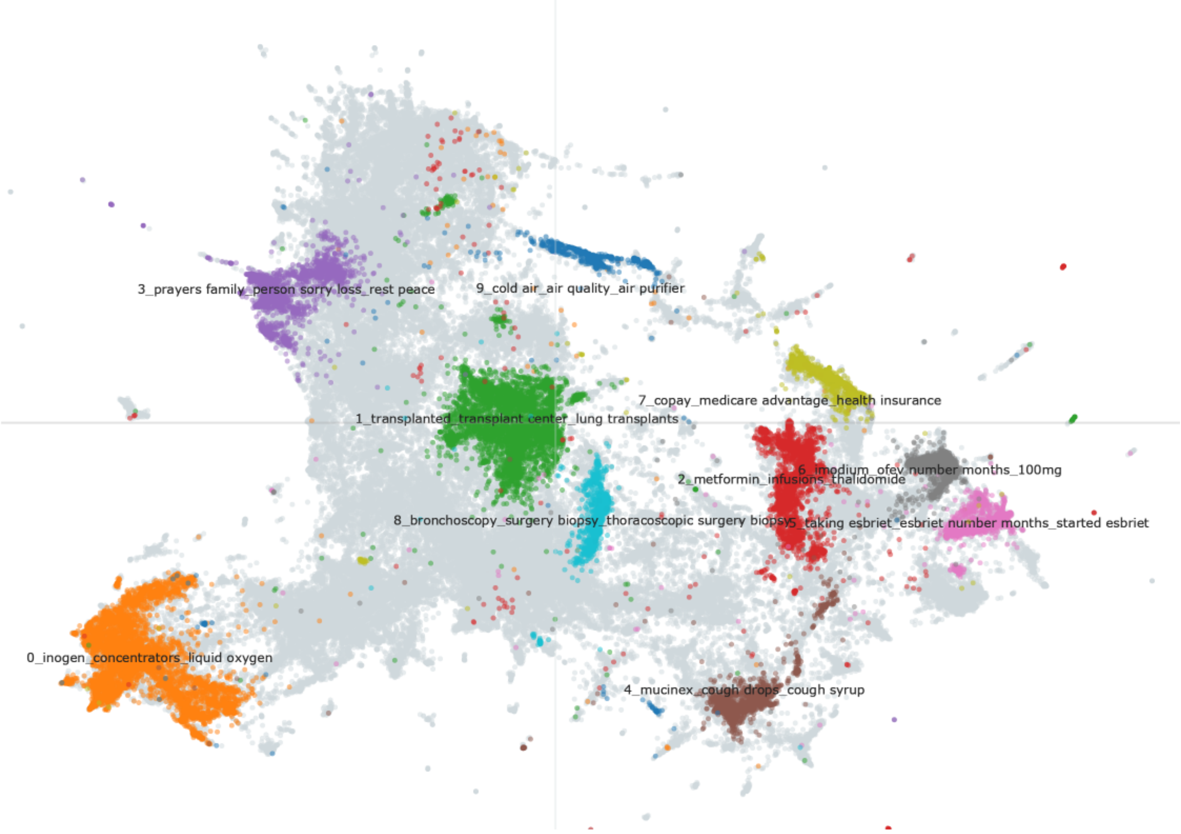


Figure 3.5: Scatter plot showing the top 10 of 100 largest distinct clusters denoted by different colors, with the top keywords denoting the cluster topics centered above the cluster. This is a two dimensional mapping of a 384 dimensional space embedding automatically provided by BERTopic, the dimensional reduction is performed with UMAP.

The number of clusters identified after the previous application of HDBSCAN was 315. A number that is reduced in the next step as it does not reduce the usefulness of the clusters for the end result, and improves the interoperability of the data for the medical staff. In Figure 3.5 you can see that the found clusters do in fact effectively make us of the HDBSCAN features as they have varying densities and irregularly shaped clusters. In Figure 3.6, a similarity matrix, we can see that almost all found clusters are varying in type and therefore the clusters are well defined.

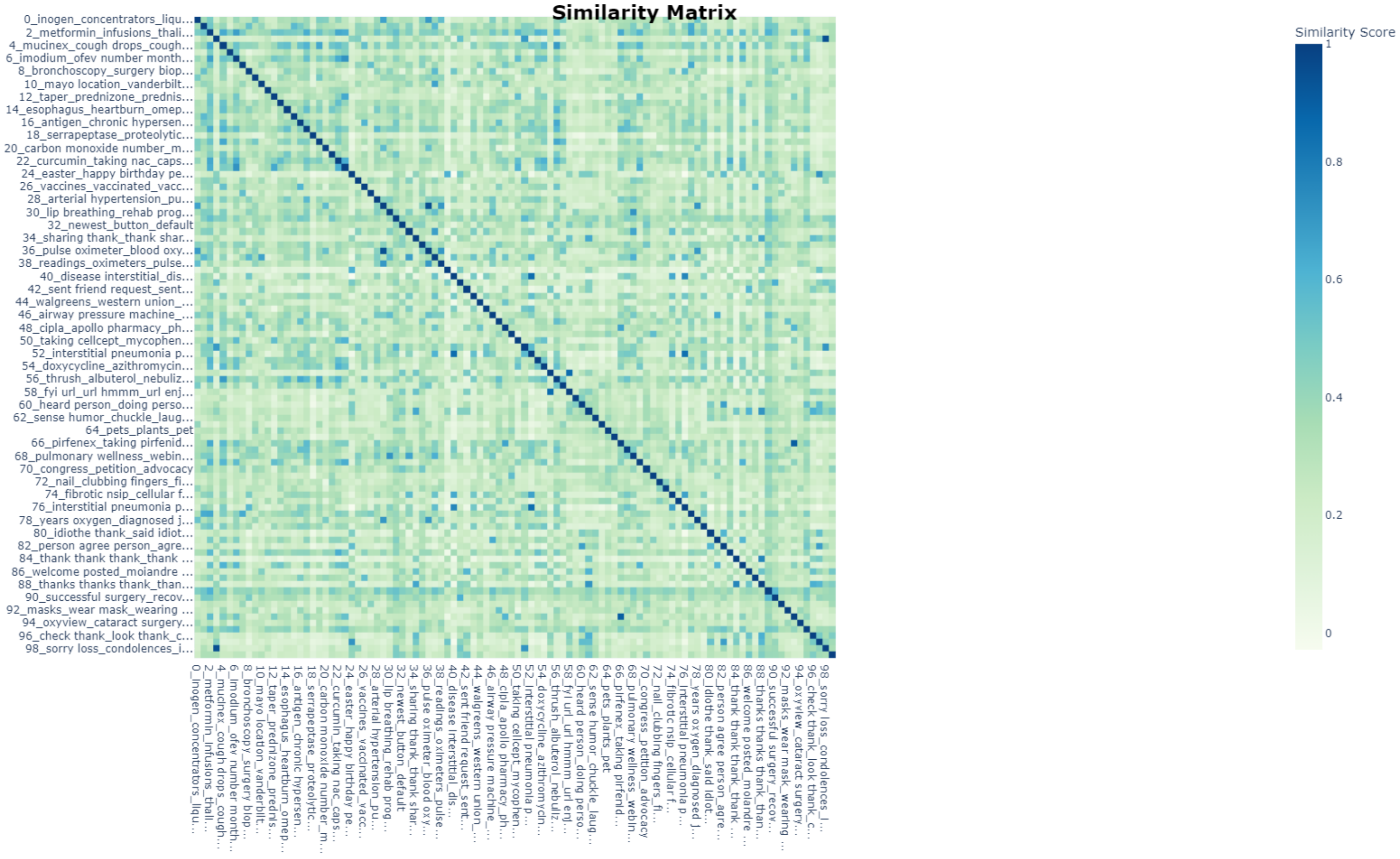


Figure 3.6: Similarity matrix, showing the similarity between all 100 topics that were found during the clustering. The metric used is the cosine similarity between the embeddings.

Cluster Merging

There is a need to reduce the number of clusters from an initial count of 300 to a smaller number. This reduction is driven by the understanding that doctors do not require overly specific information but rather seek broader insights into the usage of certain treatments or substances. For example, instead of focusing on the nuanced differences in curcumin usage, doctors are more interested in knowing that curcumin is used for various purposes.

To support this need for cluster reduction, a hierarchical plot of a selection of cluster topics from the clustering round containing approximately 300 topics can be analysed. This plot can be seen in Figure 3.7. This plot reveals that many topics exhibit significant similarities. For instance, cluster 123 discusses the potential effects of using curcumin, while cluster 314 focuses on taking curcumin as a supplement. These similarities suggest that combining clusters with comparable content can effectively reduce the number of clusters without sacrificing important information. The reduction process leverages the hierarchical data generated by BERTopic, as the employed HDBSCAN algorithm also follows a hierarchical approach. By utilising this hierarchical structure, it becomes possible to examine whether any information loss occurs when merging the most similar clusters.

Through thorough analysis, it was concluded that reducing the number of clusters from around 300 to 100 by merging the most similar ones does not result in significant information loss that would impede gaining insights into patients' experiences. To validate the absence of severe information loss, a methodology was employed to determine which clusters should be merged. This involved comparing the similarity of top posts that were likely to be associated with the clusters. By assessing the similarity of these posts, it was possible to evaluate whether merging clusters resulted in similar content being grouped together.

From analysing the hierarchical plot of cluster topics from the clustering containing 100 topics we can see that leveraging the hierarchical nature of the data, merging clusters with similar content from an initial count of 300 to 100 does not lead to substantial information loss in our use case of insight generation. Substantial information loss was seen as two clusters merging together if they did not seem closely related by two of the researchers. A section of the hierarchical plot of the reduced number of clusters can be seen in Figure 3.8. We can see in this figure that the topics are more general then with the larger number of clusters.

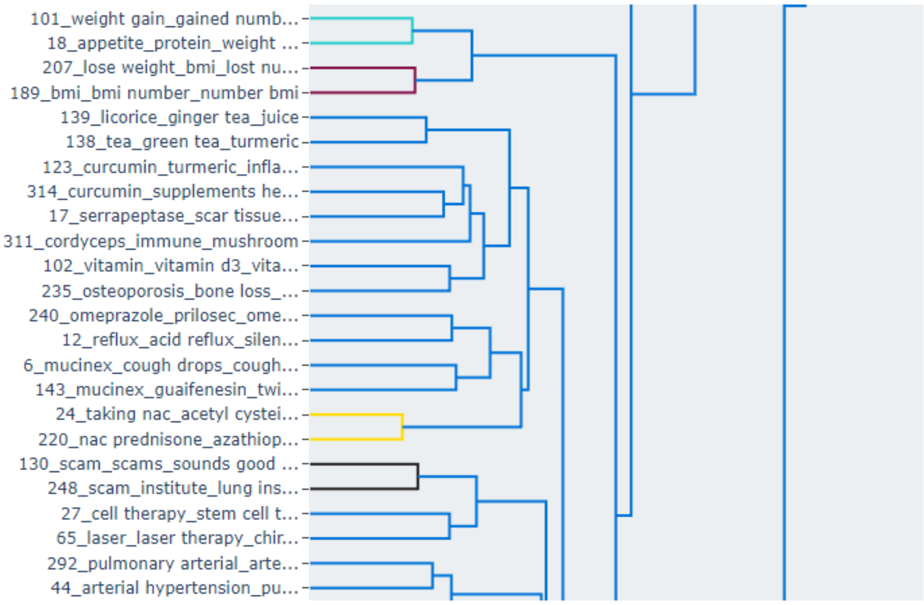


Figure 3.7: Hierarchical plot of a selection of cluster topics from the clustering round containing around 300 topics, it can be seen that many of the topics are fairly similar e.g. cluster 123 talks about the possible effects of using curcumin and cluster 314 talks about taking them as supplements.

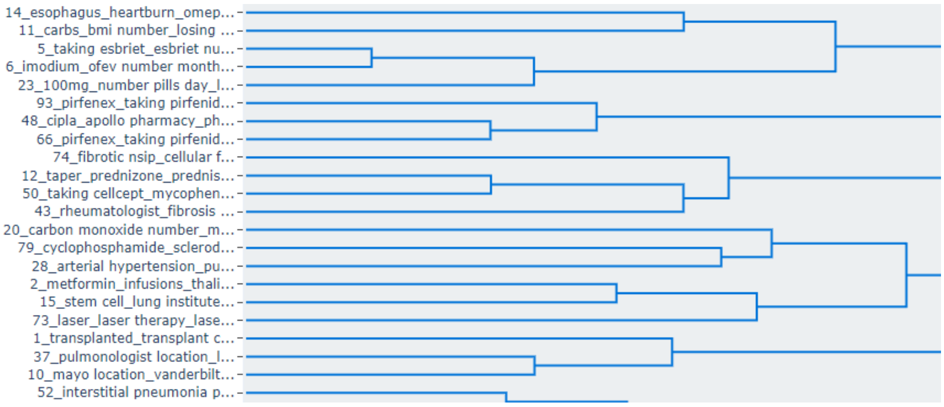


Figure 3.8: Section of the hierarchical plot of cluster topics from the clustering containing 100 topics, it can be seen that the topics are a lot more general.

Cluster Removal

There is a need to remove clusters that are more platform or country context-specific. The research focus is not on the usage of patient platforms or national healthcare by patients themselves but rather on understanding their experience with the disease in their daily lives. Information about country specific healthcare systems e.g. about American insurance plans is not useful. Thus, it is important to filter out clusters that are irrelevant to the main goal of the study.

To address this need, a methodology was employed to remove clusters that exhibited platform or country-specific content. The process involved interpreting the meaning of each cluster by reading a maximum of the top 50 posts associated with that cluster. This interpretation was then compared with the automatically generated keywords for the clusters. By performing this comparison, clusters containing irrelevant or platform-specific information were identified and subsequently removed from the analysis. Several examples of clusters that were deemed irrelevant for the research goal or specific to the American context were discovered during this process. These removed clusters included topics such as ”Educating new users and difficulties of finding the needed information”, ”Compassionate allowance from government: Supplemental Security Income (US)” and specifics about where to purchase medication. These topics were considered irrelevant as they did not contribute to the primary objective of understanding the patient’s experience with the disease.

By systematically filtering out platform-specific clusters and removing irrelevant topics, it is ensured that the topics remain focused on the core research goal of investigating the patient care trajectory experience. This approach allows for a more targeted and meaningful exploration of the patient’s experience and enhances the relevance and applicability of the research findings. After filtering 81 of the 100 clusters remained.

3.2.3. Turning the Clusters Into a Patient Journey Map

From the clusters and their meaning, that was interpreted in the last step by reading a maximum of the top 50 posts associated with that cluster, the patient journey map will be created. First the outcome of the clustering is explained, then the concept of the patient journey map is defined followed by how the clusters get turned into the patient journey map in the final result.

Outcome from the Clustering Process

The outcome of the clustering, merging and cleaning process is a total of 81 relevant topics that consist of keywords and a meaning describing a number of posts. The top 20 most frequently appearing of these topics can be seen in Table 3.1. Most of the top clusters involve topics that are about medication, managing side effects, life style changes, diagnosis and alternative or experimental treatment methods. Topic 1, the topic with the largest number of posts, is partly described by the keyword ”Inogen” which is a well known fabricator of portable oxygen concentrators and similar terms about the use of oxygen. Some of the post that are classified as being part of topic 1 are the following:

”You will need a <number> length of poly tubing thick enough to resist bending and with scissors slice it lengthwise to fit around the connecting tube. then get old rings also at store small enough not hold it securely to the tube with the slit side up as to hold off bending once in place. usually the hardware guy will help in getting you set up. hope this helps.”

-User on Inspire platform

”Has anyone had experience with the inogen one g2? if a person needs <number> liters of oxygen when mobile, would this work for them?”

-User on Inspire platform

Topic	Keywords	Meaning	#Posts
topic 1	Inogen,concentrators, liquid oxygen, oxygen concentrators, portable oxygen concentrators	Oxygen therapy: usage of oxygen concentrators	5414
topic 2	Transplanted, transplant centre, lung transplants, antibodies, lung tx	lung transplant surgery: decision making, after-care	4240
topic 3	Metformin, infusions, thalidomide, pbi number, pamrevlumab	Medication: usage of non-traditional medicine	2671
topic 4	Prayers family, person sorry loss, rest peace, condolences family, im sorry loss	Loss of loved ones, support and sympathy from peer group	1837
topic 5	Mucinex, cough drops, cough syrup, hydrocodone,coughs	Coping strategies against coughing: medication, cough syrup, supplements, physical exercises	1250
topic 6	Taking esbriet, esbriet number months, started esbriet, effects esbriet, esbriet effects	Experiences of taking esbriet: usage, effects and side effects	1199
topic 7	Imodium, ofev number months, 100mg, vomiting, effects ofev	Experiences of taking ofev: usage, effects and side effects	1127
topic 8	Bronchoscopy, surgery biopsy, thoracoscopic surgery biopsy, chest tube, open lung biopsy	Biopsy: what do patient need to prepare for it? Uncertainty of the experience of biopsy; difficulties in decision-making; Fear of the sides effects of biopsy	916
topic 9	Cold air, air quality, air purifier, hot humid, winters	Climatic condition influences on patients' syptoms: temperature, air quality, humidity	850
topic 10	Mayo location, vanderbilt, clinic location, centers excellence, closest	Choosing clinic and doctors	837
topic 11	Carbs, bmi number, losing weight, lose number, meat	Diet change, loss of appetite and losing weight Prepare for transplants surgery: maintain BMI in the required range	808
topic 12	Taper, prednizone, prednisone number, milligram prednisone, number milligram prednisone	Prednisone: dosage of use, strong side effects	744
topic 13	Cannabis, cbd oil, cannabis oil, medical marijuana, hemp oil	Supplement treatment method: CBD	706
topic 14	Esophagus, heartburn, omeprazole, esophageal, gastric	Anti-reflux: medication, keep strict diet, sleep on the bed with angle	704
topic 15	Stem cell, lung institute, cell therapy, stem cell therapy, stem cell treatments	Stem cell therapy	699
topic 16	Antigen, chronic hypersensitivity, antigens, chronic hypersensitivity pneumonitis, fungal	Hypersensitivity pneumonitis: different sources of allergens	698
topic 17	Amiodarone, pulmonary toxicity, amiodarone pulmonary, amiodarone pulmonary toxicity, atenolol	Amiodarone-induced pulmonary fibrosis	685
topic 18	Serrapeptase, proteolytic enzymes, serrapeptase nattokinase, serrapeptase number, taking serrapeptase	Usage enzyme supplements	683
topic 19	Saline, nasal spray, humidifier, saline nasal, nasal gel	Coping strategies against coughing - saline; nazal spray against dry nose	616
topic 20	Carbon monoxide number, monoxide number,vital capacity number, number percentage predicted, monoxide number percentage	Indicator of disease progression: diffuse capacity for carbon monoxide and forced vital capacity	614

Table 3.1: Top 20 of the 81 topics, sorted on frequency from most appearing to least, identified from the platform data analysis. The topics are described by the keywords automatically generated and the manually interpreted meanings.

”Hi everyone! i was wondering if anyone has found a sporty backpack for carrying a d tank? i have a small kids daypack for my tiny little tank m4 ish for running errands or hitting the dog park for a few minutes, but am looking for a way to carry a d for longer outings...”

-User on Inspire platform

We can see that the questions about the creation of a system that keeps the oxygen tubing from kinking while using, the question if a concentrator has enough oxygen output flow and a question for carrying around oxygen bottles perfectly fit with the topics interpreted meaning of *usage of oxygen concentrators*.

Patient Journey Map

Traditional journey maps have been criticised for their inability to capture the complexity of patient experiences and correlate identified touchpoints. The ”Patient Journey Map” created by Jung [63] offers a data-driven solution, leveraging natural language processing to extract patient experience stories from online communities. This allows for a comprehensive and cost-efficient analysis of a broad patient population’s self-motivations and experiences, leading to improved care pathways, products, and services.

The patient journey map approach uses text mining techniques to extract data from an existing database of patient stories from an online community. With tens of thousands of first-hand accounts, this approach enables a more representative view of the patient experience, leading to a deeper understanding of complex healthcare contexts. This visualisation of the complex care path a patient commonly experiences will be used later to leverage the doctors experience to find XAI needs. The patient journey map in Figure 3.9 is focused on the use case of IPF with messages analysed from the Inspire platform. Regular patient journey maps can include other information such as the sentiment of the patients during each stage and the events and the number of patients that deal with those during each stage of the patient care trajectory. This is however removed for our use case from the Journey Map as we only require the medical staff to focus on their experiences during these events. Where the extra information would work as a distraction to the medical staff when they are trying to get an overview of the map when reading it with a short time available to understand.

As a visual tool, patient journey maps provide insights into patients’ perspectives on their care paths, including interactions with healthcare services, staff, and organisations. By identifying potential barriers and undesirable scenarios, these maps can help inform the development of patient-centric healthcare services. In our case we can use the data mined during the platform analysis in section 3.2 to construct a patient journey map that can help elicit medical staff to think about AI solutions that are not only directly related to their direct work area.

In conclusion, patient journey maps can be an essential tool for visualising the sequence of all events and touchpoints in a patient’s care path. By leveraging natural language processing and text mining techniques, a more representative understanding of the patient experience can be gained and used to develop more patient-centric healthcare services. In this research they are used as a tool for doctors to think about the challenges that they deal with helping patients, and it will be used to identify possible opportunities for AI systems.

Placing the topics on the patient journey map

Firstly, one of the researchers connected to this thesis, who is trained in design, established an interim patient journey map by placing the identified patient topics on the timeline, aligning them with the stages outlined in the hospitals IPF care path documentation. The documentation includes three disease stages: Pre-diagnosis, Testing & Diagnosis, and Follow-up stages. The content of each patient topic is carefully considered to determine the corresponding disease stage it relates to. Then, using the standardised care path and context clues from the topics they draw the connections between the identified topics and between the existing care path to find the final picture that is the patients journey.

3.2.4. Results of the Patient Platform Data Analysis

The main outcome of the patient platform data analysis is for the first part the labeled patient posts which can be used for training AI prototypes in the following chapter and for the second part the description of the patient journey. The description of the journey in text from Figure 3.9, highlighting the different paths and decisions patients may encounter:

1. **Pre-diagnosis Stage:**
   - The patient begins experiencing symptoms associated with IPF.
   - They may start researching their symptoms and seeking additional information online.
   - Consultation with their General Practitioner for initial evaluation.
2. **Referral and Diagnosis:**
   - Depending on the initial evaluation, the GP may refer the patient to a hospital or specialist for further assessment.
   - In some cases, the patient may not receive a correct diagnosis initially and might need to go back and consult again.
   - The patient undergoes triage and receives their first appointment.
3. **Diagnostic Process:**
   - The lead practitioner reviews the patient's medical history and performs physical examinations.
   - The nurse specialist may provide the patient with information about IPF and its implications.
   - The patient undergoes further diagnostic tests such as lab tests, High-Resolution Computed Tomography, or biopsies.
   - The Multi-Disciplinary Team discusses the findings to confirm the diagnosis or conduct additional research if necessary.
4. **Receiving the Diagnosis:**
   - The patient receives a diagnosis of Pulmonary Fibrosis and must deliver the news to their loved ones.
   - They may seek additional information or consider seeking a second opinion to ensure the accuracy of the diagnosis.
5. **Treatment Planning:**
   - The patient discusses the treatment plan with their pulmonologist, who leads their care.
   - Shared decision-making occurs, involving the patient, the lead practitioner, nurse specialist, and potentially consulting online information or peer patients for more information.
   - The treatment plan may include a combination of medications, lung transplantation if applicable, pharmacological treatments, and non-pharmacological interventions like physiotherapy, rehabilitation, or oxygen therapy.
6. **Ongoing Treatment and Care:**
   - In emergency situations, the patient may be transferred to the nursing department and become an in-patient.
   - Treatment continues either at home or in the hospital, depending on the patient's condition.
   - The patient adjusts to significant changes in their life, including the impact on family, individual well-being, and social interactions.
   - Decision-making regarding the patient's quality of life becomes crucial.

7. **Disease Progression and Follow-up:**
   - The patient faces the unpredictability of how their disease progresses over time.
   - They have regular follow-up consultations, typically every three months, with their healthcare team.
   - Adjustments to the treatment plan, such as switching medications due to side effects or low effectiveness, may occur.
8. **Rapid Decline and End-of-Life Decisions:**
   - If the disease rapidly worsens, the patient may need to switch medications again.
   - Hospice care or palliative care becomes an option, and decisions regarding end-of-life care may involve consultations with the pulmonologist or nurse specialist.
   - Ultimately, the journey may end with the patient's death or, in fortunate cases, a successful lung transplant.

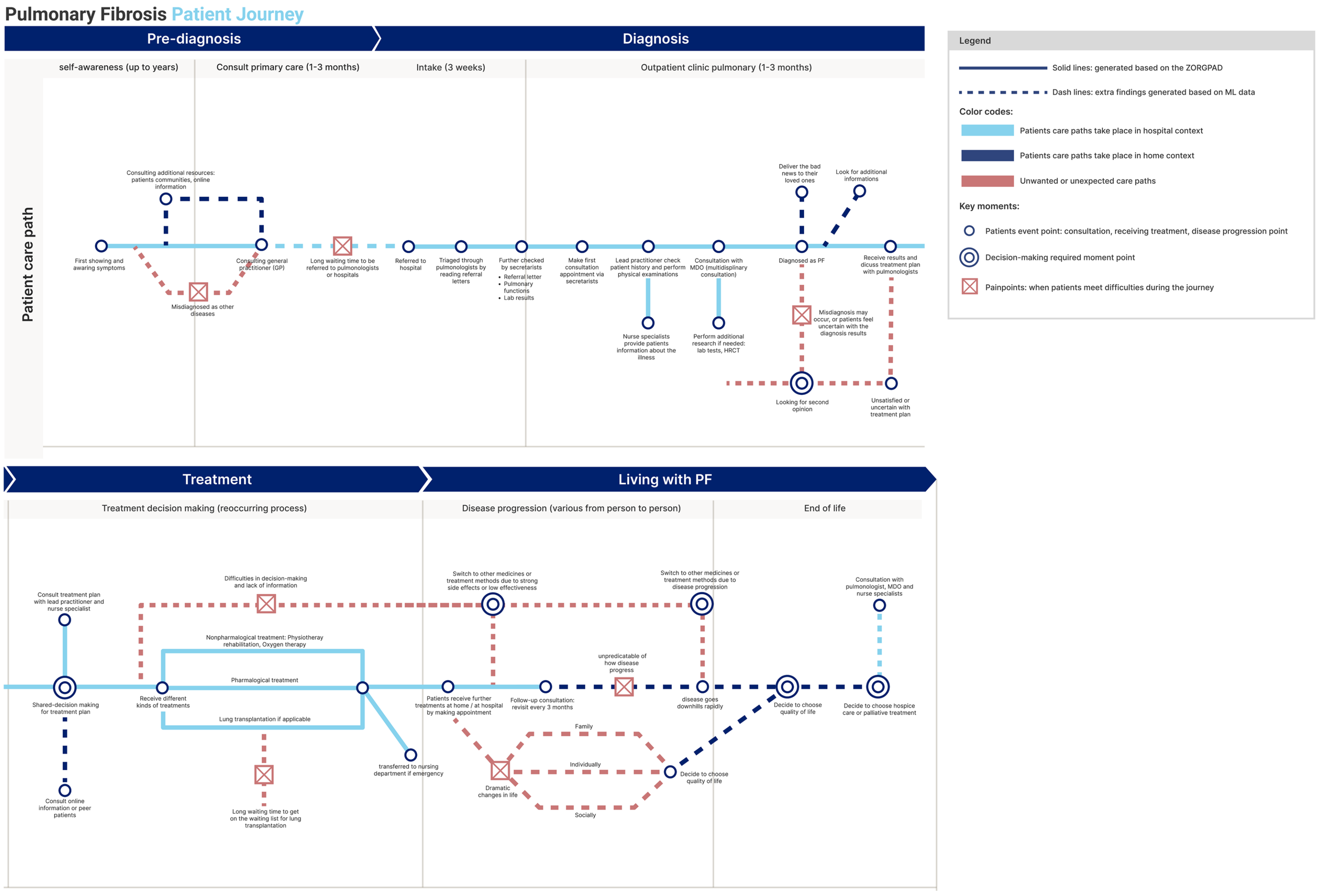

Pulmonary Fibrosis Patient Journey
Pre-diagnosis
Diagnosis
Patient care path
self-awareness (up to years)
Consult primary care (1-3 months)
Intake (3 weeks)
Outpatient clinic pulmonary (1-3 months)
Consulting additional resources: patients communities, online information
First showing and awaring symptoms
Misdiagnosed as other diseases
Consulting general practitioner (GP)
Long waiting time to be referred to pulmonologists or hospitals
Referred to hospital
Triaged through pulmonologists by reading referral letters
Further checked by secretarists
• Referral letter
• Pulmonary functions
• Lab results
Make first consultation appointment via secretarists
Lead practitioner check patient history and perform physical examinations
Nurse specialists provide patients information about the illness
Consultation with MDO (multidisplinary consultation)
Perform additional research if needed: lab tests, HRCT
Deliver the bad news to their loved ones
Look for additional informations
Diagnosed as PF
Misdiagnosis may occur, or patients feel uncertain with the diagnosis results
Looking for second opinion
Receive results and dicuss treatment plan with pulmonologists
Unsatisfied or uncertain with treatment plan
Treatment
Living with PF
Treatment decision making (reoccurring process)
Disease progression (various from person to person)
End of life
Consult treatment plan with lead practitioner and nurse specialist
Shared-decision making for treatment plan
Consult online information or peer patients
Difficulties in decision-making and lack of information
Nonpharmalogical treatment: Physiotheray rehabilitation, Oxygen therapy
Pharmalogical treatment
Receive different kinds of treatments
Lung transplantation if applicable
Long waiting time to get on the waiting list for lung transplantation
transferred to nursing department if emergency
Switch to other medicines or treatment methods due to strong side effects or low effectiveness
Patients receive further treatments at home / at hospital by making appointment
Follow-up consultation: revisit every 3 months
unpredicatable of how disease progress
Dramatic changes in life
Family
Individually
Socially
Decide to choose quality of life
Switch to other medicines or treatment methods due to disease progression
disease goes downhills rapidly
Decide to choose quality of life
Consultation with pulmonologist, MDO and nurse specialists
Decide to choose hospice care or palliative treatment
Legend
Solid lines: generated based on the ZORGPAD
Dash lines: extra findings generated based on ML data
Color codes:
Patients care paths take place in hospital context
Patients care paths take place in home context
Unwanted or unexpected care paths
Key moments:
Patients event point: consultation, receiving treatment, disease progression point
Decision-making required moment point
Painpoints: when patients meet difficulties during the journey

Figure 3.9: Part of the Patient Journey Map Idiopathic Pulmonary Fibrosis - Made by Ruixuan Zhang

3.3. Conclusion Research General Context

In this chapter, we have presented the results of our research into the medical background and context of Idiopathic Pulmonary Fibrosis. Through clinical observations and patient platform analysis, we aimed to gain a holistic understanding of IPF care and to lay a solid basis to identify concrete opportunities for explainable AI solutions in the pulmonology field.

Clinical observation session

The clinical observation sessions provided valuable insights into the challenges faced by pulmonologists. Time pressure emerged as a significant issue, impacting the doctors' ability to fully prepare for patient check-ups and address all patient questions comprehensively. Communication between medical staff was another area of concern, with excessive time required for relaying information, leading to frustration and shortcuts. Problems planning tests, patients' difficulty in remembering their questions, and conflicts between patients, caregivers, and doctors were also observed. These findings shed light on specific areas where XAI systems could offer support and improve the efficiency and effectiveness of pulmonologists' work in delivering quality care to IPF patients.

Patient platform data analysis and patient journey map

The patient platform data analysis contributed to the understanding of the patient journey in IPF. By analysing online patient stories, we gained insights into patients' experiences, coping strategies, concerns, and barriers to health behaviour changes. The patient journey mapping highlighted various stages and decisions encountered by patients, ranging from pre-diagnosis to ongoing treatment, disease progression, and end-of-life care. This patient-centric perspective complemented the existing documentation from the healthcare providers and revealed the branching nature of patient care outside the hospital setting.

The patient journey gives a different view than the current information that doctors have about the treatment of the patients as that documentation is mostly from the perspective from the doctors. These existing documents also only take the care that takes place in the hospital into account. However, a large part of the patients their care and the opportunities for the use of AI and explainability could be in the areas where the hospital has less view on the treatment of the patients. The treatment plans and existing literature are also very linear while the patient journey shows that the care path of patients can be very branching.

Combined insights

Overall, the combined insights from clinical observations and patient platform analysis provide a comprehensive understanding of IPF care, which informs the extraction of XAI needs from the pulmonology medical staff. By addressing the identified problems and considering the patient journey, we can develop XAI prototypes that could enhance collaboration between AI systems and medical staff, creating a realistic system that can help in the research towards AI-human collaboration and explainability.

The knowledge gained from this research will also inform the semi-structured interviews planned for the next stage of the study. By involving doctors with diverse backgrounds and varying levels of familiarity with AI technology and IPF patient care, we aim to develop effective strategies for explainability that meet the needs of the pulmonology field.

In conclusion, this chapter has laid the foundation for our research on explainable AI in the context of pulmonology. The insights gained from the preliminary study and patient platform analysis provide valuable input for the subsequent stages of our research, where we will further explore the needs and perspectives of the pulmonology medical staff through interviews and co-creation sessions. By understanding the challenges and complexities of IPF care, we can develop tailored XAI solutions that support medical professionals in providing high-quality, explainable care to patients with IPF and possible other respiratory disease.

4

Study 2: Research Explainability in Pulmonology Context

This chapter presents a comprehensive exploration of explainable AI use in pulmonology, divided into two key parts. The first part focuses on a preliminary study in section 4.1, which encompasses the development of a prototype, a co-creation session, and subsequent follow-up interviews. The objective of this preliminary study is to gather initial feedback and refine the understanding of explainable AI in the context of pulmonology. The insights gained from this study will play a crucial role in facilitating informative answers during the upcoming semi-structured interviews.

In the second part of the chapter section 4.2, the focus shifts to the Semi-structured Interviews. These interviews are conducted to delve deeper into the subject and elicit informative answers from the participating doctors. To facilitate this process, planned prompts and grand tour questions are employed. Planned prompts, which are formally included in the interview protocol, provide guidance and structure to the discussions. Grand tour questions, on the other hand, prompt respondents to give a comprehensive verbal tour of a subject they are well acquainted with, enabling a deeper understanding of their perspectives and insights.

Following the semi-structured interviews, the chapter proceeds with an Interview Analysis section. Here, the results of the semi-structured interviews are evaluated using thematic analysis and descriptive coding by a team of multiple researchers. Finally, the chapter concludes by summarising the research findings and insights gained from exploring explainability in the context of pulmonology. These findings contribute to the overall understanding of explainable AI and inform the development of strategies to enhance its implementation in the medical field.

4.1. Preliminary Study

This section outlines the development of a prototype, a co-creation session, and subsequent follow-up interviews that were carried out to gather initial feedback and refine the understanding of explainable AI use in pulmonology. It is important to carry out a preliminary study as we want to find out information that can help us during the semi-structured interviews later on to elicit giving informative answers.

For this cause a prototype, a co-creation session, and follow-up interviews were utilised to gather insights from medical staff regarding their experiences with explainability and their specific needs in AI and explainability solutions. The prototype aimed to facilitate realistic thinking among medical staff during the co-creation session, while the co-creation session itself involved a multidisciplinary group to obtain initial feedback and validate key factors for optimising explainable AI. The follow-up interviews provided a structured exploration to clarify participant needs and identify any gaps or missed opportunities.

By combining the development of a prototype, a co-creation session, and follow-up interviews, this preliminary research chapter aimed to gather valuable insights into the medical staff's perspective on explainability in the context of AI. These activities provide a foundation for further exploration in section 4.2 where a deeper understanding of these properties will be attempted to be found.

Figure 4.1: Image of the first co-creation session in the Erasmus MC in Rotterdam. This image shows the participants from the session. From left to right it are specialist nurse, data-science PhD, specialist doctor, specialist doctor in training, PhD in medicine, design master student, computer science master, design PhD

4.1.1. Participants

Our inclusion criteria were aimed to reach the objectives of our session by including participants with expertise in every direction necessary.

Doctor and specialised nurse For the validation of the patient journey it was important that expert participants were included that are in direct with the patients at different occasions in their care so that a wide possible view of the patient experience is included. This is why the two different types of medical staff who see and communicate with the patients most are included. The primary doctor and a nurse specialist both have a high degree of experience with patients.

Clinical researcher To maintain a wide view to answer the first aim of scoping down on the main focus of the medical staff it is also important to include non clinical staff such as researchers, therefore also a PhD in medicine is included in the participant group.

Explainability PhD For gaining a basic understanding to identify the factors that are important to optimise for using XAI in the medical domain it is also important to have a participants that are knowledge about XAI in the discussion, as medical staff often miss the knowledge about overarching problems within the field of AI such as fairness, bias, and health inequalities [100]. Therefore a PhD candidate with experience in explainability is included.

Design Facilitators The designers are included as they have extensive experience in using and leading co-creations, can think analytical about design solutions for problems and have a high degree of creativity enabling medical staff to thinking more creatively.

The 8 diverse participants in the session with expertise in patient care, design and explainable AI will result in an expert group that is able to validate and generate insights at the intersection of all these research fields. Included are 8 participants of different disciplines: pulmonologists (3), specialist nurse (1), designers (2), computer scientists (2). The participants and their individual roles can be seen in Figure 4.1. The recruitment of participants was done through personal connections within the Tu Delft and Erasmus MC. The only drawback of using this way of recruiting participants is that some of the participants have worker-boss relations which could lead to participants agreeing with their higher-ups as they might feel that they have to agree with them. This was prevented by emphasising the equal value of each participant's input during the session and following up with the individuals through the means of individual follow-up interviews.

4.1.2. Materials

In the preliminary study, both a prototype and topic cards are utilised to gather insights and feedback from medical staff. The prototype was developed based on identified needs and observations from platform data analysis, and it aimed to address the challenge of limited patient information during consultations. It incorporated an AI system relevant to doctors' expertise, providing dynamic patient context based on experiences outside the hospital setting. On the other hand, topic cards were employed as a supplementary tool during the preliminary study. These cards represented distinct topics and contained essential information in compact and organized way of presenting the information, allowing medical staff to quickly grasp the main topics and gain a broad understanding of the content to decide on the most important topics at hand.

The prototype encourages active discussion by presenting AI solutions relevant to doctors' expertise, while the topic cards serve as a quick reference guide to the identified topics, saving valuable time in the time-sensitive environment of the study. Together, these tools facilitate meaningful discussions during the co-creation session and contribute to the success of the preliminary study.

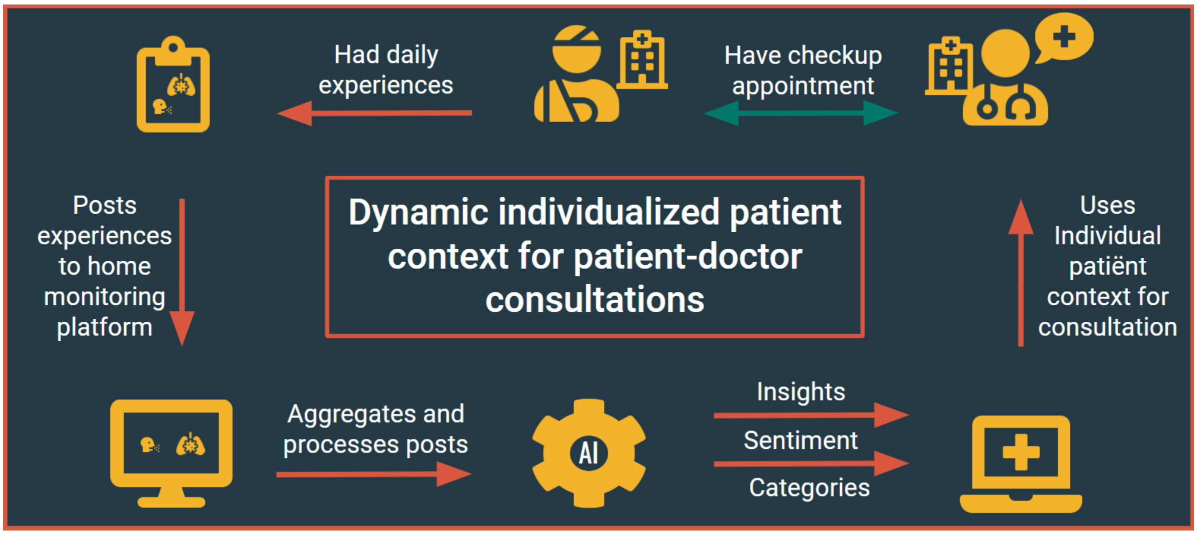
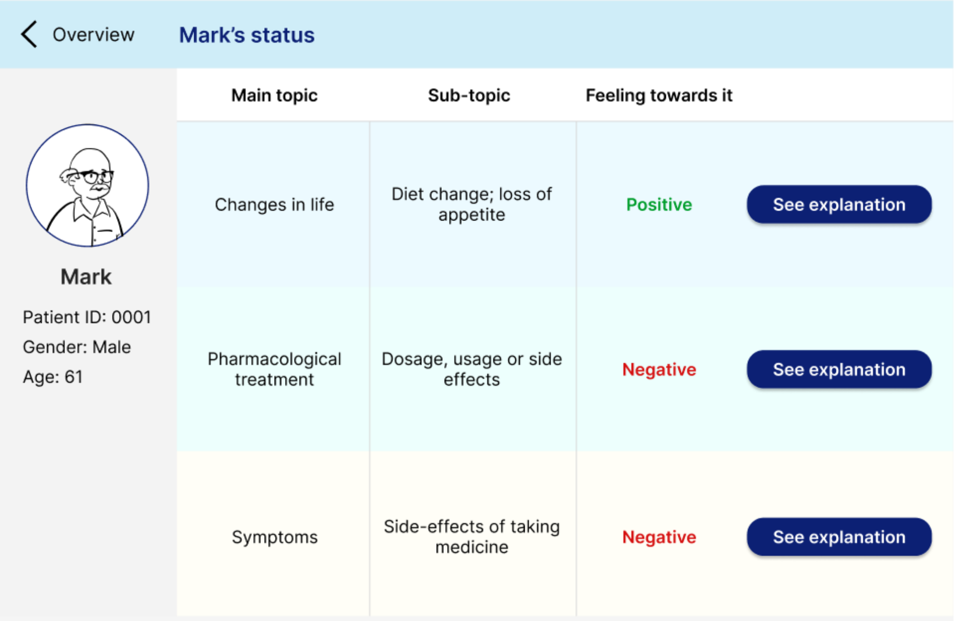

Had daily experiences
Have checkup appointment
Posts experiences to home monitoring platform
Dynamic individualized patient context for patient-doctor consultations
Uses Individual patiënt context for consultation
Aggregates and processes posts
AI
Insights
Sentiment
Categories

Figure 4.2: Internal workings of the AI powered patient needs identification prototype made to elicit thinking about explainability for the medical staff.

Prototype

To elicit realistic thinking in the medical staff about explainability a prototype was developed to address the identified needs and insights generated from the observations and platform data analysis. This prototype was designed to be utilised during a co-creation session involving a multidisciplinary group consisting of doctors, computer science experts, and design professionals. The primary objective of the co-creation session was to gather initial feedback from the medical staff regarding their use of explainability and their specific needs in AI and explainability solutions through use of the prototype and a subsequent discussion between all participants.

To encourage active discussion on explainability needs, the prototype was designed to incorporate an AI system in an application relevant to the doctors working expertise. The problem statement derived from the previous observation session was used as the basis for the prototype. Specifically, the prototype aimed to address the challenge of limited patient information available during consultations, as patients often present their best condition and healthcare professionals have limited time to prepare. The prototype tackled this problem by providing dynamic individualised patient context for patient-doctor consultations based on patient experiences outside of the hospital setting.



Mark

Patient ID: 0001
Gender: Male
Age: 61

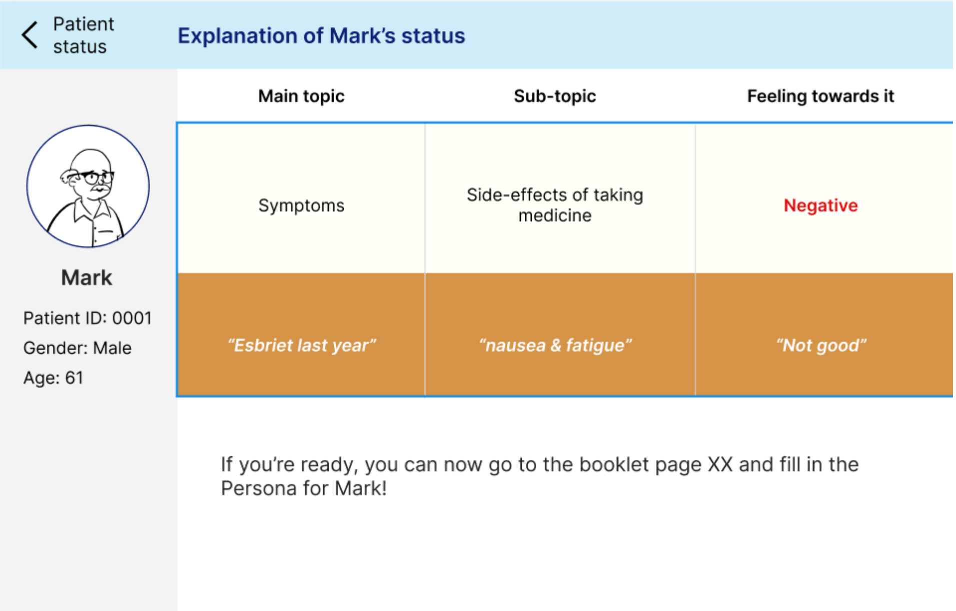
Main topic	Sub-topic	Feeling towards it
Changes in life	Diet change; loss of appetite	Positive
Pharmacological treatment	Dosage, usage or side effects	Negative
Symptoms	Side-effects of taking medicine	Negative

Figure 4.3: Example of one of the mock patients in the prototype system together with the main topics, subtopics and sentiment of that patient his messages on the home monitoring platform. These were generated by multiple AI classification methods.

Prototype Functionality

The system utilised a combination of techniques to achieve this objective, an abstract version of the internal workings can be seen in Figure 4.2. It aggregated free text messages posted by patients on the home monitoring platform about their daily experiences, aggregates them and analyses the sentiment, main and sub-topic, which corresponded to the cluster topics identified in the platform analysis. The doctor can then get a quick overview of the status of the patient before the check-up consultation with that patient. In Figure 4.3 three of these topics the patient has been writing about in the home monitoring messages can be seen together with the subtopics and their feeling towards them.

To aide the doctor into understanding why these categories were picked for this patient an explainability section has been added to each prediction. Feature attribution techniques, such as Integrated Gradients, were employed to generate explanations regarding why those main and subtopics were generated. Additionally, sentiment analysis was performed to assess the emotional tone of the patient's messages. This was added to give the medical experts, while performing the exercise during the co-creation session, some additional context to base their questions



Explanation of Mark’s status

Mark

Patient ID: 0001

Gender: Male

Age: 61

Main topic	Sub-topic	Feeling towards it
Symptoms	Side-effects of taking medicine	Negative
“Esbriet last year”	*“nausea & fatigue”*	*“Not good”*

If you’re ready, you can now go to the booklet page XX and fill in the Persona for Mark!

Figure 4.4: The main and sub-topic generated by the AI classification methods and the keywords from the patients’ messages which lead to that classification. The sentence extracts are meant as an example of using explainability for AI predictions.

on. A form of explainability, for one of the topics the patient has been talking about, can be seen in Figure 4.4. The most important words used for making the prediction of the topic symptoms and the patients’ sentiment about the topic are displayed.

Inner Workings AI models Used

The underlying AI models employed in the prototype utilised hierarchical fine-tuned BERT classification models for topic predictions. First a model is used to identify the main topic and then a specific classification model that is associated with that topic is applied to determine the subtopics. This approach was found to yield more accurate predictions during testing. To train the classification models, the platform posts and their corresponding labels obtained from unsupervised clustering were used as the training data. The hierarchical structure was taken from the manual topic analysis that was performed after the clustering. The top level topics that could be identified were *Medication, Treatment, Symptoms, Decision-making, Diagnosis, Causes, Changes in life, Emotions, Environmental influences*. Each of those main topics has an average of 10 subtopics assigned to it. For the sentiment classification a pre-trained BERT model was used to determine the sentiment that was expressed in the patient’s messages.

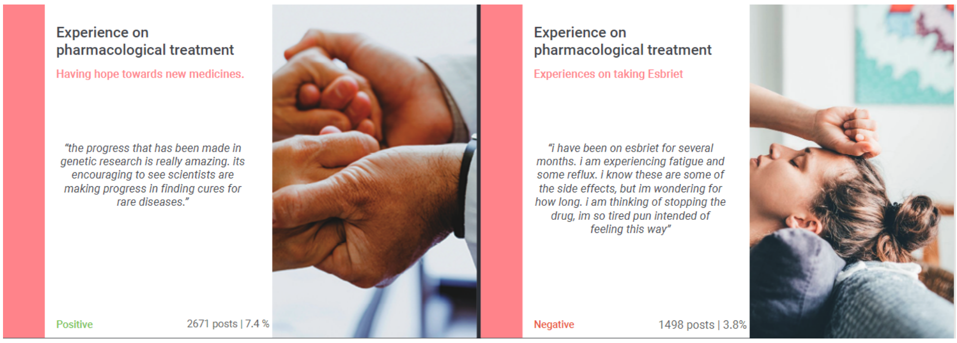
Real patient posts were collected from various sources, including IPF patient platforms not used in the data analysis, and combined with knowledge from the literature about likely combinations of symptoms to construct realistic patient cases for use during the co-creation session exercise. These patient cases were then ran through the AI models to generate the topic and sentiment classifications for the patient status dashboard seen in Figure 4.3.

To generate the explainability for the topic classifications the *Transformers Interpret*[1] package was used. This package uses integrated gradients and a variation of it called layer integrated gradients to attribute word importance scores to each word in the patient texts. The *Phrasemachine*[2] package is then used which extracts multiword phrases, such as ”not good” instead of ”good”, together with the word attribution scores to find the most important part of the sentence for the classification of that sentence. This is an important step to follow as otherwise misleading or confusing words might be shown as explanations to the medical staff. This keyword or most important multiword phrase is then used as the explanations for the medical staff as the reason

[1] Transformers Interpret Github page www.github.com/cdpierse/transformers-interpret

[2] Phrasemachine Github page www.github.com/slanglab/phrasemachine

why the patient was talking about that topic or had that sentiment.



Experience on pharmacological treatment

Having hope towards new medicines.

"the progress that has been made in genetic research is really amazing. its encouraging to see scientists are making progress in finding cures for rare diseases."

Positive 2671 posts | 7.4 %

Experience on pharmacological treatment

Experiences on taking Esbriet

"i have been on esbriet for several months. i am experiencing fatigue and some reflux. i know these are some of the side effects, but im wondering for how long. i am thinking of stopping the drug, im so tired pun intended of feeling this way"

Negative 1498 posts | 3.8%

Figure 4.5: Example of 2 topics cards containing both a a main topic, subtopic, overall cluster sentiment, number of posts classified as that subtopic and a representative patient post from that topic.

Topic Cards

Topic cards are small cards containing a theme or topic in some cases added information about that specific subject. They are often used in brainstorming or other type of ideation sessions in order to facilitate discussions and generate ideas in a certain domain. Each card represents a distinct topic, and participants can look through them randomly and select certain ones they think are interesting. The topic cards used in the preliminary study, as seen in Figure 4.5, cover the main topics and provide some added information about the topic such as the cluster sentiment, how often a sub-topic appears and it gives a representative post for that specific cluster.

The topic cards are used during the preliminary study in order to provide the participants with information which topics were found and to give them a quick overview of the contents of those topics. They offer a compact and organised way of presenting information. Which can save time for medical staff as they can quickly scan through the main topics and get a broad understanding which is useful for the large number of topics and the time-sensitive environment.

4.1.3. Co-creation

The first step in the process of gathering initial insights researching the explainability context is a co-creation session. Co-creation is a design tool for researchers which enables them to include expert stakeholders in validating and generating novel ideas and insights. Co-creation includes the passive stakeholders early in the processes [152] and makes use of the advantages that discussion between experts generates when little information is already known. Co-creation sessions have a high degree of interactivity between different stakeholders and that is why they serve great for the process of new idea generation.

Why a Co-creation Session

Despite the growing emphasis on adopting human-centric design, there remains a lack of comprehensive frameworks and guidelines to aid practitioners in effectively incorporating human-centricity into the development and deployment processes. In the context of explainability, stakeholders with varying characteristics play crucial roles. However, in many cases, these stakeholders are not adequately engaged beyond the initial requirements gathering or the testing and validation phases of projects, leading to limited integration of their valuable input. To not run into the same trap some current explanation algorithms have in regards to limitations of not taking into account of end users for the explanations, for the first expert session co-creation will be used.

Aim of the Session

The aim of the first co-created workshop was to get a basic understanding of and identify the following concepts:

1. **To scope down the focus**: What are the most important challenges that need to be addressed in doctors' perspective to improve health related quality of life of patients?
2. **Identify factors**: What factors are important to optimise for using explainable AI in the context of the medical domain?
3. **Validate Journey Map**: Does that the journey map aligns with the experiences of the experts and do all steps displayed in the journey map align with the dutch healthcare context?

From this information the consequential steps can be shaped. Besides these goals the journey map from Table 3.2.3 was also validated during this session.

Structure of the Session

The workshop was held at the Erasmus University Medical Center in Rotterdam on the 15th of March. The session was scheduled to take 1.5 hours and consisted of the multidisciplinary team that was recruited. A small introduction of the project was given, followed by sharing the research methods, limitations such as the US context of our background research and providing the schedule for the remainder of the session. In Figure 4.1 the co-creation session can be seen.

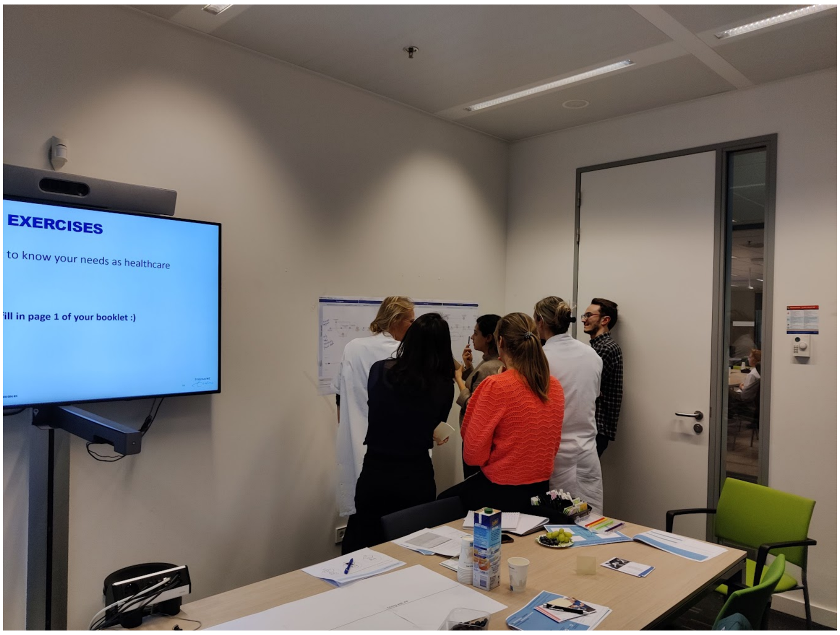

EXERCISES
to know your needs as healthcare
ill in page 1 of your booklet :)

Figure 4.6: Image of the first co-creation session in the Erasmus MC in Rotterdam. This image shows the process of validating the patient journey that was generated.

This was followed by helping participants to understand the timeline and to validate the journey map from Figure 3.9. The participants together express their thoughts when looking at the map and asked clarifying questions about the contents and the ways the data is acquired. In Figure 4.6 the validation of the journey map can be seen. Using this data the journey map was adapted to be used later during the interview stage.

Then the prototype and the questions are used to see if the medical staff understand and trust the explainability generated by the system and to see if they think it is useful. The part of the co-creation session using the explainability prototype from subsection 4.1.2 was structured as follows:

1. **Introduction prototype**: Explain how the prototype system works to the participants using the tutorial page at the start of the system and explain the assignment that they must perform.
2. **Setup**: Each participant gets a separate laptop to use, look through the three example patients that were pre-made by us, as explained in the prototype section, to display the different functions of the system.

3. **Assignment**: The participants will use the system to read the main/sub-categories and sentiment generated from the problem posts the pre-made patients are experiencing. They must create in their booklet three personas that match the three pre-made patients they explored in the system. They can use the explainability for doing this by clicking on the hint button. The extra information provided by the explanations can give them some context and help them narrow down the questions they will ask the patients.
4. **Briefing**: From these results we ask them what personas they created and see if they match the ones that we based the messages on.
5. **Discussion**: Finishing with a semi-structured discussion on if they understood the information presented and if they used and trusted the explainability hints, if they think they needed more hints and how those would look like.

After the use of the prototype explainability system a semi-structured discussion between all participants was hold. The follow up discussion points were in the same form as a semi-structured interview where four main terms were identified and from those main themes deeper questions were derived to discuss. The structure of the four main themes was taken from [34] namely, interpretability, understandability, usability and usefulness. The discussion points mostly aimed around the underlying concepts of these terms:

1. **Interpretability** The degree to which a user can intuit the cause of a decision and the ability to predict a system's results: How do they think the system made certain choices and what other types of visualisations would they use and why.
2. **Understandability** The degree to which a user can ascertain how the system works, and leads directly to confidence in the system's output: What could you add to the system to give a better view into the important parts of the life of the patients and how would you change the system to better trust its judgements.
3. **Usability** Ease with which a user can learn to operate, prepare inputs for, and interpret outputs of a system: What changes would make you use the extra information functionality of the system more often and how would you use a version of this system for tasks in a real hospital setting.
4. **Usefulness** Will one use the system because it meets a user's needs and is seen as the practical worth of a system: How could the extra information help you in gaining a more accurate understanding about the patient.

After which we focus on finding the challenges that are most important to improve on. For this purpose the participants look through the patient topic cards, which are cards denoting a topic that was identified together with a part of the most representative post and the number that were identified. Each participant picks 2 topic cards that they think is the most important to ask during their consultation with patients. Then they discuss about the choices they made. After which the participants are thanked and the session is ended.

Outcomes

Journey map

During the session the following main points of change were identified to make the journey map more valid. They were all implemented in the version seen in Figure 3.9 that was used during the semi-structured interview.

1. Distinguish the different datasets from the journey map, focus it only on the harvested platform data and not on the established care path documents.
2. Change the names of the different main phases of the patient journey, such as the diagnosis stage, since they are too vague.
3. The sentiment flow of the patient should be made according to specific key moment in the patient journey.

Participants stated that if the preceding points were implemented that the journey map would be an accurate depiction of the stages that a patient has to deal with during their care at home and at the hospital. The participants recognised the decision points and events from their own experience with patients.

Prototype

The prototype from subsection 4.1.2 was used in the co-creation session to explore how the medical staff would use explainability followed by questions asking them about elements from explainability to find out what type would fit best to their use case. The results of the task the medical staff had to perform were above expectation, showing that doctors given limited information about a patient can still fully form an image of a hypothetical patient and ask well aimed questions to find out more about those subjects.

A flaw of this was that when the medical staff recognised the information the AI gave them and it in their opinion fit a regular patient they automatically started trusting that information. They not even consider looking at the explanations and instead already started making treatment adjustments in their head before a real check-up conversation could take place with the patient. Also the lack of knowledge in some participants about AI concepts, the overarching problems AI suffers from and the trust they put in people with that expertise resulted in overtrusting the systems output. Participant P2 stated for example that *"We would assume that the data [of the AI system] was correct."* that they *"I would trust the system, I think you know what you do [building AI tools]."* and therefore they would *"I would already start biased questions [to the patient]."*. Other participants such as P4 were more interested in the underlying algorithms but would still trust the system in the end *"I would want to know what clustering etc. you used but in the end I trust you."*.

Other participants saw the use of the system P4: *"Nice to find topics that we as healthcare providers would not find."* but it would depend on the type of topic if they would use the explainability to check P1:*"Depends on the topic if I would use explainability."*. In the end the main drawback of the AI system was not that it did not play into the needs of the medical staff as P2 identified it as a need before knowing the system, as can be seen as point 2 and 3 in Table 4.1. Rather for explainability purposes it was not a suitable example as they could verify the AI's classifications with the patients themselves P2: *"We would verify the data by talking to the patients."*. This all meant that the discussion after the use of the prototype was not effective in finding the desired characteristics of explainability accurately and other steps will have to be taken to identify those.

Needs and important topics

The needs and the topics that the medical staff deemed important to tackle were also identified. These needs can be seen in Table 4.1 and correspond to the specific tasks and years of experience that each participant has in their work with patients. Medical staff who are very busy with seeing the actual patients want tools that make consultations more time efficient. Medical staff who perform research into state-of-the-art methods want automatic tools which can help with those tasks. When looking at the needs of medical staff it can be summarised that they mostly desire tools that make the work they are currently doing more efficient and less time consuming. The most important topics that the medical staff wanted to improve was for them dependent on each patient. They did however agree that the topics that they wanted to know more about were the decision making steps and daily care challenges of patients as they thought they did not have a full overview of those subjects.

From this first look into the medical staff their needs and their use of explainability prototype we look back at the aim of the session and see that a scope down towards one main important challenge to focus on is not possible. Furthermore through the use of the prototype the second aim of the session, the factors that have to be taken into account for using explainable AI in the context of the medical domain, could not yet be fully identified. As some of the points, such as needs, were not fully clear yet during the co-creation session follow-up interviews were held with two participants to find out what they exactly meant during the session. It was also concluded that a step back should be taken and that instead of a concrete prototype a wider follow up research should be performed in order to find out the explainability needs of the medical staff.

Who	Needs	Why
P1	1. I think it's important to know what patients find important in the care and what can help them to improve quality of life. 2. Besides patients I want to know what their partners need.	When you know what patients and partners need, we can adapt the care and I hope it helps to improve the quality of life.
P2	1. More time. 2. More automatic knowledge available beforehand to make. consultation efficient. 3. Clear view of patient wishes before the consultation. 4. More practical support for patients. 5. Insights on patient experiences with medicine/worries/needs. 6. Automatic data overview. 7. Psychological parameter.	
P3	1. Better tools to communicate about end-of-life decisions. 2. Know which patient to treat with which medication at which time and which place; so providing better personalized medicine at home and hospital.	Better individually targeted treatment adapted to the needs and wishes of patients would help to improve care for patients with PF and hopefully also the quality of life.
P4	1. Online reliable information. 2. Easy identification of the needs of patients. 3. Place where you refer patients, for example, rehabilitation, psychology, dietist, etc.	It would make it easier to have access to good quality information at the outpatient clinic.

Table 4.1: Needs of the medical staff during the treatment of patients with IPF expressed in the co-creation session.

4.1.4. Follow-up Interviews

Following the co-creation session a number of follow-up interviews are performed. These semi-structured interviews allow for a higher degree of structure then the discussion phase from the co-creations session, resulting in a higher probability to get answers to the posed questions. The main aim of these follow-up interviews was to ask more details about the statements made during the co-creation session, to clarify the needs they expressed and to identify that what was missed in the prototype. Which is the insights into how contact takes place between the patient and the healthcare system. Because of the limited time medical staff only two participants, which gave answers which were unclear or would require more detail, were asked for these participating in the follow-up interviews.

The follow-up interview answers strongly influenced what and how questions in the semi-structured interview would be asked. One example of this is that we won't ask questions such as *"What do you think are the current barriers for the use of AI in the medical field? Do you think those barriers can be solved by explainable AI?"* as they will likely answer with simple answers about regulatory or about who is responsible for mistakes. Rather, questions should be asked that are more aimed towards identifying concepts that are important for the explainability features. Furthermore, more care is taken into getting to know participants knowledge level and attitude before performing the interviews as for statements such as *"I would want to know what clustering you used."* it was found that want for knowledge comes more from their experience with AI than their medical knowledge, thus that influences their explainability needs.

Participant	Medical function	Experience*	Knowledge AI	Opinion AI in healthcare	Knowledge PF
P1	Pulmonologist	20 year	Yes	Positive	Yes
P2	Pulmonologist	12 year	Yes	Positive	Yes
P3	Pulmonologist	1 year	No	Positive	Basic
P4	Resident	6 year	No	Sceptic	Basic
P5	Resident, postdoc	3 year	Yes	Positive	Yes
P6	Physician-Researcher	3 year	No	Positive	No
P7	Physician-Researcher	1 year	Yes	Positive	No
P8	PhD candidate	3 year	Yes	Positive	Yes
P9	PhD candidate	2 year	Yes	Positive	Yes
P10	PhD candidate	2 year	No	Positive	Basic
P11	PhD candidate	1 year	No	Positive	No

Table 4.2: List of participants of the interviews sorted from top to bottom on their experience in the medical domain. *Experience rounded to years in the current role they have.

4.2. Semi-structured Interviews

During the co-creation session, it became evident that using a concrete prototype to assess the needs for explainability fell short of expectations. The presence of a specific use case tended to distract the attending doctors from comprehensively addressing the broader issues of AI explainability. This realisation emphasised the need for a more flexible approach that could engage healthcare professionals in a manner that encourages thoughtful reflection on the complex interplay between explainability and AI. Acknowledging the limited knowledge of some doctors regarding AI and explainability concepts, it became apparent that inclusively was paramount in developing effective strategies for explainability. Rather than solely engaging experts who possessed in-depth knowledge of AI, it was essential to involve doctors from diverse backgrounds and varying levels of familiarity with AI technology as in the end they all at some point will start having to work with it.

During the semi-structured interview we make use of planned prompts, which are prompts that are formally included in the interview protocol, and grand tour questions which are questions that ask respondents to give a verbal tour of a subject that they know well [82].

4.2.1. Method

The participants are a mix of doctors with expert medical knowledge about theoretical and practical IPF care and PhD's with less domain specific knowledge. The range of knowledge about AI and explainability varied from having no experience with it to having practical working experience. The interviews are in Dutch or English and online or in person depending on the preference of the participant. Because of the time constraint that most medical staff is under the interviews only took a minimum of 30 minutes, however when more time was available they took longer. Most interviews performed took longer than 30 minutes as participants noted they had extra time available. For scientific relevance the number of participants was set at a minimum of 12. The participants were recruited from the pulmonology department from Erasmus MC partly by giving a presentation to the PhD students asking them to participate. The more experienced pulmonologist were directly emailed to ask for their participation.

4.2.2. Aim of the Interview

The aim of the interview is to explore how medical staff can effectively collaborate with explainability to overcome adoption barriers in implementing AI in healthcare. The interviews address the limitations observed during the co-creation process, where the concrete prototype failed to capture the broader issues of AI explainability and distracted doctors with specific use cases. By adopting a flexible and inclusive approach, the focus is on explainability in a broad sense, encompassing system-level information and data. The aim is to identify opportunities for AI and explanations in respiratory medicine and understanding the needs of the medical staff.

Interviews play a crucial role as a primary data collection method in qualitative research. Their purpose is to delve into the profound experiences and insights of research participants, uncovering the significance they attach to these experiences [3]. While various qualitative research approaches are

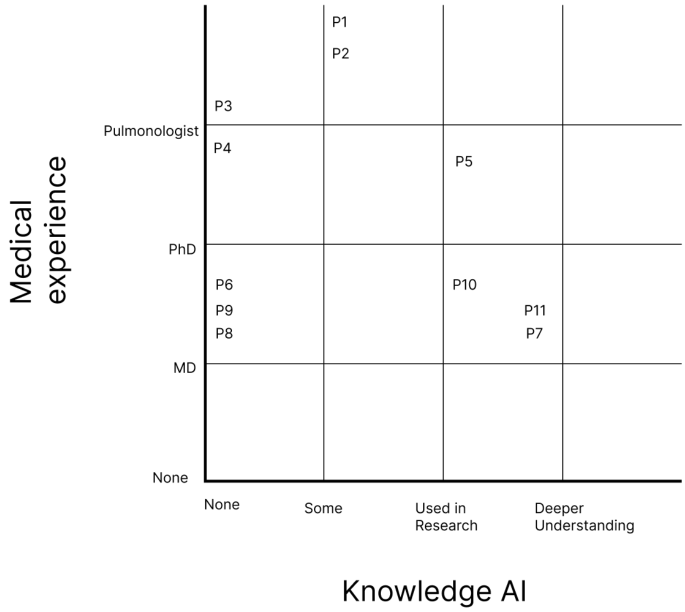


Figure 4.7: Plot denoting the medical expertise v.s. the knowledge participants from the interview have about AI.

available, and structured interviews offer consistency and ease of data analysis, semi-structured interviews offer a notable advantage. They strike a balance by allowing interviews to be targeted while providing investigators the freedom to explore relevant concepts that may arise during the interview process. This autonomy leads to a deeper comprehension of the explainability context being evaluated in the interview [5].

4.2.3. Prompts

In semi-structured interviewing, the open-ended nature of questions sets the stage for exploration, allowing both the interviewer and interviewee to delve into topics in more detail. However, there may be instances when an interviewee struggles to provide a comprehensive response or offers only brief answers. This is where planned prompts come into play, serving as valuable tools for the interviewer to encourage further reflection and generate detailed insights from the interviewee [92].

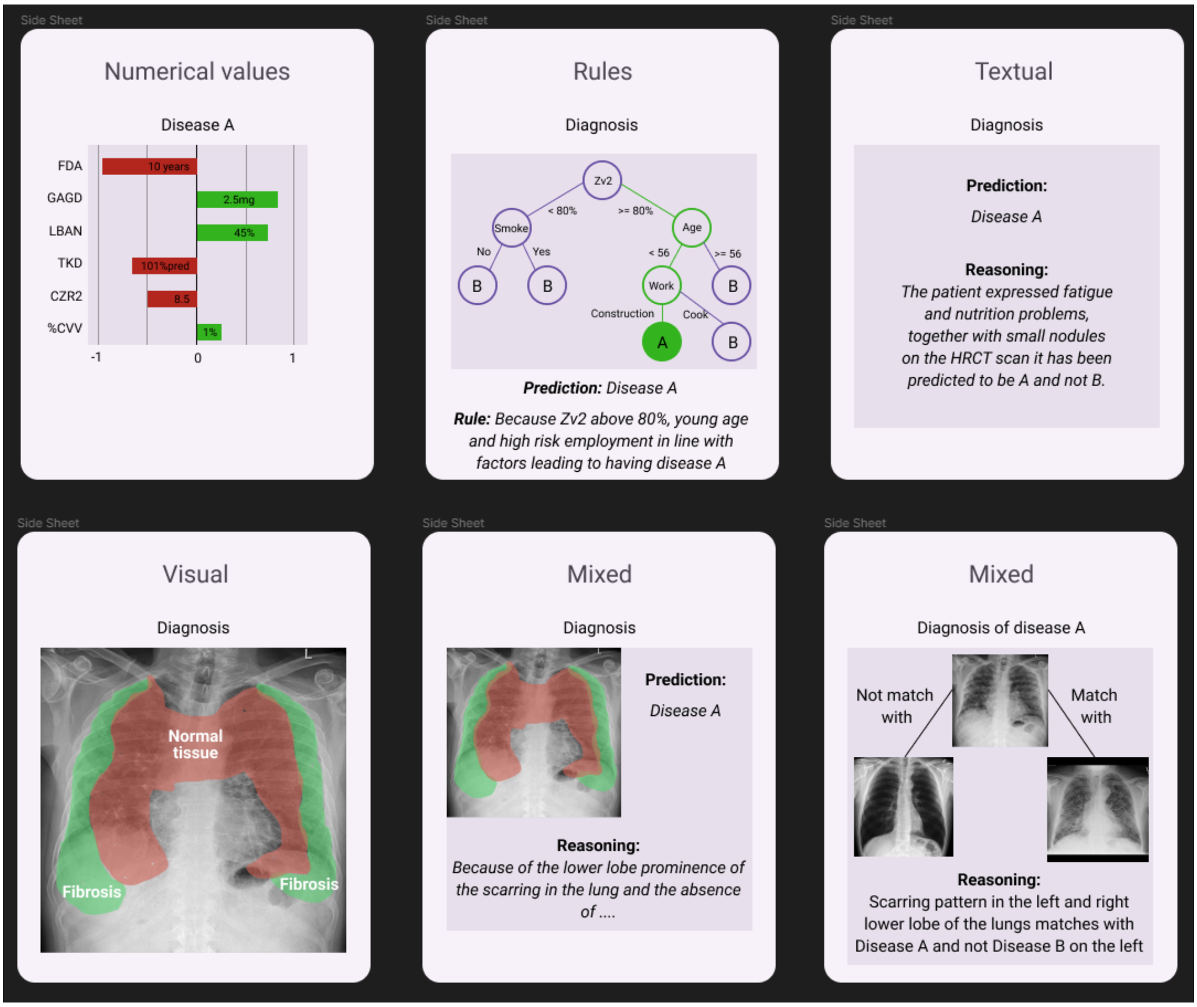
Planned prompts, also known as probes, play a crucial role in semi-structured interviews by providing direction and seeking additional information. They are formally included in the interview protocol and are as important as the questions themselves [82]. These prompts serve two key purposes: they keep the conversation flowing, ensuring that participants continue to share their thoughts and experiences, and they come to the rescue when responses become vague or incomplete [48].

Why Use Prompts

To address the challenge of involving doctors with varying degrees of AI expertise, the co-creation session identified the value of using prompts as a means to stimulate critical thinking about explainability and AI-related challenges. These prompts, carefully designed to provoke reflection and encourage discussion, offer a starting point for doctors to explore the implications of AI adoption within their professional domain. By leveraging prompts, we aim to facilitate an ongoing dialogue that empowers healthcare professionals to actively participate in shaping the future of AI in healthcare without needing to teach them about the underlying concepts of AI and explainability. Planned prompts play a pivotal role in semi-structured interviews, allowing us to elicit further explanation, delve deeper into topics, and gain a richer understanding of the interviewees' perspectives and experiences.

What are Prompts

Prompts play a crucial role in supporting participants who may encounter difficulties in understanding a question or formulating an answer. These prompts serve as additional pieces of information provided to the participant, helping to clarify the question or spark ideas that may have been temporarily inaccessible. By offering prompts, we aim to facilitate a deeper exploration of topics, fostering a more comprehensive and meaningful exchange of insights.



Prediction: Disease A

Rule: Because Zv2 above 80%, young age and high risk employment in line with factors leading to having disease A

Prediction:

Disease A

Reasoning:

The patient expressed fatigue and nutrition problems, together with small nodules on the HRCT scan it has been predicted to be A and not B.

Prediction:

Disease A

Reasoning:

Because of the lower lobe prominence of the scarring in the lung and the absence of ....

Reasoning:

Scarring pattern in the left and right lower lobe of the lungs matches with Disease A and not Disease B on the left

Figure 4.8: Six different types of mock visualisations shown to medical staff as a prompt in order to help them think about different types of forms of AI explanations.

One form of prompt we employ is visual aids, which can take various formats related to explainability. These visuals serve as supplemental information, providing participants with a visual representation of different facets of explainability. These visuals help participants grasp abstract concepts more easily, promoting a clearer understanding of the subject matter. An example of a visual prompt we use can be seen in Figure 4.8, in this figure six different configurations and formats of explanations are displayed. The categories of these different types of visualisation is taken from [149] who categorise explainability outputs into five types: numerical, rules, textual, visual and mixed. Another type of prompt we utilise is the provision of examples drawn from the journey map of doctors interacting with patients. These examples act as catalysts for participants to reflect on their own experiences and articulate their thoughts more effectively. By presenting real-life scenarios and moments of interaction, we encourage participants to contextualise their insights and explore potential implications of AI explainability within their specific domain. All prompts can be seen in item 6.3.

Reducing prompt induced bias

While prompts serve as valuable tools for facilitating thoughtful discussions, it is essential to strike a balance that allows for engagement without unduly influencing participants' perspectives. In traditional one-on-one semi-structured interviews, the risk of unintentionally introducing bias is inherent. As interviewers, bring in their own biases, conscious or unconscious, which may inadvertently influence the responses and perspectives of the medical staff. To minimise the potential bias introduced by prompts, a conscious decision was made to provide them sparingly and only when necessary. By refraining from an excessive number of prompts, it was aimed for to give participants the freedom to express their thoughts and opinions more independently, allowing their perspectives to emerge organically. This approach tries to balance between eliciting thought and the exploration of their own ideas, fostering an environment where participants could offer their unique insights without feeling confined by preconceived notions of explainability which exist in the field.

4.2.4. Interview Questions

All prepared questions of the semi-structured interview can be found in Appendix C, besides these main questions other deepening questions were asked during the session which differ for each participant and are therefore not included. Some of the main interview questions were shortly introduced with a small story to give the participants some context before answering the question. These introduction stories can also be found in this same appendix chapter. The interview questions were based on the learning from the research into the medical background and context from chapter 3 and from the gathering of initial insights section 4.1. These learning were also made into concise bullet points which were used to prevent common pitfalls while crafting and fine tuning the questions. These learning can be found in Table 4.3. The questions were made in collaboration between the author and 2 PhD's, one with experience within explainability and one with experience with holding semi-structured interviews. As validation step a number of trials with master students with experience with explainability were performed, based on these results the questions were adapted.

Structure of the Questions

The questions are based on the idea that by utilising medical staff their expertise while talking about explainability, without needing any preexisting knowledge about AI or explainability, their answers will be more interesting and valuable than when showing them state-of-the-art explainability formats of which they could have no idea what the actual advantages and drawbacks on the short and long-term are. The questions are structured in a logical order where first questions are asked about how the medical staff provide their care to patients. From these first questions about how they provide care, we look at the pain points that they encounter during their regular work. These pain points are then used to think about ”magical” AI solutions where they don't think about the actual inner workings of the systems and we start to ask them about how they would work with these solutions. From this we come to the actual more direct questions about explainability that focus on the different types, formats, interaction possibilities, trust, validation and moments that explainability can be used. The main focus and most of the time during the interviewing processes is spent on question 4, where the actual explainability gets discussed.

Questions

Current practices

Aim: The information needs of the first question is to understand the current practices followed by doctors. To achieve this, the interview begins by familiarising the doctors with the patient journey, using a validated patient journey obtained from the co-creation session. The aim is to establish a common vocabulary and ensure that doctors can relate to the patient journey being discussed. The question is a Grand Tour Question [82] which gets the medical staff talking about their experience and know-how but in a fairly focused way. These type of questions have the benefit of giving you a sense of what the normal treatment for patients is like.

Questions: The question posed to the doctors is focused on understanding their approach to work before discussing any automated systems. The question asks them to share how they would

Type	Insight	Learned from
Time	Time is limited so should not take a lot of time to use the explainability.	Observation, co-creation session follow-up interview
Verifiability	Doctors want AI that is validated with large patient cohorts, so for relevance needs not only be for validation.	Follow-up interview, paper [9]
	If the doctor can just verify the output with the patient in front of them, they won't use it.	Co-creation session
Trust	When doctors use a tool they already trust that it works otherwise it would never be in use.	Co-creation session
	If the doctor recognizes certain output from the explainability and recognize it they quickly believe it, even if it is not correct.	Co-creation session
Topic	It depends on the topic if they would use explainability. Topics they assume they know or expect they will not look into.	Co-creation session
Global v.s. Local	For exploration reasons they can be interested in more global explanation of the model to base future research off.	Co-creation session, follow-up interview
Too much info	Doctors often believe that they do not need the extra information as they normally don't have it, especially when they think the information is illogical.	Co-creation session
Knowledge	The level of knowledge of AI and medical experience influences the needs of the doctor.	Co-creation session, paper [156]
	We assume from the information gathered from the doctors that we have to design for a very low level of knowledge about AI.	Self-reported information participants

Table 4.3: List of insights gained from the researching medical background and context, co-creation session and follow-up interviews.

approach different stages of the patient journey, starting from before diagnosis until the end stage of a specific disease. Concrete examples are encouraged to provide a deeper understanding.

Prompts: Additionally, prompts are provided to assist the doctors in their responses, such as creating treatment plans with patients and understanding patients' needs before check-ups. The question also seeks to uncover whether the doctors' approach is based on clinical standards or if it is derived from their experience in the field. By addressing these aspects, the interview aims to gain insights into the doctors' current practices and their individual approaches to patient care.

Understanding pain points in current practice

Aim: In question 2 the focus shifts towards understanding the pain points in the current practices of doctors. The goal is to identify common problems they encounter, how they currently address those problems, and to prompt them to consider the importance of human collaboration and reliance in the workplace before discussing AI solutions.

Questions: The question posed to the doctors aims to uncover the most challenging or tedious tasks they face when treating patients. The term ”tedious” is clarified as tasks that are repetitive, requiring significant effort, time, or attention when it may not seem necessary. On the other hand, ”challenging” tasks refer to those that are cognitively, emotionally, or procedurally demanding.

Following that, the doctors are asked about the actions they take in such challenging or tedious situations. Examples provided include seeking advice from colleagues or nurses and spending more time reviewing the clinical records of a patient. This question seeks to understand how doctors currently address these difficulties and find potential areas where improvements or assistance could be beneficial.

By exploring these pain points and the doctors' current strategies, the interview session aims to gain insights into the practical challenges faced by medical staff and lay the groundwork for discussions around potential AI solutions.

AI in healthcare

Aim: In question 3 the focus shifts to discussing the application of Artificial Intelligence in healthcare. The goal is to explore the doctors' perspectives on AI systems and how they perceive the potential benefits and applications of AI in addressing the pain points previously discussed.

The doctors are reminded of the ongoing discussions surrounding the use of AI in healthcare, ranging from faster diagnosis to personalised treatment. Following this, they are asked to consider the pain points they mentioned earlier and whether they believe an AI or other technological solution could be beneficial in addressing those challenges. To help them envision the possibilities, they are encouraged to think of it as a 'magic wand' that could solve their problems.

Questions: The subsequent questions delve into their envisioned use of such technologies. They are asked how they would utilize these technologies, whether as collaborators, recommenders, or in other roles, and the reasons behind their choices. Additionally, the doctors are prompted to consider the factors they would take into account before deciding to rely on AI systems. Lastly, they are given the opportunity to identify any other cases or scenarios where they would like to have such systems integrated into their practice.

By engaging the doctors in this discussion, the interview aims to gather their insights, perspectives, and considerations regarding the potential use of AI systems in healthcare.

Explanations

Aim: In question 4 the focus turns towards exploring the topic of explanations and understanding what doctors consider to be 'good' explanations in the context of AI systems. The goal is to delve deeper into their thoughts on the information they would like to know about the system, when they would want it, and how it should be displayed. The interview refers back to the concerns raised in the previous section regarding the reliability and interpretability of AI systems.

Questions: The doctors are asked if they share the concerns raised by other researchers about AI and whether there is specific information they would want to know before or while using AI in their work. Examples of cases, such as IBM Watson's reliability issues and a pneumonia risk prediction system's incorrect correlation, are provided as prompts. Subsequently, the doctors are asked if having additional information would be helpful when using a hypothetical AI system and whether it is connected to the explanations or justifications they typically provide to colleagues or patients.

The questions then explore the specific kind of information the doctors would like to see covered, including capabilities and limitations of the system, features being used, data-related information, connections with medical literature, and limitations of the algorithms. The timing of receiving this information is also discussed.

The doctors are asked about their preferred format for presenting the information. The reasons behind their choices and how the chosen format helps them work on their identified pain points or interesting aspects are explored. The interview also delves into how doctors would use the provided information, whether they would trust and rely on it or consider it as an extra data point to inform their next steps.

Finally, the doctors are asked to reflect on whether knowing everything about the system would change the way they use or rely on it. By asking these questions, the interview seeks to gain insights into the doctors' perspectives on explanations, their information needs, preferences for timing and presentation, and the role of information in their decision-making and trust-building processes.

Prompts: With prompts suggesting various scenarios for the timing of explainability such as before implementation, during the early stages for learning, anytime to put the system through its paces, or based on specific triggers related to their decision-making process. Additionally, when the doctors are asked about their preferred format they receive visual examples for presenting the information, including numerical values, rules, textual/conversational formats, visual representations like heatmaps, or a combination of formats.

4.3. Interview Analysis

Thematic analysis is a tool offering valuable insights into the collective meanings and experiences of healthcare professionals. By identifying patterns and organising meaningful themes within a dataset, this method enables to understand and analyse the discussions surrounding explainability and clinical AI. Thematic analysis grounds qualitative research, ensuring its findings can be comprehended integrated into the multi-method research performed. In this study, an inductive analysis approach will be employed, aiming to generate fresh insights and theories from the collected data, without relying on pre-existing frameworks as they are lacking in completeness.

A component that can be used in thematic analysis is descriptive coding, which involves assigning labels to qualitative data passages to summarise the essential topics. This coding aids in identifying, organising and categorising the different categories of key themes and concepts. To support this analysis, ATLAS.ti is chosen, a widely used computer-assisted qualitative data analysis software. ATLAS.ti offers the flexibility to accommodate various theoretical approaches and analysis methods, making it fit perfectly to the type of analysis chosen. The analysis using ATLAS.ti will be conducted by a team of three researchers, ensuring a comprehensive and rigorous examination of the semi-structured interviews.

4.3.1. Thematic Analysis

Thematic analysis is a method used to identify and organise patterns of meaning, themes, across a dataset, focusing on collective or shared meanings and experiences. It helps researchers make sense of the commonalities in the way a topic is discussed or written about. TA is flexible, allowing researchers to analyse meaning across the entire dataset or explore specific aspects in depth, uncovering both obvious and latent meanings [21].

The two main reasons to use TA are its accessibility and flexibility, providing a systematic approach to qualitative data analysis that is suitable for newcomers to qualitative research and research teams with diverse expertise. Both of those characteristics are suitable for the team used to carry out the thematic analysis of the semi-structured interviews. TA separates qualitative research from broader theoretical debates, making its results accessible to a wider audience and facilitating its integration into multi-method research and participatory research projects [21].

Inductive v.s. Deductive

Two different types of qualitative analysis exists, deductive and inductive analysis [118]. Inductive analysis involves generating new concepts, explanations, results, or theories directly from the specific data collected in a qualitative study. It is a bottom-up approach where patterns and themes emerge from the data, leading to the development of new insights or theories. On the other hand, deductive analysis involves evaluating the extent to which qualitative data in a study align with existing general conceptualisations, explanations, results, or theories. It is a top-down approach where researchers start with pre-existing theories or concepts and then examine the data to see if they support or confirm those theories. As there is no satisfactory framework yet to place the interviews in, an explorative view will be taken. Therefore inductive analysis was picked to analyse the semi-structured interviews.

Descriptive Coding

Descriptive coding in interviews refers to the process of assigning labels, usually in the form of nouns or short phrases, to qualitative data passages [134]. The goal is to summarise the fundamental topics covered in the data. By creating descriptive codes, researchers build an inventory of topics that can be used for organising and categorising different types of data, including field notes, interview transcripts, and documents. This approach is particularly valuable for research projects that deal with social topics rather than focusing on specific social actions. In this case it was chosen to code the interviews by short phrases to make the later merging of topics into themes easier.

Tools and Team

The tool for carrying out the thematic analysis was chosen to be Atlas.ti[3]. Atlas.ti is a widely used computer-assisted qualitative data analysis software that has found application across various disciplines including healthcare. It offers flexibility in supporting different theoretical approaches and data analysis methods, such as thematic analysis, making it a valuable tool for the qualitative research performed in this thesis. The analysis using this tool is carried out by a team of three researchers who perform the first and second cycle interview analysis of all the performed semi-structured interviews.

4.3.2. Experimental Work

This section delves into the methodology and process undertaken to analyse the data collected from the coding interviews. This phase of the study is pivotal in uncovering insights and understanding the nuances embedded within the collected data. Through a systematic approach involving multiple cycles of thematic analysis, the researchers aimed to unearth underlying patterns, discern relationships, and extract meaningful clusters of information from the interview data. This section elucidates the sequence of cycles employed in the analysis, namely the first cycle focused on coding interviews, the second cycle centered on identifying clusters, and the third cycle dedicated to identifying overarching themes. Each cycle played a significant role in refining the analysis, enhancing the comprehension of the dataset, and ultimately forming the basis for the themes that are later discussed in subsequent sections of this thesis.

First Cycle - Coding Interviews

During the first cycle of thematic analysis, the researchers familiarised themselves with the interview data by transcribing, cleaning and conducting multiple readings to achieve a comprehensive understanding of the content. The data coding process commenced by identifying and assigning descriptive tags that captured the explicit content or surface-level meaning of the interviews. These codes, consisting of concise short phrases, represented specific concepts, ideas, or themes present in the interviews. The researchers manually applied the codes to relevant segments of the data, establishing an initial coding scheme. The computer-assisted qualitative data analysis software Atlas.ti was utilised to facilitate the coding process. The objective of this initial coding phase was to organise the data and identify potential themes that could emerge from the analysis. An example of statements from the interviews and codes given to specific excerpts withing those statements can be seen in Table 4.4.

Statement	Descriptive code
I think they need it some explanation and it won't be sufficient to only say, well, "it's a pneumonia".	Need for explanations
So what kind of information would I need… Yeah, the process of how it got, from the information I put into the system, to the answer and that can be very. Of course it depends on what kind of what kind of question I asked the system, but if you for example took the X-ray.	Depending on the underlying AI system
So we're used to interpreting an X-ray and getting a report with it, and now [with XAI] you can [get that too]. It's nicely combined and I think quicker than when you have to wait for the radiologist. And of course, you can interpret it yourself is also very nice to have the report of a radiologist as a sort of second opinion.	Quicker then radiologist

Table 4.4: Examples of statements from the semi-structured interviews and descriptive analysis codes assigned to the excerpts from the interview of participant P11.

From this first cycle we identified a total of 233 codes, most of these codes could be easily categorised into the interview sections they were from, as those sections covered specific themes. However, to more organically discover more in depth themes, during the second cycle, it was chosen to refrain from

[3] Atlas.ti software page www.atlasti.com

adding these to the initial descriptive codes. In total more than a 100 unique codes were found in this step.

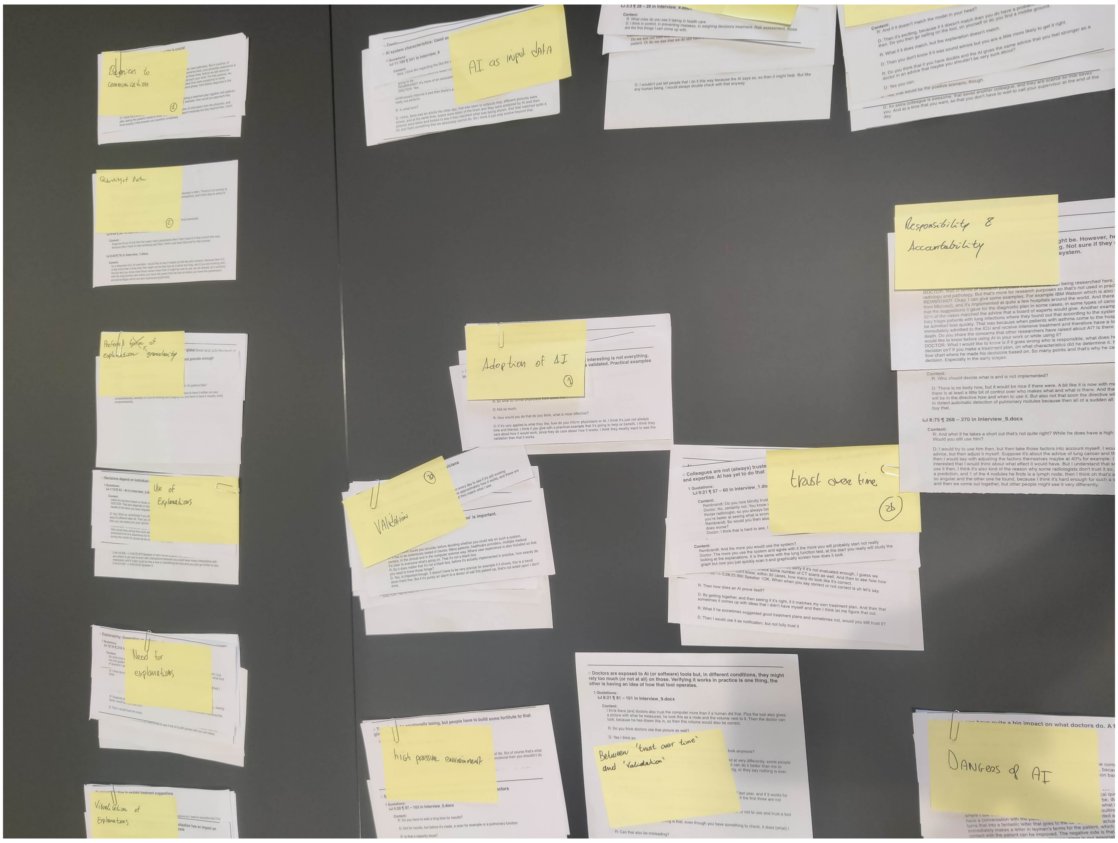

Barriers to communication
Quantity of Data
Preferred form of explanation granularity
Use of Explanations
Need for explanations
Visualisation of Explanations
AI as input data
Responsibility & Accountability
Adoption of AI
Validation
trust over time
high pressure environment
Between 'trust over time' and 'Validation'
DANGERS of AI

Figure 4.9: Second cycle of thematic analysis for the output of the semi-structured interviews. Codes and their respective quotes from the interviews are clustered on a large table by the researchers.

Second Cycle - Identifying Clusters

In the subsequent second cycle of thematic analysis, the researchers iteratively refined and expanded upon the initial coding scheme derived from the first cycle. This involved a systematic exploration of the data to uncover underlying patterns, relationships, and connections within the coded segments. The researchers engaged in a rigorous process of grouping related codes together to form potential clusters, ensuring coherence and consistency. As new clusters emerged, the coding scheme was continuously reviewed and adjusted accordingly. This iterative process involved constant comparison and analysis of the data, codes, and emerging clusters, enabling a more comprehensive and nuanced understanding of the dataset. Additionally, discussions and consultations within the research team were conducted during a meeting to ensure agreement and consensus on the identified clusters. By delving beyond descriptive coding, this thorough analysis sought to reveal deeper insights and interpretations.

During the second cycle the grouping was performed by physically placing the codes, together with their respective quotes from the interviews, into groups. These groups were then placed on large tables were their respective position on the table related to the topic of the respective groups, this can be seen in Figure 4.9. These groups were refined to subdivide the groups into more specific smaller groups and made into a hierarchical structure where some groups were subgroups to other overarching topics. This process ultimately resulted in the 22 clusters that can be seen in Table 4.5.

Third Cycle - Identifying Themes

The clusters found in the second cycle were then used by the researchers to construct 3 main themes which lead from the clusters found but also complement the research questions making it easier to

Clusters		
Uncertainty	Knowledge About AI	Responsibility & Accountability
High Pressure Environment	Visualisation of Explanations	Validation
Barriers to Communication	Preferred Granularity of Explanations	Regulations
Information Sharing	When to See an Explanation	Need for Explanations
Perceived Benefits of AI	Explainability to Collaborate	Tools
Adoption of AI	AI as Input Data	Communicating and Understanding Patients
Applications of AI	AI for Collaboration	
Dangers of AI	Trust over Time	

Table 4.5: The 22 clusters of codes found in the second cycle analysis of the output of the semi-structured interviews.

answer them in later sections of this thesis. The main themes seen in Table 4.6 are relating to where in the journey the use of AI could benefit healthcare, what information medical staff want and under which conditions they will interact with the explanations. The actual wording and description of the contents of the themes can be seen in the next section.

Theme	Groups	Sub-groups
Enhancing Clinician Decision-Making Throughout the Patient Journey	Applications of AI	Perceived Benefits of AI
		Information Sharing
		Barriers to Communication
	Dangers of AI	-
	Understanding AI Principles	-
Personalised Explanations in Explainable AI for Healthcare	Applications of AI	-
	XAI in Healthcare	Visualisation of Explanations
		Preferred Granularity of Explanations
		When to See an Explanation
		Explainability to Collaborate
		Need for Explanations
Aligning with doctor needs and values	Patient-Centric Care	Uncertainty in Care
		High Pressure Environment
		Tools of Doctors
		Communicating and Understanding Patients
	Adoption of AI	Barriers of adoption AI
		Convincing Doctors
		Friction Outside World
		Regulations
		Responsibility Accountability
		Validation
		Regulations
	Collaboration and Trust	AI for Collaboration
		Trust over Time
		AI as Input Data
		AI and Doctors

Table 4.6: The 3 main themes found in the analysis, together with the relevant groups of clusters for each theme, some of the groups are relevant for more than one theme.

4.3.3. Results Interview Analysis

The analysis of in-depth interviews with medical staff has provided valuable insights into the perceptions and potential applications of Explainable AI in the healthcare domain. Throughout the interviews, three prominent themes emerged, shedding light on the ways XAI can enhance clinical decision-making, personalise explanations for healthcare tasks, and align with the needs and values of doctors. Each of these themes represents a critical aspect of integrating AI technology into the complex landscape of medical practice. We will shortly summarise the themes, followed by an extensive view of the topics that each of these themes encompasses. This extensive explanations of themes is structured by the groups and sub-groups that it envelopes which can be seen in Table 4.6.

Theme 1 - Enhancing Clinician Decision-Making Throughout the Patient Journey

The first theme centers on the identification of specific moments in the patient journey where XAI can be instrumental in supporting clinicians. Medical professionals recognise the potential applications of AI systems to aid decision-making during crucial stages of patient care and doctors are increasingly intrigued by AI's potential to enhance patient care, envisioning its applications in automating administrative tasks, streamlining patient interactions, and supporting treatment planning and monitoring. The superiority of AI in tasks like image analysis in radiology and pathology is acknowledged, along with its potential to provide empathy and aid patient queries, particularly when time constraints hinder thorough responses. While acknowledging AI's transformative capacity, integrating it effectively into clinical workflows requires overcoming communication barriers and fostering an open attitude towards its role as a valuable supportive tool. The doctors acknowledge the associated risks and limitations of AI adoption. In Pulmonary Medicine, doctors discuss AI's potential and the challenges of implementing it, addressing concerns about time-consuming administrative tasks through automated reporting templates and transcribing consultations. AI's transformative role in patient monitoring, identifying sudden deviations in lab values for prompt intervention, is recognised, along with its capability to aid treatment planning and track side effects and lab values. The importance of ”explainability” as a means to address these concerns becomes apparent, allowing doctors to better comprehend AI-driven decisions and fostering trust in the technology.

Theme 2 - Personalised Explanations in Explainable AI for Healthcare

The second theme delves into the significance of personalised explanations in the realm of XAI for healthcare. It highlights the need for explanations that go beyond mere user types and adapt to the unique contexts of medical tasks. This adaptability is crucial in facilitating the integration of XAI into their decision-making processes. Through the analysis, it became evident that personalisation should encompass factors such as medical experience, specialisation, individual patient cases and timing. The timing of explanations, whether it's before using AI, during disagreements between the AI and doctors' decisions, or during routine patient evaluations should be corresponding to the doctors information need. This theme emphasises the crucial role of context-driven explanations in complex patient scenarios, emphasising the interplay between Explainable AI, collaboration, and trust in the healthcare setting and how personalised explanations can facilitate collaboration among healthcare teams, as well as the potential challenges related to doctors' tendencies to resist guidance, showcasing the complexity of human-AI interaction.

Theme 3 - Aligning with doctor needs and values

For the third theme collaboration between doctors and XAI systems emerges as a pivotal enabler for the successful integration of AI in healthcare. Explanations play a central role in facilitating the collaborative dynamic between doctors and AI, where explanations bridge the gap between technical outputs and medical decisions. They portray their desires for AI as a tool that streamlines workflows, allowing doctors to focus on decision-making. However, this is limited by practicality, stressing the need for evidence and tangible benefits to overcome scepticism and integrate AI effectively. Doctor's decision-making authority is seen as crucial and is seen as hard to retain if explanations are not present. The challenges of continuous validation, transparency, and clinical relevance are are stressed combined with the importance of explanations in identifying patterns in AI behaviour and reducing uncertainty. This intertwines with the broader narrative of collaboration and trust, the symbiotic relationship between doctors and AI, characterised by collaboration and trust-building over time. While underscoring the necessity to find the right balance between AI assistance and medical expertise. Aligning AI with doctors' needs, values, and trust is a fundamental aspect of enhancing AI adoption and realising its positive effects in the medical field.

Structure of remainder of the section

By exploring these three main themes, this analysis sheds light on the potential of Explainable AI in transforming healthcare decision-making, personalising explanations for medical tasks, and aligning AI technology with the needs and values of doctors. Understanding these themes provides essential insights into the challenges and opportunities surrounding AI adoption in the healthcare industry, paving the way for responsible and effective integration of AI technology in clinical practice. These themes are then further analysed during the discussion in Chapter 5.

The remainder of this section presents the results of the interview session with the medical professionals in more detail. It presents the main themes together with the groups and sub-groups that are relevant to the themes as seen in Table 4.6. These groups are described through sharing the general opinion displayed by the doctors during the interview sessions together with relevant quotes.

Theme 1- Enhancing Clinician Decision-Making Throughout the Patient Journey

During the analysis of in-depth interviews with medical staff, a significant theme emerged, revealing where in the patient journey are the main moments for clinicians where XAI can be helpful. These moments were shaped by possible applications they saw of AI systems and limited by the dangers that they see in the use of AI and the knowledge that they posses about how to use and evaluate AI. Explainability can, however, help with tackling some of these problems.

1.1 Applications of AI for Addressing Healthcare Challenges

AI has garnered the interest of doctors due to its potential to address specific challenges and enhance patient care. Doctors envision AI as a valuable tool for automating administrative tasks, streamlining patient interactions, and supporting treatment planning and monitoring. In radiology and pathology, AI's capabilities to outperform humans in tasks like image analysis are recognized. Moreover, doctors see the potential of AI in delivering empathy and supporting patient queries, especially in cases where time constraints limit their ability to respond comprehensively. While doctors acknowledge the transformative potential of AI in various areas, the successful integration of AI into clinical workflows requires addressing barriers to communication and encouraging an open mindset towards embracing AI as a supportive tool.

Automating administrative medical tasks alleviating documentation burdens Doctors have engaged in insightful discussions about the potential applications of AI in Pulmonary Medicine and the challenges they face in implementing and adopting AI solutions. One of the primary pain points identified by doctors is the burdensome and time-consuming nature of administrative tasks. As one doctor [P6] articulates, *"I think the administration, if you want to do a certain examination you have to do the requests of that, sometimes you still have to check if that is done. Writing letters as well."* They believe that automatically generated templates for patient reporting can streamline processes, saving considerable time and effort. As one doctor [P10] states, "*The documents that you have to write either for regulatory purposes or to the hospitals or for the patients or anything like that, it always there is a template that you use but it changes a bit according to the concrete drug that you are testing, the concrete PI that you are working with or yeah things like this."*. The sheer volume of patients coming through further exacerbates this issue, making it a significant challenge to handle efficiently. Additionally, AI could streamline patient interactions by transcribing conversations during consultations, alleviating doctors from the burden of documentation. As one doctor [P4] points out, *"I think it would be very useful in some specific cases such as radiologists where a lot of data is available that is also standardised and generalised."*.

Tracking patient's lab values automatically to detect and predict adverse reactions The doctors recognise the transformative potential of AI in patient monitoring and the detection of errors or concerning changes. They see an opportunity for AI to play a crucial role in alerting doctors to sudden deviations in patient values, prompting timely intervention and facilitating improved patient care. As one doctor [P8] emphasises, *"Mainly that it shows to the doctor of this has been blown this is the alarm, that it shows what is going on. So both objective things like lung function, that you can see in 1 storage this is going on, so come to the hospital or not."*. Moreover, the doctors envision AI as a valuable tool for treatment planning and tracking side effects and lab test values. By automating the process, AI can schedule treatment intervals and analyse trends in lab values, enabling informed adjustments to be made before adverse reactions occur. As one doctor [P9] explains, *"Suppose eventually you can go somewhere where you can see trends. That you know before an adverse reaction or complication occurs that you can already tell that that's going to happen. And in diagnostics itself? I'm also involved in that myself, in pathology you now have AI being run over tissue to see if you see any connections, that you can just run a program over it and the program will tell you what's there."*.

Evaluating images in radiology They also recognise the potential advantages of AI in radiology, where a wealth of standardised and generalised data is available. In the context of measuring tumour diameter, doctors discuss a specific example from biopharmaceutical processing technology. The challenge lies in evaluating the growth of cells cultivated for biological drugs, which can be attacked by bacteria and viruses. Implementing AI tools for measuring tumour diameter could be a viable solution, as it would provide faster evaluations, especially for professionals with a mathematical background.

Delivering empathy to patients One critical aspect limiting doctors care is the time constraints faced by doctors, making it difficult to deliver longer empathic responses or answer patient questions on online fora. Nevertheless, the doctors recognise that AI systems, could potentially fill this gap and provide valuable support in addressing patient queries. As one doctor [P4] shares, *"I think it is hard to see exactly where the applications of AI go. [...] As a doctor, you might not always have the time or the opportunity to deliver that empathy to patients. Longer responses, as a doctor you have no time to spend so much time for that."*

1.1.1 Perceived Benefits of Using AI in Healthcare

When asked what the perceived benefits of AI can be the following interesting points were mentioned by the doctors.

Helping with patient health data for improved personalised care In the preparation of examinations, AI can play a significant role in providing assistance. Doctors mention that when dealing with rare mutations, AI can analyse data and generate reading lists or summaries for a more informed decision-making process. This streamlines the workflow and allows doctors to focus on providing the best possible care to their patients. According to one doctor P2, stated during the co-creation session, AI can be a valuable tool to support consultations by better identifying a patient's needs, especially when providing assistance at home. The doctor explains that AI can offer tailored information provision, which is essential since there is no one-size-fits-all approach in patient care. Patients often struggle to retain all the information provided during consultations, and AI can bridge this gap by ensuring vital information is readily available.

Outperforming doctors for performing certain tasks Moreover, doctors acknowledge that AI systems may outperform humans in certain tasks, such as judging X-rays and CT scans. They recognise that AI can excel in radiology and pathology, offering precise and efficient assessments that benefit patient care.

1.1.2 Information Management About Disease and Patient

Doctors often encounter patients who have previously undergone diagnostics elsewhere, seeking a clear diagnosis or desired treatment. Shared care arrangements are established to facilitate seamless coordination between healthcare facilities, ensuring the patient receives the best possible care.

Gathering available information to enable better diagnosis Gathering and evaluating all available information is essential for accurate diagnoses and effective treatment plans. As one doctor [P11] mentions, *"the most important thing is to gather all the information you have at that point, evaluate that to come to a diagnosis."*. For patients with rare mutations, doctors need to access relevant studies and treatment options. This involves extensive preparation and collaboration with other medical institutions. As one doctor [P11] explains, *"sometimes we treat people in studies, we look for a hospital where they do treat these mutations in studies. A lot of the work is preparatory."*. One area where AI integration can significantly improve clinical workflows is the timely access to referrals and reports. As Doctor [P6] mentions, *"Referrals from other hospitals are difficult... if you could get a preliminary report it would be quicker."*. AI-powered systems can help expedite the processing and analysis of radiology reports and other test results, leading to faster decision-making and more efficient patient care.

1.1.3 Barriers Within Doctor-Patient Communication

The successful implementation and adoption of AI in clinical practice hinge on effective communication and information sharing, thus not running into any barriers within that communication process.

Reducing barrier for better communication According to Doctor [P8], *”the whole consultation, the transfer of information from the physician, and after asking the patient’s needs to know where you can support medically but also beyond that,”* are essential aspects of patient care. The implementation of AI in clinical practice raises the question of dealing with potential resistance from doctors. As Doctor [P8] points out, some doctors may find it difficult to accept that they could be wrong. It is crucial to address this issue and provide education on the history of medical errors and breakthroughs to foster a more open mindset towards embracing AI as a supportive tool.

1.2 Dangers of Using AI in Healthcare

While doctors express optimism about the potential benefits of AI in healthcare, they are also keenly aware of the associated dangers. Concerns arise from conflicting goals between healthcare providers and companies developing AI solutions, leading to questions about trusting these solutions. Over-reliance on AI and confirmation bias, especially among less experienced practitioners, poses risks, highlighting the importance of maintaining a critical attitude. Opacity in AI algorithms, leading to a lack of understanding of their inner workings, raises doubts about the reliability of AI-generated outcomes. Additionally, doctors contemplate the potential impact of AI on traditional medical roles and worry about job losses in certain areas. Striking a balance between AI's assistance and preserving human expertise remains a critical challenge in the adoption of AI in healthcare.

Trusting in companies their biased AI evaluations One of the major concerns raised by doctors revolves around the conflicting goals of healthcare providers and companies. While healthcare providers aim to enhance patient care, companies often prioritise product sales. This incongruity creates challenges in trusting the AI solutions developed by these companies. One doctor [P7] explains, *"You have to trust such a company on its blue eyes that it is correct but that is not how we work at all in healthcare."*.

Over-relying on AI Predictions Doctors also recognise the potential of AI to assist with medical tasks, such as suggesting diagnoses or interpreting images. However, they caution against over-reliance and confirmation bias, especially among less experienced practitioners. As one doctor [P4] puts it, *"If the AI then gives a suggestion that gives the correct output most of the times then it gives a correct answer but if the AI suggest the wrong prediction the inexperienced ones give far more often the wrong answer while the experienced did this a lot less."*. The potential consequences of overreliance on AI systems in healthcare are a significant worry for doctors. While AI can save time and assist in decision-making, doctors stress the importance of maintaining a critical attitude and not blindly trusting AI predictions. As one doctor [p2] highlights, *"You do have to have a certain basic knowledge to deal responsibly with the data you get."*.

Disregarding what is inside the black box The quality of data and technical advancements are also crucial factors that impact the success of AI in healthcare. Doctors are concerned about AI becoming a black box, with developers not fully understanding its inner workings. This opacity raises questions about the reliability of AI-generated outcomes, as one doctor [P4] questions, *"Even the people who make these systems don't know what is happening inside."*

Losing jobs of medical staff Despite their reservations, doctors do acknowledge the potential benefits of AI in healthcare settings. AI's ability to assist with medical questions and streamline patient interactions is seen as a positive development. However, there are concerns about AI's impact on traditional medical roles, such as radiologists and pathologists. Doctors wonder about the future role of these professionals if AI systems continue to improve. The potential effects on healthcare jobs are a matter of mixed opinions among the doctors. While AI could address the looming shortage of healthcare personnel, there are also concerns about job loss in certain areas. Finding the right balance between AI's assistance and human expertise remains a critical challenge.

1.3 Understanding the Underlying AI Principles

To navigate the dangers of the use of AI responsibly, doctors stress the importance of AI literacy among medical professionals. Being aware of AI's capabilities, limitations, and potential biases is crucial for its successful integration into clinical practice. Furthermore, fostering collaboration between healthcare providers and AI developers is vital to ensure patient safety and the delivery of optimal care.

Improving openness to AI solutions Doctors' understanding and acceptance of AI systems were influenced by their level of knowledge and exposure to AI technologies. To ensure a smooth implementation and adoption of AI in clinical practice, doctors expressed the need for proper education and practical feasibility considerations for AI systems. Some doctors expressed difficulties in imagining AI systems and understanding how AI tools might lead to different treatment plans. One doctor [P6] said, *"At this point, I can't think about how an AI tool might lead to a different treatment plan from what's there."*. AI literacy played a significant role in doctors' understanding and acceptance of AI. The level of AI knowledge among doctors influenced their perception and acceptance of AI systems. Doctors who were more involved in research showed more openness to technology changes and were willing to explore new AI tools even when they contained some errors. On the other hand, doctors with lower experience tended to be more critical of lower accuracy AI predictions.

Adding AI to the medical curriculum to improve AI literacy There were concerns about AI literacy among doctors and the need for proper education to understand AI systems. Some doctors expressed scepticism about AI systems, and their concerns were tied to their limited knowledge of AI. One doctor [P7] mentioned, *"There are certainly some things that doctors don't understand."*. Doctors acknowledged the importance of having some level of AI knowledge before using AI tools in clinical practice. While they might not need to know all the technical details, a basic understanding of how the AI system works in the background was considered helpful. One doctor [P1] emphasised, *"But it is helpful to know how the system works in the background."*. The need for including AI and XAI in the medical curriculum for future students was mentioned. Doctors believed that having a basic level of AI literacy could be beneficial for future physicians.

Theme 2- Personalised Explanations in Explainable AI for Healthcare

Explainable AI could revolutionise decision-making and patient care. However, the efficacy of XAI relies on personalised explanations that transcend mere user types and adapt to the unique contexts of medical tasks. Our analysis revealed a notable finding—personalisation of explanations in healthcare is often reliant on more than just user types. Distinctions based on experience, specialisation, and patient cases emerged as pivotal factors. Especially for complex patient cases which necessitate broader group collaboration, underlining the need for context-driven explanations. Such distinctions pave the way for a deeper understanding of the interplay between Explainable AI, collaboration, and trust in healthcare settings.

2.1 Importance of Customised Explainability for Specific Cases

The efficacy of XAI, however, hinges on its ability to provide personalised explanations that extend beyond generic user types and adapt to the distinct contexts of various medical tasks. Our analysis of in-depth interviews with medical staff has revealed a compelling finding: explanations in healthcare should not be solely reliant on user types but is heavily influenced by factors such as experience, specialisation, and the intricacies of individual patient cases and the type of AI system used.

Adapting explanations for improving their usefulness This realisation underscores the paramount importance of context-driven explanations and highlights the importance of the connection between Explainable AI and its context in healthcare settings. Within the realm of AI applications in healthcare, the diversity of tasks and responsibilities entrusted to doctors necessitates a flexible and nuanced approach to information seeking. Our investigation has unraveled how the type of information doctors seek in explanations is contingent upon the specific medical tasks they perform. Understanding the nuances of this interdependence can foster AI systems that cater to the unique needs of doctors, aligning with their decision-making processes and fostering trust in AI-driven diagnoses.

2.2 Important Characteristics for Using AI Explanations in Healthcare

Within XAI in Healthcare, the analysis explored various aspects of explainability. Visualisation of Explanations garnered attention, showcasing how personalised visual aids enhance comprehension. Preferred Granularity of Explanations revealed the need for fine-tuning explanations to match individual task requirements. Deciding When to See an Explanation highlighted the dynamic nature of explanations based on specific medical scenarios. Furthermore, Explainability to Collaborate emphasised that personalised explanations empower collaborative efforts among healthcare teams. Lastly, the demand for Need for Explanations resonated across all groups, solidifying the significance of personalised contextual explanations.

2.2.1 Visualisation of Explanations

Successful implementation and adoption of AI in clinical practice heavily rely on intelligible and informative visualisations of explanations. By tailoring these explanations to the specific needs of doctors, providing relevant and comprehensible visualisations, and offering insights into AI decision-making, healthcare practitioners can foster trust and effectively integrate AI into their workflows. As the interviews demonstrate, context-driven explanations that consider the usage context and cater to doctors' preferences can play a pivotal role in advancing the integration of AI in healthcare.

Adapting visualisations to specific tasks for deeper understanding of the AI system According to the doctors, the key lies in creating visualisations that are tailored to the specific usage context. Hence, the doctors advocate for context-driven explanations that align with the specific questions they are trying to address. They emphasise that different branches of medicine may require different visual modalities. For example, a diagnosis tool may benefit from a bar plot numeric representation, while radiology might require a mix of images and rules. As one doctor [P8] suggests, *"It very much depends on what kind of questions you ask and what kind of application."*. For radiology, they express a preference for a mixed format that combines images and rules. This approach allows them to gain a deeper understanding of the decision-making process of the AI system. As one doctor [P6] puts it, *"For radiology, I would go more for the mixed so you can see what he's looking at then and what conclusion he reached there."*.

Showing the entire processes to improve the comprehensiveness of the explanation Furthermore, the doctors stress the importance of comprehensiveness in AI explanations. They appreciate visualisations that not only show the final output but also the entire thinking process of the AI system. *"The program needs to show its thinking process,"* says one doctor [P3], particularly when explaining the rationale behind a treatment suggestion. The interviews also reveal that doctors value AI explanations that include specific examples and counterexamples from past patient cases. Having this information helps build trust and confidence in the AI system, as it provides context and reference points. *"I think a combination of the rules and the examples and counterexamples,"* suggests a doctor [P2].

Reducing complexity in explanations to prevent information overload However, the doctors also acknowledge the potential risk of information overload. They caution against overwhelming visualisations that might be too complex for busy healthcare professionals to fully engage with. *"If it's a very big decision tree,"* notes one doctor [P6], *"then I wouldn't look."*.

2.2.2 Preferred Granularity of Explanations

The doctors express the need for AI explanations that are context-driven, informative, and understandable in the context of clinical practice. They seek explanations that directly relate to their patients, a global understanding of the AI system's functioning, and comprehensive information that aids their decision-making process.

Translating explanations to clinical practise to make them more relevant The doctors emphasise the significance of connecting AI explanations to clinical practice and patient-specific scenarios. One doctor [P6] mentions, *”If there’s an immediate translation to clinical practice...where it says for this patient it’s less useful and so on.”*. They seek explanations that directly apply to their patients and provide context to known edge cases of the AI system.

Adapting the depth of explanation to suit the situation The doctors stress the significance of comprehensiveness in AI explanations, indicating that they prefer sometimes information about the system rather than individual predictions. They express interest in knowing what factors the AI uses for its predictions, as this can help build trust in the system. However, they also note that the level of detail in the explanation should be modulated to suit the situation, as excessive detail may be overwhelming. When it comes to the level of detail in explanations, the doctors highlight the need for modulation depending on the situation. They prefer not to be overwhelmed with excessive detail, as one doctor [P5] states, *”If it’s too detailed then of course people aren’t going to look anymore. It’s really per application how detailed should be.”*. Some of the doctors expressed interest in different levels at different points in time. They desired comprehensive information initially to understand the AI system’s properties and reasoning. However, in their daily practice, they appreciated more concise and focused explanations.

Providing abstractive knowledge for making informed decisions Furthermore, the doctors express a strong interest in understanding how AI systems work, even if the underlying algorithms are complex. They want to know the global structure and processes without delving into intricate mathematical details. As one doctor [P5] explains, *”Then I would still want to know, what models are used, then this comes out, this is how the scoring works, then this comes out... It’s the same with the electronic nose, it also makes sensor readings and I don’t quite know how they work either, but I do know roughly how it uses them to arrive at a reading.”*. The doctors appreciate additional context provided by AI systems, such as patient categories used for algorithms and a margin of estimation when estimating risk. They find such information helpful in making informed decisions and understanding the results better.

2.2.3 When to See an Explanation

Understanding the optimal timing for accessing AI explanations is essential to enhance decision-making and patient care. In this section, we delve into the interviewees’ perspectives on the ideal moments to seek explanations during their medical tasks. Three key contexts emerged: explanations before use, explanations with a disagreement, and explanations during use.

Providing explanations before use of the system to point out limitations The interviewees preferred explanations to be readily available, context-driven, and aligned with their decision-making process. They emphasised the need to receive information about the limitations upfront to ensure responsible usage, as one doctor [P6] stated, *”Before you use it.”* Being aware of the system’s limitations enables them to make informed decisions and build trust in the AI’s recommendations. This means avoiding the use of AI in situations where its accuracy might be compromised, particularly in certain patient groups. As one doctor [P5] pointed out, *”You have to solve that on the front end by saying we’re going to use it with these people and not with these people.”*.

Providing explanations to solve a disagreement with the system The doctors also highlighted the importance of looking at explanations when there is a mismatch between the AI’s prediction and their own decision-making. They find it crucial to inspect the explanation in such cases to understand the reasons behind the discrepancy. As one doctor [P6] explained, *”If it doesn’t match then you can look at the explanation as to why that is. And then I would do look at the explanation.”* .

Providing explanations during use to improve the integration into the decision-making process The doctors expressed their expectation for AI explanations to be readily available for inspection at any time during patient evaluation. They preferred to have access to explanations with every decision they make, even if the AI's prediction matches their own. The explanation becomes an integral part of the decision-making process rather than a separate component. Furthermore, the doctors advocated for continuous evaluation and testing of AI systems in clinical practice. They suggested using a trial-like approach initially to assess the AI's performance and refine its predictions over time. This ongoing evaluation process builds trust and confidence in the system and lowers the need for extensive explanations.

2.2.4 Explainability to Collaborate with the AI System

Collaboration between medical professionals and AI systems holds potential for improving patient care and decision-making. In this section, we explore the insights provided by the medical staff in the dynamics of explainability in fostering effective collaboration between doctors and the AI tools.

Interacting with the explanations for better understanding the AI One significant aspect raised by the doctors is the need for interactive explanations that cater to the experience of the medical professionals. As one doctor [P8] highlighted, *"I think it's important that it's visually at a glance but that if you want more information that you can zoom in for more information just. I think a doctor if he has more experience with AI that he would then want more information."*. This emphasis on interactivity reflects the doctors' desire for clear and concise overviews coupled with the option to delve deeper for a comprehensive understanding of AI-driven diagnoses. Furthermore, the capacity to engage the AI system with follow-up questions and receive meaningful responses can significantly impact trust and cooperation between doctors and the AI tool. The doctors pointed out that this interactive feature could lead to improved patient care, as it allows them to gain a better understanding of the predictions and make necessary corrections. One doctor [P1] elaborated, *"And if it would improve care, that you could see that this prediction does not work, and that your cooperation with the AI would provide better care?"*.

Collaborating to circumvent tendency of not listening to advice Nevertheless, the interviews also uncovered a human tendency among doctors to be opinionated and occasionally resist the guidance offered by AI or their colleagues. Some doctors admitted that they might disregard explanations if their personal opinions override the insights presented by the AI system. As one doctor [P2] candidly acknowledged, *"We are terribly opinionated, of course... So we are stubborn after all, and I wonder if... there is a discussion between colleague A and that B, C, D are on 1 line, and that colleague A hears that and that he still does what he himself wants and not what B, C, and D advise."*.

Learning AI's internal model to collaborate The doctors also mentioned the value of having additional background information to check if the AI system's factors align with the patient's condition. One doctor [P7] explained that if the rules from the AI system displayed by the explainability did not match their own mental model they would doubt its outcomes *"Yes then I would definitely doubt it."*.

2.2.5 The Need for Explanations within Healthcare

Explanations in the context of AI adoption are paramount to establishing trust and fostering collaboration between medical professionals and AI systems. In this section, we delve into the multifaceted nature of the need for explanations and its implications in clinical practice.

Providing explanations for improving trust into the system Trust in AI is intricately linked to the presence of explanations for its outcomes. As one doctor [P6] stated, "if the information of how it got to his decision is not there then I would trust it less," highlighting the importance of transparency and understanding. One key aspect that emerged from the interviews is the doctors' expectation of explanations in the clinical setting, with one doctor [P5] being confused about the question as they already assumed explanations being build in AI systems by default "Yes, that's just in there anyway. This I assumed at all, that you don't just get A or B but also this". This is also corroborated by another doctor [P7] who mentioned the value of having additional background information to check if the AI system's factors align with the patient's condition *"To check if the tool is correct, per prediction I would take the additional background information of how the prediction was arrived at to check if the basic factors considered match the patient I am seeing."*.

Knowing what you use for responsible patient care The doctors also focus on their duty of knowing the characteristics of what they are using for the care of patients, with one doctor [P1] stating *"I think we kind of do owe it. Especially if we want to do more with it, that we have to see what are the ins-and-outs of such a system."*.

Using explainability to enable collaboration The interviews also shed light on the collaborative aspect of Human-AI interaction in clinical settings. Doctors mentioned using AI predictions as data points for discussion during Multidisciplinary Team Meetings. They emphasised the importance of being able to explain AI predictions, particularly when discussing cases with other doctors or explaining decisions to patients. One doctor [P5] elaborated, *"If you are going to explain it to your patients then you can say, for example, within a year you are going to die. Of course, you're not going to say it like that. But if you were to say to a patient or to predict something early you have to be able to say why that is."*.

Theme 3- Aligning with doctor needs and values

Collaboration between doctor and AI emerges as a crucial enabler for advancing the integration of AI in healthcare. Explanations serve as a powerful vehicle for integrating AI into clinical workflows, bridging the gap between technical AI outputs and medical decision-making. Participants stressed the importance of explanations in identifying and understanding patterns of AI behaviour, utilising concepts such as the uncertainty reduction theory. Furthermore, collaboration extends beyond one-on-one interactions, promoting collaborative efforts within medical teams. It strengthens patient-centric care by empowering healthcare professionals to make informed decisions tailored to each patient's unique needs. Moreover, it reinforces Trust and accountability, essential attributes of Responsible AI in healthcare.

But to actually get to a point where these positive effects can be seen, the AI first has to align with the doctors their needs, values and trust. This theme explores the insights doctors gave about the alignment between doctor and AI that are relevant to explanations such that adoption of AI can be improved.

3.1 Features of Patient-centric Care Important for Aligning with XAI

Amidst the high-pressure environment of healthcare, Patient-centric Care plays a pivotal role. Uncertainty remains an inevitable aspect of medical practice, and the cluster further delves into the use of Tools for managing uncertainty and fostering better patient communication and understanding. The integration of AI in the clinical workflow holds promise for enhancing patient-centric care. However, doctors should ensure that advancements in technology do not disrupt the core principles of providing the best care possible to patients. Continual adaptation and personalised decision-making will remain essential elements of patient-centric care, even in an AI-driven healthcare landscape.

3.1.1 The Problem of Uncertainty Within Medical Practise

The unpredictable nature of certain diseases adds complexity to patient care. Lung diseases, for example, can progress differently in each patient, making it difficult to predict outcomes accurately. Doctors need to be flexible and continually reevaluate treatment approaches to meet the unique needs of each patient.

Contradicting tests making diagnosis difficult Uncertainty often arises when interpreting test results or making diagnoses. Conflicting test outcomes and low-quality test results can make it difficult for doctors to determine the most likely diagnosis. As Doctor [P3] states, *"Sometimes the results of one test contradict the other, or it contradicts the story of the patient. So then you have to think, what is the most likely diagnosis."*.

Patient heterogeneity causing the need for personalised treatment Moreover, patient variability plays a crucial role in healthcare, with responses to treatments varying widely among individuals. The doctors noted that patients can experience different outcomes, with some showing adverse effects from therapies and others not responding to standard treatments. To effectively integrate AI into clinical practice, it must be capable of understanding and accounting for this variability, providing personalised treatment recommendations based on individual patient characteristics.

Needing to communicate to reduce the uncertainty When faced with uncertainty, doctors emphasised the importance of seeking input from colleagues and patients before making critical decisions. Another doctor [P7] mentioned, *"In the clinical standard is that the opinion of patients is very important because it's a bit of a grey area, so a patient who is a little bit more concerned is different than one who prefers not to be in the hospital as often. So with a patient who is a little bit more concerned, more scans are probably done. But other things are very much on protocol."*. The interviews also touched upon the importance of effective communication and maintaining the patient-clinician relationship in the face of uncertainty. Communicating uncertainty to patients can be challenging, and AI's integration should not interfere with these clinician-patient interactions. As Doctor [P1] highlighted, *"the challenge is to tell the patient in a nice way that we think it's something else."*.

Healthcare is a dynamic field, and medical knowledge and practices constantly evolve. As Doctors emphasised, AI systems need to continually learn from new data and adapt to changes in medical practices to effectively address evolving uncertainties in patient care.

3.1.2 Dealing with the High Pressure Environment of Healthcare

The high-pressure environment also affects the responsiveness of doctors, often due to staffing and capacity issues. Delays in obtaining test results can impact the efficiency of patient care. To address such challenges, AI can be instrumental in automating routine tasks and streamlining workflows, enabling clinicians to focus more on patient interaction and critical decision-making.

Importance of dealing with emotional pressure Emotional resilience is crucial for doctors working in patient-centric care, given the nature of the diseases they handle. Doctors acknowledged that discussing end-of-life matters with patients can be emotionally challenging. However, having empathy and building emotional fortitude is essential for providing compassionate care. As doctor [P5] emphasised, *”You shouldn’t try to bring that up, you should just feel that, otherwise you shouldn’t be a doctor.”*. The emotional toll on healthcare providers, as described by doctor [P5], highlights the importance of AI in reducing the burden and stress on doctors. AI can assist in handling administrative tasks, such as scheduling appointments and managing patient records, allowing clinicians to allocate more time to patient care and empathetic communication.

3.1.3 Tools of Doctors for Dealing with the Uncertainty and High Pressure

To deal with the complexities in the high pressure environment the medical staff makes use of a few different options they have to reduce that complexity. These include using personal experience, asking colleagues, holding MDO’s and using the literature.

- **Experience** Doctor [P11] notes, *”It completely depends on how someone presents, what kind of information I’m already told, the most important thing is to gather all the information you have at that point, evaluate that to come to a diagnosis.”* The combination of experience and training plays a vital role in this process, as another doctor [P11] mentions, *”When I started, I had to think about it for a long time, but now you know when a patient comes in of this is what I have to do.”*.

- **Literature review** To cope with the complexity of patient care, doctors often engage in extensive preparation and literature review. AI can greatly aid in this aspect by providing automated analysis and knowledge sharing. As Doctor [P11] suggests, *”That’s where I think AI can really save time, having that done for you.”*.

- **Colleagues and MDO’s** Doctor [P3] explains, *”If I’m unsure, then I contact the pulmonologist... because they are the most experienced.”*. The need for further tests may arise to ensure a proper diagnosis, as Doctor [P3] points out, ”Very often we do an extra test to make sure that we rule the pulmonary embolism out.”. Simple cases may be addressed individually, but more complex ones often require the input of a Multi-Disciplinary Meeting for collaborative decision-making [P6] *”...that MDO gives an advice, and then as the lead clinician you make a decision.”*.

3.1.4 Communicating and Understanding Patients

Effective communication and understanding the unique needs of patients form the cornerstone of patient-centric care. In the pursuit of providing tailored treatment plans and good patient care, doctors recognise the crucial role played by good communication between them and the patient.

Importance of the first interview In patient-centric care, effective communication and understanding patients’ individual needs are crucial for providing personalised treatment plans. Doctors emphasise the significance of the initial interview, during which they gather essential information from patients. As Doctor [P6] explains, *”I think you mainly find that out at the initial interview at the outpatient clinic... it’s from the treating physician,* *through consultation of the patient, that should be determined which route to take."*. The patient journey is complex and not always linear, and doctors need to adapt their approach accordingly.

Communicating bad news Doctor [P11] highlights the difficulties in communicating bad news to patients, stating, *"Good news always goes well, bad news you always need a lot of time. Cancer is back, you can't do that in 10 minutes."*. Addressing patients' needs requires empathy and the ability to sense what they require, as Doctor [P11] points out, *"There is a certain amount of empathy in that you have to sense, of what does or doesn't someone need, have a need for, what we think they have a need for."*.

Dealing with heterogeneous patients Dealing with different patients' personalities and coping strategies is a significant challenge, as patients may not always follow treatment plans as prescribed. Doctor [P2] mentions, *"The biggest problem we have is people who behave difficult on a personal level, so personality issues, or coping strategy... it's very difficult to demonstrate there that someone there is not doing what you say."*.

Understanding patients through experience To ensure consistency in patient care and optimise resource utilisation, doctors emphasise the importance of tracking past decisions and patient outcomes. By maintaining a comprehensive database of patient experiences, doctors can make informed decisions in similar situations. As Doctor [P2] suggests, *"To see, what did we do, what were the steps and what were the numbers and what was the outcome of that [...] then you can say, we are in this situation again, what is the possible chance of rejection, or should we now think of another problem."*.

3.2 Factors Influencing the Adoption of AI in healthcare

While some doctors are hopeful about the potential of AI to enhance patient care, others express reservations due to the lack of critical evaluation and concerns about the impact on their autonomy and job roles. Doctors acknowledge that barriers to AI adoption exist, including legal and regulatory issues, and the conflicting goals between healthcare providers and profit-driven companies, which may influence trust in AI solutions. Over-reliance on AI and potential confirmation bias are also concerns, prompting doctors to emphasise the importance of maintaining a critical mindset.

As AI has the potential to transform medical roles and alleviate personnel shortages, doctors grapple with finding the right balance between AI's assistance and preserving their expertise. Explainability and adaptability are also crucial factors, as doctors seek to adjust treatments based on new information while integrating AI suggestions.

In this section, the challenges and perspectives surrounding the adoption of AI in healthcare mentioned during the interviews by the medical staff are shown. We will explore how doctors can be convinced of its benefits, the friction faced when integrating AI with traditional practices, the importance of regulations and validation, and the considerations regarding responsibility and accountability in utilising AI technology in clinical settings.

3.2.1 Barriers for Adoption in Healthcare

The adoption of Artificial Intelligence has sparked a mix of optimism and concern among healthcare professionals. While some doctors express hope for the promising potential of AI in improving patient care, others voice reservations about the lack of critical evaluation regarding its actual impact and are scared of losing their autonomy or even their job. In one interview, a doctor [P7] remarks, *"Promising, but little critical evaluation yet of what it can really bring to patients."*. The enthusiasm for AI's capabilities is tempered by the realisation that significant barriers exist in the healthcare landscape. These barriers include not only legal and regulatory issues but also the contrasting goals of healthcare providers and profit-driven companies. One doctor [P7] notes, *"A lot of obvious ones, like laws and regulations, some less amenable things like what doctors think and think about it."*.

Dangers of overreliance of predictions without explanations Furthermore, AI is seen as a useful suggestion rather than a replacement for doctors' expertise. There is a growing awareness of confirmation bias in relying too heavily on AI suggestions. As one doctor [P4] explains, *"The unexperienced radiologist trust those, If they AI then gives a suggestion that gives the correct output most of the times then it gives a correct answer, but if the AI suggests the wrong prediction the unexperienced ones give far more often the wrong answer while the experienced did this a lot less."*. While some doctors acknowledge its potential benefits, others remain sceptical and cautious. Trust in AI systems is tempered by concerns about losing critical thinking and over-relying on technology. As one doctor [P4] admits, *"I think the danger is also the other way around where people trust the computer too much."*. And despite the promising aspects, doubts persist regarding the ability to fully comprehend AI's inner workings. One doctor [P4] comments, *"Also, with AI being a black box, when you only put data in that algorithm, and it gives only the outcome. I think even the people who make these systems don't know what is happening inside."*.

Possibility of losing their autonomy, final say and job The implementation of AI in healthcare also raises questions about its impact on job roles. While doctors ponder the possibility of AI potentially replacing some jobs, there is also hope that it could alleviate personnel shortages in certain departments. As one doctor [P1] explains, *"We will soon have a looming shortage of healthcare personnel, and that's going to get worse and worse."*. In terms of AI system characteristics, discussions revolve around empowering patients to use AI tools to verify their doctors' diagnoses. However, it is essential for doctors to remain mindful of biases in the data that could influence the machine's decision-making process. As one doctor [P2] emphasises, *”You have to keep that critical attitude.”*.

Losing the ability to explain the treatment decisions to patients Doctors also express their concerns about explainability and adaptability. They emphasise that treatments should be adjusted based on new information, combining AI suggestions with their own knowledge for better patient outcomes. The ability to integrate new information into the AI system is crucial to maintaining effective and up-to-date medical practices. Nonetheless, the overall attitude towards AI in healthcare remains positive, particularly with regard to its potential for aiding in diagnosis and optimising patient care. As one doctor [P1] states, *”I have a patient with this and that, what could it be, diagnosis. Can also be in terms of I have a CT scan here with a weird image and can’t figure it out, what does the AI think about it.”*.

3.2.2 Convincing Doctors of the Benefits and Validity

Initially, many doctors might approach AI with scepticism, hesitant to fully integrate it into their clinical workflow without established evidence of its effectiveness. To encourage adoption, physicians stress the significance of providing practical examples that demonstrate how AI can enhance medical practice and improve patient outcomes.

Importance of overcoming initial scepticism for adoption For the successful adoption op AI in the clinical workflow it is important to convince the doctors of the benefits. One doctor states [P11], *”Until it has an established place in the order, I will still look at it fairly sceptically.”* showing the hesitance to accept it immediately. To increase adoption, the doctors emphasised the importance of providing practical examples that demonstrate how AI can benefit medical practice. As one doctor [P7] stated, *”The idea is that it adds something in practice, and so even if there are pitfalls, you still have to try to convince them with the results.”*.

Importance of accuracy before willingness to adopt The doctors stressed the critical role of accuracy in gaining trust and adoption of AI systems. One doctor [P5] expressed the desire for a very high accuracy threshold, *”So then you shouldn’t use the system because then it’s not good enough. It has to have a really low threshold that it’s wrong. It has to be really really good enough. Otherwise, it’s of no use to you.”*. Other doctors [P9] have a more relaxed threshold comparable to the success of existing diagnostic tests *”I think it could be very good for the patient, but you have to like other diagnostic tests that are not AI related, there has to be a certain success rate.”*. This underscores the importance of ensuring that AI systems demonstrate a high level of accuracy and reliability to win the confidence of a wider range of physicians and increase their willingness to embrace AI in their clinical workflow.

3.2.3 Friction With the Outside World

The adoption of AI in healthcare can be hindered by friction between the healthcare and other domains. These challenges include doctors’ resistance to altering judgement based on technology, the importance of multidisciplinary collaboration with external experts, ensuring AI’s compatibility with healthcare processes, and aligning values with external AI developers.

Resistance to change judgement based on technology Doctors highlighted the challenge of changing their judgement based on technologies, particularly when it comes to the use of AI in medical decision-making. They mentioned that physicians often rely on visual observations and clinical progression to make decisions regarding patient care, which might hinder immediate acceptance of AI-based recommendations. Convincing doctors to adopt AI involves overcoming their ingrained practices and traditional approaches, as one doctor [P10] stated, *”It is sometimes very interesting and difficult to change the opinion of doctors because they do things according to how they did it all the time and how they were taught.”*. Addressing this friction requires showcasing the benefits and reliability of AI in medical settings.

Multidisciplinary collaboration with experts outside the medical world Secondly, multidisciplinary collaboration emerged as a key factor in promoting AI adoption among physicians. One doctor [P8] stated for example *”First in the technical validation phase, more physicians should be involved, and then in the clinical validation phase more technicians.”*. Involving physicians in the technical and clinical validation phases of AI development allows them to see the potential opportunities and contribute to AI solutions that align better with their needs. Collaborating with technical experts can lead to improvements in AI systems, which may further facilitate physicians’ acceptance and trust in AI technology.

AI not harming other healthcare processes Thirdly, maintaining the consistent quality of healthcare delivery remains paramount, regardless of technological advancements like AI. Doctors emphasised the importance of ensuring that AI’s role in healthcare does not compromise the overall quality of patient care. The focus should be on providing the best care and communicating this commitment to patients. This patient-centric approach aligns with Responsible AI principles, ensuring that AI technology serves as a valuable tool in delivering high-quality healthcare.

Aligning values with external AI developing companies Lastly, doctors emphasised the significance of maintaining a balanced relationship with companies developing AI tools. While collaboration with companies can drive innovation, doctors highlighted the need to ensure that AI tools are trustworthy, validated, and reliable. It is stated that their priorities are different [P7] *”They have a very different goal, making a profit and touting their products as nicely as possible.”*. Overhyping AI products with unvalidated claims could lead to scepticism among physicians, hindering the adoption of AI in healthcare.

3.2.4 The Importance of Regulations and Committees

Adoption of AI systems hinges on comprehensive regulations, extensive clinical validation, and governing bodies that can ensure the reliability and safety of AI applications. The doctors express concerns about the development phase of AI systems, highlighting the need for rules and checks.

Moving the burden of adoption from individual doctors to expert committees Medical staff acknowledges the importance of certifying the AI like medical devices, subjecting AI systems to extensive testing and approvals by organisations before specific uses. Asked *”Are there other factors influencing whether you trust it?”* [P1] responded *”Yes so if such a committee has looked at it and approved it.”*. Therefore not putting the responsibility of use on individual doctors but on knowledgeable committees as stated by [P5] *”That would not be something that as an individual doctor should look at, that should be higher up.”*. They view the involvement of committees as helpful in critically evaluating AI applications from different perspectives, ensuring that they meet high standards.

3.2.5 The Importance of Clear Responsibilities and Accountability

As AI technology continues to make its way into clinical practice, the question of responsibility and accountability becomes paramount for medical professionals. While doctors acknowledge the potential benefits of AI in providing valuable references and support when combined with clear guidelines and transparency, they assert the ultimate decision-making authority rests with them.

Doctors keeping ultimate decision-making authority The doctors emphasise the importance of their ultimate decision-making authority, even if the AI provides interesting references. They express their willingness to consult colleagues when unsure, suggesting that the AI’s role is supportive rather than directive as noted by [P11] *”Ultimate decision I make myself anyway, so then I’m not afraid of that. Otherwise I would rather consult my own colleagues.”*. They express their preference for being able to adjust AI advice themselves based on their expertise. One doctor mentions the challenge of trusting AI predictions, particularly in the case of medical images like lung nodules, where interpretations may vary among radiologists and AI systems. Overall, the doctors advocate for a cautious and well-informed approach to AI integration in clinical practice, prioritising transparency, accountability, and the ultimate responsibility of medical professionals as one doctor [P11] said, *"Ultimate decision I make myself anyway, so then I'm not afraid of that."*.

Having clear guidelines and transparency of the AI implementation The issue of responsibility and accountability for AI decisions is a major concern for the doctors. They raise questions about who should be held responsible if the AI makes a wrong decision and what factors the AI bases its decisions on such as [P6] states *"What I would like to know is if it goes wrong who is responsible, what does he base his decision on?"*. They stress the need for transparency in AI algorithms and decision-making processes to better understand and interpret the results. The doctors believe that having clear guidelines and control over AI implementation is necessary to ensure its appropriate use and to avoid potential risks.

3.2.6 Medical and Technical Validation of the System

The insights gathered from the interviews highlight the necessity for both clinical validation and expected behaviour in AI systems. By involving a multidisciplinary team and ensuring robust validation, doctors can confidently embrace AI technologies that enhance patient care and clinical outcomes.

Challenges With Validation in Healthcare

The validation of AI models poses several challenges that impact their acceptance and adoption in clinical practice. Doctors express concerns about the constant adjustments made by AI models and the lack of transparency in understanding these changes. Additionally, passing clinical validation alone is deemed inadequate, as AI systems must demonstrate accurate behaviour without biases or shortcuts in their predictions. The level of trial and validation significantly influences doctors' willingness to embrace AI technology, as they seek assurances regarding patient safety and system reliability. Ultimately, doctors believe that the level of trial and validation plays a significant role in their willingness to adopt AI systems. They suggest that as validation and security progress, their usage and trust in the systems naturally increase. *"But that in the steps that you've taken before that you have to have all the validation and the securing of the patients already taken care of,"* concludes a doctor [P9]. However, some challenges exist when validating these systems. Despite the risk of reproducing past errors, doctors show interest in AI's potential benefits. They are open to AI systems displaying slightly better performance than existing methods in a controlled test environment. *"Yes, you could, because you always have with new test you always recreate what was before, and if that's a little bit better then you know I'm on the right track,"* another doctor [P7] points out.

Difficulties with validation of evolving models They express concern about some AI models constantly adjusting themselves and the lack of transparency in understanding what factors are being adjusted. One doctor [P8] remarks, *"We haven't talked about deep learning constantly adjusting itself, which to me is really kind of a dark black box. That there has to be transparency, what factors are being adjusted."*. They emphasise that this constant adjustment could lead to the need for re-validation, as another doctor [P8] adds, *"Suppose you do an adjustment once a year then yes, but you shouldn't have to want to adjust it every day."*.

Only clinical validation is not enough Another significant concern raised by doctors is the distinction between clinical validation and expected behaviour. They emphasise that passing clinical validation alone is not enough. The system must demonstrate accurate behaviour and avoid shortcuts or biases in predictions. *"Then it would be a bad model, that would not be a model that could then be validated. Then you're going to have a problem of bias where 99 out of 100 people have a syndrome, and he says 100 out of 100 times it's that syndrome then that system just doesn't work,"* a doctor [P5] explains.

How To Validate Healthcare AI

The validation of AI systems in healthcare is essential for ensuring their reliability and safe integration into clinical practice. Doctors emphasise the importance of a multidisciplinary team of experts behind AI algorithms, with thorough testing and validation involving bioinformaticians and domain experts. Explanations play a vital role in the validation process, enabling doctors to trust and act upon AI predictions. Extensive clinical testing with diverse patient populations and both technical and clinical validation is crucial for the reliability and clinical relevance of AI technology.

Importance of multidisciplinary experts When it comes to the validation process, doctors stress the importance of a solid team of experts behind the AI algorithms. They believe thorough testing and validation, involving bioinformaticians and experts, should be undertaken before deploying the system with patients. *"I think if going into AI then, there are all bioinformaticians behind people that really understand those algorithms, that write them, and I think you have to make sure that they really have a solid team that knows what they're doing. Then if you're really going to apply a system like that on people to people then it has to have been checked three or four times before you really apply it,"* says a doctor [P9].

Using explanations for continuous validation They state that part of the validation comes from the explanations and draw a parallel with lab values, where doctors need to interpret the results and make decisions accordingly. As one doctor [P9] states, *"Always in medicine you have to be able to make a good case for why you make certain decisions. If the AI says this factor I take negative and therefore I decide this, we already do that with lab values."*. Doctors stress that AI systems should provide explanations for their predictions to facilitate better decision-making, enabling doctors to trust and act upon the results.

Solid clinical testing with diverse patient populations They stress the importance of extensive testing involving diverse patient populations. One doctor [P8] states, *"It has to be extensively tested of course. Many patients, healthcare providers, multiple medical centers. In the clinical and in the computer science way."*. They also highlight the need for both technical and clinical validation as a doctor [P8] remarks, *"If it's not technically validated then it's of no use to you clinically."*.

Extensive documentation and transparency in the implementation The lack of documentation and transparency in some AI tools raises concerns for doctors. In the absence of clear information, they resort to historical data and prospective testing to assess the tool's reliability and performance. Doctors stress the need for sharing underlying algorithms for proper validation, even if it means overcoming competitive sensitivity. As one doctor [P7] mentions, *"So now we are going to validate their tools ourselves in a study. Actually now we have the tool, we're going to use it in the prospective sense and we're going to pretend we had it 6 years ago and unleash it on all CT scans to see what the difference is."*. Validating AI tools within a clinical setting is essential to ensure their suitability and effectiveness in patient care.

3.3 Collaboration and Trust with XAI

Collaboration and trust are essential components in the integration of AI within the clinical workflow. Doctors emphasise the role of AI as a valuable collaborative tool, and state that trust develops with repeated interactions. Explanations promote collaboration between healthcare professionals and AI systems, which enabling the combination of human expertise with AI-driven insights. The doctors emphasise that AI should not replace human expertise but rather collaborate with it which can not be done effectively with only predictions. Trust and reliance on AI are crucial for successful implementation and adoption in clinical practice, ultimately leading to improved patient care and outcomes. As most doctors [P1] agree *"doctors with AI are going to replace doctors without AI."*.

3.3.1 Collaboration with Automated Systems

The collaboration between doctors and AI is seen as a complementary relationship, where XAI serves as a valuable colleague providing evidence-based insights and risk assessments. Doctors appreciate XAI as an extra tool, offering a different perspective and acting as a backup system for complex diagnoses. The success of this collaboration relies on AI's ability to align with the doctor's mental model and expertise, with explanations that reinforce trust.

Complimentary tool that helps like a colleague Collaboration between doctors and XAI is viewed as complementary, where AI excels in risk assessment and provides less subjective, evidence-based insights. Doctors value XAI as an extra colleague that acts as a backup system and helps them navigate complex diagnoses, as another doctor [P7] notes, *"an extra colleague is awesome, that saves another colleague."*. The doctors appreciate XAI as an additional source of input, especially when the human and AI have different outcomes. It helps them identify potential issues they might have overlooked, as one doctor [P6] explains, *"picking out what AI has seen but what I haven't seen."*. While XAI is seen as a valuable companion that provides additional input, the doctors emphasise the importance of human expertise and empathy in patient care. They believe AI should be used supportively, as one doctor [P11] puts it, *"an assistant in your diagnostics."* AI is particularly useful for surfacing additional information, especially when doctors may be prone to stubbornness or overlook certain aspects of patient care. Doctors view XAI as an additional sounding board that ultimately leads to better patient care, as one doctor [P2] admits, *"we're just stubborn, and yourself you still think, you go to your colleagues because you don't want to miss something."*.

Aligning with their own mental model to improve trust The success of AI-doctor collaboration depends on how well the AI system aligns with the doctor's mental model and expertise. Doctors are more likely to trust XAI if it provides explanations that align with their own knowledge and experience. One doctor [P1] highlights, *"you trust someone's knowledge and ability. If you consult someone who you know is very knowledgeable about something, that is of course more difficult in such a large automated system."*. XAI is considered an evolving tool that constantly improves and becomes more integrated into medical practice. Doctors acknowledge its potential to take on more responsibilities, but they also stress the need for transparency and substantive discussions.

3.3.2 Trusting the XAI over Time Through Interaction

The establishment of trust in AI systems within the medical field is for most doctors a gradual process, heavily influenced by the AI's reliability, explainability, and alignment with doctors' decision-making processes. First-hand validation plays a crucial role in building trust over time, as doctors actively compare AI recommendations with their own clinical judgements. Accepting the limitations and errors of AI is also part of the trust-building process. Just like any tool or human intervention, AI might not work perfectly with all patients, and doctors understand that. It is crucial to have realistic expectations while using AI in clinical practice. Ultimately, the doctors recognise that trust in AI is a gradual process, influenced by the AI system's characteristics, their own experience, and the alignment between AI suggestions and their clinical judgement. To fully integrate AI into the clinical workflow, AI systems need to prove their reliability, explain their decisions, and consistently demonstrate their effectiveness over time.

Actively comparing XAI to their own judgements during repeated interactions First-hand validation plays a significant role in building trust over time. One doctor [P7] explains, *"if it works for the first three patients, then you think, great, we are going to use this tool, if the first three are not correct, then it is suddenly difficult."*. If the AI consistently proves to be correct, people start believing in it. However, it might be challenging to gain immediate trust in AI, as another doctor [P8] points out, *"for now there is too little confidence"* initially. Similar to accepting the effectiveness of medications like Paracetamol, the AI must demonstrate its efficacy over time to gain trust and acceptance. For doctors to trust AI, it needs to demonstrate reliability and align with their own clinical judgements. One doctor [P6] explains, *"Just the reliability... by comparing to what would you have done yourself without that AI tool being there."*.

Explainability and ability to align with doctors' decision-making processes The doctors acknowledge that their experience and clinical judgement heavily influence their reliance on AI. While some physicians might over-rely on AI, others might not use it enough, and there is a range of approaches in between. Trust in AI is also dependent on the AI's characteristics and performance. An AI system's explainability and ability to align with doctors' decision-making processes contribute to building trust. Doctors view AI as an additional tool, akin to a colleague, but not at the same level of trust. The AI's ability to consistently align with the doctor's decisions and suggest reliable treatment plans over time is essential for building reliance. As one doctor [P11] envisions, *"you get a proposal and then you kind of trust it, but that's with everything, it has to prove itself first."*.

3.3.3 XAI as Validation Tool for Doctors

Some doctors want full interactivity on collaborative decision making with AI, however, other medical staff hammer on AI solely as a validation tool. Doctors strike a balance between harnessing AI's strengths and relying on their own knowledge and experience to provide the best possible patient care. Caution against over-reliance on XAI is evident, with doctors treating AI suggestions as part of the puzzle and not the sole determinant of diagnoses, recognising the importance of critical thinking and human expertise in the decision-making process. The collaborative approach between doctors and AI is fundamental to the successful adoption of AI in clinical practice while acknowledging its limitations.

Supportive tool assisting in tasks The doctors make it clear that AI is not meant to replace their expertise. It is considered a supportive tool, assisting in predictive tasks like identifying patient trajectories and detecting potential exacerbations. Nonetheless, they emphasise that AI's role is to complement their judgement, not take over as a leading force. *"No because then why did I study. Some things you just know too right?"* one doctor [P6] explains, reaffirming the continued importance of their own knowledge and experience. AI is welcomed as an additional data point and a supporting tool but is not meant to replace doctors' judgement. Transparency and explanations for AI outcomes are crucial for building trust. By maintaining control over the decision-making process and using AI as a validation tool, doctors can effectively leverage AI's strengths while relying on their expertise to provide the best possible care to patients. This collaborative approach is fundamental to the successful implementation and adoption of AI in clinical practice.

Caution against over-reliance In the collaborative decision-making process, doctors do not rely solely on AI suggestions. They treat AI as a validation tool, confirming their findings and conclusions. They consider it alongside other test results, using it as part of the puzzle, rather than as the sole determinant of diagnoses. As one doctor [P5] puts it, *"it's not like only, not something that single-handedly determines what you have, you* *can't have that and you never want to have that."*. The doctors are cautious about the pitfalls of over-reliance on AI and blindly adopting its recommendations. They believe in keeping an open mind but assert that decisions must be grounded in their own expertise. *"I would always use it as a kind of help and never as a guiding thing. That you then go blindly on that,"* says one doctor [P2], highlighting the importance of maintaining control and critical thinking. Despite recognising the strengths of AI, such as its ability to predict patient progress in the ICU, the doctors acknowledge that AI has limitations and must be used with discernment. For instance, in radiology, where AI may detect numerous nodules, it can be overwhelming to navigate through the results. In such cases, AI is seen as a helpful tool, but doctors again maintain their essential role in the decision-making process.

3.3.4 Collaboration Between AI and Doctors for Mutual Learning and Growth

The collaboration between doctors and AI is characterised by the potential for mutual learning and growth, envisioning an evolving partnership where AI systems learn from doctors' interactions. The role of AI is not fixed but adaptable, varying depending on the specific usage context, serving as an aid in certain scenarios and assuming auxiliary roles in others. Doctors highlight the importance of understanding AI's thought process to enhance their confidence in using it. However, they also express caution about potential negative effects, such as overreliance on AI leading to complacency and reduced critical thinking. Striking a balance between leveraging AI's capabilities and maintaining independent analysis remains a key focus in this dynamic collaboration.

Mutual learning from interactions The collaboration between doctors and AI is viewed as an evolving partnership, with the potential for mutual learning and growth. Doctors expressed interest in AI systems that learn from their interactions and adapt based on the questions they ask. One doctor [P2] envisioned a future where AI continuously learns and gains insights from doctors, remarking, *"As you grow with the system, that does change, that it learns prospectively from the questions you ask it."*.

Varying role of AI depending on the situation The role of AI is not fixed, but rather varies depending on the specific usage context. In outpatient settings, AI may serve as an aid in providing instant summaries, while in prediction models for ICU stays, it assumes more of an auxiliary role. This adaptability of AI's function is acknowledged by the doctors, as one [P5] of them points out, *"You can't really say it takes 1 role since it's different for each thing."*.

Interactivity to learn its thought process They believe that a better understanding of how AI thinks can enhance their confidence in using it. One doctor [P8] emphasised, *"It seems nice also to be able to at least indicate that in the system, what if this assumption is not right and I convert it to the one I think is right, what happens to the result."*.

Potential negative effects of working together However, doctors are also cautious about the potential negative effects of overreliance on AI. They worry that excessive dependence on AI might lead to complacency and reduced efforts to think critically and analyse information independently. *"If you start consulting ChatGTP, you no longer have to figure anything out for yourself,"* expressed a doctor [P2], indicating a potential pitfall.

4.4. Conclusion Research Explainability in Pulmonology Context

The conclusion section comprises of five subsections, the first one highlighting critical findings from the preliminary study. The preliminary study is followed by the three main themes found in the semi-structured interviews: enhancing clinician decision-making, providing personalised explanations in healthcare, and aligning AI technology with doctor needs and values. This is concluded with a section about the general conclusion of this second study performed.

Performing the Preliminary Study

During the co-creation session, it became evident that relying solely on a concrete prototype to assess the needs for explainability fell short of expectations. The specific use case presented in the prototype tended to divert the attention of attending doctors, limiting their ability to comprehensively address the broader issues surrounding AI explainability. This realisation emphasised the need for a more flexible approach that could engage healthcare professionals in a manner that encourages thoughtful reflection on the intricate interplay between explainability and AI. It also highlighted the importance of inclusivity, considering the varying levels of knowledge and familiarity with AI and explainability concepts among doctors. To develop effective strategies for explainability, it became essential to involve doctors from diverse backgrounds and varying levels of expertise, recognising that they will all encounter AI technology in their work.

Performing the Semi-structured Interviews

Following this realisation the semi-structured interviews were hold. The analysis of in-depth interviews with medical staff has illuminated three fundamental themes that underscore the potential of Explainable AI in the healthcare domain. The first theme emphasises the importance of enhancing clinician decision-making throughout the patient journey through the implementation of explainable AI systems. The second theme highlights the significance of personalised explanations in healthcare tasks, transcending user types to cater to the unique contexts of medical practice. Finally, the third theme underscores the necessity of aligning AI technology with doctor needs and values, promoting collaborative efforts for patient-centric care.

These themes collectively underscore the transformative potential of XAI in healthcare decision-making, while also addressing the challenges and limitations of its integration. By providing transparent and interpretable insights, XAI can foster trust among medical professionals and patients alike, ultimately revolutionising healthcare practices for the better. As the field of Explainable AI continues to advance, it is essential for researchers, developers, and healthcare professionals to collaborate in harnessing its power responsibly, ensuring a future where AI technology complements and enhances the expertise of doctors, leading to improved patient outcomes and more effective medical practices. The main points learned for each theme are described in the paragraphs below.

Enhancing Clinician Decision-Making Throughout the Patient Journey

The integration of Explainable Artificial Intelligence into the healthcare sector, particularly in enhancing clinician decision-making throughout the patient journey, presents both significant opportunities and challenges. Doctors perceive AI as a valuable tool that can automate administrative tasks, streamline patient interactions, support treatment planning and monitoring, and even deliver empathy in time-constrained situations. The potential of AI to outperform humans in certain tasks, such as image analysis in radiology and pathology, is also recognised.

However, the successful adoption of AI in healthcare is contingent upon overcoming barriers to communication, fostering an open mindset towards AI, and ensuring effective information sharing. Doctors also express concerns about the dangers of AI, including the potential for over-reliance, opacity of AI algorithms, and the impact on traditional medical roles. To navigate these challenges, doctors emphasise the importance of understanding AI principles and fostering AI literacy among medical professionals. This includes the need for proper education and practical feasibility considerations for AI systems. The inclusion of AI and XAI in the medical curriculum is seen as a crucial step towards ensuring future physicians are equipped with the necessary knowledge to effectively utilise AI in their practice.

Overall, while the journey towards fully integrating AI into healthcare is complex, the potential benefits for patient care and clinician decision-making are substantial. With careful navigation of the associated challenges and a commitment to education and understanding, AI can, with the help of explainability, become a transformative tool in the healthcare landscape.

Personalised Explanations in Explainable AI for Healthcare

But to get the XAI working for the healthcare landscape more information is needed to find out how to integrate it into that complex ecosystem. Our analysis of in-depth interviews with medical staff revealed that the efficacy of XAI hinges on its ability to provide personalised explanations that go beyond generic user types and adapt to the unique contexts of various medical tasks.

One compelling finding is that the personalisation of explanations in healthcare relies on more than just user types. Factors such as experience, specialisation, and individual patient cases emerged as pivotal in tailoring explanations to meet the distinct needs of doctors. Especially in complex patient cases requiring broader group collaboration, the need for context-driven explanations becomes apparent, underscoring the importance of the interplay between Explainable AI, collaboration, and trust in healthcare settings.

Within the realm of AI applications in healthcare, doctors' diverse tasks and responsibilities necessitate a flexible and nuanced approach to information seeking. Understanding the nuances of this interdependence can lead to AI systems that cater to the unique needs of medical professionals, aligning with their decision-making processes and fostering trust in AI-driven diagnoses.

Moreover, the doctors emphasised the significance of interactive explanations that cater to their experience and align with their decision-making processes. Such interactivity provides doctors with clear and concise overviews while allowing them to delve deeper for a comprehensive understanding of AI-driven diagnoses. This interactive feature can lead to improved patient care, as doctors can gain a better understanding of the predictions and make necessary corrections.

The interviews also highlighted the importance of explanations in fostering collaboration between medical professionals and AI systems. Doctors stressed that interactive explanations could empower collaborative efforts among healthcare teams and enable informed decision-making during Multidisciplinary Team Meetings. Having additional background information to check if the AI system's factors align with the patient's condition was valued by the doctors, as it bolstered trust in the system.

Overall, the theme of ”Personalised Explanations in Explainable AI for Healthcare” demonstrates that personalised, context-driven explanations play a critical role in advancing the integration of AI in healthcare settings. By providing transparent and interpretable insights, XAI can bridge the gap between AI outputs and medical decision-making, ultimately leading to improved patient-centric care and enhanced trust and accountability—essential attributes of Responsible AI in healthcare.

Aligning with doctor needs and values

From the interviews with doctors about the use of explainable AI and AI in healthcare highlights the crucial role of collaboration between doctors and AI in advancing healthcare practices. The integration of AI into clinical workflows is seen as a powerful tool for improving patient-centric care by empowering healthcare professionals to make informed decisions tailored to each patient's unique needs. Explanations play a vital role in bridging the gap between technical AI outputs and medical decision-making, helping doctors understand AI behaviour and patterns. Moreover, collaboration extends beyond individual interactions, promoting teamwork and collaborative efforts within medical teams.

However, to achieve the positive effects of AI integration, it is essential that AI systems align with doctors' needs, values, and trust. The interviews shed light on various aspects of this alignment, such as dealing with uncertainty in patient care, managing the high-pressure environment of healthcare, and understanding patients' needs through effective communication. Doctors rely on their experience, consult colleagues, and utilise literature to cope with the complexity and uncertainties in medical practice.

The adoption of AI in healthcare faces certain barriers, with some doctors expressing concerns about the lack of critical evaluation, overreliance on AI, and potential changes in their roles as medical professionals. Convincing doctors to embrace AI technology requires providing practical examples of AI's benefits and ensuring a high level of accuracy and reliability in AI systems. Another one of the critical factors for AI adoption is addressing the friction with the outside world, particularly the resistance to change based on technology. Doctors often rely on their traditional approaches and visual observations for medical decision-making. Convincing them to adopt AI involves showcasing its benefits and reliability in the context of their needs and practices. Multidisciplinary collaboration emerges as essential in promoting AI adoption among physicians. Involving doctors in the technical and clinical validation phases of AI development allows them to contribute to AI solutions that align better with their needs. Moreover, maintaining the consistent quality of healthcare delivery is paramount, ensuring that AI's role in healthcare enhances patient care without compromising overall quality.

The subject of regulations highlights the importance of comprehensive regulations and external evaluation committees to ensure the reliability and safety of AI applications. Doctors emphasise the need for extensive testing and approvals by knowledgeable committees to certify AI as medical devices before specific uses. This approach shifts the responsibility of AI use from individual doctors to expert committees, ensuring higher standards are met. Regarding responsibility and accountability, doctors assert their ultimate decision-making authority, even with the support of AI. They emphasise clear guidelines, transparency, and accountability for AI decisions. They express concerns about the responsibility and factors AI uses for its decisions, which highlights the need for transparency in AI algorithms and decision-making processes. The validation subject emphasises the necessity for both clinical validation and expected behaviour in AI systems. Doctors express concerns about the constant adjustments made by AI models and the importance of distinguishing clinical validation from expected behaviour. Involving a multidisciplinary team and conducting extensive clinical testing are crucial for the reliability and clinical relevance of AI technology. Transparent documentation and sharing of underlying algorithms are essential for proper validation.

Finally, collaboration and Trust underscores the significance of AI as a collaborative tool rather than a replacement for human expertise. Doctors view AI as a complementary colleague, providing evidence-based insights and risk assessments. Trust in AI develops over time through first-hand validation and alignment with doctors' decision-making processes. Doctors stress caution against over-reliance on AI and the need to strike a balance between leveraging AI's capabilities and maintaining independent analysis.

By achieving alignment between doctors and AI is pivotal for the successful integration of AI in healthcare. By understanding and addressing doctors' needs, values, and concerns, AI can be harnessed as a powerful ally in providing patient-centric care, improving medical decision-making, and enhancing overall healthcare outcomes. To fully realise the potential of AI in healthcare, ongoing collaboration and a continuous dialogue between doctors and AI developers are essential to ensure responsible and effective use of AI technology in clinical settings. In conclusion, successful integration and adoption of AI in healthcare depend on aligning with doctors' needs and values, fostering collaboration and trust between doctors and AI systems, adhering to robust regulations and validation processes, and upholding responsibility and accountability in AI use. By embracing these considerations, AI can be harnessed as a valuable tool to enhance patient care and clinical outcomes, while human expertise remains at the forefront of medical decision-making.

General Conclusion

In conclusion, the research on explainability in the context of pulmonology underscores the transformative potential of explainable AI in healthcare decision-making. The study highlights the need for a flexible approach that engages healthcare professionals in the development of effective strategies for explainability. By involving doctors from diverse backgrounds and expertise levels, AI solutions can be tailored to meet their unique needs, ultimately leading to improved patient-centric care.

The interviews with medical staff reveal that personalised explanations and context-driven AI systems play a pivotal role in fostering trust and collaboration between doctors and AI technology. The successful integration of AI in healthcare hinges on addressing friction with traditional approaches, comprehensive regulations, and validation processes that ensure the reliability and safety of AI applications. Moreover, maintaining transparency, accountability, and the inclusion of doctors in decision-making processes are essential to building trust and promoting responsible AI use. By embracing these considerations and responsibilities, AI can become a valuable tool in enhancing medical decision-making, leading to improved patient outcomes and more effective medical practices.

5

Discussion

To harness the full benefits of XAI, it is essential to understand the circumstances under which it can be most helpful for clinicians. In this discussion section, we address three sub-research questions:

1. Where in the patient journey are moments for clinicians where XAI can be helpful?
2. What information do doctors seek in explanations?
3. Under which conditions do doctors engage with explanations?

This will be done by combining the data from both studies that were performed and connecting it with the recent literature on explainability and AI in healthcare. By exploring these questions, we aim to gain insights into the specific scenarios where XAI can be most beneficial and the factors that influence the acceptance and utilisation of AI explanations by medical staff. Concluding the chapter by answering the main research questions how pulmonologists' needs and intents shape the design of XAI solutions.

To begin, we presented a comprehensive analysis of the entire healthcare journey, involving both the perspective of the medical professionals and patients dealing with a specific disease, namely Idiopathic Pulmonary Fibrosis. Through this research, we have gained a holistic understanding of the moments where XAI can prove to be of significant value for clinicians. These instances are influenced by a chain of interrelated factors, including healthcare challenges related to patient-centric care and communication barriers. Moreover, the perceived benefits and challenges of adopting AI in healthcare play a crucial role in shaping the potential applications of XAI.

Next, we explore the type of information doctors seek in explanations when utilising XAI in clinical settings. The specific information required by doctors is influenced by the tasks they are performing, the need for explanations, and the properties of the explanations, such as the type of visualisation, preferred granularity, and collaboration with the AI system. By identifying the specific information needs of doctors, we can design XAI solutions that cater to their preferences and enhance the effectiveness of medical decision-making. Central to successful AI adoption is understanding the conditions under which doctors engage with AI explanations. When doctors actively engage with explanations, the potential advantages of XAI are realised. However, if doctors refrain from engaging with explanations, the benefits of XAI may be forfeited. Therefore, we investigate the factors that influence doctors' acceptance and utilisation of AI explanations, focusing on collaboration dynamics, trust-building, and the system's requirements for effective implementation in the hospital setting.

In the following sections, we will delve deeper into each of these sub-research questions, examining the links within the identified chains of factors. We will explore the characteristics of instances where XAI is helpful, the specific information sought by doctors, and the conditions that drive engagement with explanations. By thoroughly understanding these elements, we can provide valuable insights into shaping the design of XAI solutions that align with the needs and intents of pulmonologists and contribute to improved patient care in the dynamic healthcare landscape.

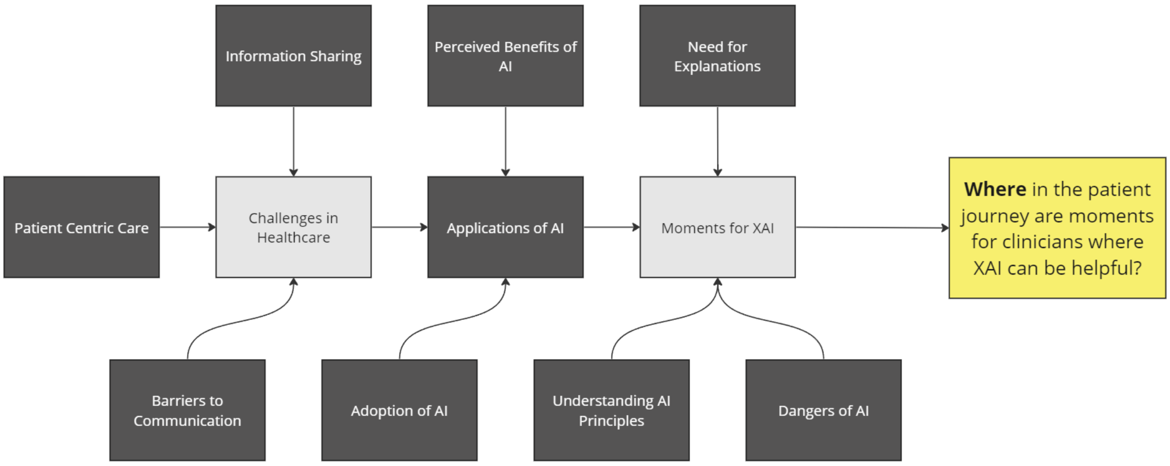


Figure 5.1: Chain of factors identified by the performed research which step-by-step influence where in the patient journey XAI can be helpful.

5.1. Where in the patient journey are moments for clinicians where XAI can be helpful?

Key takeaways:

1. The chain of factors, starting from healthcare challenges such as communication barriers and patient-centric care, influences the potential applications of AI in healthcare.
2. The adoption of AI, its perceived benefits, the knowledge of medical staff, and the dangers associated with AI are the main factors that lead to the practical implementation of XAI.
3. Explanations play a pivotal role in addressing the dangers of AI and fostering trust between medical professionals and AI systems.
4. The need for (X)AI literacy among medical staff is essential for the successful integration of AI and XAI, and the concerns about job roles and over-reliance on (X)AI must be carefully navigated.

By conducting comprehensive research into the entire healthcare journey, involving both the medical staff and patients we acquired a holistic understanding of the type of moments where XAI can prove beneficial for clinicians, for in this case the specific disease IPF.

From the research we concluded that where XAI can be useful is contingent on a chain of interrelated factors, as illustrated in Figure 5.1. This chain begins with various healthcare challenges, including issues related to patient-centric care and communication barriers. These challenges, in turn, influence the potential applications of AI in healthcare the possible applications are strongly influenced by the perceived benefits of using AI in healthcare and the complex topic of the adoption of AI in healthcare. The knowledge of medical staff about AI together with the dangers of the use of AI systems in healthcare turns those AI applications to the practical applications of XAI. Where in conclusion it could offer valuable assistance to clinicians in their interactions with patients.

In the remainder of this section, we will delve deeper into the links within this chain to precisely identify the factors that influence the instances where XAI proves beneficial. By thoroughly examining these connections, we can gain insights into the specific circumstances that call for the application of XAI. Additionally, we will assess the validity of the outcomes and provide a rationale for the decision not to pinpoint exact moments in this particular use case. Instead, we will emphasise the significance of identifying the characteristic features of these specific instances, as it offers greater value.

5.1.1. Finding the Moments From Challenges in Healthcare

From the results of the semi-structured interviews, observation session and the earlier patient research we see that the challenges of barriers to communication, information sharing, and patient-centric care have significant implications and present some of the main challenges in healthcare faced by doctors currently. These challenges highlight the need for innovative solutions, and the integration of AI in healthcare offers promising avenues to address these issues effectively. By looking at these challenges, opportunities can be found where AI and thus also XAI solutions could be implemented leading to the highest benefit for medical staff and patient.

The benefits of using AI for these challenges are widely stated. AI can play a pivotal role in overcoming communication barriers for doctors with patients [136] or by facilitating timely access to referrals and reports. As doctor [P6] highlighted, obtaining referrals from other hospitals can be challenging, leading to delays in patient care. AI-powered systems can streamline this process by expediting the processing and analysis of medical reports, ensuring faster decision-making and more efficient patient management. Collaboration and access to relevant studies and treatment options are essential for complex cases involving rare mutations, as mentioned by doctor [P11]. Here, AI can be instrumental in facilitating seamless information exchange and knowledge sharing among medical institutions, promoting effective collaboration and informed decision-making. Moreover, AI's ability to process vast amounts of medical data efficiently can support doctors in gathering and evaluating comprehensive patient information, improving diagnostic accuracy and treatment outcomes [153]. Patient-centric care is a cornerstone of healthcare, aiming to address individual patient needs and deliver personalised treatment plans. However, the high-pressure environment and uncertainties in medical practice can make it challenging to consistently provide patient-centric care. AI integration offers innovative solutions to manage uncertainty and improve patient-centric care [62]. By analysing large datasets and patterns, AI can assist doctors in navigating the complexities of patient cases with varying responses to treatments (patient heterogeneity). This enables medical staff to make personalised treatment recommendations, enhancing patient care and outcomes.

5.1.2. Applications of AI Tackling Healthcare Challenges

The findings from the performed research sheds light on the main applications of AI in healthcare for medical staff, focusing on the adoption of AI and the perceived benefits of AI systems. The integration of AI in healthcare holds tremendous potential for improving patient care and empowering medical staff to overcome the main challenges they face in delivering high-quality care. However, not all challenges faced within healthcare can be realistically solved within a short time-span. The applications of AI are limited by its current benefits it can provide but the adoption is also dependent on regulations, friction of the outside world, validation needs, accountability and the need to convince the medical staff. In this section these factors will be discussed more in detail.

Perceived Benefits of AI in Healthcare

The benefits of the use of AI in healthcare are known to include its availability, ease of use, and potential to improve efficiency and reduce the cost of health care service delivery [31]. During this research these and other benefits were identified. Doctors perceived several benefits of AI that can revolutionise medical practice. By promoting effective communication and information sharing, AI can facilitate seamless coordination between healthcare facilities and streamline workflows, leading to more efficient patient care, reducing the medical staff work load and the staff shortage [37, 106, 74]. AI's ability to process vast amounts of medical data can aid doctors in gathering comprehensive patient information, enhancing diagnostic accuracy and treatment outcomes. Moreover, AI can play a pivotal role in managing uncertainty by providing data-driven insights. This enables medical staff to provide personalised treatment recommendations, ensuring patient-centric care remains a priority. Overall, the integration of AI in healthcare holds tremendous potential to improve healthcare practices, enhance patient outcomes by combined human-AI decision making [132], and empower medical staff in overcoming the main challenges they face in delivering high-quality care to their patients. However, it is essential that technological advancements do not compromise the fundamental principles of patient-centric care. As healthcare continues to evolve, the continual adaptation of AI systems to changes in medical practices will be crucial in effectively addressing the challenges faced by medical staff.

Limitations in Adoption

While AI's potential benefits in healthcare are evident, doctors express reservations and concerns regarding its adoption which could limit the moments it can be successfully implemented. How to improve the adoption of AI has been only limited studied in the context of healthcare [74, 85]. The barriers for adoption stem from the lack of critical evaluation, uncertainties about AI's impact on their autonomy and job roles, and concerns about over-reliance on AI suggestions and potential confirmation bias. Additionally, the conflicting goals between healthcare providers and profit-driven companies, along with legal and regulatory issues, may influence trust in AI solutions. There are gaps in current national and international regulations regarding who should be held responsible for errors or failures of AI systems, particularly in medical AI [84]. The complexity of roles and responsibilities among various actors involved in the process, such as healthcare professionals and AI developers, makes it challenging to define accountability clearly. This lack of clarity can leave clinicians and other healthcare professionals in a vulnerable position, especially when using non-transparent AI models.

The integration of AI in healthcare requires finding the right balance between AI's assistance and preserving doctors' expertise. Explainability and adaptability are crucial factors, allowing doctors to adjust treatments based on new information while incorporating AI suggestions. Moreover, the adoption of AI raises questions about its impact on job roles, with doctors grappling with the possibility of AI potentially replacing some jobs while alleviating personnel shortages in certain departments [106]. To encourage the adoption of AI in clinical practice, it is essential to convince doctors of its benefits and practical utility. Doctors approach AI with scepticism initially, but practical examples showcasing how AI can enhance medical practice and improve patient outcomes can increase adoption. The level of accuracy achieved by AI systems plays a significant role in winning doctors' trust, and the validation process. The validation of AI systems in healthcare is critical to ensure their reliability and safe integration into clinical practice, however, agreed on reporting standards for AI in healthcare are still lacking [59]. Doctors stress the importance of multidisciplinary teams of experts behind AI algorithms, thorough testing, and validation involving bioinformaticians and domain experts. Explanations play a vital role in the validation process, enabling doctors to trust and act upon AI predictions. Extensive clinical testing with diverse patient populations and both technical and clinical validation are crucial for the reliability and clinical relevance of AI technology.

As AI technology makes its way into clinical practice, responsibility and accountability become paramount for medical professionals. Doctors emphasise that while AI can provide valuable references and support, the ultimate decision-making authority rests with them as otherwise it could undermine their epistemic authority [55]. Explanations enabling transparency in AI algorithms and providing the ability to easier integrate AI advice based on their expertise are essential for doctors to trust and act upon AI predictions and is increasingly being incorporated into binding legal frameworks, for instance, requiring the provision of explanations for automated decision-making processes to patients [85].

5.1.3. Identifying Moments for XAI from AI Applications

The insights into the main applications of AI in healthcare for medical staff from previous section provide a foundation for identifying potential moments for the use of XAI in healthcare. By including the dangers of the use of AI together with the knowledge of AI systems of medical staff and by addressing the challenges and barriers faced by them, such as the need for transparency, explainability, and adaptability of AI systems, XAI can provide interpretable explanations for AI predictions reducing some of these problems.

Need for Explanations

Explanations play a central role in building trust and fostering collaboration between medical professionals and AI systems. Transparency in AI decision-making is vital to gain the confidence of doctors in AI-generated outcomes. Especially as many previous implementations of AI in healthcare performed by large trustworthy companies have failed in the past [50]. The interviews reveal that medical staff expect and value explanations in the clinical setting. For doctors, knowing how AI arrives at its decisions is essential for trusting and accepting its recommendations. Furthermore, explanations are crucial for doctors to comprehend the factors influencing AI predictions fully. This understanding allows medical professionals to make informed decisions and confidently adjust AI-generated advice based on their expertise and knowledge of individual patient cases. Explanations enable doctors to validate the AI system's alignment with patient conditions and verify the relevance and reliability of AI-generated insights. In multidisciplinary team meetings, explainable AI offers valuable insights that facilitate collaborative decision-making. By providing interpretable explanations for AI predictions, doctors can discuss cases with their peers, debate potential treatment plans, and collectively validate the best course of action for improved patient care. These needs match with Adadi et al. [2] findings that XAI can be used to control, justify, improve and to discover AI systems.

Understanding AI Principles

To successfully incorporate XAI into healthcare, actionable strategies should be developed. To navigate the dangers of AI responsibly, medical staff stress the importance of AI literacy among healthcare professionals. Understanding AI principles, capabilities, limitations, and potential biases is crucial for the successful integration of AI in clinical practice. The level of AI knowledge among doctors influences their perception and acceptance of AI systems. AI-literate doctors are more open to technology changes and are willing to explore new AI tools, even tools that contain some errors. On the other hand, less experienced doctors may be more critical of AI predictions with lower accuracy. The need for proper education and practical feasibility considerations for AI systems is emphasised by medical staff. AI literacy is seen as beneficial for future physicians, and the inclusion of AI and also XAI in the medical curriculum is recommended. Even though some literature claims doctors would have too limited time [58] and other subjects would need to make way for it. The inclusion of XAI is not only needed for the showing the benefits but also the drawbacks as the improper application of XAI can be deceiving or have other negative consequences [154]. By familiarising medical staff with AI and XAI concepts, potential pitfalls, and the advantages, the integration process becomes more seamless and efficient and the need and uses of XAI become also apparent for doctors.

Dangers of AI

While optimism exists about AI's potential benefits, the dangers associated with its adoption in healthcare cannot be ignored. Major concerns between healthcare providers and companies developing AI solutions is not only the ethics of patient data sharing [13] but also the conflicting goals such as quick monetization [46]. This incongruity creates challenges in trusting AI solutions, as medical staff may question the motivations and biases behind AI-generated outcomes. Over-reliance on AI, especially among less experienced practitioners or doctors who are more clinical susceptible [49], poses risks such as confirmation bias, where doctors may unquestioningly follow AI recommendations without critically evaluating them. The potential consequences of overreliance on AI systems in healthcare are concerning, as it may compromise patient safety and the quality of care provided [114]. Research has been performed to reduce this overreliance, Buccina et al. [24] reduce overreliance successfully by forcing the participants to use their cognition while looking at explanations instead of forming heuristics when to use the AI's decisions. However, this also leads people to dislike using the explanations more. Vasconcelos et al. [148] show that by using explanations strategically overreliance can also be reduced.

The opacity of AI algorithms presents another significant danger. Doctors express worries about AI becoming a ”black box,” with developers not fully understanding its inner workings. This lack of transparency raises doubts about the reliability of AI-generated outcomes and makes it challenging for medical staff to trust and utilise AI systems effectively. Additionally, doctors contemplate the potential impact of AI on traditional medical roles and express concerns about job losses in certain areas. Additionally, some doctors are afraid of specialist losing skills when they are aided by AI, this is a primary concern under healthcare leaders [123]. Striking a balance between AI's assistance and preserving human expertise remains a critical challenge in the adoption of AI in healthcare.

Doctors' limited understanding hinders the complete identification of all relevant moments

Because most doctors have a limited understanding about the underlying mechanisms and problems of AI systems [158, 1] and lack understanding of what AI implementations could represent [10], it is arguable that the doctors did not have enough understanding to suggest all type of moments where AI could be helpful. Because XAI moments arise from these moments where AI could be helpful combined with various other factors such as knowledge about general challenges in AI, limitations of AI, adoption considerations, and the benefits of the use of AI explainability it could be argued that we did not find all possible moments. However, we argue that because of the inclusion of prompts and working from the expertise of doctors, instead of asking them directly about these moments we got a diverse set characteristics that these moments should adhere to which reached saturation later in the interview sessions.

Doctors' limited understanding of patients' home life hinders the comprehensive identification of all relevant moments.

It is furthermore arguable that medical staff do not have a general view of the home life of patients therefore missing out on possible moments to apply XAI. However, by the inclusion of the journey map in the interviews, generated on thousands of patients their complete experiences dealing with a severe disease, we argue that doctors did have all those moments available to them. Researchers have described the fields and given examples of moments where AI systems could be used Davenport et al. found that diagnosis and treatment, patient engagement and adherence, and administrative applications could be beneficial [37]. Administrative tasks are often quickly identified by doctors as an easy and useful way to implement AI systems in care [68]. Meskó et al. states that AI is most advantageous for tasks characterised by high repetition and that involve the analysis of quantifiable data [105]. They give examples ranging from improving in-person and online consultations till research activities like drug creation.

The challenges faced within healthcare, including communication barriers and patient-centric care, shape the opportunities where AI and finally XAI solutions could be implemented to benefit medical staff and patients. AI can play a pivotal role in overcoming these challenges, promoting effective communication and information sharing, streamlining workflows, and improving diagnostic accuracy. While the benefits of AI in healthcare are evident, adoption is limited by barriers such as a lack of critical evaluation, concerns about its impact on job roles, and over-reliance on AI suggestions. Additionally, conflicting goals between healthcare providers and profit-driven companies, along with legal and regulatory issues, may influence trust in AI solutions. Explanations are crucial for solving these problems building trust, fostering collaboration, and validating AI-generated insights. AI literacy among medical staff is essential to ensure the successful integration of AI and shape those applications into XAI resulting into the integration into clinical practice.

Although the identification of all relevant moments for XAI may be limited by doctors' understanding of (X)AI and the complex nature of patient home life, by delving deeper into the links within the chain of factors, this study has provided valuable insights into specific circumstances that call for the application of XAI in healthcare. Furthermore, the research emphasises the identifying characteristic features of these instances, offering value for researchers wanting to implement XAI in healthcare.

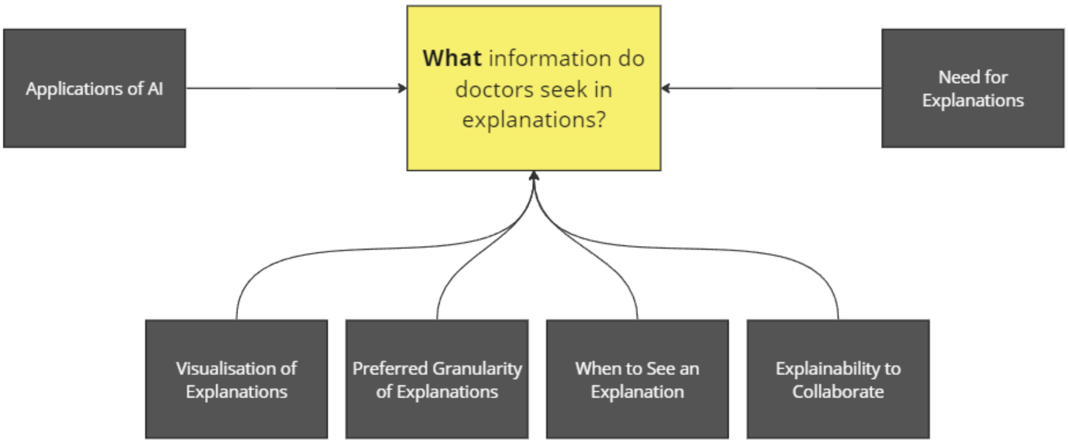

Applications of AI
What information do doctors seek in explanations?
Need for Explanations
Visualisation of Explanations
Preferred Granularity of Explanations
When to See an Explanation
Explainability to Collaborate

Figure 5.2: Factors identified by the performed research which influence what doctors seek in the information provided by explanations.

5.2. What information do doctors seek in explanations?

Key takeaways:

1. Visualisations of explainability should combine patient-specific context, show the AI's thinking process, and avoid overwhelming doctors with extraneous details.
2. Type of information influenced by the context of their tasks and their desire for comprehensive, relevant, and actionable insights
3. Preferred granularity of explanations varies from patient-specific insights for decision-making to a broad understanding of AI functioning for exploration and research.
4. Timing of explanations plays a pivotal role, with doctors preferring access before implementation, during discrepancies, and continuously during use.
5. Effective collaboration with AI systems is fostered through interactive explanations, aligning AI factors with doctors' mental models, and addressing human biases and tendencies to disregard AI advice.

When we have identified where XAI can be helpful it follows that we need to know what the doctors want to see when they are using it. This follows from the applications of AI as it is, according to the doctors, dependent on the task they are performing which type of information they want to see. Together with the need for explanations and the properties of explanations such as the type of visualisation, preferred granularity, when to see it and how to collaborate with the AI we can derive what type of information to show to the doctors at which moment. In Figure 5.2 a systematic overview is given of the factors what factors influence the information doctors want to see from explanations. In the remainder of this section a detailed look will be given into the properties and the external factors which influence the information that needs to be displayed to the doctors towards providing them with the optimal explainability at each moment.

5.2.1. Personalised Explainability Beyond User Types

The successful implementation and widespread adoption of Artificial Intelligence in clinical practice heavily rely on intelligible and informative visualisations of explanations. Tailoring these explanations to meet the specific needs of doctors and providing relevant and comprehensible visualisations can play a crucial role in effectively integrating AI into healthcare workflows. The evaluations conducted by Schoonderwoerd et al. [138] with clinicians revealed their strong demand for explanations of the AI-output, specifically to mitigate false positive diagnoses and minimise false negative where individual differences in the rating of explanations highlighted the potential need for personalisation in the explanation process. This is consistent with our interviews with doctors were it was stated that context-driven explanations that consider the usage context and cater to doctors' preferences can be pivotal in advancing the integration of AI in healthcare.

Even though past research has demonstrated that the type of explanations is dependent on the user [145] often authors of explainability methods flaunt their method as being task agnostic [155] and model agnostic [133] thus being universally applicable to every situation. However, in practise the generated explanations forms are too low-level to be to use to non-experts and even machine learning practitioners with different levels of experience with computer vision can have a hard time envisioning uses of certain explanations [12]. Task-specific visualisations with medical relevant features understandable to medical practitioners emerge as a key characteristic that doctors want to see in explanations of medical AI.

5.2.2. Visualisations of Explainability

The doctors emphasise the importance of creating visualisations tailored to the specific usage context and the particular questions they are trying to address. They point out that different branches of medicine may require different visual modalities. For instance, a diagnosis tool may benefit from a bar plot numeric representation, while radiology might require a mix of images and rules. The preference for a mixed format combining images and rules in radiology allows doctors to gain a deeper understanding of the decision-making process of the AI system, enabling them to see what the AI system is analysing and the conclusions it reaches.

Show the Thinking Processes with Additional Context

Doctors stress the significance of comprehensiveness in AI explanations. Visualisations that not only present the final output but also the entire thinking process of the AI system are highly valued. When explaining the rationale behind a treatment suggestion, doctors believe that the program needs to show its thinking process. In this regard, the inclusion of specific examples and counterexamples from past patient cases is seen as highly beneficial. Providing this additional context and reference points helps build trust and confidence in the AI system, enabling doctors to better understand and validate the AI's recommendations. Branley-Bell found that different visualisations also lead to notable difference in user trust [20]. Which indicates that factors beyond understanding and explainability influence user trust in AI.

The Risk of Information Overload

It is important to consider the potential risk of information overload, as doctors acknowledge that overwhelming visualisations can be too complex for busy healthcare professionals to fully engage with. A balance must be struck between providing comprehensive insights and avoiding excessively complicated visualisations. Keeping the visualisations clear and focused on the relevant information is crucial in ensuring that doctors can effectively interpret and use the AI explanations in their practice. Information overload could be solved by further personalisation [112] or interactive explanations. In addition to these characteristics, our follow-up interviews with doctors from the co-creation session revealed another important aspect: the doctors' belief that they do not always need extra information, especially when they perceive it as illogical or beyond their typical access to information. This finding highlights the importance of providing information that aligns with doctors' existing knowledge and reasoning process. AI explanations should avoid overwhelming doctors with extraneous or irrelevant details that could potentially undermine their trust in the AI system.

5.2.3. Preferred Granularity of Explainability

The findings from both the semi-structured interview and the co-creation session with doctors shed light on the topics that influence the granularity of explanations preferred by doctors in the context of XAI. These topics include the need for context-driven, informative, and clinically relevant explanations that aid decision-making, the importance of connecting explanations to clinical practice and patient-specific scenarios, the adaptive depth of explanation, the desire for abstractive knowledge of AI systems, and the value of global explanations for exploration and future research purposes.

Patient Specific Explanations for Helping in Decisions

Often in the field of XAI the granularity of explanations is described as a dichotomy, either an explanations is global or local in structure and by doing some complex procedures you can often try to switch between those [94]. If we keep in line with this line of reasoning then in the medical case these local explanations are patient specific. The doctors' expressed the desire that for patient specific context-driven explanations which highlights their emphasis on relevance and applicability in their clinical practice. They seek explanations that directly relate to their patients, as it allows them to make informed decisions tailored to the individual cases they deal with. This aligns with the notion that AI explanations should be at least partly patient-specific and grounded in the clinical context to be truly valuable in the healthcare setting.

Adaptive Depth Global Explanations for Comprehension and Exploration

Furthermore, the doctors' interest in the global functioning of AI systems demonstrates their need for a broader understanding beyond individual predictions. They seek insights into the factors that AI models utilise for their predictions, which helps build trust in the system. However, they also recognise the need for adaptive depth of explanation. While comprehensiveness is valued, they acknowledge that overwhelming levels of detail can hinder the effectiveness of explanations. Modulating the level of detail based on the specific situation is essential to strike the right balance between comprehensiveness and usability. The desire for abstractive knowledge reveals the doctors' interest in gaining a high-level understanding of AI systems without delving into intricate mathematical details. They appreciate explanations that provide a global structure and processes of the AI model, allowing them to grasp its functioning without being overwhelmed by complex technicalities. This indicates that the doctors value the broader implications and insights provided by AI systems rather than the intricate workings behind them.

Additionally, as seen in [2] explanations can be used for discovery. The doctors' interest in exploration purposes for future research suggests that global explanations of the AI model play a significant role in their quest for deeper understanding. By gaining insights into the overall behaviour and mechanisms of the AI system, they can base their future research on more comprehensive grounds such as the clinical indicators the AI based its decisions on. This emphasise the importance of providing doctors with explanations that support exploration and enable them to uncover novel insights that can benefit patient care and medical practice.

5.2.4. When to See Explainability

Understanding the optimal timing for accessing explanations is crucial to enhance decision-making and improve patient care. In this section, we delve into the perspectives of the doctors interviewed, exploring the characteristics that describe when they prefer to seek explanations during their medical tasks. Three key contexts emerged: explanations before use, explanations with a disagreement, and explanations during use.

Explanations Before Implementation in the Hospital

The interviewees expressed a strong preference for having explanations readily available before using AI in their medical practice. They emphasised the importance of context-driven explanations that align with their decision-making process. Knowing the limitations of the AI system upfront was considered essential to ensure responsible usage. By being aware of potential shortcomings, such as decreased accuracy in certain patient groups, they can make informed decisions and avoid compromising patient care. The statements from doctors highlights the need for transparency and preemptive awareness of the system's capabilities and limitations. Furthermore, it connects to the research done by McDermid et al. who found that in healthcare global explanations should be employed before deployment to enhance confidence, ensure compliance, and to facilitate continuous development through local explanations post-hoc [102].

Explanations when Conflicting Decisions

Moreover, the doctors stressed the significance of inspecting explanations when there is a disagreement between the AI's prediction and their own clinical judgement. This aspect of XAI is seen as a valuable tool to investigate the reasons behind the discrepancy. By understanding the rationale for the AI's prediction, doctors can validate the accuracy of the model and identify potential areas for improvement. As one doctor pointed out that if it doesn't match then you can look at the explanation as to why that is. This emphasises the value of explanations as a means of enhancing trust and fostering collaboration between human experts and AI systems.

Continuous Evaluation of the System

The doctors also expressed their expectation for explanations to be available during use, regardless of whether the AI's prediction aligns with their own decision. Having access to explanations with every decision allows them to incorporate AI insights as an integral part of their decision-making process. Markus et al. [99] showed that explanations become necessary in situations where the AI system's performance has not been fully established in practice, and there is a need to build user trust, satisfaction, and acceptance. This continuous integration of explanations facilitates a deeper understanding of the AI's functioning and enables doctors to validate its recommendations in real-time. The doctors' desire for ongoing evaluation and testing of AI systems reflects their commitment to refining predictions and ensuring optimal performance in clinical practice. Adopting a trial-like approach initially allows for iterative improvement, ultimately leading to increased trust and confidence in the AI system over time.

5.2.5. Explainability to Collaborate

Collaboration between medical professionals and AI systems holds significant promise for improving patient care and decision-making in the healthcare domain. Especially as past research has shown that the information requirements from explanations closely resemble what clinicians require when engaging with medical peers to deliberate on a patient case [50]. In this section, we explore the insights provided by medical staff on the characteristics of explainability that foster effective collaboration between doctors and AI tools.

Collaboration by Interactive Explanations

One essential characteristic highlighted by the doctors is the need for interactive explanations that cater to the unique experiences of medical professionals. A considerable number of the current state-of-the-art explainability approaches are static, one-off systems that do not consider user input or preferences beyond the initial configuration and parameterisation [141]. The doctors expressed a desire for explanations that provide clear and concise overviews while also offering the option to delve deeper for a comprehensive understanding of AI-driven diagnoses. This interactivity allows doctors to tailor the level of detail to their specific needs and expertise. As one doctor pointed out that its important that everything is visual at a glance, but that if you want more information, you can zoom in for more information. Such interactivity enhances the doctors' understanding of AI predictions and provides them to be dynamic with the time they spent interacting with the explanations. Additionally, the ability to engage the AI system with follow-up questions and receive meaningful responses fosters cooperation between doctors and the AI tool. This dynamic interaction can lead to improved patient care, as doctors can gain better insights into the AI's predictions and make necessary corrections. However, research from Liu et al. has shown the actual improvement in decision making performance might be limited while reinforce human biases [89] .

Matching Mental Models

Moreover, the doctors mentioned the value of having additional background information to check if the AI system's factors align with the patient's condition. A doctor carefully judges a patient's condition based on symptoms and examination before making an explanation. To be accepted in healthcare, AI must try and mimic human judgement and interpretation skills [29]. This desire to understand the internal model of the AI system reflects their intent to validate the AI's predictions against their own mental model. Having access to such information helps doctors gain confidence in the AI's recommendations and ensures that the AI aligns with their medical expertise. This notion however ignores the

Rashomon effect where different prediction models achieve similar performance but construct the prediction relying on different features for their predictions [110]. Anderson et al. [7] showed furthermore that forming accurate mental models of AI systems is cognitively demanding and might not accurately match the underlying model. Druce et al. [39] go a step further and claim that human subjects often form incredibly inaccurate mental models of AI's, and these models can be challenging to dispel. These researchers show that the wish of doctors for gaining an accurate mental model of the AI is very difficult to achieve.

Resistance to Listening to Advice

Despite the valuable insights that could be provided by AI, some doctors admitted that they might disregard them if their personal opinions override the AI's recommendations. It is not only with AI recommendations this happens but similarly happens with their colleagues' advice. This highlights the importance of considering human factors in the collaboration between doctors and AI systems. However, more significant relevance is given to AI advice when explanations are provided [117] then when they are not. Addressing this tendency requires designing AI explanations in a way that ensures that doctors view AI as a supportive tool rather than a replacement for their expertise and that when ignoring the decision they still remain thoughtful of the provided advice in order to build in safety checks to their own treatment plan.

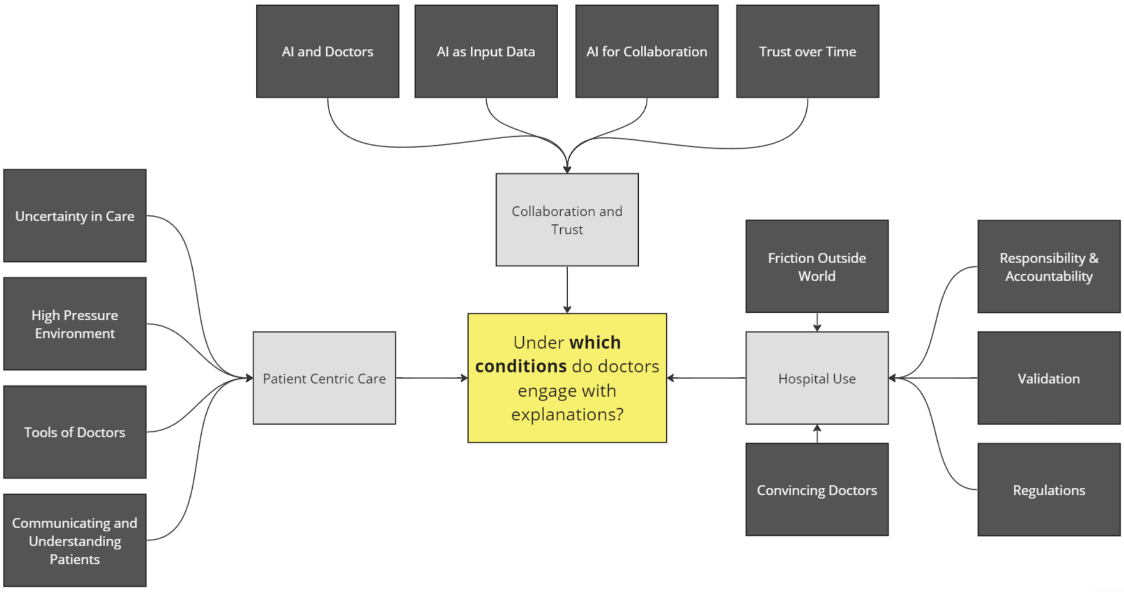


Figure 5.3: Different groups of factors identified by the performed research show under which conditions doctors state they will engage with the explanations.

5.3. Under which conditions do doctors engage with explanations?

Key takeaways:

1. Doctors emphasise the importance of thorough testing, multidisciplinary expertise, and extensive clinical validation to ensure the reliability and safety of AI technology. Pilot testing should run alongside actual care to avoid any negative influence on patient outcomes.
2. Patient-centric care drives doctors' engagement with XAI, particularly in navigating uncertainty and the high-pressure healthcare environment.
3. Collaboration and trust between doctors and AI systems are foundational, where AI is seen as a valuable complement that augments clinical decision-making.
4. Effective integration of XAI systems requires alignment with doctors' tools and decision-making processes, while also ensuring accountability, where the final authority is the doctor.

Central to the successful adoption of AI systems in clinical practice is the understanding of the conditions under which doctors engage with explanations provided by these AI systems. It can be logically deduced that when doctors refrain from engaging with the explanations, the potential advantages of these explanations are forfeited and thus it is important to find out how to engage the doctors with the explanations. This is especially useful as it was seen during the co-creation that a bad implementation of the system can lead to doctors feeling like the added explanations are unnecessary and thus the possible benefits are lost. This section delves into the crucial factors that influence doctors' acceptance and utilisation of AI explanations, shedding light on the dynamics of collaboration, trust, as well as the requirements for the system to be used in the hospital.

5.3.1. Collaboration and Trust

Collaboration and trust play pivotal roles in shaping the conditions under which doctors engage with explanations in healthcare. In this section, we explore how AI and doctors collaborate, build trust over time and understand AI as input data to enhance patient care. Despite the significant research interest in XAI, its application in healthcare for guiding clinical decision-making poses challenges when looking at trust [157].

Collaborate With the System

Doctors view AI as a valuable complementary colleague, providing evidence-based insights and risk assessments. As an extra tool, AI offers a different perspective and acts as a backup system for complex diagnoses. XAI builds trust among clinicians by providing explanations for prediction outcomes, enabling them to understand how to apply predictive modeling in practical situations rather than blindly relying on the predictions [41]. The collaborative relationship between doctors and AI is seen as mutually beneficial, leveraging AI's strengths while retaining the vital role of human expertise and empathy in patient care. Research has shown that collaboration between AI and doctors can lead to better detection of cancer and reduce the workload of the doctors considerably. The implementation of AI-assisted screening led to the detection of 20% more cancers compared to conventional screening methods with a considerably reduced workload [76]. Doctors appreciate AI as an additional sounding board, which could lead to improved decision-making and better patient outcomes [137]. The success of this collaboration lies in AI's ability to align with the doctor's mental model and expertise, reinforcing trust through clear and relevant explanations.

Doctors strike a balance between harnessing AI's strengths and relying on their own knowledge and experience in providing the best patient care. AI is viewed as a supportive tool, assisting in predictive tasks and offering valuable insights. However, doctors treat AI as a validation tool, using it alongside other test results to form a comprehensive patient diagnosis. The collaborative approach between doctors and AI is fundamental to the successful adoption of AI in clinical practice. While AI's role is adaptable and varies depending on the specific usage context, doctors emphasise the importance of maintaining control over the decision-making process. The fear of doctors being replaced is not entirely ungrounded as many AI researchers purely look at creating AI systems that perform a specific task better than the current state-of-the-art, which is often the current doctors, instead of looking at integration with current clinicians to improve their performance. AI systems should however look more into a AI-augmented health system instead of a AI-dominated one [131].

The collaboration between doctors and AI is characterised by mutual learning and growth. Doctors express interest in AI systems that learn from their interactions and adapt based on the questions they ask. The role of AI is not fixed; it evolves to suit different scenarios. Doctors acknowledge the strengths of AI, but they remain cautious about potential pitfalls of over-reliance, preserving their critical thinking and independent analysis. Understanding AI's thought process enhances doctors' confidence in using it effectively. Striking a balance between leveraging AI's capabilities and maintaining independent analysis remains a key focus in this dynamic collaboration. These collaboration effects are rarely studied as post-adoption user perception, experience and collaboration remain understudied [150].

Trusting the System Over Time

Trust in AI is not an instantaneous process; rather, it develops over time through repeated interactions. First-hand validation is vital in building this trust, as doctors actively compare AI recommendations with their own clinical judgements. As AI proves its reliability and consistency, doctors gradually gain confidence in its abilities. The transparency of AI's decision-making and its alignment with doctors' clinical judgement contribute to building trust. However, doctors are cautious about potential negative effects, such as over-reliance on AI leading to complacency and reduced critical thinking. Trust in AI is an ongoing process that depends on its performance and the continued alignment with doctors' decision-making processes. When AI systems fail it will not only be harmful to the adoption of AI in medical care but also to general patient trust in medicine and technologies used within this field [57].

Research from Nourani et al. [113] finds that individuals who noticed the system's strengths at an early stage exhibited a higher susceptibility to automation bias, leading to a significant increase in errors due to their positive initial impression of the system. However, as they developed a more accurate understanding of the system's capabilities, their mental model improved. In contrast, those who encountered weaknesses in the system early on made considerably fewer errors as they tended to rely more on their own judgement. Despite this, they also underestimated the model's competencies due to their negative first impression of the system.

5.3.2. Patient-Centric Care

Patient-centric care is a fundamental aspect of medical practice, and doctors' primary focus is on providing personalised and compassionate treatment to their patients. In the integration of XAI, doctors seek explanations that align with this patient-centric approach. They value AI systems that can offer evidence-based insights and recommendations tailored to individual patient needs. AI's ability to provide transparent and contextually relevant explanations becomes crucial in gaining doctors' trust and fostering collaboration with AI systems

Navigating Uncertainty with AI Explanations

Uncertainty is an inherent part of medical practice [56], and the unpredictable nature of certain diseases poses challenges for doctors. As time progresses and the individual's condition changes, their response to treatment also fluctuates. Drug selection and dosage decisions heavily rely on the medical protocol and the experience of the healthcare provider, where experienced doctors are more at ease with more uncertainty [77], leading to inherently varied and sometimes suboptimal outcomes. When faced with uncertainty, doctors often seek input from colleagues and rely on their experience and critical thinking skills. Here, AI explanations play a pivotal role in augmenting doctors' decision-making processes. To tackle these challenges, AI approaches have emerged as valuable decision-making support tools [19]. Transparent and interpretable explanations from AI systems can help doctors understand the reasoning behind AI-driven predictions, enabling them to make more informed decisions in uncertain situations. By providing insights into the factors influencing AI predictions and potential limitations, explanations can bolster doctors' confidence in incorporating AI recommendations into patient care.

Impact of the High Pressure Environment

The high-pressure environment in healthcare, characterised by but not limited to time constraints and emotional stress [22], shapes doctors' responsiveness to AI explanations. Doctors must juggle multiple tasks, making it essential for AI explanations to be concise, easy to comprehend, and quickly accessible. In the observation sessions the time pressure was already perceived as high by doctors, in other countries this pressure can be perceived even higher as some doctors need to take care of 75 patients in a typical day [150]. AI can aid doctors by automating routine tasks and streamlining workflows, allowing more time for patient interaction and critical decision-making. In this context, AI explanations that can be readily integrated into the clinical workflow become more likely to be engaged with by doctors, facilitating their effective utilisation in high-pressure situations.

5.3.3. Hospital Use Requirements

The successful integration of Artificial Intelligence in healthcare relies heavily on doctors' engagement with AI explanations. Our semi-structured interviews and co-creation session with medical professionals unveiled various conditions and considerations that significantly impact their acceptance and utilisation of AI technology in clinical practice.

Managing Uncertainty and High Pressure Environment with AI Tools

Amidst the high-pressure environment of healthcare, patient-centric care stands as a pivotal factor influencing doctors' engagement with AI explanations. Uncertainty remains an inherent aspect of medical practice, with diseases often progressing differently in each patient. Dealing with uncertainty requires doctors to be flexible, continually reevaluate treatment approaches, and seek input from colleagues and patients before making critical decisions. AI's integration must not disrupt effective communication and maintaining the patient-clinician relationship, as effectively conveying uncertainty to patients is essential. The high-pressure environment in healthcare affects doctors' responsiveness and emotional resilience. AI can play a crucial role in alleviating the burden by automating routine tasks and streamlining workflows, enabling clinicians to focus more on patient interaction and critical decision-making. By reducing administrative tasks, AI helps doctors cope with time pressures and devote more time to compassionate patient care.

The Role of Doctors' Tools in Decision-Making

Doctors employ various tools to cope with complexities in the high-pressure environment. Experience, literature review, and consulting colleagues in multi-disciplinary meetings are common practices. The specific utilisation of these tools relies on various factors among physicians, including their level of expertise, aversion to medical ambiguity, resource constraints, and time limitations [52]. AI can significantly aid doctors in dealing with uncertainty, providing automated analysis, knowledge sharing, and valuable insights for informed decision-making. Doctors expect AI to serve as a supportive tool, providing valuable references and insights, rather than replacing their expertise. As often is stated AI is not meant to replace medical professionals, but the doctors using AI will probably replace those who don't [106]. Effective communication and understanding patients are essential for patient-centric care [52]. Doctors emphasise the significance of the initial patient interview, during which crucial information is gathered. Dealing with heterogeneous patients requires doctors to adapt their approach to meet individual needs. By tracking past decisions and patient outcomes, doctors can utilise a comprehensive database to make informed decisions and provide tailored treatment plans.

Responsibility and how to validate for Trustworthiness

The adoption of AI systems in healthcare is contingent upon extensive testing and approvals, as highlighted by doctors. AI validation serves as evidence that these systems can consistently and effectively deliver value [26]. They emphasise the need for certification by knowledgeable committees to ensure that AI is treated as a medical device, shifting responsibility from individual doctors to expert committees. This approach aims to uphold higher standards and enhance trust in AI technologies. A key aspect of adoption concerns responsibility and accountability in the decision-making process. Despite the support of AI, doctors assert their ultimate authority in making medical decisions. To foster trust in AI systems, doctors emphasise the significance of clear guidelines, transparency, and accountability for the decisions made by AI. This demand for transparency extends to the underlying algorithms and decision-making processes to understand the factors influencing AI decisions thoroughly. Transparency is crucial because even if the results of an AI are trustworthy, several ethical concerns persist with black box algorithms [42]. The validation of AI models poses several challenges that can impact their acceptance and integration into clinical practice. Doctors express concerns about the constant adjustments made by AI models and the lack of transparency in understanding these changes. The lack of documentation and transparency in some AI tools is a concern for doctors, highlighting the need for sharing underlying algorithms for proper validation, even if it involves overcoming competitive sensitivity. Validating AI tools within a clinical setting is essential to ascertain their suitability and effectiveness in patient care.

To successfully validate AI systems for adoption in healthcare, doctors emphasise the importance of a multidisciplinary team of experts behind AI algorithms. Despite vendors making claims about improving efficiency and quality of care, it remains uncertain whether these claims are actually being fulfilled in clinical practice [83]. Thorough testing and validation involving bioinformaticians and domain experts are vital before deploying AI systems with patients. Additionally, explanations play a crucial role in the validation process, enabling doctors to trust and act upon AI predictions. However, the lack of prerequisite skills among medical practitioners for interpreting AI-based systems hinders the availability of expertise to support the validation, iteration, and improvement of AI-based healthcare solutions [119]. Extensive clinical testing involving diverse patient populations [51] with multi-centre data collection [85] is also essential to ensure the reliability and clinical relevance of AI technology. Both technical and clinical validation are imperative, as clinical validation alone is considered insufficient. It is crucial for AI systems to demonstrate accurate behaviour without biases or shortcuts in their predictions, going beyond mere clinical validation. The level of trial and validation significantly influences doctors' willingness to embrace AI technology, as they seek assurances regarding patient safety and system reliability. Conducting small-scale on-site pilot testing serves as an effective method to validate an AI application [30] but doctors mention it should be run besides actual care and not negatively influence it during the pilot testing.

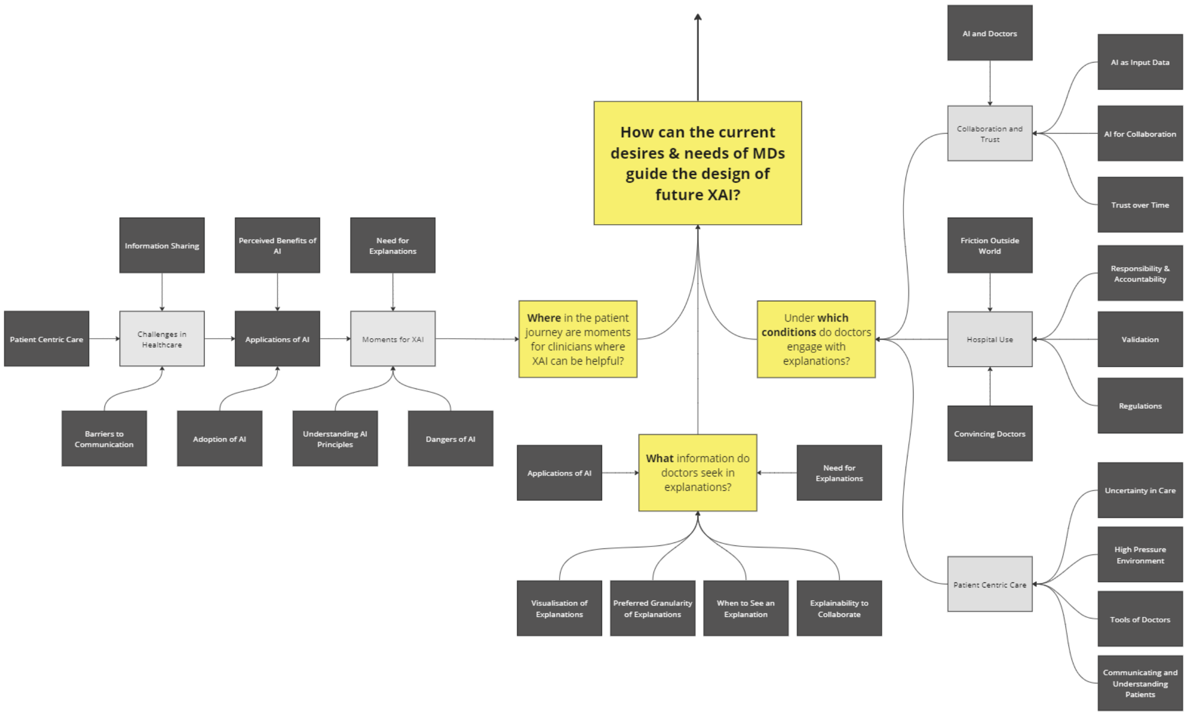

How can the current desires & needs of MDs guide the design of future XAI?
Information Sharing
Perceived Benefits of AI
Need for Explanations
Patient Centric Care
Challenges in Healthcare
Applications of AI
Moments for XAI
Barriers to Communication
Adoption of AI
Understanding AI Principles
Dangers of AI
Where in the patient journey are moments for clinicians where XAI can be helpful?
Under which conditions do doctors engage with explanations?
Applications of AI
What information do doctors seek in explanations?
Need for Explanations
Visualisation of Explanations
Preferred Granularity of Explanations
When to See an Explanation
Explainability to Collaborate
AI and Doctors
Collaboration and Trust
AI as Input Data
AI for Collaboration
Trust over Time
Friction Outside World
Hospital Use
Responsibility & Accountability
Validation
Regulations
Convincing Doctors
Patient Centric Care
Uncertainty in Care
High Pressure Environment
Tools of Doctors
Communicating and Understanding Patients

Figure 5.4: Systematic view of the factors answering the main research question of how the current desires and needs of MDs guide the design of future XAI.

5.4. How do pulmonologists' needs and intents shape the design of XAI solutions?

In the healthcare sphere, integrating explainable artificial intelligence solutions is gaining prominence as a means to elevate medical practice. This section delves into the alignment between the needs of pulmonologists, respiratory health experts, and the design of XAI solutions. The investigation revolves around three key subquestions for the explainability: Moments to Design For, What to Show at These Moments, and How to Make Them Use It. These inquiries serve as the bedrock for comprehending the intricate interplay that shapes the design of XAI solutions in response to the requisites of pulmonologists. In Figure 5.4 it is visible how these factors build up to answer the main research question.

The research embarks with an analysis of the clinical pathway, pinpointing pivotal junctures where XAI integration can strategically bolster pulmonologists' efforts. These junctures are closely tied through a chain of factors ranging from healthcare challenges to the complexities of AI adoption. These elements lay the groundwork for augmenting patient care effectiveness by incorporating AI insights. This initial exploration sets the stage for an in-depth examination of specific moments that necessitate the inclusion of XAI. Shifting the focus towards the content to be delivered during these crucial moments, the research delves into pulmonologists' preferences for custom explanations, appropriate visualisations, and a seamless interaction with AI systems. By outlining the sought-after information, this phase of the study offers a roadmap for tailoring XAI solutions to meet the specific needs of respiratory healthcare professionals.

Designing for Specific Moments

The moments that necessitate the implementation of XAI solutions emerge from various healthcare challenges and the perceived benefits of AI. Communication barriers and patient-centric care are pivotal challenges that XAI can address through improved communication, information sharing, and accurate diagnosis. Additionally, the potential benefits of enhanced efficiency, reduced costs, and better patient outcomes motivate the creation of XAI solutions. However, the design must also tackle limitations in AI adoption, ensuring interpretable explanations to foster trust and overcome concerns. Furthermore, moments arise from the need to educate medical staff about AI principles and navigate the dangers associated with AI adoption, such as ethical conflicts and patient safety concerns. By recognising characteristic features and applying them to journey maps, valuable insights into moments for XAI implementation can be gained, despite potential limitations in doctors' understanding.

1. **Healthcare Challenges:** Moments for XAI arise from challenges faced within healthcare, such as communication barriers and patient-centric care. XAI can help overcome these challenges by facilitating effective communication and information sharing, streamlining workflows, and improving diagnostic accuracy.

2. **Perceived Benefits of AI in Healthcare:** XAI can be helpful in moments where AI is perceived to offer significant benefits in healthcare. These benefits include improving efficiency, reducing costs, and enhancing patient outcomes.

3. **Limitations in Adoption:** Moments for XAI occur in the context of addressing the limitations and barriers to AI adoption. XAI can provide interpretable explanations to build trust and overcome concerns related to over-reliance on AI and the opacity of AI algorithms.

4. **Doctors' Limited Understanding:** XAI moments arise from the need for AI literacy among medical staff. Understanding AI principles, capabilities, and limitations is crucial for successful integration and adoption of XAI in clinical practice. The identification of all relevant moments for XAI may be limited by doctors' limited understanding of (X)AI. However, by considering characteristic features and journey maps, valuable insights into specific circumstances for XAI implementation can be gained.

5. **Dangers of AI:** Moments for XAI also emerge from the need to navigate the dangers associated with AI adoption, such as conflicting goals between healthcare providers and profit-driven companies and concerns about job roles and patient safety.

Tailoring Explanations for Different Moments

In addressing the question of what to show at these moments, it becomes apparent that the presentation of explanations in XAI systems should cater to pulmonologists' distinct needs. Task-specific visualisations, comprehensive explanations, and preferred granularity are essential components. Doctors demand explanations tailored to their medical tasks, accommodating different visual modalities for various medical branches. Comprehensive explanations, showcasing the AI's thinking process, help establish trust and confidence. The level of granularity should adapt to provide patient-specific insights and overall system functioning without overwhelming medical professionals. The timing of explanations is crucial, with readily available explanations before using AI and continuous access during use. Interactive explanations, enabling customisation and follow-up questions, foster cooperation between doctors and AI. Additionally, mimicking human judgement and interpretation skills, matching mental models, and overcoming resistance to AI advice play crucial roles in convincing doctors to embrace AI recommendations.

1. **Task-Specific Visualisations:**Doctors want explanations that are tailored to their specific medical tasks. Different branches of medicine may require different visual modalities. For example, a diagnosis tool might benefit from numeric representations like bar plots, while radiology might require a mix of images and rules to understand the decision-making process of the AI system.

2. **Comprehensive Explanations:** Doctors emphasise the importance of visualisations that present not only the final output but also the entire thinking process of the AI system. Including specific examples and counterexamples from past patient cases is seen as highly beneficial. Comprehensive explanations help build trust and confidence in the AI system.

3. **Preferred Granularity:** Doctors desire patient-specific explanations that align with their clinical practice and decision-making. They also value global explanations that provide insights into the overall functioning of AI systems without delving into intricate mathematical details. The granularity of explanations should be adaptable to strike the right balance between comprehensiveness and usability.

4. **Timing of Explanations:** Doctors prefer explanations to be readily available before they start using AI in their medical practice. This allows them to be aware of the AI system's limitations up-front and make informed decisions. They also seek explanations when there is a disagreement between the AI's prediction and their own clinical judgement. Moreover, having access to explanations with every decision during use enables ongoing evaluation and testing of AI systems, leading to increased trust and confidence over time.
5. **Interactive Explanations:** Doctors value interactive explanations that allow them to tailor the level of detail to their specific needs and expertise. The ability to engage the AI system with follow-up questions and receive meaningful responses fosters cooperation between doctors and the AI tool.
6. **Matching Mental Models:** Doctors seek additional background information to check if the AI system's factors align with the patient's condition. They want the AI to mimic human judgement and interpretation skills, which helps build confidence in the AI's recommendations.
7. **Overcoming Resistance to AI Advice:** Some doctors admit that they might disregard AI recommendations if their personal opinions override them. Effective explanations are crucial in ensuring that doctors view AI as a supportive tool rather than a replacement for their expertise. Properly designed explanations can help doctors take AI advice into consideration even when they make the final decision themselves.

Facilitating Engagement and Adoption

The transition from design to actual usage hinges on facilitating doctors' engagement and adoption of XAI solutions. Collaboration and trust emerge as significant factors, with doctors viewing AI as a valuable colleague rather than a replacement. Trust evolves over time through repeated interactions and validation, resulting in mutual learning and growth. Patient-centric care drives engagement, with AI explanations augmenting doctors' decision-making in uncertain situations. Meeting hospital use requirements is vital, ensuring AI's integration into existing tools and alignment with doctors' mental models. Ultimately, AI serves as a supportive reference, allowing doctors to retain decision-making authority and patient care responsibility. Trust is fortified through transparency, accountability, and validation, reinforcing the collaborative relationship between doctors and AI.

1. **Collaboration and Trust:** Doctors engage with explanations when they view AI as a valuable complementary colleague, providing evidence-based insights and risk assessments. Trust in AI develops over time through repeated interactions and validation. The collaborative relationship between doctors and AI is characterised by mutual learning and growth.
2. **Patient-Centric Care:** Doctors engage with explanations when they are patient-centric, aligning with the individual needs of each patient. AI explanations play a pivotal role in augmenting doctors' decision-making processes in uncertain situations and navigating the high-pressure environment of healthcare.
3. **Hospital Use Requirements:** Trust in AI systems also relies on validation, certification, transparency in algorithms, and accountability for decisions made by AI. The decision to use an AI tool should not be made by an individual doctors but decided by an expert committee. Doctors engage with explanations when AI is integrated into their existing decision-making tools, serves as a supportive tool, and does not disrupt effective communication with patients. Accountability for decisions should be with doctors, they expect ultimate decision-making authority and retain responsibility for patient care, while AI serves as a valuable reference and provides insights for informed decision-making.

6

Conclusion

This chapter marks the culmination of an intricate exploration into the integration of explainable artificial intelligence solutions in healthcare, specifically focusing on pulmonology. It unveils the pivotal role XAI plays in bridging the gap between complex AI outputs and human understanding for informed clinical decisions. The first section illuminates key findings, emphasising content delivery, explanations, and collaborative engagement as cornerstones of successful AI adoption. These findings will be presented in the form of a checklist which enables future researchers to easily use this when building XAI systems for the medical domain. The second section revisits the study's contributions: pinpointing opportunities in respiratory medicine and crafting human-centered XAI experiences for practitioners. Acknowledging limitations, the third section recognises the contextual bounds, sample size, and ethical considerations inherent in the research. In the recommendations section, it propels future exploration, from broadening speciality engagement to ethical frameworks. Lastly, the future work section calls for rigorous validation, patient involvement, and collaborative frameworks as we embark on the next chapter of XAI's integration in healthcare.

6.1. Key Findings

The integration of explainable artificial intelligence solutions into healthcare, particularly in the field of pulmonology, holds immense promise for improving patient care and enhancing medical practice. The concept of XAI emerged as a solution, serving as a conduit to bridge the gap between the intricate outputs of AI algorithms and the requisite human comprehension for sound clinical decision-making. This study aimed to explore the intricate interplay between the needs and intents of pulmonologists and the design of XAI solutions, focusing on key moments, content delivery, and facilitation of engagement and adoption. By presenting this in a summarised checklist form it is possible for future researchers to use this while designing XAI solutions for the medical domain. This list can also be found in a checklist format in the Appendix E.

- **Design XAI systems from the medical needs**

  Moments that require XAI implementation arise from healthcare challenges, perceived benefits of AI, limitations in adoption, doctors' limited understanding of AI, and the need to navigate the potential dangers of AI integration. These moments serve as entry points for designing tailored XAI solutions that can address communication barriers, promote patient-centric care, and overcome challenges associated with AI adoption.

- **Involve doctors to fit the system into their patient centric workflow**

  The research underlined the imperative of co-designing solutions, engaging in collaboration with domain experts, and embedding XAI within established clinical workflows. These concluding remarks resound the significance of the continuous pursuit of aligning AI technologies with the foundational values of the medical realm, patient-centric care, and alignment with existing workflows. The examination of XAI algorithms, the in-depth comprehension of medical contexts, and the user-centric design approach collectively compose a narrative that can lead to a transformation in healthcare practices for the betterment of both medical professionals and their patients.

- **Each task has different explainability needs**

  In addressing the question of what to show at these moments, the study found that XAI solutions should provide task-specific visualisations, comprehensive explanations, and preferred granularity. These elements are vital for catering to pulmonologists' distinct needs, enabling them to trust and effectively engage with AI systems. The explanations should also make sure to not overwhelm the doctor with information when they have limited time.

- **Trust through repeated interaction**

  Facilitating engagement and adoption of XAI solutions among pulmonologists involves building trust and collaboration. This collaborative relationship evolves over time through repeated interactions and validation showing the system can perform well with their patient group, leading to more collaboration and ultimately leading to mutual learning and growth.

- **Treat it as a helpful addition, not their replacement**

  Doctors' willingness to engage with AI as a colleague, rather than a replacement, is pivotal. Maintaining doctors' decision-making authority is also a critical factor that foster engagement with XAI. They want the system to show a human like judgement and interpretation skills that matches theirs which might limit their resistance to using the AI advice in their own judgements.

- **Doctors keep in charge and responsible for their patients**

  Trust in AI systems hinges on validation, certification, transparent algorithms, and accountability for AI-driven decisions, with the decision to use AI tools ideally made by expert committees rather than individual doctors. While AI serves as a valuable reference for insights in decision-making, ultimate accountability remains with physicians, who retain decision-making authority and patient care responsibility.

- **What explanations to show matters on where in the medical processes its used**

  The timing of explanations is crucial, with doctors preferring access to explanations before engaging with AI to understand its limitations and make informed decisions. Real-time explanations during disagreements between AI predictions and clinical judgement foster ongoing evaluation, building trust over time. Interactive explanations, tailored to doctors' needs, enhance cooperation, while aligning AI-generated insights with patient conditions and overcoming resistance to AI advice through effective explanations further bolster integration.

- **Align the explanation algorithms with the doctors not the model**

  The research highlighted that the alignment between pulmonologists' needs and XAI design is a complex process driven by various factors. The incorporation of XAI within the medical domain necessitates more than just algorithmic prowess; it hinges upon comprehending the unique requisites and perspectives of clinicians. While the algorithms themselves hold import, their efficacy is intrinsically linked to their alignment with end-users – the medical practitioners.

- **Doctors don't need to be convinced by anything other then results showing improved patient care**

  It is essential to persuade healthcare personnel to adopt the system in order to unlock the system's potential benefits. Getting medical staff to adopt the system might seem difficult, however by showing them that the XAI can be helpful offer significant benefits in healthcare makes them quickly adopt such a system. These benefits include improving efficiency, reducing costs, and most importantly enhancing patient outcomes and should be displayed with extensively validated results.

- **Educate medical staff on the dangers of (X)AI**

  Understanding (X)AI principles, capabilities, and limitations is crucial for successful integration and responsible adoption of XAI in clinical practice. The continues monitoring and collaboration of the XAI system is important to guarantee its safety. As (X)AI is still mostly limited understood by the doctors and it is not adequately thought during their education it is important to display the intrinsic problems and sometimes deceiving outcomes of (X)AI.

In conclusion, the design of XAI solutions for pulmonologists is a multifaceted process that requires careful consideration of specific moments, content delivery preferences, and strategies for engagement and adoption. By addressing these aspects, XAI can become an invaluable tool that enhances pulmonologists' decision-making processes, augments patient care, and reinforces the collaborative partnership between medical professionals and AI systems. The insights gained from this study contribute to the ongoing development and implementation of XAI solutions in the healthcare domain.

6.2. Contributions

The central contributions of this study revolve around identifying avenues within the domain of respiratory medicine and establishing the parameters necessary to establish a human-centered experience of Explainable AI for practitioners in this field and presenting them in a checklist form. The overarching goal is to pave the way for future research aimed at reducing barriers to the adoption of AI applications in the medical realm. The specific contributions can be categorised as follows:

The main contributions this work makes are related to finding opportunities in the respiratory medicine domain and about finding the parameters that are related to creating a Human-centred XAI experience for the practitioners in this field. With the key ambition of trying to work towards enabling future research into lowering the barrier for the implementation of AI applications in the medical domain. The exact contributions will be classified as follows:

Contribution 1 — *Opportunity finding: where and how AI and explanations could be applied in the pulmonology domain.*

This study contributes by identifying the characteristics of strategic entry points within the realm of pulmonology where AI and explanatory models can be seamlessly integrated. By pinpointing these specific areas, the research lays the groundwork for the effective implementation of AI solutions within the domain of respiratory medicine. This contribution holds the potential to revolutionise medical practices by leveraging AI to address critical challenges and optimise patient care in pulmonology.

Contribution 2 — *Identifying: Needs, wants, and goals of medical staff about explanations.*

The value of this research is further underscored by its comprehensive examination of the nuanced needs, preferences, and objectives of medical personnel concerning the integration of explanatory AI. By shedding light on these intricacies, this contribution offers invaluable insights into the human-centric requirements that are pivotal for the successful design and adoption of Explainable AI tools within the medical community. Through this understanding, the study establishes a foundation for fostering collaborative and effective utilisation of XAI technologies, aligning them with the aspirations and demands of medical professionals.

6.3. Limitations, Recommendations and Future Work

Acknowledging the inherent constraints in the research's scope and methodology, the limitations highlight contextual specificity and potential transferability challenges. Additionally, considerations about sample size, prototype dynamics, and the study's timeline are explored. These limitations collectively underscore the need for cautious interpretation of the study's outcomes. Moving beyond limitations, the recommendations section charts a course for future research. It advocates for broader engagement across medical specialties, emphasising diverse perspectives for robust insights. Integrating AI literacy and longitudinal studies are suggested to bridge gaps and capture the evolving impact of AI adoption. Furthermore, ethical dimensions are illuminated, urging further exploration into bias mitigation, patient privacy, and accountability. Incorporating patient voices and transitioning to dynamic AI solutions are also foregrounded. The future work portion looks ahead, envisioning a rigorous validation of XAI solutions in clinical environments for practicality and user satisfaction. Longitudinal exploration promises insights into sustained influence, while ethical frameworks ensure responsible AI integration. The role of education in fostering seamless AI interactions is discussed, along with the imperative of patient empowerment and addressing technological barriers. Collaborative frameworks are poised to guide the symphony of AI integration in the healthcare landscape.

Limitations

It's important to acknowledge the limitations inherent in the study, which influence the scope and applicability of its findings. One notable limitation lies in the contextual specificity of the research. While the study's outcomes provide valuable insights for the field of pulmonology and IPF, the transferability of its conclusions to other medical domains or different lung conditions might not be straightforward. Each medical speciality comes with its own distinct complexities that could influence the suitability and effectiveness of the proposed XAI solutions.

Another limitation pertains to the sample size and diversity of participants in the user study involving clinicians from the pulmonology department. Although the study's insights are insightful, the limited sample size and potential lack of diversity in terms of medical professionals' expertise and backgrounds could hinder the broader generalisability of the findings. Additionally, the prototypes used in the study to assess clinicians' preferences and requirements for XAI explanations are static in nature. While they offer valuable insights, they might not fully capture the dynamism and interactivity of actual XAI systems in real clinical settings. The study's timeline also presents a potential limitation, as it might restrict the depth of understanding that could be attained. The iterative nature of effective XAI integration demands more extended timeframes to fully explore and validate the real-world impact of XAI adoption.

Furthermore, the study's focus on technical and usability aspects might not comprehensively address the ethical concerns associated with AI adoption in healthcare. Ethical considerations concerning patient privacy, bias mitigation, and AI accountability are significant but might not be fully explored in this study. The acceptance and sustainability of XAI solutions in clinical practices over the long term could also be a potential limitation. While the study offers insights into initial perceptions and short-term responses, the evolving nature of XAI's integration within existing workflows and its long-term impact warrant further investigation.It's important to note that the study primarily engages medical professionals and lacks direct input from patients, particularly those diagnosed with IPF. Incorporating the patient perspective could offer additional insights into tailoring XAI solutions to address diverse stakeholder needs. Lastly, the successful deployment of XAI solutions relies on the presence of a specific level of technological infrastructure and seamless integration within existing healthcare systems. The study might not fully delve into potential challenges and barriers associated with this technological integration.

Recommendations

Venturing beyond its current boundaries, future research could traverse a broader spectrum of medical specialties and conditions. This expansion would unlock a deeper understanding of how XAI solutions can seamlessly adapt across diverse healthcare landscapes. To infuse the study's findings with greater credibility, a more expansive and diverse assembly of medical professionals, each contributing their unique expertise, is essential. This broader participant pool promises more perspectives that enrich the understanding of XAI's integration and usability. Acknowledging the potential gap in participants' AI literacy, integrating concise AI training or educational materials becomes an indispensable facet. By equipping participants with the necessary AI concepts and terminologies, this approach facilitates more meaningful interactions with XAI solutions. Pioneering a longitudinal approach, a study that tracks the enduring impact of XAI adoption within clinical practices is a compelling direction. Such an endeavour would unveil insights into the sustenance, acceptance, and evolving demands of medical professionals as they navigate the tides of time. Beyond the surface, delving into the ethical labyrinth of XAI integration within healthcare beckons further exploration. Navigating issues like bias mitigation, patient privacy, and AI accountability crafts a comprehensive perspective on the challenges and opportunities this convergence presents. In a bid to humanise the trajectory, involving patients, especially those grappling with IPF, in the co-creation and evaluation of XAI solutions assumes significance. This patient-centric approach ensures that the XAI systems resonate with the unique needs and concerns of those they intend to serve. Evolving beyond static prototypes, the future lies in interactive and dynamic XAI solutions that mirror real-time clinical scenarios. This transition promises an authentic portrayal of how medical professionals seamlessly engage with AI explanations in the fabric of their daily routines.

Future work

Yet, the story doesn't end with recommendations; Stepping beyond theory, real clinical environments call for rigorous usability tests of the proposed XAI solutions. This pragmatic validation promises insights into their practicality, efficiency, and user satisfaction, grounding them in the realities of healthcare. A longitudinal exploration of XAI's adoption promises to unveil its sustained influence on clinical decision-making, patient outcomes, and the overarching efficiency of healthcare delivery. The ethical dimension calls for dedicated attention. Crafted frameworks tailored to AI's integration in healthcare offer a moral compass, guiding the responsible design and deployment of XAI solutions, while addressing the ethical quandaries they may bring. Forging partnerships with medical institutions to infuse AI education within medical curricula prepares future healthcare professionals to seamlessly navigate interactions with AI technologies. Yet, other challenges must not be overlooked. An exploration of the technological barriers in integrating XAI solutions within established healthcare systems completes the picture showing the implementation realities. Pioneering comprehensive frameworks that delineate the roles, responsibilities, and harmonious interactions between medical practitioners and AI systems sets the stage for a symphony of efficient collaboration through explanations.

Appendix

Appendix A

This appendix section contains the questions discussed during the semi-structured discussion session during the co-creation session following the use of the explainability prototype.

Questions Co-creation

Interpretability

1.1 How do you think the system made certain choices?

1.2 How can the systems predictions be explained visualy?

Understandability

2.1 What could you add to the system to give a better view into the life of the patients?

2.2 How would you change the system to better trust its judgements?

Usefulness

3.1 How could the extra information help you in gaining a more accurate understanding about the patient?

Usability

4.1 What changes would make you use the extra information functionality of the system more often?

4.2 How would you use a version of this system for tasks in a real hospital setting?

Appendix B

This appendix section contains the questions asked during the follow-up interviews.

Follow-up Interviews

Interview 1

1. What is meant with better tools to communicate about end-of-life decisions?
2. Communication within the team or with the patient or caregiver?
3. How is current communication?
4. How would knowing which patient to treat with which medication at which time and which place; so, providing better personalised medicine at home and hospital look like?
5. You stated ”I would want to know what clustering you used.”, how deep would you want to know the overall make-up of the model? Is that coming from your experience as a doctor or with your experience with AI yourself
6. For your need ”Better individually targeted treatment adapted to needs and wishes of patients would help to improve care for patients with PF and hopefully also quality of life.” how would you normally probe for this?
7. What do you think are the current barriers for the use of AI in the medical field? Do you think those barriers can be solved by explainable AI?
8. Under which conditions do medical expert engage with explanations?
9. Do you think medical experts with the help of the journey map could pin-point moments where certain AI enabled systems could be used?

Interview 2

1. What are all the contacts between the hospital and the patient?
2. What is the purpose of each individual contact between the hospital and the patient?
3. How do they structure the check-ups?
4. How do they normally (i.e., without automated system) decide the questions to ask?
5. How can AI help you in improved decision making during those contact moments, looking at adding information catering to the needs?
6. What are the contacts between the hospital and other facilities that are not specialized but do provide the patient with healthcare?

Appendix C

This appendix section contains the design prompts and the main interview questions which were used during the semi-structured interview.

Design Prompts

Prompts Question 1.1:

- creating treatment plan with patients
- understanding patients' needs before check-ups

Prompts Question 2:

- Tedious: repetitive. Taking a lot of effort, time, or attention when you feel it should not.
- Challenging: cognitively, emotionally, or procedurally

Prompts Question 4:

- Tell short anecdote about: IBM Watson shows reliability problems with only 49% concordance with experts in some countries, even though the theoretical cases were great. "IBM Watson, Heal Thyself"
- Tell short anecdote about: Pneumonia risk prediction case where the AI system learned a correlation between 'people with Asthma' and 'having a lower chance of death' which was not true in real life. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission."

Prompts Question 4.2.1:

- Capabilities and limitations (try to connect with their pain points), e.g., known edge cases of the system Features being used, e.g., gender or ethnicity
- Data-related: descriptive (the characteristics of the patients), who collected it, how it was collected, pre-processing
- Connection with medical literature?
- Limitations of the algorithms (e.g., wrong correlations)

Prompts Question 4.3

Process

- Before implementation
- Early stages, for learning
- Anytime, put the system through its paces

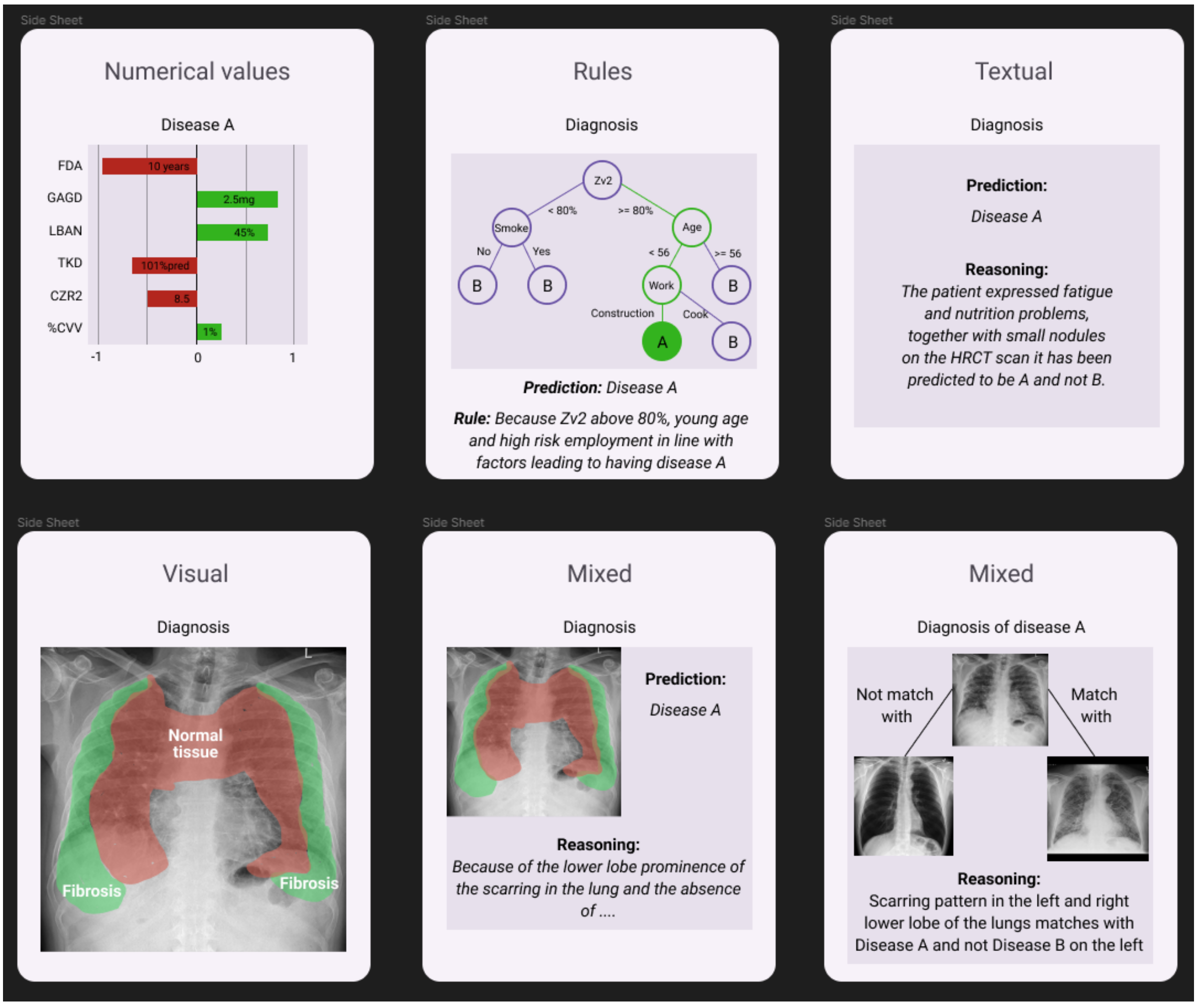
Decisions

- When the AI decision aligns with yours
- When the AI decision does not align with yours
- When you know the patient is an outlier (for whatever reason) based on your experience

Trigger

- Depends on the above, doctors keep the agency
- Always visible

Prompts Question 4.4



Prediction: *Disease A*

Rule: *Because Zv2 above 80%, young age and high risk employment in line with factors leading to having disease A*

Prediction:

Disease A

Reasoning:

The patient expressed fatigue and nutrition problems, together with small nodules on the HRCT scan it has been predicted to be A and not B.

Prediction:

Disease A

Reasoning:

Because of the lower lobe prominence of the scarring in the lung and the absence of ....

Reasoning:

Scarring pattern in the left and right lower lobe of the lungs matches with Disease A and not Disease B on the left

Semi-structured Interview Questions

Current practices

1 Considering from before diagnosis till the end stage from your specific disease, as a doctor how would you approach the different stages in it?

1.1 Could you give me some concrete examples?

1.2 Is that something that you were trained for, some clinical standard, or did it come from experience?

Understanding pain points in current practice

2 In this setting, which are the most challenging, or tedious, tasks you encounter when you are treating a patient?

2.1 Which actions do you take in those situations?

AI in healthcare

3 Considering the pain-points you just mentioned, do you think an AI (or another technological) solution could be beneficial to address those? If it helps, you can think about it as a ‘magic wand’ for your problems.

3.1 How would you use those technologies? As a collaborator, recommender, ...?

3.2 What factors would you consider before deciding to rely on it?

3.3 Are there any other cases you would like to have such systems in your practice?

Explanations

4 Do you share the concerns other researchers have raised about AI? Is there something you would want to know before, or while, using AI in your work?

4.1 Do you think having additional information, would help you when using this hypothetical AI system?

4.1.1 Is this connected to the explanations, or justifications, you may usually be expected to give (to colleagues or patients)?

4.2 Is this connected to the explanations, or justifications, you may usually be expected to give (to colleagues or patients)?

4.2.1 Would you be interested in knowing information about the system as a whole? Or only about specific decisions/recommendations about patients?

4.3 When would you like to receive that information?

4.4 When would you like to receive that information?

4.4.1 Why do you like it?

4.5 How would you use that information?

4.5.1 “Trust and use it” or as an extra data point to inform the next steps?

4.5.1.1 Trust the same way as you do with your colleagues?

4.6 Assume you can know everything about the system, would this change anything in the way you use or rely on it?

Appendix D

This appendix section contains the final checklist that was constructed to help the future development of XAI systems in healthcare.

XAI Healthcare Checklist

□ **Design XAI systems from the medical needs**

Moments that require XAI implementation arise from healthcare challenges, perceived benefits of AI, limitations in adoption, doctors' limited understanding of AI, and the need to navigate the potential dangers of AI integration. These moments serve as entry points for designing tailored XAI solutions that can address communication barriers, promote patient-centric care, and overcome challenges associated with AI adoption.

□ **Involve doctors to fit the system into their patient centric workflow**

The research underlined the imperative of co-designing solutions, engaging in collaboration with domain experts, and embedding XAI within established clinical workflows. These concluding remarks resound the significance of the continuous pursuit of aligning AI technologies with the foundational values of the medical realm, patient-centric care, and alignment with existing workflows. The examination of XAI algorithms, the in-depth comprehension of medical contexts, and the user-centric design approach collectively compose a narrative that can lead to a transformation in healthcare practices for the betterment of both medical professionals and their patients.

□ **Each task has different explainability needs**

In addressing the question of what to show at these moments, the study found that XAI solutions should provide task-specific visualisations, comprehensive explanations, and preferred granularity. These elements are vital for catering to pulmonologists' distinct needs, enabling them to trust and effectively engage with AI systems. The explanations should also make sure to not overwhelm the doctor with information when they have limited time.

□ **Trust through repeated interaction**

Facilitating engagement and adoption of XAI solutions among pulmonologists involves building trust and collaboration. This collaborative relationship evolves over time through repeated interactions and validation showing the system can perform well with their patient group, leading to more collaboration and ultimately leading to mutual learning and growth.

□ **Treat it as a helpful addition, not their replacement**

Doctors' willingness to engage with AI as a colleague, rather than a replacement, is pivotal. Maintaining doctors' decision-making authority is also a critical factor that foster engagement with XAI. They want the system to show a human like judgement and interpretation skills that matches theirs which might limit their resistance to using the AI advice in their own judgements.

□ **Doctors keep in charge and responsible for their patients**

Trust in AI systems hinges on validation, certification, transparent algorithms, and accountability for AI-driven decisions, with the decision to use AI tools ideally made by expert committees rather than individual doctors. While AI serves as a valuable reference for insights in decision-making, ultimate accountability remains with physicians, who retain decision-making authority and patient care responsibility.

□ **What explanations to show matters on where in the medical processes its used**

The timing of explanations is crucial, with doctors preferring access to explanations before engaging with AI to understand its limitations and make informed decisions. Real-time explanations during disagreements between AI predictions and clinical judgement foster ongoing evaluation, building trust over time. Interactive explanations, tailored to doctors' needs, enhance cooperation, while aligning AI-generated insights with patient conditions and overcoming resistance to AI advice through effective explanations further bolster integration.

□ **Align the explanation algorithms with the doctors not the model**

The research highlighted that the alignment between pulmonologists' needs and XAI design is a complex process driven by various factors. The incorporation of XAI within the medical domain necessitates more than just algorithmic prowess; it hinges upon comprehending the unique requisites and perspectives of clinicians. While the algorithms themselves hold import, their efficacy is intrinsically linked to their alignment with end-users – the medical practitioners.

□ **Doctors don't need to be convinced by anything other then results showing improved patient care**

It is essential to persuade healthcare personnel to adopt the system in order to unlock the system's potential benefits. Getting medical staff to adopt the system might seem difficult, however by showing them that the XAI can be helpful offer significant benefits in healthcare makes them quickly adopt such a system. These benefits include improving efficiency, reducing costs, and most importantly enhancing patient outcomes and should be displayed with extensively validated results.

□ **Educate medical staff on the dangers of (X)AI**

Understanding (X)AI principles, capabilities, and limitations is crucial for successful integration and responsible adoption of XAI in clinical practice. The continues monitoring and collaboration of the XAI system is important to guarantee its safety. As (X)AI is still mostly limited understood by the doctors and it is not adequately thought during their education it is important to display the intrinsic problems and sometimes deceiving outcomes of (X)AI.

Bibliography

[1] Rana Abdullah and Bahjat Fakieh. “Health Care Employees’ Perceptions of the Use of Artificial Intelligence Applications: Survey Study”. In: *J Med Internet Res* 22.5 (May 2020), e17620. ISSN: 1438-8871. DOI: 10.2196/17620. URL: http://www.ncbi.nlm.nih.gov/pubmed/32406857.

[2] Amina Adadi and Mohammed Berrada. “Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)”. In: *IEEE access* 6 (2018), pp. 52138–52160.

[3] Eike Adams. “The joys and challenges of semi-structured interviewing”. English. In: *Community Practitioner* 83 (July 2010). Article, pp. 18+. ISSN: 14622815. URL: https://link.gale.com/apps/doc/A229720174/HRCA?u=anon~b6a91b33&sid=googleScholar&xid=e9a35cb0.

[4] Julius Adebayo et al. “Sanity Checks for Saliency Maps”. In: *CoRR* abs/1810.03292 (2018). arXiv: 1810.03292. URL: http://arxiv.org/abs/1810.03292.

[5] Omolola A. Adeoye-Olatunde and Nicole L. Olenik. “Research and scholarly methods: Semi-structured interviews”. In: *JACCP: JOURNAL OF THE AMERICAN COLLEGE OF CLINICAL PHARMACY* 4.10 (2021), pp. 1358–1367. DOI: https://doi.org/10.1002/jac5.1441. eprint: https://accpjournals.onlinelibrary.wiley.com/doi/pdf/10.1002/jac5.1441. URL: https://accpjournals.onlinelibrary.wiley.com/doi/abs/10.1002/jac5.1441.

[6] Zubair Ahmad et al. “Artificial intelligence (AI) in medicine, current applications and future role with special emphasis on its potential and promise in pathology: present and future impact, obstacles including costs and acceptance among pathologists, practical and philosophical considerations. A comprehensive review”. In: *Diagnostic Pathology* 16.1 (Mar. 2021), p. 24. ISSN: 1746-1596. DOI: 10.1186/s13000-021-01085-4. URL: https://doi.org/10.1186/s13000-021-01085-4.

[7] Andrew Anderson et al. “Mental Models of Mere Mortals with Explanations of Reinforcement Learning”. In: *ACM Trans. Interact. Intell. Syst.* 10.2 (May 2020). ISSN: 2160-6455. DOI: 10.1145/3366485. URL: https://doi.org/10.1145/3366485.

[8] Michael V Angrosino. *Naturalistic observation*. Routledge, 2016.

[9] Anna Markella Antoniadi et al. “Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review”. In: *Applied Sciences* 11.11 (2021). ISSN: 2076-3417. DOI: 10.3390/app11115088. URL: https://www.mdpi.com/2076-3417/11/11/5088.

[10] Yuri YM Aung, David CS Wong, and Daniel SW Ting. “The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare”. In: *British medical bulletin* 139.1 (2021), pp. 4–15.

[11] John W Ayers et al. “Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum”. In: *JAMA Internal Medicine* (2023).

[12] Agathe Balayn et al. “How Can Explainability Methods Be Used to Support Bug Identification in Computer Vision Models?” In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517474. URL: https://doi.org/10.1145/3491102.3517474.

[13] Jean Baric-Parker and Emily E. Anderson. “Patient Data-Sharing for AI: Ethical Challenges, Catholic Solutions”. In: *The Linacre Quarterly* 87.4 (2020). PMID: 33100395, pp. 471–481. DOI: 10.1177/0024363920922690. eprint: https://doi.org/10.1177/0024363920922690. URL: https://doi.org/10.1177/0024363920922690.

[14] Shaney Barratt et al. “Idiopathic Pulmonary Fibrosis (IPF): An Overview”. In: *Journal of Clinical Medicine* 7.8 (Aug. 2018), p. 201. ISSN: 2077-0383. DOI: 10.3390/jcm7080201. URL: http://dx.doi.org/10.3390/jcm7080201.

[15] Mohamed Karim Belaid et al. “Compare-xAI: Toward Unifying Functional Testing Methods for Post-hoc XAI Algorithms into an Interactive and Multi-dimensional Benchmark”. In: ().

[16] Visar Berisha et al. “Digital medicine and the curse of dimensionality”. In: *npj Digital Medicine* 4.1 (Oct. 2021), p. 153. ISSN: 2398-6352. DOI: 10.1038/s41746-021-00521-5. URL: https://doi.org/10.1038/s41746-021-00521-5.

[17] Subrato Bharati, M Rubaiyat Hossain Mondal, and Prajoy Podder. “A Review on Explainable Artificial Intelligence for Healthcare: Why, How, and When?” In: *IEEE Transactions on Artificial Intelligence* (2023).

[18] Umang Bhatt et al. “Explainable Machine Learning in Deployment”. In: *CoRR* abs/1909.06342 (2019). arXiv: 1909.06342. URL: http://arxiv.org/abs/1909.06342.

[19] Agata Blasiak, Jeffrey Khong, and Theodore Kee. “CURATE. AI: optimizing personalized medicine with artificial intelligence”. In: *SLAS TECHNOLOGY: Translating Life Sciences Innovation* 25.2 (2020), pp. 95–105.

[20] Dawn Branley-Bell, Rebecca Whitworth, and Lynne Coventry. “User trust and understanding of explainable ai: Exploring algorithm visualisations and user biases”. In: *International Conference on Human-Computer Interaction*. Springer. 2020, pp. 382–399.

[21] Virginia Braun and Victoria Clarke. *Thematic analysis.* American Psychological Association, 2012.

[22] Patrick J Bridgeman, Mary Barna Bridgeman, and Joseph Barone. “Burnout syndrome among healthcare professionals”. In: *The Bulletin of the American Society of Hospital Pharmacists* 75.3 (2018), pp. 147–152.

[23] Olesia Brill and Eric Knauss. “Structured and unobtrusive observation of anonymous users and their context for requirements elicitation”. In: *2011 IEEE 19th International Requirements Engineering Conference*. IEEE. 2011, pp. 175–184.

[24] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. “To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making”. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1 (2021), pp. 1–21.

[25] Adrian Bussone, Simone Stumpf, and Dympna O’Sullivan. “The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems”. In: *2015 International Conference on Healthcare Informatics*. 2015, pp. 160–169. DOI: 10.1109/ICHI.2015.26.

[26] Federico Cabitza and Jean-David Zeitoun. “The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence”. en. In: *Ann Transl Med* 7.8 (Apr. 2019), p. 161.

[27] Carrie Cai et al. “”Hello AI”: Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making”. In: *Proceedings of the ACM on Human-Computer Interaction* 3 (Nov. 2019), pp. 1–24. DOI: 10.1145/3359206.

[28] Rich Caruana et al. “Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission”. In: *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 2015, pp. 1721–1730.

[29] Ahmad Chaddad et al. “Survey of Explainable AI Techniques in Healthcare”. In: *Sensors* 23.2 (Jan. 2023), p. 634. ISSN: 1424-8220. DOI: 10.3390/s23020634. URL: http://dx.doi.org/10.3390/s23020634.

[30] Mei Chen and Michel Decary. “Artificial intelligence in healthcare: An essential guide for health leaders”. In: *Healthcare Management Forum* 33.1 (2020). PMID: 31550922, pp. 10–18. DOI: 10.1177/0840470419873123. eprint: https://doi.org/10.1177/0840470419873123. URL: https://doi.org/10.1177/0840470419873123.

[31] Han Shi Jocelyn Chew and Palakorn Achananuparp. “Perceptions and needs of artificial intelligence in health care to increase adoption: scoping review”. In: *Journal of medical Internet research* 24.1 (2022), e32939.

[32] Giovanni Cinà et al. “Why we do need explainable ai for healthcare”. In: *arXiv preprint arXiv:2206.15363* (2022).

[33] Julien Colin et al. “What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 2832–2845.

[34] Carlo Combi et al. “A manifesto on explainability for artificial intelligence in medicine”. In: *Artificial Intelligence in Medicine* 133 (Oct. 2022), p. 102423. DOI: 10.1016/j.artmed.2022.102423.

[35] Nilakash Das et al. “Collaboration between explainable artificial intelligence and pulmonologists improves the accuracy of pulmonary function test interpretation”. In: *European Respiratory Journal* (2023). ISSN: 0903-1936. DOI: 10.1183/13993003.01720-2022. eprint: https://erj.ersjournals.com/content/early/2023/03/15/13993003.01720-2022.full.pdf. URL: https://erj.ersjournals.com/content/early/2023/03/15/13993003.01720-2022.

[36] Khishigsuren Davagdorj et al. “Explainable artificial intelligence based framework for non-communicable diseases prediction”. In: *IEEE Access* 9 (2021), pp. 123672–123688.

[37] Thomas Davenport and Ravi Kalakota. “The potential for artificial intelligence in healthcare”. en. In: *Future Healthc J* 6.2 (June 2019), pp. 94–98.

[38] Weiping Ding et al. “Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey”. In: *Information Sciences* 615 (2022), pp. 238–292. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2022.10.013. URL: https://www.sciencedirect.com/science/article/pii/S002002552201132X.

[39] Jeff Druce et al. “Brittle AI, Causal Confusion, and Bad Mental Models: Challenges and Successes in the XAI Program”. In: *CoRR* abs/2106.05506 (2021). arXiv: 2106.05506. URL: https://arxiv.org/abs/2106.05506.

[40] Yuhan Du et al. “An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus”. In: *Scientific Reports* 12.1 (2022), p. 1170.

[41] Jamie Duell et al. “A comparison of explanations given by explainable artificial intelligence methods on analysing electronic health records”. In: *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE. 2021, pp. 1–4.

[42] Juan Manuel Durán and Karin Rolanda Jongsma. “Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI”. In: *Journal of Medical Ethics* 47.5 (2021), pp. 329–335. ISSN: 0306-6800. DOI: 10.1136/medethics-2020-106820. eprint: https://jme.bmj.com/content/47/5/329.full.pdf. URL: https://jme.bmj.com/content/47/5/329.

[43] Roman Egger and Joanne Yu. “A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts”. In: *Frontiers in sociology* 7 (2022).

[44] Upol Ehsan et al. “Human-Centered Explainable AI (HCXAI): beyond opening the black-box of AI”. In: *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 2022, pp. 1–7.

[45] Upol Ehsan et al. “Human-Centered Explainable AI (HCXAI): Coming of Age”. In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–7.

[46] Farshad Firouzi et al. “AI-driven data monetization: The other face of data in IoT-based smart and connected health”. In: *IEEE Internet of Things Journal* 9.8 (2020), pp. 5581–5599.

[47] Maximilian Förster et al. “Fostering human agency: A process for the design of user-centric XAI systems”. In: (2020).

[48] Nick Fox. "Using interviews in a research project". In: *The NIHR RDS for the East Midlands/Yorkshire & the Humber* 26 (2009).

[49] Susanne Gaube et al. "Do as AI say: susceptibility in deployment of clinical decision-aids". In: *NPJ digital medicine* 4.1 (2021), p. 31.

[50] Julie Gerlings, Millie Søndergaard Jensen, and Arisa Shollo. "Explainable ai, but explainable to whom? an exploratory case study of xai in healthcare". In: *Handbook of Artificial Intelligence in Healthcare: Vol 2: Practicalities and Prospects* (2022), pp. 169–198.

[51] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. "The false hope of current approaches to explainable artificial intelligence in health care". In: *The Lancet Digital Health* 3.11 (2021), e745–e750.

[52] Amit K. Ghosh and Shashank Joshi. "Tools to manage medical uncertainty". In: *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 14.5 (2020), pp. 1529–1533. ISSN: 1871-4021. DOI: https://doi.org/10.1016/j.dsx.2020.07.055. URL: https://www.sciencedirect.com/science/article/pii/S1871402120303039.

[53] Sherif Gonem et al. "Applications of artificial intelligence and machine learning in respiratory medicine". en. In: *Thorax* 75.8 (May 2020), pp. 695–701.

[54] Maarten Grootendorst. *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. 2022. arXiv: 2203.05794 [cs.CL].

[55] Thomas Grote and Philipp Berens. "On the ethics of algorithmic decision-making in healthcare". In: *Journal of medical ethics* 46.3 (2020), pp. 205–211.

[56] Paul KJ Han, William MP Klein, and Neeraj K Arora. "Varieties of uncertainty in health care: a conceptual taxonomy". In: *Medical Decision Making* 31.6 (2011), pp. 828–838.

[57] Navid Hasani et al. "Trustworthy artificial intelligence in medical imaging". In: *PET clinics* 17.1 (2022), pp. 1–12.

[58] Dennis M. Hedderich et al. "AI for Doctors—A Course to Educate Medical Professionals in Artificial Intelligence for Medical Imaging". In: *Healthcare* 9.10 (2021). ISSN: 2227-9032. DOI: 10.3390/healthcare9101278. URL: https://www.mdpi.com/2227-9032/9/10/1278.

[59] Tina Hernandez-Boussard et al. "MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care". In: *Journal of the American Medical Informatics Association* 27.12 (June 2020), pp. 2011–2015. ISSN: 1527-974X. DOI: 10.1093/jamia/ocaa088. eprint: https://academic.oup.com/jamia/article-pdf/27/12/2011/34838637/ocaa088.pdf. URL: https://doi.org/10.1093/jamia/ocaa088.

[60] Andreas Holzinger et al. "Explainable AI Methods - A Brief Overview". In: Apr. 2022, pp. 13–38. ISBN: 978-3-031-04082-5. DOI: 10.1007/978-3-031-04083-2_2.

[61] Piotr Janowiak, Amelia Szymanowska-Narloch, and Alicja Siemińska. "IPF Respiratory Symptoms Management - Current Evidence". en. In: *Front Med (Lausanne)* 9 (July 2022), p. 917973.

[62] Kevin B Johnson et al. "Precision medicine, AI, and the future of personalized health care". In: *Clinical and translational science* 14.1 (2021), pp. 86–93.

[63] Jiwon Jung. "Developing Data-enabled Design in the Field of Digital Health". In: (2023).

[64] Harmanpreet Kaur et al. "Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–14. ISBN: 9781450367080. DOI: 10.1145/3313831.3376219. URL: https://doi.org/10.1145/3313831.3376219.

[65] Christopher J. Kelly et al. "Key challenges for delivering clinical impact with artificial intelligence". In: *BMC Medicine* 17.1 (Oct. 2019), p. 195. ISSN: 1741-7015. DOI: 10.1186/s12916-019-1426-2. URL: https://doi.org/10.1186/s12916-019-1426-2.

[66] Sunnie S. Y. Kim et al. “”Help Me Help the AI”: Understanding How Explainability Can Support Human-AI Interaction”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Apr. 2023. DOI: `10.1145/3544548.3581001`. URL: `https://doi.org/10.1145%2F3544548.3581001`.

[67] Anastasiya Kiseleva, Dimitris Kotzinos, and Paul De Hert. “Transparency of AI in healthcare as a multilayered system of accountabilities: Between legal requirements and technical limitations”. In: *Frontiers in Artificial Intelligence* 5 (2022), p. 82.

[68] A Baki Kocaballi et al. “Envisioning an artificial intelligence documentation assistant for future primary care consultations: A co-design study with general practitioners”. In: *Journal of the American Medical Informatics Association* 27.11 (Aug. 2020), pp. 1695–1704. ISSN: 1527-974X. DOI: `10.1093/jamia/ocaa131`. eprint: `https://academic.oup.com/jamia/article-pdf/27/11/1695/34363907/ocaa131.pdf`. URL: `https://doi.org/10.1093/jamia/ocaa131`.

[69] Abigél Margit Kolonics-Farkas et al. “Differences in Baseline Characteristics and Access to Treatment of Newly Diagnosed Patients With IPF in the EMPIRE Countries”. In: *Frontiers in Medicine* 8 (2021). ISSN: 2296-858X. DOI: `10.3389/fmed.2021.729203`. URL: `https://www.frontiersin.org/articles/10.3389/fmed.2021.729203`.

[70] Hans-Peter Kriegel et al. “Density-based clustering”. In: *WIREs Data Mining and Knowledge Discovery* 1.3 (2011), pp. 231–240. DOI: `https://doi.org/10.1002/widm.30`. eprint: `https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.30`. URL: `https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.30`.

[71] Satyapriya Krishna et al. “The Disagreement Problem in Explainable Machine Learning: A Practitioner’s Perspective”. In: *CoRR* abs/2202.01602 (2022). arXiv: `2202.01602`. URL: `https://arxiv.org/abs/2202.01602`.

[72] Shinjini Kundu. “AI in medicine must be explainable”. In: *Nature Medicine* 27.8 (Aug. 2021), pp. 1328–1328. ISSN: 1546-170X. DOI: `10.1038/s41591-021-01461-z`. URL: `https://doi.org/10.1038/s41591-021-01461-z`.

[73] Samuli Laato et al. “How to explain AI systems to end users: a systematic literature review and research agenda”. In: *Internet Research* 32.7 (2022), pp. 1–31.

[74] Yi Lai, Atreyi Kankanhalli, and Desmond Ong. “Human-AI collaboration in healthcare: A review and research agenda”. In: (2021).

[75] Himabindu Lakkaraju et al. “Rethinking Explainability as a Dialogue: A Practitioner’s Perspective”. In: *CoRR* abs/2202.01875 (2022). arXiv: `2202.01875`. URL: `https://arxiv.org/abs/2202.01875`.

[76] Kristina Lång et al. “The Safety of an Artificial Intelligence Supported Screen-Reading Procedure Versus Standard Double Reading in the Mammography Screening with Artificial Intelligence (MASAI) Trial: A Randomised, Controlled, Screening Accuracy Study”. In: *Controlled, Screening Accuracy Study* ().

[77] Rebecca Lawton et al. “Are more experienced clinicians better able to tolerate uncertainty and manage risks? A vignette study of doctors in three NHS emergency departments in England”. In: *BMJ Quality & Safety* 28.5 (2019), pp. 382–388.

[78] David J. Lederer and Fernando J. Martinez. “Idiopathic Pulmonary Fibrosis”. In: *New England Journal of Medicine* 378.19 (2018). PMID: 29742380, pp. 1811–1823. DOI: `10.1056/NEJMra1705751`. eprint: `https://doi.org/10.1056/NEJMra1705751`. URL: `https://doi.org/10.1056/NEJMra1705751`.

[79] Augustine S Lee et al. “The burden of idiopathic pulmonary fibrosis: an unmet public health need”. en. In: *Respir Med* 108.7 (Apr. 2014), pp. 955–967.

[80] Jang Ho Lee et al. “Epidemiology and comorbidities in idiopathic pulmonary fibrosis: a nationwide cohort study”. In: *BMC Pulmonary Medicine* 23.1 (Feb. 2023), p. 54.

[81] Minyoung Lee, Joohyoung Jeon, and Hongchul Lee. “Explainable AI for domain experts: a post Hoc analysis of deep learning for defect classification of TFT–LCD panels”. In: *Journal of Intelligent Manufacturing* 33.6 (Aug. 2022), pp. 1747–1759. ISSN: 1572-8145. DOI: 10.1007/s10845-021-01758-3. URL: https://doi.org/10.1007/s10845-021-01758-3.

[82] Beth L. Leech. “Asking Questions: Techniques for Semistructured Interviews”. In: *PS: Political Science and Politics* 35.4 (2002), pp. 665–668. DOI: 10.1017/S1049096502001129.

[83] Kicky G van Leeuwen et al. “Clinical use of artificial intelligence products for radiology in the Netherlands between 2020 and 2022”. In: *European Radiology* (July 2023).

[84] Karim Lekadir et al. “Artificial Intelligence in Healthcare-Applications, Risks, and Ethical and Societal Impacts”. In: *European Parliament* (2022).

[85] Karim Lekadir et al. “FUTURE-AI: guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging”. In: *arXiv preprint arXiv:2109.09658* (2021).

[86] Q. Vera Liao, Daniel Gruen, and Sarah Miller. “Questioning the AI: Informing Design Practices for Explainable AI User Experiences”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Apr. 2020. DOI: 10.1145/3313831.3376590. URL: https://doi.org/10.1145%2F3313831.3376590.

[87] Q. Vera Liao and Kush R. Varshney. “Human-Centered Explainable AI (XAI): From Algorithms to User Experiences”. In: *CoRR* abs/2110.10790 (2021). arXiv: 2110.10790. URL: https://arxiv.org/abs/2110.10790.

[88] Q. Vera Liao et al. “Question-Driven Design Process for Explainable AI User Experiences”. In: *CoRR* abs/2104.03483 (2021). arXiv: 2104.03483. URL: https://arxiv.org/abs/2104.03483.

[89] Han Liu, Vivian Lai, and Chenhao Tan. “Understanding the Effect of Out-of-Distribution Examples and Interactive Explanations on Human-AI Decision Making”. In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW2 (Oct. 2021). DOI: 10.1145/3479552. URL: https://doi.org/10.1145/3479552.

[90] Xiao Liu et al. “Doctor-Patient communication: a comparison betweenTelemedicine consultation and face-to-face consultation”. In: *Internal Medicine* 46.5 (2007), pp. 227–232.

[91] Alex John London. “Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability”. en. In: *Hastings Cent Rep* 49.1 (Jan. 2019), pp. 15–21.

[92] Jacqueline Low. “Unstructured and Semi-structured interviews in Health Research”. In: *Researching Health: Qualitative, Quantitative and Mixed methods. London: Sage publications* (2019), pp. 123–41.

[93] Scott M Lundberg et al. “Explainable machine-learning predictions for the prevention of hypoxaemia during surgery”. en. In: *Nat Biomed Eng* 2.10 (Oct. 2018), pp. 749–760.

[94] Scott M Lundberg et al. “From local explanations to global understanding with explainable AI for trees”. In: *Nature machine intelligence* 2.1 (2020), pp. 56–67.

[95] Toby M Maher and Mary E Strek. “Antifibrotic therapy for idiopathic pulmonary fibrosis: time to treat”. In: *Respiratory research* 20.1 (2019), pp. 1–9.

[96] Toby M Maher et al. “Identifying barriers to idiopathic pulmonary fibrosis treatment: a survey of patient and physician views”. In: *Respiration* 96.6 (2018), pp. 514–524.

[97] Toby M. Maher et al. “Global incidence and prevalence of idiopathic pulmonary fibrosis”. In: *Respiratory Research* 22.1 (July 2021), p. 197. ISSN: 1465-993X. DOI: 10.1186/s12931-021-01791-z. URL: https://doi.org/10.1186/s12931-021-01791-z.

[98] Aniek F. Markus, Jan A. Kors, and Peter R. Rijnbeek. “The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies”. In: *Journal of Biomedical Informatics* 113 (2021), p. 103655. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2020.103655. URL: https://www.sciencedirect.com/science/article/pii/S1532046420302835.

[99] Aniek F. Markus, Jan A. Kors, and Peter R. Rijnbeek. “The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies”. In: *Journal of Biomedical Informatics* 113 (2021), p. 103655. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2020.103655. URL: https://www.sciencedirect.com/science/article/pii/S1532046420302835.

[100] Andreia Martinho, Maarten Kroesen, and Caspar Chorus. “A healthy debate: Exploring the views of medical doctors on the ethics of artificial intelligence”. In: *Artificial Intelligence in Medicine* 121 (2021), p. 102190. ISSN: 0933-3657. DOI: https://doi.org/10.1016/j.artmed.2021.102190. URL: https://www.sciencedirect.com/science/article/pii/S0933365721001834.

[101] John A. McDermid et al. “Artificial Intelligence explainability: The Technical and Ethical Dimensions”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379.2207 (2021), p. 20200363. DOI: 10.1098/rsta.2020.0363.

[102] John A. McDermid et al. “Artificial intelligence explainability: the technical and ethical dimensions”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379.2207 (2021), p. 20200363. DOI: 10.1098/rsta.2020.0363. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2020.0363. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2020.0363.

[103] Leland McInnes and John Healy. “Accelerated Hierarchical Density Based Clustering”. In: *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. 2017, pp. 33–42. DOI: 10.1109/ICDMW.2017.12.

[104] Leland McInnes, John Healy, and Steve Astels. “hdbscan: Hierarchical density based clustering.” In: *J. Open Source Softw.* 2.11 (2017), p. 205.

[105] Bertalan Meskó and Marton Görög. “A short guide for medical professionals in the era of artificial intelligence”. In: *npj Digital Medicine* 3.1 (Sept. 2020), p. 126. ISSN: 2398-6352. DOI: 10.1038/s41746-020-00333-z. URL: https://doi.org/10.1038/s41746-020-00333-z.

[106] Bertalan Meskó, Gergely Hetényi, and Zsuzsanna Győrffy. “Will artificial intelligence solve the human resource crisis in healthcare?” In: *BMC Health Services Research* 18.1 (July 2018), p. 545. ISSN: 1472-6963. DOI: 10.1186/s12913-018-3359-4. URL: https://doi.org/10.1186/s12913-018-3359-4.

[107] Amanda I. Messinger, Gang Luo, and Robin R. Deterding. “The doctor will see you now: How machine learning and artificial intelligence can extend our understanding and treatment of asthma”. In: *Journal of Allergy and Clinical Immunology* 145.2 (2020), pp. 476–478. ISSN: 0091-6749. DOI: https://doi.org/10.1016/j.jaci.2019.12.898. URL: https://www.sciencedirect.com/science/article/pii/S0091674919326053.

[108] Carlo Metta et al. “Exemplars and Counterexemplars Explanations for Image Classifiers, Targeting Skin Lesion Labeling”. In: *2021 IEEE Symposium on Computers and Communications (ISCC)*. 2021, pp. 1–7. DOI: 10.1109/ISCC53001.2021.9631485.

[109] Tim Miller. “Explanation in artificial intelligence: Insights from the social sciences”. In: *Artificial Intelligence* 267 (2019), pp. 1–38. ISSN: 0004-3702. DOI: https://doi.org/10.1016/j.artint.2018.07.007. URL: https://www.sciencedirect.com/science/article/pii/S0004370218305988.

[110] Christoph Molnar et al. “General pitfalls of model-agnostic interpretation methods for machine learning models”. In: *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*. Springer. 2020, pp. 39–68.

[111] Vincenzo Moscato, Antonio Picariello, and Giancarlo Sperlí. “A benchmark of machine learning approaches for credit score prediction”. In: *Expert Systems with Applications* 165 (2021), p. 113986.

[112] Mohammad Naiseh et al. “Personalising explainable recommendations: literature and conceptualisation”. In: *Trends and Innovations in Information Systems and Technologies: Volume 2 8*. Springer. 2020, pp. 518–533.

[113] Mahsan Nourani et al. “Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems”. In: *26th International Conference on Intelligent User Interfaces*. IUI ’21. College Station, TX, USA: Association for Computing Machinery, 2021, pp. 340–350. ISBN: 9781450380171. DOI: 10.1145/3397481.3450639. URL: https://doi.org/10.1145/3397481.3450639.

[114] Sandip S. Panesar et al. “Promises and Perils of Artificial Intelligence in Neurosurgery”. In: *Neurosurgery* 87.1 (2020). ISSN: 0148-396X. URL: https://journals.lww.com/neurosurgery/Fulltext/2020/07000/Promises_and_Perils_of_Artificial_Intelligence_in.4.aspx.

[115] Cecilia Panigutti, Alan Perotti, and Dino Pedreschi. “Doctor XAI: An Ontology-Based Approach to Black-Box Sequential Data Classification Explanations”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 629–639. ISBN: 9781450369367. DOI: 10.1145/3351095.3372855. URL: https://doi.org/10.1145/3351095.3372855.

[116] Cecilia Panigutti et al. “Co-design of human-centered, explainable AI for clinical decision support”. In: *ACM Transactions on Interactive Intelligent Systems* (2023).

[117] Cecilia Panigutti et al. “Understanding the Impact of Explanations on Advice-Taking: A User Study for AI-Based Clinical Decision Support Systems”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3502104. URL: https://doi.org/10.1145/3491102.3502104.

[118] Michael Quinn Patton. *Qualitative research & evaluation methods: Integrating theory and practice*. Sage publications, 2014.

[119] Urja Pawar et al. “Incorporating Explainable Artificial Intelligence (XAI) to aid the Understanding of Machine Learning in the Healthcare Domain.” In: *AICS*. 2020, pp. 169–180.

[120] Urja Pawar et al. *Incorporating Explainable Artificial Intelligence (XAI) to aid Understanding of Machine Learning in the Healthcare Domain*. Oct. 2020. DOI: 10.13140/RG.2.2.24754.02246.

[121] Seyedeh Neelufar Payrovnaziri et al. “Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review”. en. In: *J Am Med Inform Assoc* 27.7 (July 2020), pp. 1173–1185.

[122] Junfeng Peng et al. “An Explainable Artificial Intelligence Framework for the Deterioration Risk Prediction of Hepatitis Patients”. In: *Journal of Medical Systems* 45 (May 2021). DOI: 10.1007/s10916-021-01736-5.

[123] Lena Petersson et al. “Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in Sweden”. In: *BMC Health Services Research* 22.1 (2022), pp. 1–16.

[124] Hilary Pinnock et al. “Prioritising the respiratory research needs of primary care: The International Primary Care Respiratory Group (IPCRG) e-delphi exercise”. In: *Primary Care Respiratory Journal* 21.1 (2012), pp. 19–27. DOI: 10.4104/pcrj.2012.00006.

[125] Milda Poceviciute, Gabriel Eilertsen, and Claes Lundström. “Survey of XAI in digital pathology”. In: *CoRR* abs/2008.06353 (2020). arXiv: 2008.06353. URL: https://arxiv.org/abs/2008.06353.

[126] Milda Pocevičiūtė, Gabriel Eilertsen, and Claes Lundström. “Survey of XAI in digital pathology”. In: *Artificial intelligence and machine learning for digital pathology: state-of-the-art and future challenges* (2020), pp. 56–88.

[127] Forough Poursabzi-Sangdeh et al. “Manipulating and measuring model interpretability”. In: *Proceedings of the 2021 CHI conference on human factors in computing systems*. 2021, pp. 1–52.

[128] Thomas P. Quinn et al. “The three ghosts of medical AI: Can the black-box present deliver?” In: *Artificial Intelligence in Medicine* 124 (2022), p. 102158. ISSN: 0933-3657. DOI: https://doi.org/10.1016/j.artmed.2021.102158. URL: https://www.sciencedirect.com/science/article/pii/S0933365721001512.

[129] Hamon R, Junklewitz H, and Sanchez Martin JI. "Robustness and Explainability of Artificial Intelligence". In: KJ-NA-30040-EN-N (online) (2020). ISSN: 1831-9424 (online). DOI: 10.2760/57493(online).

[130] Ganesh Raghu et al. "Diagnosis of idiopathic pulmonary fibrosis. An official ATS/ERS/JRS/ALAT clinical practice guideline". In: *American journal of respiratory and critical care medicine* 198.5 (2018), e44–e68.

[131] Sandeep Reddy, John Fox, and Maulik P Purohit. "Artificial intelligence-enabled healthcare delivery". In: *Journal of the Royal Society of Medicine* 112.1 (2019). PMID: 30507284, pp. 22–28. DOI: 10.1177/0141076818815510. eprint: https://doi.org/10.1177/0141076818815510. URL: https://doi.org/10.1177/0141076818815510.

[132] Carlo Reverberi et al. "Experimental evidence of effective human–AI collaboration in medical decision-making". In: *Scientific Reports* 12.1 (Sept. 2022), p. 14952. ISSN: 2045-2322. DOI: 10.1038/s41598-022-18751-2. URL: https://doi.org/10.1038/s41598-022-18751-2.

[133] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision model-agnostic explanations". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018.

[134] Hans-Gerd Ridder. *Book Review: Qualitative data analysis. A methods sourcebook*. Vol. 28. 4. Sage publications Sage UK: London, England, 2014.

[135] Avi Rosenfeld. "Better metrics for evaluating explainable artificial intelligence". In: *Proceedings of the 20th international conference on autonomous agents and multiagent systems*. 2021, pp. 45–50.

[136] Padhraig Ryan et al. "Using artificial intelligence to assess clinicians' communication skills". In: *Bmj* 364 (2019).

[137] Thomas Schaffter et al. "Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms". In: *JAMA network open* 3.3 (2020), e200265–e200265.

[138] Tjeerd A.J. Schoonderwoerd et al. "Human-centered XAI: Developing design patterns for explanations of clinical decision support systems". In: *International Journal of Human-Computer Studies* 154 (2021), p. 102684. ISSN: 1071-5819. DOI: https://doi.org/10.1016/j.ijhcs.2021.102684. URL: https://www.sciencedirect.com/science/article/pii/S1071581921001026.

[139] Suzanna Sia, Ayush Dalmia, and Sabrina J Mielke. "Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too!" In: *arXiv preprint arXiv:2004.14914* (2020).

[140] Alberto Signoroni et al. "BS-Net: Learning COVID-19 pneumonia severity on a large chest X-ray dataset". en. In: *Med Image Anal* 71 (Mar. 2021), p. 102046.

[141] Kacper Sokol and Peter Flach. "One explanation does not fit all: The promise of interactive explanations for machine learning transparency". In: *KI-Künstliche Intelligenz* 34.2 (2020), pp. 235–250.

[142] Kaitao Song et al. "MPNet: Masked and Permuted Pre-training for Language Understanding". In: *arXiv preprint arXiv:2004.09297* (2020).

[143] Paolo Spagnolo et al. "Idiopathic pulmonary fibrosis: An update". In: *Annals of Medicine* 47.1 (2015). PMID: 25613170, pp. 15–27. DOI: 10.3109/07853890.2014.982165. eprint: https://doi.org/10.3109/07853890.2014.982165. URL: https://doi.org/10.3109/07853890.2014.982165.

[144] Eliza Strickland. "IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care". In: *IEEE Spectrum* 56.4 (2019), pp. 24–31.

[145] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. "Visual, Textual or Hybrid: The Effect of User Expertise on Different Explanations". In: *26th International Conference on Intelligent User Interfaces*. IUI '21. College Station, TX, USA: Association for Computing Machinery, 2021, pp. 109–119. ISBN: 9781450380171. DOI: 10.1145/3397481.3450662. URL: https://doi.org/10.1145/3397481.3450662.

[146] Sana Tonekaboni et al. “What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use”. In: *CoRR* abs/1905.05134 (2019). arXiv: 1905.05134. URL: http://arxiv.org/abs/1905.05134.

[147] Eric J. Topol. “High-performance medicine: the convergence of human and artificial intelligence”. In: *Nature Medicine* 25.1 (Jan. 2019), pp. 44–56. ISSN: 1546-170X. DOI: 10.1038/s41591-018-0300-7. URL: https://doi.org/10.1038/s41591-018-0300-7.

[148] Helena Vasconcelos et al. “Explanations Can Reduce Overreliance on AI Systems During Decision-Making”. In: *Proc. ACM Hum.-Comput. Interact.* 7.CSCW1 (Apr. 2023). DOI: 10.1145/3579605. URL: https://doi.org/10.1145/3579605.

[149] Giulia Vilone and Luca Longo. “Classification of Explainable Artificial Intelligence Methods through Their Output Formats”. In: *Machine Learning and Knowledge Extraction* 3.3 (2021), pp. 615–661. ISSN: 2504-4990. DOI: 10.3390/make3030032. URL: https://www.mdpi.com/2504-4990/3/3/32.

[150] Dakuo Wang et al. ““Brilliant AI Doctor” in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445432. URL: https://doi.org/10.1145/3411764.3445432.

[151] Dakuo Wang et al. “From human-human collaboration to Human-AI collaboration: Designing AI systems that can work together with people”. In: *Extended abstracts of the 2020 CHI conference on human factors in computing systems*. 2020, pp. 1–6.

[152] Lars Witell et al. “Idea Generation: Customer Co-Creation versus Traditional Market Research Techniques”. In: *Journal of Service Management* 22 (Apr. 2011), pp. 140–159. DOI: 10.1108/09564231111124190.

[153] Victoria Wurcel et al. “The value of diagnostic information in personalised healthcare: a comprehensive concept to facilitate bringing this technology into healthcare systems”. In: *Public health genomics* 22.1-2 (2019), pp. 8–15.

[154] Oskar Wysocki et al. “Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making”. In: *Artificial Intelligence* 316 (2023), p. 103839. ISSN: 0004-3702. DOI: https://doi.org/10.1016/j.artint.2022.103839. URL: https://www.sciencedirect.com/science/article/pii/S0004370222001795.

[155] Yaochen Xie et al. “Task-agnostic graph explanations”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 12027–12039.

[156] Xuhai Xu et al. *XAIR: A Framework of Explainable AI in Augmented Reality*. Mar. 2023.

[157] Christopher C. Yang. “Explainable Artificial Intelligence for Predictive Modeling in Healthcare”. In: *Journal of Healthcare Informatics Research* 6.2 (June 2022), pp. 228–239. ISSN: 2509-498X. DOI: 10.1007/s41666-022-00114-1. URL: https://doi.org/10.1007/s41666-022-00114-1.

[158] Thomas James York et al. “Clinician and computer: a study on doctors’ perceptions of artificial intelligence in skeletal radiography”. In: *BMC Medical Education* 23.1 (Jan. 2023), p. 16. ISSN: 1472-6920. DOI: 10.1186/s12909-022-03976-6. URL: https://doi.org/10.1186/s12909-022-03976-6.

Acknowledgments

Laying in front of you is a tangible embodiment of 273 days of work, with the deep fingerprints of mentors, peers, and loved ones distinctly present, shaping the narrative of this thesis. As my thesis grew to nearly double the word count of Kafka's novella The Metamorphosis, I often jokingly thought about the parallels between myself and its protagonist. In The Metamorphosis, the central character awakens one fateful morning to find himself subjected to a peculiar and disconcerting transformation – he has morphed into a sizeable insect. This transformation necessitates his navigation of the profound shifts, both internal and external, that this metamorphosis imposes on his life. In the process of crafting my thesis, I occasionally experienced a similar sensation of transformation, albeit the one of my life turning into the realm of my own academic work; me becoming my thesis. Fortunately, the narrative of my story concludes with a more positive outcome then Kafka's. This I attribute to the unwavering support of the individuals that have stood by me during this phase of my life.

In no order of importance I want to thank the people who gave their time and compassion to help me emotionally by providing encouragement, listening to my complaints, understanding the problems I faced and being part of a positive environment in which I could make the most of myself. But also supporting me practically by providing research assistance, feedback, helping with my time management, providing me with resources, technical assistance, opening up their medical consultations for me, and sharing their contacts. By lack of space and not wanting to sound like an award show ceremony I will only say some short positive notes about the persons I'm thankful to. I want to convey my thanks to Jie and Jiwon, my daily and external supervisors. The memory of my arrival at Building 28 last year, ready to present my own thesis topic, only to find a different path outlined by the two of you, is enduring. In retrospect, I can confidently assert that making the most of that chance was, unquestionably, a wise decision. Especially as that got me very committed and no matter the circumstances always enthusiastic and supportive supervisors. Supportive not only for my thesis, but also for me as a person after that. I also want to thank Lorenzo who joined the project at exactly the right time where his extensive knowledge and opinions about explainability gave me immeasurable guidance and support on the topic and will undoubtedly lead him towards his PhD title. I also want to thank Agathe for her expert input at the start of the project. These people together with the expertise they offered did really make the difference. I thank the doctors, nurses and supportive staff at Erasmus MC for providing me with the opportunity to explore the medical domain and get an insight into their practises. The unwavering optimism, genuine warmth, and compassionate nature displayed by these individuals, despite grappling with the harsh realities of caring for people with a terminal illness like IPF, left me amazed and motivated. I want to personally thank Marlies, Gizal and Karen for their continuous input and readiness to help. I am deeply grateful for the warm reception I received from all the patients. Your openness and willingness to include me without hesitation meant a lot to me and I hope that your further journey dealing with this disease might be met with the same positiveness and perseverance showed during those moments we met. I'm grateful for my friends and the people around me who, also working on their thesis, shared with me bonding moments forged through shared challenges. I could write paragraphs about each, but I will just shortly thank some of them Maurits, Jan, Koen, Arjen, Sreeparna, Ruixuan, Ayush your advise, collaboration, speedminton & spikeball matches kept me able to work throughout this time. From the dedication and perseverance each of them showed I am certain that each and any of them will promptly successfully finish their master thesis. Inês the structure you provided in my life, together with the moments and experiences we shared were a driving force for me to start, continue and keep continuing day after day, every day to finish this achievement in life. I have been more abroad with you, on our escaping trips this year, then any other person in my whole life, and I wouldn't have liked it any other way. The comments you made about my habits have had a lasting impression on my life and I know wherever you will choose to go next in the world you'll make a positive impact on those people too. Lastly I want to thank my family, even though for them what I was exactly doing for my thesis often was an enigma, they still stood there for me every time I needed advice or motivation to go on. I knew and I still know that I can always depend on you guys, and of that I am ever grateful. But in the end I am the most delighted for the knowledge that I had to adjust the font size on this page to include everyone and anything I am thankful for.

Forever grateful,
Rembrandt
Delft, August 2023