Designing XAI-based Computer-aided Diagnostic Systems: Operationalising User Research Methods

Elsa Oliveira[1,†], Cristiana Braga[1,†], Ana Sampaio[1], Tiago Oliveira[2], Filipe Soares[1,*] and Luís Rosado[1,*]

[1]Fraunhofer Portugal AICOS, Rua Alfredo Allen 455/461, 4200-135, Porto, Portugal
[2]First Solutions - Sistemas de Informação, S.A., Rua Conselheiro Costa Braga, Matosinhos, Portugal

Abstract

AI technology has the potential to support humans' processes and tasks by augmenting human capabilities and effectiveness. Computer-aided systems have been implemented in healthcare mainly to support clinical decisions. As in other areas, the impact, complexity, and opacity of AI operations have led to the establishment of guidelines for trustworthy AI, which implies being understandable. This study describes the user research work carried out by a multidisciplinary team composed of ML engineers, design researchers, and medical experts, to inform the design of algorithms and user interfaces for two XAI-based clinical decision support tools targeted at Cervical cancer and Glaucoma screening. In particular, we sought to leverage and bridge individual and collective expertise to understand the context, decision-making processes and criteria, and values that frame the respective clinical decisions. The article describes how we operationalised the research activities with expert users and what strategies we followed for subsequent content analysis, ending with the sharing of lessons learned as valuable insights for other research teams interested in designing computer-aided diagnostic systems based on human-centred XAI approaches.

Keywords

Explainable AI, Computer-aided detection, Decision Support System, Ophthalmology, Glaucoma, Cytology, Cervical cancer, Retinal Imaging, Microscopy,

1. Introduction

Despite its potential, AI has struggled to be understandable. This requirement has been critical in several areas, mainly in healthcare [1, 2], where AI can support clinical decisions. There has been consensus on the need to promote accountable and trustworthy AI. The European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG) says that whenever an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process [3]. These considerations have led AI towards Explainable AI (XAI), which in turn leveraged human-centred design methods to uncover what to explain, why, how, and for whom [4, 5, 6, 7]. We share a study of how we operationalised Human-Centred Design (HCD) methods to inform the design of algorithms and user interfaces for two XAI-based clinical decision support tools for Cervical cancer and Glaucoma screening. We were concerned with grasping medical experts' mental models and reasoning processes. While mental models are mental constructs that represent a distinct possibility and derive a conclusion from them, reasoning implies a process to derive a conclusion and depends on envisaging the possibilities (mental models) consistent with a starting point [8]. So, to access the diagnosis' reasoning and identify the decision-making data and the explanations structures to apply in the design of XAI-based clinical decision support tools, we needed to get inside the diagnosis process with those who practice it - the medical experts. [9, 10]

This paper is structured into 6 sections. First section introduces the demand for XAI systems. Section 2 identifies the objectives and design of the study, subdivided into three phases: contextualisation, elicitation, and validation. Section 3 briefly introduces the medical context of Cervical cancer and Glaucoma, on which the work was focused. Section 4 describes how we operationalised the research work focusing on the research activities with the users and the analysis of the collected content. Finally, in section 5 we share lessons learned from this study, and section 6 indicates the main conclusions and

Elsa Oliveira, Cristiana Braga, Ana Sampaio, Tiago Oliveira, Filipe Soares and Luís Rosado. 2023. Designing XAI-based Computer-aided Diagnostic Systems: Operationalising User Research Methods. Joint Proceedings of the ACM IUI Workshops 2023, March 2023, Sydney, Australia, 11 pages.

*Corresponding author.
†These authors contributed equally.
elsa.oliveira@aicos.fraunhofer.pt (E. Oliveira);
cristiana.braga@aicos.fraunhofer.pt (C. Braga);
ana.sampaio@aicos.fraunhofer.pt (A. Sampaio);
tiago.oliveira@first-global.com (T. Oliveira);
filipe.soares@aicos.fraunhofer.pt (F. Soares);
luis.rosado@aicos.fraunhofer.pt (L. Rosado)
0000-0002-7105-9654 (E. Oliveira); 0000-0002-9384-2252 (C. Braga); 0000-0003-1770-4429 (A. Sampaio); 0000-0002-2881-313X (F. Soares); 0000-0002-8060-831X (L. Rosado)

future work.

2. Goals and Study Design

As a multidisciplinary team, composed by Machine Learning (ML) engineers, design researchers, and medical experts, we sought to leverage and bridge individual and collective expertise, especially from the medical area for which the systems were conceived, to inform the design of algorithms and user interfaces for explainable decision support software targeted at Cervical cancer and Glaucoma screening. We based our study on the results of the user research activities, which aimed to understand the context, processes, and values that frame clinical decisions in the above-mentioned health areas. The research process was guided by three phases: contextualisation, elicitation, and validation. The user research methods applied in each phase (described below) returned a considerable amount of fieldwork materials, i.e., written, verbal and visual content, that researchers needed to analyse to enhance understanding of the data. In analysing these data, we initially focused on codifying what the clinicians said (written transcription) about their decision-making process. However, most of their explanations evoked visual aspects of the images. As we are non-experts, we quickly realised that we needed to match what the doctors were saying with the respective visual elements they were characterising in their explanations. For example, when clinicians explained that a cell was abnormal "because it had a halo around the nucleus", HCD and ML researchers could not understand what a halo was without a visual reference of that cell containing a halo. The content analysis process based on Transcription, Coding, and Systematisation, paved the way for the decision-making data and inherent reasoning structure.

2.1. Contextualisation

As a first step, the researchers sought to become familiarised with the jargon, clinical practices, and decision-making processes used by health professionals. Initially, the researchers made more superficial research in online medical articles, also to acquire the basic knowledge to prepare for the interviews with medical experts. In fact, the contextualisation was accomplished mainly through semi-structured interviews which script applied Task Reflection and Retrospection methods, to prompt participants to reflect and describe their daily clinical tasks and diagnostic practices. The interviews gave us an overview of clinical practices, decision-making processes, values, and a quick window into participants' mental models as they gave examples of clinical cases and how they decided on them.

2.2. Elicitation

The elicitation phase asked for more detail on the decision-making process, decision-making data, and on the explanation structures that support it. To this end the research team relied on referenced methods for mental models' elicitation [11], such as Semi-structured interviews, Observation, and Think-Aloud [12, 13], together with co-creation practices - that made use of imaging data and other design materials to facilitate participants in demonstrating the processes of analysis and decision-making. Nielsen refers Think-Aloud method as effective in giving us insights into users' mental models regarding a given task. The study also drew on the procedures of a field study method based on Observation and interviews to understand work practices and behaviors - Contextual inquiry [14, 15]. Kim Salazar on Nielsen Norman Group website highlights the value of the contextual inquiry method – to inquiry in context, which results in a collaborative interpretation between researchers and expert users about work practices and behaviors, with a more in-depth understanding of experts' reasoning. With these references in mind, the research team set up workshops to observe, and question medical experts analysing and deciding on clinical cases and from clinical data.

2.3. Validation

The validation stage allowed us to discuss, correct, complete, and refine with medical experts the research findings. Through co-creation design practices, researchers designed group and individual workshops, in both remote and in-person versions, in order to display the decision-making criteria within the respective structures, to be discussed and easily edited and iterated in real-time. For some questions, we used A/B testing method for participants to select the best option.

3. Cervical cancer and Glaucoma

As mentioned in the introduction, this study addresses two distinct health areas, Anatomical Pathology and Ophthalmology, more specifically, Cervical cancer and Glaucoma. The main goal of the study was to design an explainable decision support software per area, both based on imaging screening, to be used by medical experts, and physicians in training.

Cervical cancer screening will mainly rely on cytological microscopic images, while Glaucoma screening on retinal images. The research team needed to go deep into each clinical practice to define the systems' requirements. Table 1 lists the main aspects that characterise the two health areas under study, considering the analysis that medical experts carry out per patient. This knowledge

Table 1
Characteristics of Cervical cancer and Glaucoma screening

	Cytology: Cervical cancer	**Ophthalmology: Glaucoma**
Purpose	Screening for reversible gynaecological disease	Screening for irreversible eye disease
Main Imaging Data collection	Cervical cytology specimen (liquid-based)	Color Fundus Photography
Complementary exams	HPV diagnosis	Ocular anatomy (e.g., narrow angle), IOP (intraocular pressure), Corneal tomography for pachymetry, volume and depth, Retinal Nerve Fiber Layer Thickness (RNFLT) measured via OCT, and visual field tests.
Patient data	(Not mandatory) Age, last menstruation, contraceptive method, relevant medical therapeutics, e.g., hormonal, chemotherapy	Age, ethnicity, family history of the condition, associated pathologies (e.g., diabetes, cataracts) and risk medication (e.g., antidepressant)
Digitalisation outcome - Artefact of analysis	Approximately 100 microscopic images per sample	Between 1 and 7 retinal fundus images per eye [16, 17]
Variation	Each image represents a small section of the entire sample	Each image can vary in eye laterality (left or right) or field of view
Image navigation	The expert browses, image by image, zooming in and out, to identify cells with abnormal aspect	The expert checks an image individually, zooming in and out, to look for abnormalities in the main structures
Criteria of adequacy	Representation of the Transformation Zone. Minimum of 5000 squamous cells [18] (average of 3.8 cells/image). Good image quality.	Visibility and sharpness of optic nerve and macula. Completeness of temporal arcade. Clear visibility of small vessels. Field of view well illuminated (min. 80%) [19].
Case classification	Grading by lesion level, according to the Bethesda System's convention [20]	Staging of Glaucoma [21]
Comparative analysis	Experts take intermediate squamous cells as a reference for comparison	Experts check the symmetry between the left eye and the right eye
End-users	Cytopathologists, cytotechnicians (diagnose), and physicians in training	Glaucomatologists, (diagnose), ophthalmologists, and physicians in training

was built-up throughout the contextualisation and elicitation phases. While both areas share common aspects, they also have some significant differences.

4. Operationalisation of research activities

In this section, we describe how we operationalised the research activities for the contextualisation, elicitation, and validation phases. At the beginning of the study, all participants received a general informed consent that gave an overview of the user research agenda, being further provided a detailed informed consent per activity. The study counted with up to 5 participants per health area - in Cervical cancer, 3 cytopathologists and 2 cytotechnologists; in Glaucoma, 4 ophthalmologists specialised in Glaucoma (glaucomatologists). Because of COVID-19, in particular the restrictions on in person group meetings and on normal access to hospitals and clinical settings, most user research activities took place remotely through digital and online platforms. Through these, participants were able to access anonymised screening images, as well as other clinical data, to demonstrate their decision-making process, and reasoning, while observed and questioned by the research team.

4.1. Contextualisation interviews

After some basic research through online medical articles, the research team draw the interview script addressed to the medical experts. The interview script aimed at:

understanding clinical procedures, i.e., from the first consultation up to and after diagnosis, eliciting medical experts' values, i.e., their motivation for the medical field, examples of impactful cases, and, very important, medical expectations regarding the introduction of AI systems in the clinical practice. The semi-structured interviews were carried out remotely through video call by Microsoft Teams software. To note that in Cervical cancer, the research team took advantage of the results of a previous and related study with cytopathologists and cytotechnicians [22, 20] that had conducted in-person semi-structured interviews with the same participants. These interviews enabled us to understand the processes involved in cytological analysis, from the reception of the sample to the diagnosis.

4.1.1. Interviews analysis

Once the interviews were completed, we transcribed them using oTranscribe software [23]. We then organised the participants' insights into the main themes raised during the interviews.

4.2. Workshops for eliciting diagnostic processes

Familiarised with both medical areas, we inspired ourselves in the contextual inquiry method to design the workshops that would enable us to elicit experts' diagnostic assessment process. Our goal was to understand what experts look at when they analyse a clinical case, specifically, an imaging examination, and what criteria they use to assess whether it is a pathological change.

4.2.1. Designing remote workshops

The analysis of imaging examinations was a requirement for the diagnosis assessment, thus, we needed to observe experts analysing such images. Usually, we would visit the experts' workplace and observe them in a real clinical setting. However, due to COVID-19, the workshops had to be remote, and so, we mimicked this observation remotely.

For Cervical cancer, we asked a cytologist to provide us with images of liquid-based cytological samples. For Glaucoma, we asked a glaucomatologist to provide us with retinal images. However, this was not all. From the interviews, we learned that both medical fields complemented images' interpretation with clinical data, which we were attending to. But we also learned that Glaucomatous pathology was more complex to diagnose, because glaucomatologists often had to integrate complementary diagnostic exams to reach a diagnosis.

With this in mind, we asked the glaucomatologist to provide us with a set of anonymised complementary diagnostic exams with diverse diagnosis: *unconfirmed*, *borderline*, *early stage Glaucoma* and *advanced stage Glaucoma*.

To ensure participants' unbiased decisions, we first conducted individual workshops. Afterwards we ran a group workshop for both medical fields to help identify consensual criteria and foster discussions around the least consensual ones.

4.2.2. Conducting individual workshops

Each workshop consisted of one main task: the participant, as a medical expert, would assess imaging examinations in real-time and think aloud about their analysis. This way, we could follow the assessment process and ask questions whenever needed to better understand it. Moreover, we asked participants to annotate relevant findings whose appearance suggested a pathological change and to provide the respective diagnosis classification. Experts in Cervical cancer classified cytological images according to the Bethesda System's convention. Experts in Glaucoma classified retinal images according to the four stages mentioned in section 4.2.1. Each participant analysed from three to seven images consisting of liquid-based cytological samples (in Cervical cancer) or from four to sixteen images consisting of eight pairs of retinal images (in Glaucoma). Figure 1 shows a visual field of a cytological sample with two cells annotated by a participant for their abnormality, and figure 2 a retinal image being analysed by a participant.

Given the interdependence with other examinations in Glaucoma diagnosis, and the wider range of diagnostic factors outside imaging data, Glaucoma workshops comprised an additional task. Each participant was asked to list the steps of a usual medical procedure, from the first consultation to the diagnosis, describing other relevant examinations beyond the retinal image. Figure 3 shows the timeline filled in by 1 of the 4 participants considering the examinations performed throughout the analysis of a given clinical case (for example, José, 62 years old with high intraocular pressure) - from the first consultation to diagnosis, and, where necessary, in the patient follow-up. In the second part of the workshop, the participant accessed anonymised eye examinations, corresponding to different diagnoses, from *non-Glaucoma* to *advanced stage Glaucoma*, to then select and analyse the most representative of a specific clinical case. Figure 2 is one of the retinal images that a participant has zoomed in and centred on the optical disc to show which image features reflect the state of the eye's structures and should therefore be considered as a criterion for decision making.

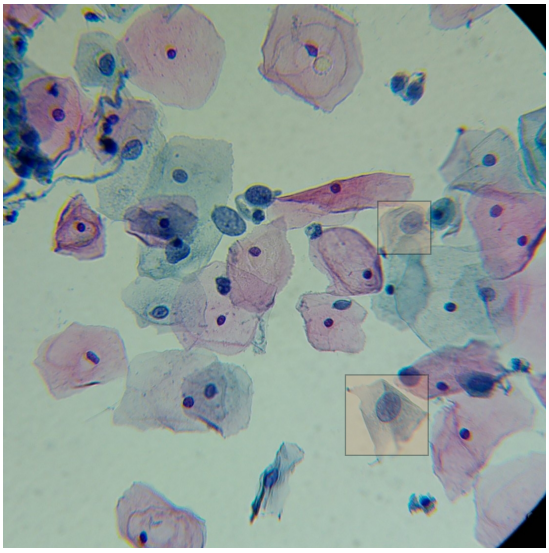


Figure 1: Screenshot from the individual workshop for elicitation, conducted remotely through a digital platform showing a digital liquid-based cervical sample, with two of several cells noted by the participant for their abnormality

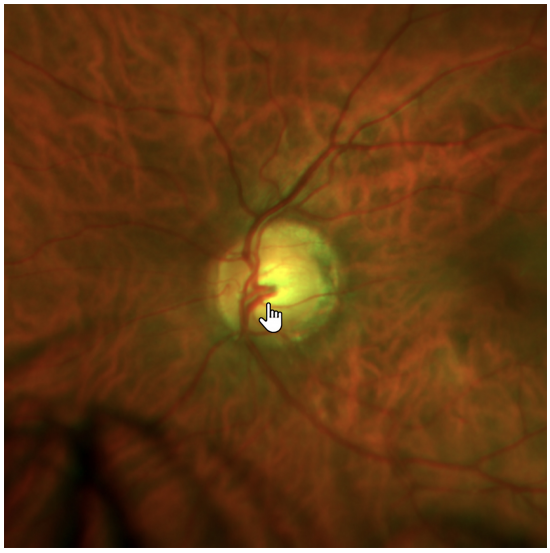


Figure 2: Screenshot from the individual workshop for elicitation, conducted remotely through a digital platform showing a retinal image being analysed with the participant pointing to the optical nerve to explain what is a pathologic optic disc cupping (excavation)

4.2.3. Transcribing and analysing

As we transcribed the workshops, it became evident that we should assign textual excerpts to image cut-outs, as most of experts' explanations consisted of descriptions of visible characteristics in the analysed images. Thus, mapping the object of analysis with the respective transcription enabled us to keep a correspondence between what was said and what was being observed in the image (Figure 4).

We did this for each participant. Almost all participants, from both medical fields, mentioned how the analysis and conclusions of some clinical cases were subjective. For instance, a glaucomatologist said: "Sometimes it's not black and white, it's grey", meaning that the same examinations and clinical data may lead experts to different decisions. This happens when the available elements for diagnosis are unclear, due to either image characteristics that hinder experts' analysis (e.g. blurry image), or to characteristics of the anatomical structures, which can be themselves confusing (when the same visual appearance can be the result of different possible causes), which requires more tests and more time.

Moreover, Cervical cancer experts highlighted the subjectivity intra- and inter-observer, explaining that not only the decision may vary between experts, as the same expert could give a different classification to the same sample at different moments in time. Therefore, we sought this subjective dimension by comparing the analysis of each participant to the same object of analysis, and in fact, we were able to verify this. Below (Figure 5) is an example of the same cytological field analysed by the five Cervical cancer experts. Both annotations of suspicious or abnormal cells and final classifications varied across analysts. While three experts classified the cytological field as ASC-US - an official classification for uncertainty regarding an Abnormal Squamous Cell(s), two of the five experts classified the sample as LSIL - an official classification comparable to ASC-US, but that assigns a Low grade of Intraepithelial Lesion to the Squamous cell(s).

4.2.4. Coding and systematisation

Once we completed the transcripts, we created a categorisation matrix in Excel to code the data into a set of categories that constitute the building blocks of the explanations, which allowed us to uncover a generic explanation structure suitable for both use cases. We used the columns' headings for the categories, and the rows to list the image that has triggered the explanation together with the textual explanation (quote) and the set of categorisable criteria (Figure 6).

As we went on with the codification, we iteratively refined the categories into Key structure(s) examined, Key feature(s) concerned, Risk factor, Not Cervical cancer/Glaucoma factor, Doubt factor, Result attributed, and finally, Key expression used by the expert. As the Excel's content increased, we identified that the criteria we filled in the categories would repeat. So, we created an Excel

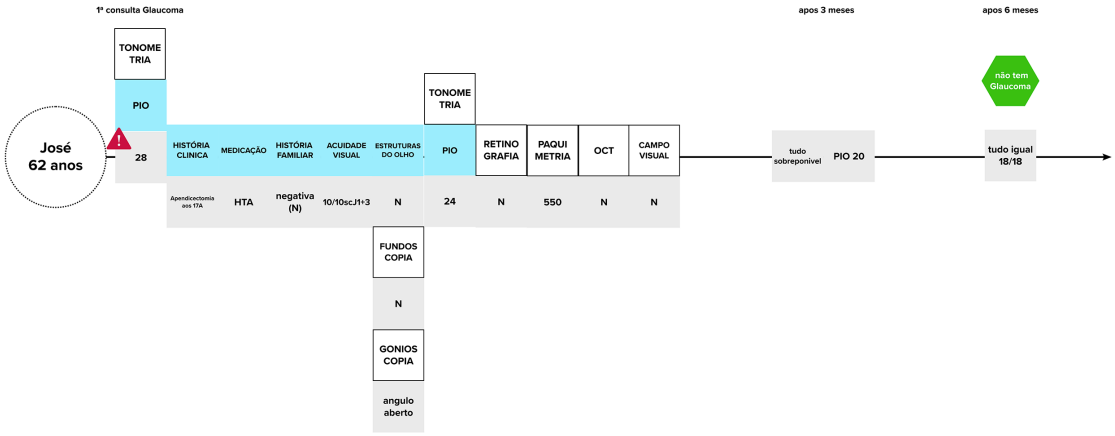


Figure 3: Screenshot from the individual workshops, conducted through Mural platform, listing a sequence of steps (required eye examinations) until reaching a diagnosis for a given hypothetical patient

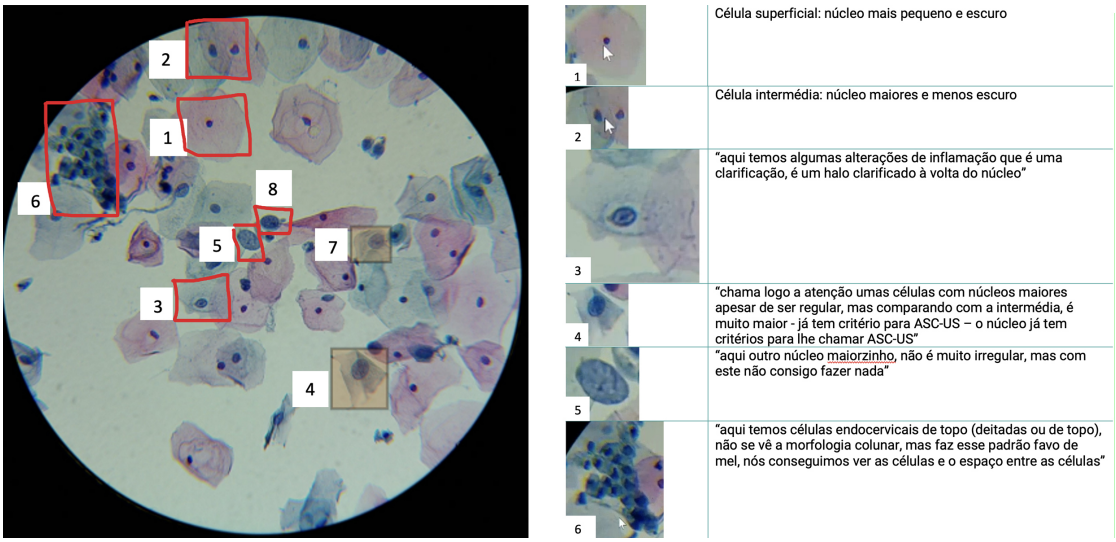


Figure 4: Word template setup for transcription aiming at mapping visual content - cells, or other structures - with textual excerpts, while keeping the order of analysis. The process included annotating on the image the object/area analysed and associating the number that corresponds to its order of analysis by the participant.

tab to list the criteria for each category as they emerged throughout the process. We ended up gathering a list of options that enabled us to streamline the filling-in process. To avoid subjectivity and/or interpretation errors in the process of codification, we organised an internal panel of three coders composed of researchers involved in these activities. All transcriptions were assigned to this panel, varying who would be the first coder. While the first coder would codify the transcription from scratch, the following two would validate the first codification. Taking the following quote as an example from Cervical cancer, we would describe as table 2 shows.

> *It has a darker nucleus, but with this resolution, when I try to zoom in, I can't see the characteristics.*

Figure 6 shows the variability of decision criteria by category that was raised throughout the analysis.

By the end of the analysis, we uncovered the most relevant criteria used by experts in each medical field to analyse and explain their decisions. And we could standardise that most of the explanations followed this structure:

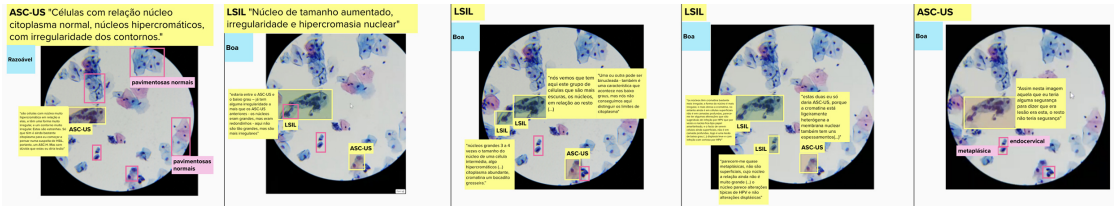
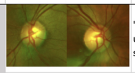


Figure 5: An example of the inter-observer comparative analysis carried out in the Cervical cancer study. The same microscopic field of view was analysed by 5 medical experts resulting in different annotations and classifications.

ID	Objeto de explicação	Explicação	Estrutura-chave examinada	Característica-chave examinada	Fator Potencial Glaucoma	Fator Não Glaucoma	Insatisfatório para análise	Fatores confundidores
001		"tenho **dois discos que me parecem corados**, que têm um anel que é regular, uma escavação que é pequena, que será igual ao inferior a 0,4, que é simétrica (...)	Discos	Cor		Corado		
002		"tenho dois discos que me parecem corados, **que têm um anel que é regular**, uma escavação que é pequena, que será igual ao inferior a 0,4, que é simétrica (...)	Anel neurorretiniano	Forma/Contorno				
003		"tenho dois discos que me parecem corados, que têm um anel que é regular, uma **escavação que é pequena, que será igual ao inferior a 0,4**, que é simétrica (...)	Discos	Geometria				
004		"tenho dois discos que me parecem corados, que têm um anel que é regular, uma **escavação** que é pequena, que será igual ao inferior a 0,4, **que é simétrica** (...)	Discos	Simetria OD/OE				
		"apesar de uma **escavação 05 não ser uma escavação obrigatoriamente patológica, o facto de um dos olhos não ter nenhuma escavação e outro ter uma escavação já considerável é muito sugestivo de Glaucoma** (...)	Discos	Geometria	Assimetria OD/OE			
005		"**este vaso faz uma curva** (A) e este também um bocadinho (B) (...) isto **já é um bocadinho mais suspeito** (...) este par de olhos é o menos suspeito de todos, apesar **desta deflexão aqui poder ter algum significado**, mas depois para isso é que vamos pedir os exames"	Vasos	Trajetória	Com deflexão ("joelhos")			

Corado
Escavação simétrica -> fisiológica
Escavação < 0.4/0.5
Normal
ISNT preservado
Regular
Simetria (anel)
Atrofia peripapilar
Sem deflexão (retilíneo)
2. Alteraçoes por medição em zona menos adequada
3. Simetria inter-olhos
3. Sem escotomas

Figure 6: Screenshot of the explanations' systematisation in Excel for Glaucoma. In the columns' headings, we may read the categories, left to right: Explanation object, Explanation, Key structure examined, Key feature examined, Glaucoma potential factor, Not Glaucoma factor, Unsatisfactory for analysis, and Confounding factors. Each category was fed according to the criteria identified in the explanations given by the medical experts.

The **[Key feature concerned]** of the **[Key structure(s) examined]** is **[Risk factor]** OR **[Not Cervical cancer/Glaucoma]**.

> *e.g. Cervical cancer: The **[colour]** of the **[nucleus]** is **[hyperchromatic]**. Glaucoma: The **[optic disc]** has an **[excavation greater than 0.4]**.*

Moreover, we found that sentences stating a "not Cervical cancer/Glaucoma factor" or "doubt factor" could follow the Key feature concerned. Experts used them to suggest a plausible contradiction that prevented them from providing a classification of which they were confident.

> *e.g., Cervical cancer: The **[colour]** of the **[nucleus]** is **[hyperchromatic]**, however, **[there are overlapping cells]**. Glaucoma: The **[optic disc]** has an **[excavation greater than 0.4]**, however, **[is symmetric]**.*

Table 2
Explanation categorisation example

Key area of image examined	*It has a darker nucleous* (Part of a cell)
Key structure(s) examined	*nucleous* (Nucleous)
Key feature(s) concerned	*a darker nucleous* (Colour intensity)
Risk factor	*a darker nucleous* (Hyperchromasia)
Not Cervical cancer / Glaucoma factor	Not applicable
Doubt factor	*but with this resolution,... I can't see the characteristics* (Image quality - Blurred)
Assigned result	*... I can't see the characteristics* Insufficient / No classification

In these explanations, the experts point out a structure that he/she observed and characterise its aspect – reflecting a well-known and established risk factor in the domain knowledge, i.e., Cervical cytology: [hyperchromatic], Glaucoma: [excavation greater than 0.4]. Nevertheless, the explanations also stress - through the contrastive expression 'however' - other characteristics that complement and contrast the first ones, i.e., Cervical cytology: [there are overlapping cells], Glaucoma: [is symmetric]. And this prevents the experts from discerning with certainty whether the first observed characteristic is an anomaly or not.

4.3. Workshops for validation

Based on the results of previous user research activities, researchers designed validation workshops to: (i) ensure no conflicting information among the knowledge shared by each participant, (ii) remove possible imprecision from researchers' interpretation and consequent analysis outcomes, and (iii) get insights on a first version of the graphical user interface (GUI) designed from scratch to attend the elicited diagnostic processes.

4.3.1. Conducting group workshops

The first validation session was carried out through the Mural platform, from where participants accessed and interacted (by editing, deleting, or adding content) with the list of decision-making criteria raised so far in order to ensure their correctness and completeness. In Glaucoma workshops, participants were also asked to analyse several examinations, mainly retinal images, and to choose the applicable criteria for each one from the list elicited by researchers. We asked participants to position the selected criteria in one of three possibilities: non-Glaucoma, Glaucoma, or borderline (Figure 7).

4.3.2. Validating content and container - an informed GUI prototype

In the second validation session, researchers presented the validated decision-making criteria integrated into a Graphical User Interface (GUI) prototype. The aim was to get feedback on the criteria and on the UI components presenting it. According to participants' availability, the Cervical cancer session took place in person (Figure 8), and the Glaucoma session took place remotely (Figure 9).

Some categories and criteria seemed to have more than one possible way to name or present in the interface. Thus, to assess the correctness and completeness of the data as well as the system's components and related features, we applied A/B testing for participants to choose the best options.

In Glaucoma study, we conducted a remote session through which we shared a PowerPoint presentation with images of the prototype of the GUI together with its content (the elicited decision-making criteria) listed in an editable text box, as shown in Figure 9. The content was discussed in real-time and, whenever necessary, easily edited.

5. Lessons Learned

L1: Multidisciplinary team The design of XAI-based clinical decision support tools requires extensive knowledge from various domains. It is paramount that teams ensure an iterative communication that keeps all in the loop, i.e., design researchers, medical experts, ML Engineers, etc. Let us highlight ML Engineers' guidance on the feasibility of the required functionalities, their support in defining the needed data, i.e., quantity and quality, and the infrastructure for implementation. Many systems based on supervised learning require annotated data, analysed by experts in terms of elements needed to guide the models' learning process. In the case of medical XAI systems, this requires close cooperation with clinical experts to ensure the annotation of the data instances objectively and uniformly. This way, ML Engineers guarantee that the final data set comprises cases sufficiently representative of the different data properties that may arise in practical scenarios.

L2. Contextual inquiry method as a basis for elicitation The contextual inquiry method inspired the study to observe experts performing a task as close to reality as possible by having them verbalise their thoughts while analysing imaging examinations and providing diagnostic classifications for them. We conclude that, when in-loco sessions are not possible, researchers can simulate the method remotely using digital and online platforms that enable to: video call, screen sharing, display relevant data for analysis and discussion, and freely write. We asked the experts for analysis materials from their daily work, e.g., anonymised imaging examinations, and then used the online platform Mural to present analysis tasks using these materials. While sharing the screen, experts analysed, selected, and annotated the digital images, and researchers asked timely questions that arose from observing what participants were doing and saying (think-aloud).

L3. Mapping text with images helped associate features to structures From the elicitation to the content analysis, we found it elementary to map the textual transcripts with the image that experts were analysing. We cropped, framed, and sketched over the images to correlate what experts were saying with what they were seeing. In doing this, some categories emerged transversely among both experts and images analysed, so this mapping led to discovering a standard structure of the explanations.

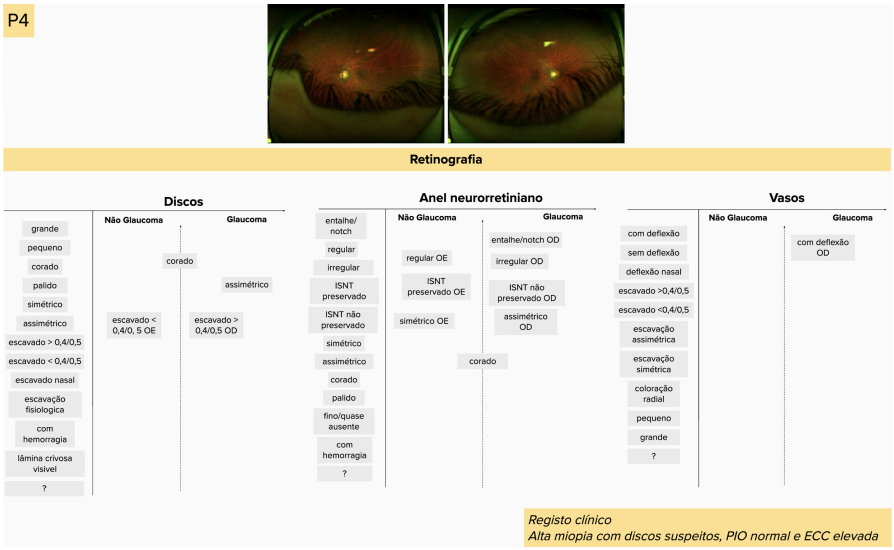


Figure 7: Screenshot of the first validation of the clinical decision criteria by the glaucomatologists. At the top, a magnification of a retinal image (right and left eye) centred on the optical disc. Below, 3 tables, one per eye structure: Discs, Neuroretinal ring, and Vessels. To the left of each table, there is the respective list of criteria from which participants were asked to select those observed in the retinal image and associate them to a non-Glaucoma, Glaucoma, or borderline diagnosis. The criteria positioned between the two columns would be considered borderline case criteria.

Figure 8: In-person workshop with three cytopathologists and one cytotechnologist to validate the decision-making criteria integrated into a GUI prototype.

L4. Categorisation matrix for multidisciplinary analysis As the categories emerged, we used Excel's functionalities, such as drop-down lists to streamline the process of matching features to structures facilitating the systematisation of the analysis across more team members, i.e., design researchers and ML engineers.

6. Conclusions and Future Work

This paper describes the user research activities carried out by a multidisciplinary team to inform the design of Machine Learning algorithms and user interfaces for two XAI-based computer-aided diagnostic systems for Cervical cancer and Glaucoma. We shared what we think might be useful for other teams involved in the design of Explainable AI systems, namely, ways to operationalise

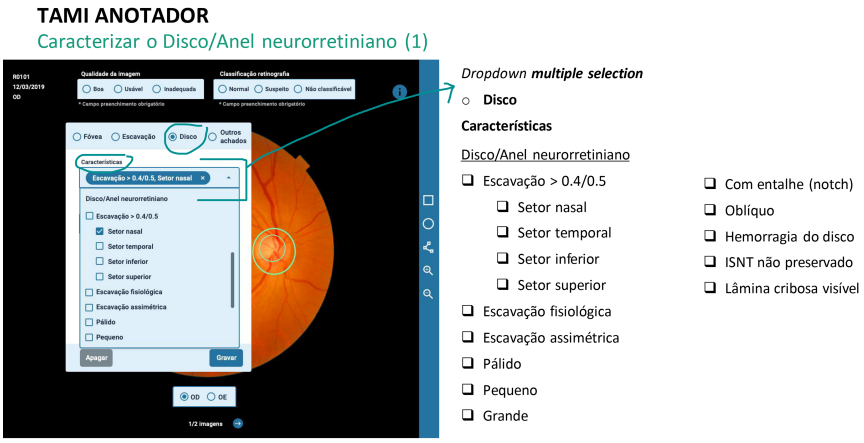


Figure 9: A PowerPoint slide showing on the right the decision-making criteria list regarding the optical disc, and on the left side a GUI prototype showing an annotated retinal image with an open dropdown menu component showing part of the decision-making criteria list.

human-centred design methods considering the objectives of Contextualisation, Elicitation, and Validation of such systems. In that scope, we demonstrate transcription, coding, and systematisation strategies that facilitated our content analysis, in particular, a categorisation matrix that helped uncover decision-making criteria and respective explanations' structure to inform the design of AI-generated explanations. Future work will focus on further developing the graphical user interface (GUI) to adapt it to an AI-based classification system to support experts' decision-making process.

Acknowledgments

We would like to thank the medical experts from the Anatomical Pathology Service of the Portuguese Oncology Institute - Porto (IPO-Porto) and from the University Hospital Centre of Porto (CHPorto), who participated in the user research sessions. A special thanks to our senior colleagues at Fraunhofer Portugal AICOS, Ana Barros and Francisco Nunes, who mentored us during the writing of the article. Finally, this work was financially supported by the project Transparent Artificial Medical Intelligence (TAMI), co-funded by Portugal 2020 framed under the Operational Programme for Competitiveness and Internationalization (COMPETE 2020), Fundação para a Ciência and Technology (FCT), Carnegie Mellon University, and European Regional Development Fund under Grant 45905.

References

[1] F. K. Došilović, M. Brčić, N. Hlupić, Explainable artificial intelligence: A survey, in: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018, pp. 0210–0215. doi:10.23919/MIPRO.2018.8400040.

[2] J.-M. Fellous, G. Sapiro, A. Rossi, H. Mayberg, M. Ferrante, Explainable artificial intelligence for neuroscience: Behavioral neurostimulation, Frontiers in Neuroscience 13 (2019). URL: https://www.frontiersin.org/articles/10.3389/fnins.2019.01346. doi:10.3389/fnins.2019.01346.

[3] F. E. Commission, Ethics Guidelines for Trustworthy AI - FUTURIUM - European Commission, 2021. URL: https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html, [Online; accessed 13. Oct. 2022].

[4] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (xai), IEEE Access 6 (2018) 52138–52160. doi:10.1109/ACCESS.2018.2870052.

[5] N. Burkart, M. F. Huber, A Survey on the Explainability of Supervised Machine Learning, J. Artif. Intell. Res. 70 (2021) 245–317. doi:10.1613/jair.1.12228.

[6] Q. V. Liao, M. Pribić, J. Han, S. Miller, D. Sow, Question-driven design process for explainable ai user experiences, 2021. URL: https://arxiv.org/abs/2104.03483. doi:10.48550/ARXIV.2104.03483.

[7] P. Lopes, E. Silva, C. Braga, T. Oliveira, L. Rosado, Xai systems evaluation: A review of human and

computer-centred methods, Applied Sciences 12 (2022). URL: https://www.mdpi.com/2076-3417/12/19/9423. doi:10.3390/app12199423.

[8] P. N. Johnson-Laird, Mental models and human reasoning, Proc. Natl. Acad. Sci. U.S.A. 107 (2010) 18243–18250. doi:10.1073/pnas.1012933107.

[9] C. Rickheit, Gert; Habel, Mental Models in Discourse Processing and Reasoning, Elsevier Science B.V., Amsterdam, 1999. URL: https://books.google.pt/books?hl=pt-PT&lr=&id=96jBqz_ar8AC&oi=fnd&pg=PP1&dq=mental+models+versus+reasoning+process&ots=Ou3b1SOv77&sig=r5NouxMzR56klQTrokyvHSclJuQ&redir_esc=y#v=onepage&q=mental%20models%20versus%20reasoning%20process&f=false.

[10] Z. Liu, J. Stasko, Mental models, visual reasoning and interaction in information visualization: A top-down perspective, IEEE Transactions on Visualization and Computer Graphics 16 (2010) 999–1008. doi:10.1109/TVCG.2010.177.

[11] J. S. Holtrop, L. D. Scherer, D. D. Matlock, R. E. Glasgow, L. A. Green, The Importance of Mental Models in Implementation Science, Front. Public Health 9 (2021). doi:10.3389/fpubh.2021.680316.

[12] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, N. Shadbolt, 'it's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 1–14. URL: https://doi.org/10.1145/3173574.3173951. doi:10.1145/3173574.3173951.

[13] T. Kulesza, S. Stumpf, M. Burnett, W.-K. Wong, Y. Riche, T. Moore, I. Oberst, A. Shinsel, K. McIntosh, Explanatory debugging: Supporting end-user debugging of machine-learned programs, in: 2010 IEEE Symposium on Visual Languages and Human-Centric Computing, 2010, pp. 41–48. doi:10.1109/VLHCC.2010.15.

[14] S. Jalil, T. Myers, I. Atkinson, M. Soden, Complementing a Clinical Trial With Human-Computer Interaction: Patients' User Experience With Telehealth, JMIR Human Factors 6 (2019) e9481. doi:10.2196/humanfactors.9481.

[15] T. Dagdelen, Modernizing the User Interface of a Legacy System at the Swedish Police Authority : Collaborative Mental Model: A New Participatory Design Method, 2019. URL: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1366483&dswid=8599.

[16] D.-G. of Health of Portugal, Rastreio da retinopatia diabética - portal das normas clínicas, 2018. URL: https://normas.dgs.min-saude.pt/2018/09/13/rastreio-da-retinopatia-diabetica/.

[17] L. P. Aiello, I. Odia, A. R. Glassman, M. Melia, L. M. Jampol, N. M. Bressler, S. Kiss, P. S. Silva, C. C. Wykoff, J. K. Sun, D. R. C. R. Network, Comparison of Early Treatment Diabetic Retinopathy Study Standard 7-Field Imaging With Ultrawide-Field Imaging for Determining Severity of Diabetic Retinopathy, JAMA Ophthalmol. 137 (2019) 65–73. doi:10.1001/jamaophthalmol.2018.4982. arXiv:30347105.

[18] Eurocytology, Criteria for adequacy of a cervical cytology sample | Eurocytology, 2022. URL: https://www.eurocytology.eu/en/course/1142, [Online; accessed 13. Oct. 2022].

[19] S. Rêgo, M. Monteiro-Soares, M. Dutra-Medeiros, F. Soares, C. C. Dias, F. Nunes, Implementation and evaluation of a mobile retinal image acquisition system for screening diabetic retinopathy: Study protocol, Diabetology 3 (2022) 1–16. URL: https://www.mdpi.com/2673-4540/3/1/1. doi:10.3390/diabetology3010001.

[20] T. Conceição, C. Braga, L. Rosado, M. J. M. Vasconcelos, A Review of Computational Methods for Cervical Cells Segmentation and Abnormality Classification, Int. J. Mol. Sci. 20 (2019). doi:10.3390/ijms20205114.

[21] D. A. De Jesus, L. S. Brea, J. B. Breda, E. Fokkinga, V. Ederveen, N. Borren, A. Bekkers, M. Pircher, I. Stalmans, S. Klein, T. van Walsum, OCTA Multilayer and Multisector Peripapillary Microvascular Modeling for Diagnosing and Staging of Glaucoma, Trans. Vis. Sci. Tech. 9 (2020) 58. doi:10.1167/tvst.9.2.58.

[22] CLARE: Computer-aided cervical cancer screening, 2023. URL: https://www.aicos.fraunhofer.pt/en/our_work/projects/clare.html, [Online; accessed 15. Feb. 2023].

[23] E. Bentley, oTranscribe: A free web app to take the pain out of transcribing recorded interviews., 2023. URL: https://otranscribe.com/, [Online; accessed 15. Feb. 2023].