"That Would Have Been Bad": How Radiologists Interact with Voice User Interface Systems When Authoring Reports

RORY STUART CLARK, University of Bristol, UK
TOM OWEN, Swansea University, UK
MATT JONES, Swansea University, UK
MARTIN PORCHERON, Bold Insight, UK
PHILLIP WARDLE, National Imaging Academy Wales, UK
THOMAS MICIC, National Imaging Academy Wales, UK
BETHANY DELAHAYE, Swansea University, UK

This paper presents an exploration of how NHS Wales radiologists interact with Voice User Interface (VUI) systems and peripherals when authoring diagnostic reports. We conducted a laboratory study with 10 practicing clinical radiologists to investigate the ways in which radiologists utilise speech-based technology to construct, edit and proof their work by having them report on real-world anonymised medical studies on camera. A sample of the participants also participated in interviews in which their data was collaboratively analysed and examined to offer deeper insight into the realism and generalisability of our findings and conclusions. We conclude that better training should be given to radiologists on how VUI systems work, and further investigation should be carried out on the best ways to interact with Speech To Text systems in safety critical environments.

CCS Concepts: • **Human-centered computing** → **HCI design and evaluation methods**; **Laboratory experiments**; **Empirical studies in interaction design**; • **Applied computing** → *Health care information systems*; *Health informatics*.

Additional Key Words and Phrases: VUI, ethnography, ethnomethodology, radiology, speech, healthcare

ACM Reference Format:
Rory Stuart Clark, Tom Owen, Matt Jones, Martin Porcheron, Phillip Wardle, Thomas Micic, and Bethany Delahaye. 2025. "That Would Have Been Bad": How Radiologists Interact with Voice User Interface Systems When Authoring Reports. *Proc. ACM Hum.-Comput. Interact.* 1, 7, Article CSCW263 (November 2025), 23 pages. https://doi.org/10.1145/3757444

1 Introduction

Understandably, the primary focus when designing tools and systems for safety critical environments is that they offer the least risk to humans and the surrounding environment [73]. This often creates an emphasis on formal methods and mathematical analysis to remove as many technical faults as is possible [16]. However, this can be at the expense of considering usability, and human factors engineering is often an afterthought when evaluating the efficacy of a safety critical device – this can lead to a design disconnect in the "Work As Imagined" vs "Work As Done" space, where

Authors' Contact Information: Rory Stuart Clark, University of Bristol, Bristol, UK, rory.clark@bristol.ac.uk; Tom Owen, Swansea University, Swansea, UK, t.owen@swansea.ac.uk; Matt Jones, Swansea University, Swansea, UK, aways@acm.org; Martin Porcheron, Bold Insight, London, UK, martin@boldinsight.co.uk; Phillip Wardle, National Imaging Academy Wales, Pencoed, UK, phillip.wardle@wales.nhs.uk; Thomas Micic, National Imaging Academy Wales, Pencoed, UK, thomas.micic@wales.nhs.uk; Bethany Delahaye, Swansea University, Swansea, UK, b.c.f.delahaye@swansea.ac.uk.

ACM 2573-0142/2025/11-ARTCSCW263
https://doi.org/10.1145/3757444

devices are built with the assumption that they will be used perfectly every time, as opposed to taking into account the reality that human error permeates all aspects of workflow [12].

Specifically, tools in medical environments are often subject to evaluation and assessment based upon performance metrics in a laboratory or vacuum environment, as opposed to adopting a holistic perspective that takes into account how they would be implemented in-situ. This can lead to small usability errors that compound over time, often attributed to human error as opposed to a lack of pragmatic design considerations [31]. Radiology is a branch of diagnostic medicine highly reliant on digital systems to receive, analyse and comment on medical scans, with modern radiologists often separate from the patient through their entire interaction with their treatment [60]. Voicing their clinical opinions aloud has always been the method of choice for diagnostic clinicians, but modern medical settings have done away with the tape-recorder and transcriptionist in favour of using digital Speech To Text systems that allow for better integration into the modern digitised hospital workflow [14]. Radiologists in the [Redacted Country] have to interact with Voice User Interface (VUI) systems in order to provide proper clinical opinions to referring clinicians, and whilst these systems are all deemed to be "safe", there is little evaluative work on their usability and the ways in which they operate in-situ.

In this paper, we examine the ways in which a radiologist and Voice User Interface work together to author a diagnostic report by constructing a facsimile reporting office and providing genuine anonymised patient data, and recording the interaction process on video. We frame our research around the general question: *How do radiologists interact with the VUI systems at their disposal?* with supplementary questions exploring *How do interaction methods affect error rates when reporting?* and *How do users perceive the efficacy and accuracy of VUI systems?*.

We utilise an ethnographic and ethnomethodological perspective to investigate and analyse these interactions with a focus on the "mundane" aspects that make up observable behaviours that are otherwise difficult to obtain through other methods of researching practitioners [32]. Following this, we adopt a collaborative approach, involving participants in the analytic stages as a means of ensuring an accurate representation of the reporting process and gathering domain expert opinion on the efficacy of the tools examined. Our video analysis and collaborative discussion with participants come together to construct a vignette of how a radiologist behaves in a quasi-naturalised environment when constructing their medical opinions, with domain expert insight into common failures of technology and the mitigation techniques that have been adopted to cope with them. We offer practitioner's perspectives on the state and usability of VUI systems in radiology, as well as our own commentary on how systems and training can be better suited to the practical aspects of "being a radiologist".

We present an image of a radiologist who is no longer simply a doctor that reports on medical studies - the introduction of VUIs into their environment means they are now responsible for the whole transcription and editing of their work. The practitioner has to work with black-box systems that suffer frequent mis-transcriptions and require them to split their thinking between the high cognitive load of applying their medical expertise to a patient's data and the difficult task of proof-reading complex plain text. The introduction of computing technologies has changed the nature of the practitioner's relationship to their work, but the job description and training remains the same.The VUI implemented to replace the medical typist has increased the workload and scope of the radiologist, but has not been designed or deployed with this in mind. We believe the ways in

which speech-based text input is perceived in healthcare environments should be re-evaluated by both designed and educators.

1.1 Motivation

An example: In their office, radiologist has to examine a Computed Tomography scan of a patient's thorax on a vertical computer monitor. The scan comprises of hundreds of individual images compiled together into a slide deck that the practitioner scrolls through, back and forth to give the illusion of movement *through* the patient – as they are studying the image, they are speaking aloud their clinical findings into a handheld microphone-controller connected to a separate computer monitor with a text box, transcribing their words in real time. Reporting on the scan takes over 4 continuous minutes of concentration and description. Upon finishing the report, they spend roughly a minute proof-reading the text on the screen, correcting 7 mis-transcribed words or grammatical errors. One of these errors is critical; the transcription software has failed to register the word *no* when discussing the presence of lymphadenopathy, a disease of the lymph nodes that could be a sign of anything from mild infection to a serious condition such as cancer. A reported presence of this condition could send a patient for additional and unnecessary tests. Audibly, the radiologist groans *"That would have been bad, wouldn't it"*, highlights the sentence, and verbally inserts the negative determinism into the text.

Had the mistake been left uncorrected, it would be almost impossible to determine who's responsibility the error was without the presence of a camera, human or system. Luckily, this case occurred in our laboratory study and so no patients were at risk, but it highlights the real presence of potentially sentinel events in the reporting process – errors such as these occur with high frequency in the radiologist's office, and they have become adept at proof reading and editing their reports to ensure that they are as accurate as possible before they are sent to the referring clinician [5]. The perpetual risk of a small mistake such as this means that they must focus equally on free-form text as they do on the complex medical studies they are responsible for analysing. In the past, radiologists would have access to medical secretaries and trained typists that would transcribe their work for them, but the introduction of computer-powered automatic Speech To Text systems in the end of the 20thcentury has meant heavier emphasis on the radiologist to be responsible for the authoring and editing of their work.

The tools and systems involved in the reporting process have to be heavily scrutinised for their safety and reliability from governing bodies before they are able to be implemented, but are often not evaluated post-implementation in-situ. Despite some ethnographic study on how work is done in-situ [19, 24] there is little camera based study on how radiologists interact with these systems in their day-to-day, despite their importance to patient safety and correct treatment. Devices in healthcare environments are typically not subjected to usability analysis until there has been a significant failure that has resulted in patient harm, and "invisible errors" can compound quickly [31]

2 Background

2.1 Safety Critical Environments and VUIs

Evaluating the efficacy of speech-based text input compared to traditional hardware (such as keyboards) has been a mainstay of the HCI world since the introduction of speech recognition technology. As early as the 1990s work in HCI has argued that keyboard based entry "felt more

natural", was markedly quicker and more accurate than speech-to-text based input [47]. Demonstrating the longevity of the area of research, but displaying a significant departure from this work, a 2017 study by Ruan found speech recognition is now almost 200 percent faster at transcribing short messages on a smartphone than keyboard entry, with the accuracy gap rapidly closing [69]. This research does not necessarily deal with the direct comparison between different methods of data entry, but these book-ended papers demonstrate the continued interest in the effectiveness of speech recognition, and its subsequent advancement over the last 30 years. The cost-benefit of speech to text software is well-trodden ground in the realm of machine learning and HCI – in safety critical environments such as cars, cockpits, construction and healthcare, it has been recognised as an efficient and powerful method of data entry for over 30 years [46]. The common conclusion, inclusive of clinical diagnostics, is that whilst speech recognition systems are more error-prone, the speed and cost efficiency mean that it is more often than not the preferable option [3]. This, however, does not tell the complete story: a considerable amount of work has been done on the efficiency and productivity elements of speech recognition in radiology, but little work has been done with a specific emphasis on cause and mitigation in clinical and radiological reporting [10, 44, 68].

Due to its relative ubiquity, utilisation of speech recognition and VUIs in medicine is a common topic of research. MedSpeak was introduced in the late 1990s as an alternative to the traditional recording-to-typist route that was costly and time consuming [49]. There was some reluctance amongst clinicians to view SR technology as the new normal, but it was categorically a quicker method of writing reports; emphasis was placed on the speed and efficiency of SR systems compared to medical typists, despite the association to a higher error-rate [55]. More recently, there have been multiple investigative literature reviews that collate existing research into speech-to-text as a method of data entry compared to hand transcription, with a consensus that speech based systems are faster and cheaper, but more error prone [3, 45]. In a meta-analysis of speech recognition in medicine, it was found that 39% (the individual highest group) of all papers published had a focus on productivity, with few placing importance on *cause* of error [10]. Studies that did investigate causal relationships between error and speech recognition suggest recency bias and length of reports may affect how well a practitioner is able to proof read their own work [7, 58]. Addressing these causes has resulted in some papers offering basic heuristic requirements, but it is clear from the research conclusion (and the fact that they are both over 10 years old) that existing systems do not meet sufficient standards, and previous work acknowledges the difficulty in getting large 3rd party developers to acquiesce to requirements [71, 75]. An opinion piece on "error in radiology" emphasised the importance of acknowledging that reporters and reports will not be perfect, and mistakes to not equate to negligence but strategies to mitigate mistakes must be implemented; amongst those discussed, meta-awareness of the systems involved and regular inquiries into quality control should be made - these systems should be repeatedly examined for the effectiveness and accuracy in-situ [17].

2.2 Healthcare and Ethnography in CSCW

Ethnography is a common method of assessing realistic *Work As Done* as opposed to *Work As Imagined* [38, 39] with regards to interaction with devices in safety critical environments, with applications in technical manufacturing, sociology, communication design and training of domain experts [9, 23, 56, 57]. Ethnographic methods of investigating interactions have been a common aspect of CSCW research, as it provides a solid method of allowing designers to understand the tacit knowledge of experts [13, 26, 41]. Utilising ethnography to supplement technical design has been described as enabling ontological convergence [4, 53], and healthcare presents itself as a complex

environment that can be relatively inaccessible to outsiders, relying heavily on years of training and practice to reach competency, meaning it is a safety critical environment that lends itself well to observational and dialogue-based methods of study and analysis [9, 34, 66].

Specifically, CSCW-based ethnographic research into health often examines the holistic relationship between the practitioner and their environment. Examinations into temporal and spatial organisations in hospitals has revealed the need for digital interventions to follow the natural "rhythm" of medical work [66, 67], with similar work outlining the necessity to understand existing social structures between clinicians and nurses or clinicians and their patients before beginning to suggest changes [29, 40].

2.2.1 "Invisible" Work. An increasing area of focus in this domain is the emphasis on examining so-called "invisibility", referring to the aspects of clinical work that are overlooked, but are essential to the proper running of the hospital environment. This can come in the form of studying often-marginalised aspects of the patient treatment process such as orderlies and patient transfer [1, 74], or by investigating how unnoticed or minor errors and mistakes can compound into serious sentinel events [11, 31]. The need to understand Work As Done as opposed to Work As Imagined [12] seeks to highlight the practical details of conducting work in a healthcare environment, and has brought attention on the ways in which peripherals such as paper based records [6, 15, 40] or infusion pumps [2, 42, 70] are used in a real-world setting and how it may differ from pre-conceived notions of "proper use".

2.2.2 Diagnostic Medicine. Specific focus on diagnostic medicine (especially reporting radiology) is less prevalent in CSCW than other disciplines, but some work has been done on the interactions between diagnostic clinicians and digital systems; Briedis' sociological and phenomenological endeavours provide an overview of how American radiology offices are laid out, and the hardware and software found within [19, 20], and Clark et al's work on how NHS radiologists communicate provides supplemental information on how practitioners utilise technology when analysing medical studies [24]. Similarly, Hartswood et al conducted work on British radiologists examining how paper-based annotations of digital documents provide an insight into hybrid documentation should be implemented in-situ [36].

Radiologists themselves provide the most detail on digital interactions and their perspectives on it, with pieces on implementation of systems ranging from overtly positive on the possibilities of autonomous and computer-aided reporting [14, 35, 52] to pessimistic about the need for training and tendency to create errors [54, 61, 76], with inconsistent recommendations over best practice and most viable options for the future.

2.3 Summary

In the literature review, we have demonstrated that placing an ethnomethodological lens on healthcare practice can often reveal aspects of behaviour and practice that are not considered or known by designers. We have also demonstrated research with VUIs in safety critical environments has been thoroughly examined with general recognised trends of accuracy and efficiency in clinical environments. There has not, however, been a dedicated examination study of of VUI *interaction techniques* in a diagnostic setting that explores the underlying reasons behind supposed accuracy or efficiency - by examining how practitioners interact with VUI systems in a quasi-in-the-wild environment, we can begin to offer more realistic and suitable guidance that helps to bed statistical and quantitative data in a contextual understanding. By studying how diagnostic clinicians interact with VUI systems when authoring reports, we provide a different perspective on where the VUI fits in terms of clinical workflow.

3 Method

Ethical approval for this research was provided by Swansea University Faculty of Science and Engineering Board review, and by NHS Research Ethics Committee. A note on terminology – *"medical studies"* refers to the scan or collection of scans of a patient. This can refer to a set of X-Rays or an MRI and CT slide deck. In order to preserve the terminology used when explaining the study to participants, we have chosen to describe these scans as studies as radiologists would.

3.1 Approach

We adopted an ethnomethodological approach for this study, where our intention was to focus on the "mundane" aspects of practitioner behaviour that would be difficult for experienced clinicians to identify due to their experience and familiarity with the systems and tools in [32]. This emphasis on the observable aspects of routine behaviour allows us to reconstruct events and activities that are otherwise inaccessible to non-members of the clinical community. Our aim is not to rigidly codify all instances and methods of interaction, but to instead illustrate the otherwise tacit "common sense" for analysis and exploration.

In addition to this ethnomethodological approach, we also employed a direct "fact-checking" method of collaborative analysis [50]. The research team behind this study had a level of vulgar competency from previous investigations that allowed them to understand the basic interactions and terminologies in play surrounding radiology, but to avoid extrapolating information that was unique to our laboratory environment (as healthcare environments are distinctly dynamic) we invited participants to view their own reporting techniques and methods alongside researchers to offer explanations as to "why they did what they did" and how generalisable this behaviour was to a realistic reporting session. 4 participants took part in this collaborative exercise. This collaborative analysis was done after footage was initially reviewed and analysed by the authors, so was used as supplementary "colour-commentary" to add contextual information and weight to findings and conclusions.

3.2 Participants

Participants in this study were all active practitioners in radiology who work within NHS Wales. This means that they are qualified doctors who would regularly report on patients and were familiar with the systems and tools used in the study. 5 of the participants were Consultant Radiologists and the other 5 were Trainee Radiologists - trainees still report on real patients and have experience in the domain, but have not yet completed their radiological studies. Participants had a range of experience, with the least experienced having worked as a reporter for 3 years and the most having more than 25 years. 10 participants took part in the lab study, with 4 participants returning to collaboratively analyse their data. 3 of those who participated in the collaborative analysis were Consultants. All participants were recruited from the same health board, meaning they all used the same version of software that is replicated for this study.

Difficulties with active clinical practitioners as participants in research has been addressed frequently, and this was something that was taken into account when recruiting and conducting for our study [30, 64]. Furthermore, we refer to the ethnographic principle that human activities contain their own means of generalisation, and as such attach significance to our findings from these sessions [28].

Participant Number	Role	Plain Films Reported Upon	CT Scans Reported On	Took Part in Analysis
1	Consultant	19	1	Yes
2	Trainee	11	1	Yes
3	Consultant	13	1	No
4	Consultant	16	1	Yes
5	Consultant	17	1	No
6	Trainee	8	0	No
7	Trainee	19	1	No
8	Trainee	18	1	No
9	Consultant	9	1	Yes
10	Trainee	11	1	No

Table 1. Participant Demographic Information

3.3 Set Up

3.3.1 Room Set Up. The study operated out of National Imaging Academy Wales Campus, on a site that regularly admitted patients for medical evaluation and reporting. This meant that we were able to accurately recreate and deploy a reporting room. Participants were presented with a room that was a facsimile of their normal working environment: the room had a desktop computer, 2 computer monitors, a keyboard, mouse and dictaphone, as would be found in any standard reporting office [60]. The desktop computer was equipped to run the same Picture Archive Communication System (PACS) and dictation software that would be found in the participant's standard reporting environment. We then set our research equipment up around these devices as so to mould ourselves *into* the space as opposed to influencing it. We set up 2 cameras - a webcam on top of a computer monitor facing the participants directly and connected to a locally recorded Zoom call, and another DSLR camera placed over-the-shoulder behind the participant at a 45 degree angle to see both reporter and workstation. These two angles together allowed us to see both the focus of the reporter and how they utilised the dictaphone face-on, and how they interacted with their workspace and screen content. In addition to these camera angles, we utilised a screen recording program that allowed us to examine how a report was made word-by-word in real time for deeper analysis. This screen recording program provided an additional benefit of allowing us to take a frame of the finished report for analysis before the participant moved on to the next report - due to ethical and technical limitations we were not able to export the reports once they were finished, but we were able to pause the screen recording at the moment at which a report was finished to view it in its completed state.

3.3.2 Tools and Software. Participants used an edited version of NUANCE, a Speech To Text software package that is specifically trained to learn and adapt to radiological lexicon, paired with an anonymised set of medical studies on a PACS built by Fujifilm, designed to store and organise health data and records. In a real-world situation, radiologists are assigned individual medical studies and the results are exported directly to the referring clinician who requested the original scan – this means that the standard digital workflow does not allow for multiple practitioners to be assigned to the same set of studies. Here, we wanted to have all participants examine the same set of medical studies to eliminate unwanted variables, but also wanted to preserve as much of the existing and familiar workflow as possible.

As such, the technical manager of the site adapted the NUANCE software to only operate locally using a collection of downloaded, anonymised studies. This allowed all participants to work with the same corpus of medical studies on the software found in the real world, but with the caveat that they could not be exported to the cloud-based storage system, as we were operating outside of

Fig. 1. Study Setup of Office, with Dictaphone Highlighted; Monitors Show Reporting Window on Left and PACS on Right

the genuine reporting workflow. The only differences participants would experience would be that all patient history and identifiable information was redacted, and they would have to manually erase their reports once finished and select a new study, as opposed to simply exporting and being automatically assigned a study to analyse. By having screen recording software running, we were able to see the full report before it was manually erased and examine it after the session had finished.

The technical manager would explain these differences to participants, providing them with a visual aide in the form of a graphic print-out, and would stay with them whilst they familiarised themselves with the system to answer questions. We would not commence recording until the participant verbally confirmed they were comfortable using this adapted software.

3.4 Sample Reports and Procedure

When participants entered our mock reporting set-up, they would be provided with consent and information documentation, as well as a quick briefing on how our system differed from the one that they were familiar with using (such as being locked down with some reduced functionality). The briefing was provided by a technical manager of the site who would also be on hand to assist with any failures or difficulties that participants experienced.

On the Picture Archive Communications System, we had pre-loaded a collection of medical studies from real patients that had been anonymised, with the patient data redacted and any other identifying features removed. This provided us with genuine and appropriate data for reporting, and allowed for us to set up the PACS and Speech-To-Text platforms properly. The majority of the sample medical studies were Chest and Abdomen X-Rays, comprising of a single still X-Ray image of a patient's lungs and ribs, with 3 CT scans also available for reporting on. The reason for fewer CT scans than X-Rays was practical – when discussing the set up with the clinically-based authors, we came to the conclusion that it would be best to see as many individual cases of interaction in a session as possible, and as CT scans can take upwards of 5 minutes in some more complex cases,

we decided that it would be best to encourage participants to review both a series X-Rays and at least one CT scan to provide as much data as possible.

A session would last for 30 minutes, as literature on radiologists suggests that this comprises of, on average, the longest a reporter would sit and continually report on studies with no breaks or interruptions [24]. Participants would be invited to sit and report in the manner that felt most natural to them, whilst a member of the research team would sit away from them, out of their field of vision, to ensure that the cameras and software operated correctly and throughout the entire session. Participants would be encouraged to report as closely as they could to a normal work session, and repeatedly informed that we were not surveying the accuracy or quality of their diagnostic opinions and abilities.

Once a session had finished, participants would be invited to be part of a collaborative analysis exercise either in person or remotely to view and analyse their own data. We would collect contact details and organise a 1:1 session with participants that agreed to take part – this collaborative exercise comprised of viewing several fragments of video that contained different elements of interaction, with the participant offering commentary and contextual deconstruction of what each fragment contained.

4 Analysis

Video analysis took the form of 3 key passes: initial viewing, analytic viewing, and collaborative viewing. Video data would be compiled in editing software to show both camera angles simultaneously, and an initial viewing of each session would be conducted by two members of the research team independently to orient themselves with the nature of the data. Rough, unstructured notes would be taken on interaction methods, areas of friction and errors where both human and technology were responsible. Discussion would then take place on areas to focus on, and an analytic viewing would take place that catalogued errors and mitigation techniques, time taken to proof and correct reports, and detailed notes to reconstruct the interaction process [37].

Further discussion would then take place on the most representative fragments to review with participants that wished to take part in the collaborative analysis. The fragments chosen would be an attempt to recreate the full 30 minute session in small, manageable vignettes [27, 63] – the emphasis would be placed on aspects of behavior and interaction that occurred frequently and made up intrinsic parts of how work was done by that particular participant [72].

During the collaborative analysis sessions, between 5 and 7 fragment clips ranging from 15 seconds to 2 minutes comprising of an *event* would be shown to participants. Examples of these fragments included a participant reporting on a study without having pressed the record button, a participant correcting the term *Atelectasis* from a mis-transcription by the software of *8 Electricians* and a participant editing a CT report in full using only keyboard as opposed to the dictaphone (as was standard practice). The video would be played several times for participants, and unstructured discussion on the nature of the event, how frequently it may occur in standard daily workflow, and in occasions of error or failure what they believe the cause and solution to be. Discussion would be prompted and stimulated by open ended questions by the researcher, such as *"Would you like to explain what you think happened there?", "Is that a standard occurrence in day to day practice"*, and *"How do you think you would go about accommodating for that?"*. Due to scheduling restrictions from participants, these sessions lasted roughly 30 to 45 minutes all took place over video conferencing, and were recorded.

The resulting transcription was subjected to basic thematic analysis, based upon the Braun and Clarke method [18]. Transcripts would be read in full by two authors separately, before open coding was conducted to examine participant's perspectives and explanations for patterns of behaviour. Discussion would then take place on these open codes with a view across all of the sessions, before further coding was conducted to codify patterns into recurring themes. A second and final round of discussion was held to extract and agree upon key quotes and align participant discussion themes with patterns and themes identified in video fragments to construct a multidimensional image. Whilst not the primary focus of analysis, these transcripts will be the source for commentary and quotations throughout our results and discussion sections.

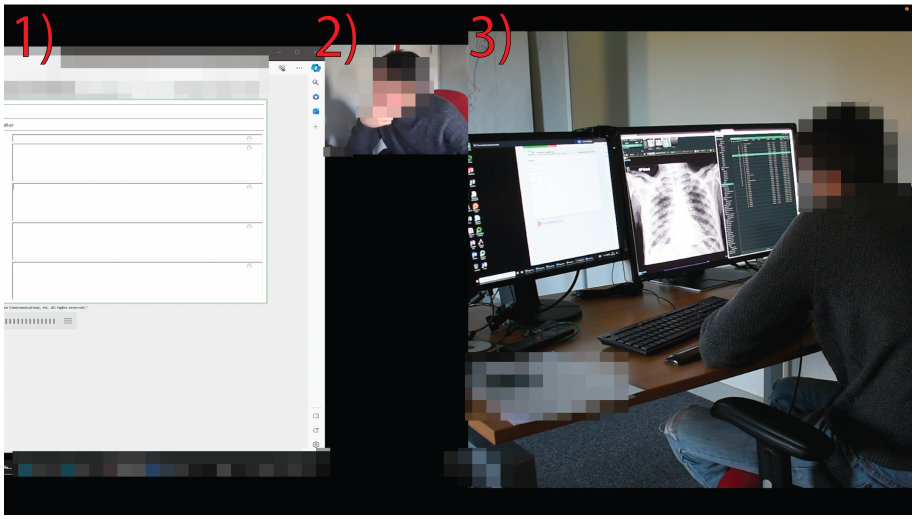

Fig. 2. Compiled View Of Data For Analysis Showing: 1) Reporting Window; 2) Participant Front On View; 3) Participant Over-Shoulder View

5 Results

Here, we present key findings from our analysis of the video data with added depth and commentary from participant's own analysis. Whilst we will sometimes refer to numeric data, this is not for statistical analysis but instead to add context and illustration to the "shape" of the data collected. Throughout, we present vignettes from the video data reconstructed in "thick description" style [33].

5.1 Examining Reporting Methods and the Interaction Process

Despite all participants having different preferences with reporting style and composition (some preferred to keep their reports short, only reporting actionable findings, whereas others commented on all aspects of the medical study regardless of importance to the referring clinician), the interaction *methods* and behaviours were consistent across the participant pool.

When reporting, participants would hold the dictaphone in one hand, with a mouse in the other to control parameters of the PACS system and manipulate the scan for better analysis. Whilst the dictaphone was in-hand through virtually all of the reporting time only put down when keyboard-based interactions required use of both hands, when participants were audibly recording their diagnostic

P	User Errors	System Errors	Unfixed Errors
1	21	7	5
2	10	11	0
3	7	42	1
4	8	4	3
5	8	8	1
6	19	4	1
7	17	9	0
8	5	18	10
9	12	8	1
10	3	4	3

Table 2. Summary of Instances of Error Each Participant Experienced Across Reporting Session

findings, it would be brought in extremely close proximity to their mouth and their speech would become quieter and more monotone to enter a "reporting mode" - by this, we mean the inflections of speech no longer carry specific emphasis on certain words, and instead all sentences are spoken with equal emphasis on every word. This mirrors the anecdotal findings of Clark [24]. The VUI would be triggered into recording mode with a button on the dictaphone, and participants would self-gate their own input by continually triggering the record function to speak, then turning it off when they had finished.

They are sat in front of two screens, one of which contains a medical study and patient information and the other contains the reporting window, comprised of a series of rectangular text boxes with headings such as *Clinical Findings*, *Impression* and *Plan*, as well as a volume meter at the bottom of the screen colour-coordinated to the level of input of the microphone - appropriate volume of talking is signalled with a green meter, whereas too quiet or too loud is signalled by yellow or (in terms of input that leads to a peak) a red meter. Participants offered that the talking with a yellow or red meter may lead to more instances of mis-transcription, but there was a level of uncertainty in this claim, which will relate to further discussion points later in this paper.

The reporting window is hosted on a web page, with white boxes on a gray background and black Calibri font text appearing on screen. In turn, as they are speaking, their words appear in the selected text box in real time word by word; if there is an extended period of speaking (i.e a sentence of 10 or more words) there will be a brief delay and the sentence will appear all at once in full. This meant that after a long sentence or full paragraph of speech, participants would often pause to check that the system had recorded everything, as the longer the continuous speech the longer the delay would be before the text appeared on-screen. This, in turn, took their concentration away from the medical study they were analysing. Due to having one hand on their mouse and one hand filled with the dictaphone, punctuation and organisation is done verbally with participants vocalising terms like *full stop, comma, colon* as well as *new line* or *new paragraph*. This means speech is not naturalised, with a participant verbalising *"Chest new line AP erect full stop the patient is mildly rotated full stop mild proximal thoracic scoliosis comma convex to the right full stop"*. This understandably led to common issues with the medical term "Colon" being confused for the punctuation of a colon, which occurred frequently. In addition, there were common instances of the software simply transcribing the punctuation as plain text, with no seeming explanation or cause as to why this mistake would occur.

These factors, such as actively having the volume of speech fed back to them and having to vocalise punctuation, mean that reporting on a medical study differs significantly from everyday conversation and even from traditional tape-based recording of medical observations. Participants talk into the dictaphone in a different way to interacting with another human being because there are several parameters that are impacted by the nature of casual and everyday speech, which could have an impact on the overall accuracy of their end reports.

5.2 Proofing and Errors

P	Corrections When Reporting	Corrections After Reporting	Keyboard Corrections	Microphone Corrections
1	21	5	6	20
2	10	10	3	17
3	6	42	23	25
4	7	5	3	9
5	10	6	2	14
6	7	14	8	15
7	16	8	7	17
8	16	7	3	20
9	4	16	18	2
10	4	3	2	5

Table 3. : Summary Of Correction Techniques and When They Are Employed Across Reporting Sessions

A key element of the interaction process is how participants proof-read, identified and corrected errors in the text to ensure that it was as accurate as possible. Here, we examine this behaviour, as well as errors on behalf of the participant - we define user error as Reason does in their work on human error as *the failure of a planned sequence of mental or physical activities to achieve its intended outcome when these failures cannot be attributed to chance* [65], whilst system error is defined simply as a malfunction that results in an unintended outcome. Table 2 provides a brief overview of these errors across the participant pool, demonstrating the variance on behalf of both users and the reporting software; also included is the amount of errors that a participant did not notice and correct before moving on to the next medical study to report, referred to as an *unfixed error*.

Common (occurring multiple times across different participants) examples of user errors we saw in sessions included not properly triggering the record function and reporting on a scan without their words being transcribed, pausing in the middle of a word resulting in a mis-transcription, and getting the orientation of a medical study wrong resulting in having to change "left" to "right" and similar. Examples of system error comprised of mis-transcription, where the VUI incorrectly identified the participant's words. The NUANCE system was trained on a radiological lexicon, so was familiar with medical language, but would still often get medical and common phrases or terms wrong and required frequent corrections.

We observed two distinct ways of participants identifying errors in the end report, regardless of human or system origin. The first was the *frequent glance* whereby participants would be continuously reading the most recent piece of information inputted into the system whilst reporting, and if they identified a mistake they would correct it instantly. Participants would typically be sat at an angle, facing with their body and eyes pointed towards the computer monitor with the PACS and medical study on it, and would shift their eyes over to the reporting window at the end of a sentence or when they took a pause mid-way through speaking. This method typically occurred

once a participant had already experienced a failure or setback such as not triggering the recording function and wasting time having to restart their report, or having to spend a considerable amount of time correcting one report, implying that the participant had lost "trust" that the system was performing accurately. Due to the personalised and unstructured nature of report writing, it is difficult to comment on whether this *frequent glance* affecting the time taken to finish a report, but it is noticeable that participants would employ it as a proofing method more often if they were having to correct and amend their reports frequently *"I spend a lot more time checking because I realized the reports are a lot more inaccurate"* (P1).

The second was the end proof read - whilst not all participants utilised the *frequent glance* to fully look for errors, every participant engaged in a final proof read of the finished report. This would comprise of readjusting their position in front of the monitors to more directly face the reporting window and placing the dictaphone down. Participants would then read the finished report start to end, correcting errors as and when they came across them. P9 expressed that they would *"rather speak concentrating on the scan in the knowledge that what's coming up on the screen isn't exactly what I want*". We will address specific correction techniques further in the paper, but it is of note that participants were aware and comfortable with the high number of errors that they would find in their finished reports, and seemed resigned to the idea that they were not going to the able to catch every mistake "*It's something you're acutely aware of every time you report*" (P1); *"I know I'm going to have to delete things anyway"* (P2). Of note is the tendency for participants to prefer utilising one method - only a single participant (P2) corrected as many errors *during* a report as *after* it was finished, with the majority of practitioners heavily favouring one or the other. There was not, however, a cohort wide preference for which method was more heavily used, as the end results showed an average of 10.1 corrections made *during* reports and 11.6 corrections made *after* the report was finished across all 10 sessions.

It is also worth noting that many participants left their reports with visible errors left unfixed, that were identified in the analytic stages of reviewing footage; 8 participants had at least one unfixed error, with the highest being P8 with 10 instances. These unfixed errors were identified in the initial video coding analysis sessions – when the finished report was about to be erased, the screen recording would be paused and read through by the researcher. Unfixed errors ranged from spelling mistakes and grammatical errors through to incorrect words, such as *mildly* being left in the report although *markedly* had been said.

5.2.1 Correction Techniques. We saw two key methods of correcting errors found in reports - using the dictaphone to audibly repeat the intended word or sentence (*re-vocalising*), and using the keyboard and mouse to instead erase the incorrect term and replace it. Table 3 provides an overview of participant's preferences on when to conduct a correction and the method most preferred - it is interesting to note that, on the whole, participant's displayed a preference for one method of identifying and correcting errors to another; only P2 corrected as many reports during as after, and we see that P3 has a balanced ratio of keyboard correction to verbal ones. The process of using the dictaphone comprised of the user noticing the error, double clicking on the incorrect word or sentence with the mouse to highlight it, and re-vocalising the correct term - this would replace the highlighted text and return the cursor to the end of the new input. This method was most common way of correcting errors, but was prone to failure along the process; on several occasions, participants would have to repeat the correct word several times, as the system would continually mis-transcribe the term. This meant continually having to re-click and highlight, re-vocalise and examine to ensure the sentence was correct. Additionally, due to the fact that re-vocalising would return the cursor to the end of the most recent input, there would be cases of participants forgetting

to return the cursor to the end of the report and continue reporting in the middle of an unrelated sentence. This would take time to erase the misplaced text and manually move the cursor to where was intended.

Keyboard-based correction followed a more commonly found method of correcting text-based errors, where participants would use the mouse to place the cursor after the mis-transcription, erase it, and type out by hand the intended term. Using the keyboard and mouse to correct errors was significantly less common, apart from P9 as seen in Table 3 - out of 20 corrections counted in their session, only 2 were made using the dictaphone re-vocalisation method, the rest being hand corrected with the keyboard. P9 was the only participant who utilised the keyboard more than the dictaphone that we observed, making them a significant outlier, but notable for their demonstrable preference for keyboard-based correction. Keyboard correction also resulted in fewer failures, as participants seemingly had more control over what the intended input was going to be.

Of note is that, whilst re-vocalisation was overwhelmingly the preferred method of correcting mistakes, keyboard based correction was seen as a "fail safe" - in most cases, if a participant attempted to correct an error with their dictaphone and it failed on more than two attempts, they would resort to hand-typing the intended term - "*If I just type it, I know it's going to be right*" (P9). The inference here is that despite re-vocalisation being the recognised method of correcting text errors in the reports, it was recognised that utilising a keyboard was more likely to render the intended result first time. This often also occurred with verbalised punctuation, as participants would vocalise instructions such as *new line* or *full stop* which would frequently be recorded as plain text as opposed to the intended punctuation - due to the fact that the correction was often a single key stroke, participants would more often simply type in a comma or return with the keyboard instead.

5.2.2 Types and Severity of Errors: No vs Known. Here, we wish to draw attention to an aspect of the proofing and correcting process identified by participants as of peak significance; the semantic nature of "errors". What is meant by this is how the *appropriateness* of the mistake affects the overall legibility of the sentence, and thus the likelihood it will be caught on a first-pass of the report. On occasions, the system would substitute a common medical term with a semantically unrelated word - a key example being "*8 electricians*" being transcribed when the user said "*atelectasis*", which happened several times to different participants across the sessions. That mis-transciption renders the sentence in the final report semantically illegible, and participants opined this is an easy error to find and correct in the proofing process. The dangerous errors, however, were when the system mis-transcribes the participant's words with a semantically appropriate term - a homo-phonically similar term that is still appropriate but incorrect. We saw this with cases of *mildly* being confused with *markedly*, *liver* with *lung* and *hypoinflated* with *hyperinflated*, and participants opined the most frequent and dangerous of these was *no* vs *known* - "*Its time consuming and also this patient may have completely different management based on that single word* [no] *missing from the report*" (P1). The presence alone of this mis-transcriptions is dangerous, as it can lead to inappropriate or even lack of treatment to a patient if not caught and corrected, but additionally it became clear in collaborative discussion that concern on participant's behalf was due to a lack of understanding and transparency between user and system as to *why* mis-transcription occurs, and *how* they can do their best to mitigate it.

Discussing the prevalence of these homo-phonic errors, participants often claimed it was their own fault for using terms that can be confused easily by the VUI system. When presented with

examples, participants suggested the best mitigation is for themselves to change their technique: *"Its possibly something I should train myself not to do"*(P4); *"Some consultants will say don't ever use that word… because it comes up wrong"*, (P2); "*I have a feeling if I if I managed to get in the habit of pronouncing it differently, it would be less problematic*" (P9). The implication of these comments is that practitioners believe the most sustainable method of mitigating these important semantic errors is for them to adjust to a system, as opposed to improving performance of the VUI.

6 Discussion

6.1 The Radiologist as an Auteur

The presence of unfixed errors in submitted reports even in an environment with no distractions and interruptions (which is often unrealistic for a clinician in a hospital [25, 61]) demonstrates how difficult it is for mistakes to be consistently spotted. In the vacuum left by implementing VUIs in the place of a medical typist, it is clear that the radiologist has taken on a role as their own transcriptionist and editor, and have to spend a considerable amount of their time editing and correcting both user and system errors, ultimately becoming solely responsible for all aspects of the diagnostic process, save for capturing the scan themselves. It also became increasingly clear from discussion with participants that this was *not* covered in their training, and that operating the VUI devices and systems was taken to be intuitive enough not to merit dedicated study and teaching. Previous work has highlighted that the primary benefit for implementing VUIs in medical settings is often to increase the practitioner's ability to multitask [51], but in this situation increasing workload appears to be detracting from the user's main focus.

This indicates the importance of improving both practice and technology, as both play a role in the time taken to complete a report on a patient's data - if the radiologist is to continue in their "auteur" role [1], it should be integrated into pedagogical practice how best to utilise an VUI system and good literary practice in proofing an editing. These would be expected of a dedicated typist or editor in other contexts, so a paradigm shift in perspective of the duties of a radiological practitioner should follow suit. Similarly, systems should be developed with this in mind; word processors and other dedicated safety critical tools are designed to highlight potential mistakes and dangers to the user, and radiological VUIs should be no different.

6.2 System Design

6.2.1 One Way Communication and Non-Collaboration. Of note when examining the ways in which practitioners interact with these VUIs is that they are unlike much of what is the current focus of deploying Machine Learning systems in clinical environments. The rise of Clinical Decision Support tools has resulted in a codified framework of what clinicians want to have and to see when interacting with machine learning in their work – common requests revolve around wanting reasoning behind choices made by Artificial Intelligence, in order to form a mental model of how the tool works in attempts to align it to a human colleague [21, 22] .

The system being examined here, however, does not allow for any two-way communication or collaboration. The practitioner has no choice to ask why the system may have mis-transcribed a word, and cannot explore the model's inner workings and preferences like they could with a medical typist. As such, these systems are not Conversational like a digital assistant, nor are they Decision Support as they offer no insight or opportunity for choice. Instead, we are presented with a black box at a crucial stage of the diagnostic and treatment process.

[1]A term used mostly in media industries to refer to an individual who excises complete creative control over a project

Similarly, participants were frustrated and often confused by the nature of how the systems worked - "*I can't figure out when it decided to do that… I don't know how it works*" (P1); "[transcription issues] *happen all the time… I don't know how to fix that*" (P2). The reason behind mis-transcription choices were unknown, with participants expressing that often mis-transcribed words had no semantic relationship to the rest of a sentence, with frustration or confusion over word choices: "*Nobody in their right mind would say that*" (P9); "*It's not as though it hasn't heard what I said*" (P2), demonstrating the black-box pitfalls of utilising such a simple GUI with VUIs – it was noted by several participants that they wanted to see some form of context or explanation as to why a word has been mis-transcribed, often so that they could understand how to mitigate it themselves. This may go some way to demonstrate the variances seen in Tables 2 and 3; participants may develop particular habits and preferences of identifying and correcting reports as a way to preserve high levels of accuracy in a system they know little about. As we have already established, medical practitioners often want to construct a mental model of how these systems work [22], and it is clear that style is an important element of how radiologists construct and communicate findings. In line with Cai, participants wanted to know (and had established beliefs about) when the system was likely to trip up over a long sentence or particularly hard to pronounce medical term, and wanted to be able to adjust accordingly.

In these instances of mis-transcription, participants often believed the way that they interacted with the microphone and dictation software to be at fault, as opposed to there being an issue with the system: *"Sometimes it does pick up the wrong words, and maybe that's when I've not said it clearly enough"* (P2); *"I would usually assume it's user error"* (P4). This reflects the findings of Furniss and Blandford that, when "unremarkable" or "invisible" errors occur, the blame will often be laid at the feet of the human [11, 31]. In other cases, participants claimed that external factors such as background noise or microphone choice "sometimes" affect the ways in which the dictaphone responds to them, but that settings are not available to them as users: "*I don't think there's enough input to optimise the system*" (P4); "*if I'm sat at my computer, where there is no one around, there's no background noise going on... I can sit there and report the whole report really and have very minimal changes*" (P1). It became clear during the collaborative analysis sessions that radiologists are not specifically taught how to interact with these systems, and instead develop interaction techniques through folk theories and "best guesses", leading to issues of digital literacy.

6.2.2 Visual Feedback. When addressing the current design of the Speech to Text system and associated areas of friction, feedback is a crucial aspect of interacting with any digital system, but is found wanting here. We see highlighting as a method of displaying errors, but in a way clearly intended for keyboard input over voice input – incorrectly spelled words are given a red underline, as one would find on a standard word processor, but this was often applied to specific medical terms that the NUANCE system recognised (as the word would be inputted appropriately) but clashed with the inbuilt dictionary. Practically, there is little need for an underlining of incorrectly spelled words, as the system should correctly spell the vocal input. As a result, participants by and large ignored any underlining. Highlighting of terms is recognised as an effective method of drawing attention when skim reading, but is applied inappropriately here – instead, implementing highlighting with contextual awareness (via an AI co-pilot or similar) could provide semantic assistance; highlighting words or sentences that deviate from the norm, such as actionable findings in an otherwise standard report, could work to draw the radiologist's eye to areas that require double-checking.

In a similar vein, the Speech to Text would often cause grammatical and layout areas with its assumptions of normal speech patterns: when reporting, participants would pause before confirming

certain medical findings. After an establishing phrase such as "heart size is…" a cursory glance would be given to the medical study. This pause would often lead to a full stop, an automatic function of the software, and the next words would be on a new sentence. As a counter measure, we would see some participants even trigger the recording off to examine the medical study in depth to avoid unwanted punctuation and breaks in their text. This is a small aspect, a so-called "unremarkable error" but one that nonetheless requires a correction, and brought frequent annoyance to participants. It again demonstrates that the system had not taken into account that dictation is an unnatural activity when compared to standard speech and conversation.

6.2.3 Reporting Preferences. When it came time to clean and analyse our data, it became clear that making overarching and sweeping changes to the ways in which VUIs are designed and implemented in these environments is challenging, due to the personalised and unique way that each participant reported on their medical studies. Whilst all participants had the same equipment and the same medical studies to report upon, the length and detail of finished reports varied, as did the location of actionable information. For example, P6 spent on average 120 seconds reporting on X-Ray plain films, whilst P1, P3 and P8 spent little over 30 seconds. Due to the smaller sample size of CT scans reported it is harder to extrapolate definite conclusions, but the longest time spent reporting on a CT was 435 seconds (P4) whereas the shortest was 120 (P3) where both participants were consultants. This variability in the data makes it hard to draw deterministic comments about stylistic trends. More relevant for designers, the ways in which the system and the user made mistakes was inconsistent throughout - as Table 2 shows; P3 experienced 41 individual separate cases of the Speech To Text software mis-transcribing their words, where the next highest was 18 (P8). P3 spoke fluent English with a South Asian accent, but was not the only participant with a non-Anglophone accent in the study – the requirement for systems to acknowledge participants without native accents has been highlighted since 2007, as demonstrated by an opinion piece from McGurk et al [54]. This is demonstrated in Table 3: the average number of transcription errors left uncorrected in a finished report was 2.5, with P2 and P7 not leaving a single report with a mistake in the text, whilst P8 left 10 reports with errors unfixed, a significant outlier and departure for practice.

The issue of style having an impact on digital interventions in radiology is not new – Wallis' 2011 paper highlights the concern medical practitioners have over voice recognition systems producing more errors for junior members of staff who have not developed their unique voice [76], and it has been established that automated and digitally structured reports are preferable to the clinician who has to read them, but radiologist resistance is a key factor in their lack of implementation [43, 48, 62]. It is apparent from these results that there needs to be higher levels of customisation and flexibility in these systems available to participants, for them to be able to tailor to their specific preferences.

6.3 Implications For Designers

This discussion segment has demonstrated the ways in which practitioners interact and perceive the digital systems at their disposal. It is also important to acknowledge how these factors impact the future development and design of tools in this space:

6.3.1 Radiologists are self-taught transcriptionists. Acknowledging that these clinicians have to conduct thorough proof reading and editing of their reports but that this is not something they are trained for has ramifications for the way in which VUI systems are presented to them. Whilst changes in training and pedagogical method would go some way to assist in this, it is unreasonable to expect the already intensive and time-consuming radiology training program to accommodate more elements into its syllabus.

Instead, non-invasive methods of assisting the clinician should be tested in-situ. These systems should ensure that they do not affect the clinician's existing workflow to a noticeable extent, a factor highlighted by participants in their own commentary, and so investigations should be undertaken to consider ways in which clinicians can receive support akin to a trained typist or transcriptionist assistant that does not interrupt their already time-sensitive duties. Systems should be contextually aware of the situation in which they are operating, as we have highlighted in our results that simply adapting methods employed for traditional word processors are not appropriate. For example, AI-based digital assistant systems are already found in other medical contexts [8, 22] and applying a trained and context-aware program in this situation is something that may provide a safety net for the difficult to spot but potentially serious errors discussed in this study.

6.3.2 Appropriate Visual Feedback. We have demonstrated that simply taking the methods of feedback found in traditional word processing systems is inappropriate in a context with such a high level of unique and specific language. Instead, much like with providing proofing assistance, visual feedback should have an emphasis on context-awareness, highlighting semantic errors over grammatical or spelling ones. The emphasis should not be on ensuring that a report has perfect legibility, and instead should be placed on ensuring that the finished report echoes the sentiment of the author. Participants indicated that overloading the author with information constantly throughout the reporting process would quickly lead to it being ignored, and so should be used sparingly. Feedback in the form of highlighting unusual findings or summarisations of paragraphs could go some way to allowing the participant to report and proof as normal, but again provides the safety net in checking the semantic meaning of the report remains consistent.

6.3.3 The Two-Way Relationship. Previous literature has highlighted the importance of allowing the clinician to form an understanding of the system they are "collaborating" with [21], but when working with VUIs we have shown that the relationship is very much "one-way" – the system itself is not offering a diagnostic or medical opinion, but that does not mean that it is not a collaborative element of the report writing process. We see from participants that they attempt to form a mental sketch of what each system does or doesn't "like" the user doing, but there is still considerable confusion over how the system makes certain transcription or formatting errors.

Allowing for more querying, personalisation and explainable decision making in what is transcribed would allow participants to get a firmer mental model of how their particular VUI system works, and as such form a stronger relationship with it. These systems were designed to replace a human colleague, but with a typist the radiologist has an opportunity to discuss the output and their personal preferences – with these VUI systems, there is no such chance. Further efforts and investigations should be undertaken into how to make diagnostic interfaces more collaboratively friendly even when they are not making a clinical decision.

7 Conclusion

7.1 Limitations and Future Work

We acknowledge that our participant recruitment is small, and whilst it represents a significant number of active practitioners in this field, (our pool of 10 is 4.4% of the population of the entire workforce in Wales [59]) this work is illustrative of the issues facing diagnostic clinicians in the workforce as opposed to an all-encompassing summary of the landscape [26]. This follows other examples of ethnomethodological study of clinicians in a similar field [19, 20, 24]. We previously

highlighted the recognised difficulties in having active clinicians as participants in research, especially non-clinical, and as such future work should look to replicate and expand upon our findings with a larger longitudinal study. It is also worth noting that this study took place in a country with a Nationalised healthcare system, and one that relies on practitioners interacting with VUIs to complete their work. This may not be representative of radiologists across all countries, and as such we have decided to focus our scope.

Participants also opined that it would be pertinent to repeat a study of this style with scheduled or organised interruptions; radiologists frequently experience interruptions in the form of phone calls, peers asking for advice or requests for referral, and the presence and ubiquity of these disturbances is likely to have an effect on the ways in which they carry out their work in-situ, and so it would be a worthwhile endeavour to examine how these interactions are changed by the presence of an interruption.

7.2 Conclusions

This paper has presented a reconstruction of the ways in which a radiologists interacts with a Voice User Interface and peripheral devices when authoring a report. We have demonstrated how the reporting process does not mirror naturalised speech due to the requirements of the systems involved, and how practitioners go about identifying and correcting mistakes in their plain text reports, some of which could have catastrophic ramifications if not spotted and fixed. These results present an image of a radiologist acting as the doctor, transcriptionist and proof editor of their work all whilst having to maintain a focus on a black-box system that they have only empirical knowledge of. Methodologically, we have managed to construct a study that offers insight that could be extrapolated to real in-the-wild practice – our use of actual tools and software found in a radiology office, combined with a realistic setting and genuine patient data means that we provided participants with a setting they would be familiar enough with to provide accurate interactional data. By also conducting collaborative discussion and analysis with a subset of the participant pool, we have also made sure that our findings and conclusions are representative and accurate as a means of "fact-checking" ourselves.

Overall, we have provided developers and educators with a demonstration that, although interacting with VUIs is of critical importance to the work of a radiologist in the NHS Wales, their use in-situ is built upon the assumption that using these systems is straightforward and safe. We have highlighted the gaps in design, implementation and training that provide friction to the reporting process, and it is clear that more work should be done on the real-world use of tools in safety critical medical environments. VUI systems have been deployed in radiology departments as a way of innovating and improving digital radiology, but their lack of simplicity and high maintenance requirements mean that clinicians often have to spend large amounts of their time wrangling with corrections and administrative issues that could be spent examining patient data. Proper ways to interact with VUI systems should be part of a radiologist's training, but acknowledgement of a radiologist's duties should also form part of the fundamental design underpinning the development of VUI and speech based text input technology.

8 Acknowledgements

This work was funded in part by EPSRC grant EP/S021892/1

References

[1] Joanna Abraham and Madhu C. Reddy. 2008. Moving patients around: a field study of coordination between clinical and non-clinical staff in hospitals. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work* (New York, NY, USA, 2008-11-08) *(CSCW '08)*. Association for Computing Machinery, 225–228. doi:10.1145/1460563.1460598

[2] Carlos Monroy Aceves, Patrick Oladimeji, Harold Thimbleby, and Paul Lee. 2013. Are prescribed infusions running as intended? Quantitative analysis of log files from infusion pumps used in a large acute NHS hospital. 22 (2013), 15–21. Issue Sup13. doi:10.12968/bjon.2013.22.Sup13.15 Publisher: Mark Allen Group.

[3] Ahmad Al-Aiad, Ahmad K. Momani, Yazan Alnsour, and Mohammad Alsharo. 2020. The Impact of Speech Recognition Systems on The Productivity and The Workflow in Radiology Departments: A Systematic Review. *AMCIS 2020 TREOs* (2020). https://aisel.aisnet.org/treos_amcis2020/62

[4] Tariq O. Andersen. 2013. Medication management in the making: on ethnography-design relations. In *Proceedings of the 2013 conference on Computer supported cooperative work* (New York, NY, USA, 2013-02-23) *(CSCW '13)*. Association for Computing Machinery, 1103–1112. doi:10.1145/2441776.2441901

[5] Barun Aryal, Derek A. Khorsand, and Theodore J. Dubinsky. 2018. The Clinical and Medicolegal Implications of Radiology Results Communication. 47, 5 (2018), 287–289. doi:10.1067/j.cpradiol.2017.09.009

[6] Jørgen Bansler, Erling Havn, Troels Mønsted, Kjeld Schmidt, and Jesper Hastrup Svendsen. 2013. Physicians' Progress Notes: The Integrative Core of the Medical Record. In *ECSCW 2013: Proceedings of the 13th European Conference on Computer Supported Cooperative Work, 21-25 September 2013, Paphos, Cyprus*, Olav W. Bertelsen, Luigina Ciolfi, Maria Antonietta Grasso, and George Angelos Papadopoulos (Eds.). Springer London, 123–142. doi:10.1007/978-1-4471-5346-7_7

[7] Sarah Basma, Bridgette Lord, Lindsay M. Jacks, Mohamed Rizk, and Anabel M. Scaranelo. 2011-10. Error rates in breast imaging reports: comparison of automatic speech recognition and dictation transcription. *AJR. American journal of roentgenology* 197, 4 (2011-10), 923–927. doi:10.2214/AJR.11.6691

[8] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2020-04-23) *(CHI '20)*. Association for Computing Machinery, 1–12. doi:10.1145/3313831.3376718

[9] R. Bentley, J. A. Hughes, D. Randall, T. Rodden, P. Sawyer, D. Shapiro, and I. Sommerville. 1992. Ethnographically-informed systems design for air traffic control. In *Proceedings of the 1992 ACM conference on Computer-supported cooperative work - CSCW '92* (Toronto, Ontario, Canada, 1992). ACM Press, 123–129. doi:10.1145/143457.143470

[10] Suzanne V. Blackley, Jessica Huynh, Liqin Wang, Zfania Korach, and Li Zhou. 2019-04-01. Speech recognition for clinical documentation from 1990 to 2018: a systematic review. *Journal of the American Medical Informatics Association: JAMIA* 26, 4 (2019-04-01), 324–338. doi:10.1093/jamia/ocy179

[11] Ann Blandford, Dominic Furniss, George Buchanan, Harold Thimbleby, and Paul Curzon. 2010. Who's looking? Invisible problems with interactive medical devices. *WISH* (2010).

[12] Ann Blandford, Dominic Furniss, and Chris Vincent. 2014. Patient safety and interactive medical devices: Realigning work as imagined and work as done. 20, 5 (2014), 107–110. doi:10.1177/1356262214556550

[13] Jeanette Blomberg and Helena Karasti. 2013. Reflections on 25 Years of Ethnography in CSCW. 22, 4 (2013), 373–423. doi:10.1007/s10606-012-9183-1

[14] G. Boland. 2007. Enhancing the radiology product: the value of voice-recognition technology. 62, 11 (2007), 1127. doi:10.1016/j.crad.2007.05.014

[15] Claus Bossen and Lotte Groth Jensen. 2014. How physicians 'achieve overview': a case-based study in a hospital ward. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (New York, NY, USA, 2014-02-15) *(CSCW '14)*. Association for Computing Machinery, 257–268. doi:10.1145/2531602.2531620

[16] Marco Bozzano and Adolfo Villafiorita. 2010. *Design and safety assessment of critical systems.* CRC press.

[17] Adrian P. Brady. 2017-02. Error and discrepancy in radiology: inevitable or avoidable? *Insights into Imaging* 8, 1 (2017-02), 171–182. doi:10.1007/s13244-016-0534-1

[18] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa Publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa.

[19] Mindaugas Briedis. 2019. Phenomenological ethnography can lead to the improvement of radiology diagnostics. *Adaptive Behavior* 27 (2019), 347–350. doi:10.1177/1059712319861663

[20] Mindaugas Briedis. 2020. Phenomenological ethnography of radiology: expert performance in enacting diagnostic cognition. *Phenomenology and the Cognitive Sciences* 19, 2 (2020), 373–404. doi:10.1007/s11097-019-09612-x

[21] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. 5 (2021), 188:1–188:21. Issue CSCW1. doi:10.1145/3449287

[22] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. 3 (2019), 104:1–104:24.

Issue CSCW. doi:10.1145/3359206
[23] Paulo Victor R. de Carvalho, Angela W. Righi, Gilbert J. Huber, Caio de F. Lemos, Alessandro Jatoba, and José Orlando Gomes. 2018. Reflections on work as done (WAD) and work as imagined (WAI) in an emergency response organization: A study on firefighters training exercises. 68 (2018), 28–41. doi:10.1016/j.apergo.2017.10.016
[24] Rory Stuart Clark, Tom Owen, Matt Jones, Martin Porcheron, and Phillip Wardle. 2023. It Works Better When I Do That: Interaction and Communication In Radiology Departments. In *Proceedings of the 41st ACM International Conference on Design of Communication* (Orlando FL USA). ACM, 55–62. doi:10.1145/3615335.3623011
[25] E. Coiera and V. Tombs. 1998. Communication behaviours in a hospital setting: an observational study. 316, 7132 (1998), 673–676. doi:10.1136/bmj.316.7132.673
[26] Andy Crabtree, Steve Benford, Chris Greenhalgh, Paul Tennent, Matthew Chalmers, and Barry Brown. 2006. Supporting ethnographic studies of ubiquitous computing in the wild. In *Proceedings of the 6th ACM conference on Designing Interactive systems - DIS '06* (University Park, PA, USA, 2006). ACM Press, 60. doi:10.1145/1142405.1142417
[27] Andrew Crabtree, Mark Rouncefield, and Peter Tolmie. 2012. *Doing Design Ethnography*. Springer London. doi:10.1007/978-1-4471-2726-0
[28] Andy Crabtree, Peter Tolmie, and Mark Rouncefield. 2013. "How Many Bloody Examples Do You Want?" – Fieldwork and Generalisation. doi:10.1007/978-1-4471-5346-7_1
[29] Gunnar Ellingsen and Kristoffer Røed. 2010. The Role of Integration in Health-Based Information Infrastructures. 19, 6 (2010), 557–584. doi:10.1007/s10606-010-9122-y
[30] Dominic Furniss. 2014. HCI Observations on an Oncology Ward: A Fieldworker's Experience. In *Fieldwork for Healthcare: Case Studies Investigating Human Factors in Computing Systems*. Springer International Publishing, 19–25. doi:10.1007/978-3-031-01596-0_3
[31] Dominic Furniss, Ann Blandford, and Astrid Mayer. 2011-07-01. Unremarkable errors: low-level disturbances in infusion pump use. doi:10.14236/ewic/HCI2011.47
[32] Harold Garfinkel. 1967. Studies in Ethnomethodology. *Studies in Ethnomethodology* (1967).
[33] Clifford Geertz. 2008. Thick description: Toward an interpretive theory of culture. In *The cultural geography reader*. Routledge, 41–51.
[34] Kristina Groth and Jeremiah Scholl. 2013. Coordination in highly-specialized care networks. In *Proceedings of the 2013 conference on Computer supported cooperative work companion* (New York, NY, USA, 2013-02-23) *(CSCW '13)*. Association for Computing Machinery, 143–148. doi:10.1145/2441955.2441992
[35] J L Hart, A Mcbride, D Blunt, P Gishen, and N Strickland. 2010. Immediate and sustained benefits of a "total" implementation of speech recognition reporting. 83, 989 (2010), 424–427. doi:10.1259/bjr/58137761
[36] M. Hartswood, R. Procter, M. Rouncefield, and R. Slack. 2002. Performance Management in Breast Screening: A Case Study of Professional Vision. 4, 2 (2002), 91–100. doi:10.1007/s101110200008
[37] Christian Heath, Jon Hindmarsh, and Paul Luff. 2010. *Video in Qualitative Research: Analysing Social Interaction in Everyday Life*. SAGE Publications, Inc. doi:10.4135/9781526435385
[38] Erik Hollnagel. 2015. Why is work-as-imagined different from work-as-done? In *Resilient Health Care*, Robert L Wears, Erik Hollnagel, and Jeffrey Braithwaite (Eds.). Ashgate Studies in Resilience Engineering, Vol. 2. Ashgate, 249–264.
[39] Erik Hollnagel and Robyn Clay-Williams. 2022. Work-as-Imagined and Work- as-Done. In *Implementation Science*. Routledge. Num Pages: 3.
[40] Steven Houben, Mads Frost, and Jakob E. Bardram. 2015. Collaborative Affordances of Hybrid Patient Record Technologies in Medical Work. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (New York, NY, USA, 2015-02-28) *(CSCW '15)*. Association for Computing Machinery, 785–797. doi:10.1145/2675133.2675164
[41] John Hughes, Val King, Tom Rodden, and Hans Andersen. 1994. Moving out from the control room: ethnography in system design. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work* (New York, NY, USA, 1994-10-22) *(CSCW '94)*. Association for Computing Machinery, 429–439. doi:10.1145/192844.193065
[42] Ioanna Iacovides, Ann Blandford, Anna Cox, Bryony Dean Franklin, Paul Lee, and Chris J. Vincent. 2014. Infusion device standardisation and dose error reduction software. 23, 14 (2014), S16, S18, S20 passim. doi:10.12968/bjon.2014.23.sup14.s16
[43] Fernando De Castro Guimarães Rios Ignácio, Luis Ronan Marquez Ferreira De Souza, Giuseppe D'Ippolito, and Mayara Martins Garcia. 2018. Radiology report: what is the opinion of the referring physician? 51, 5 (2018), 308–312. doi:10.1590/0100-3984.2017.0115
[44] Annette J. Johnson, Michael Y. M. Chen, J. Shannon Swan, Kimberly E. Applegate, and Benjamin Littenberg. 2009-10. Cohort Study of Structured Reporting Compared with Conventional Dictation. *Radiology* 253, 1 (2009-10), 74–80. doi:10.1148/radiol.2531090138
[45] Maree Johnson, Samuel Lapkin, Vanessa Long, Paula Sanchez, Hanna Suominen, Jim Basilakis, and Linda Dawson. 2014-10-28. A systematic review of speech recognition technology in health care. *BMC Medical Informatics and*

Decision Making 14, 1 (2014-10-28), 94. doi:10.1186/1472-6947-14-94

[46] Dylan Jones, Clive Frankish, and Kevin Hapeshi. 1992-03-01. Automatic speech recognition in practice. *Behaviour & Information Technology* (1992-03-01). doi:10.1080/01449299208924325 Publisher: Taylor & Francis Group.

[47] Clare-Marie Karat, Christine Halverson, Daniel Horn, and John Karat. 1999. Patterns of entry and correction in large vocabulary continuous speech recognition systems. In *Proceedings of the SIGCHI conference on Human factors in computing systems the CHI is the limit - CHI '99* (Pittsburgh, Pennsylvania, United States). ACM Press, 568–575. doi:10.1145/302979.303160

[48] M. Lafortune, G. Breton, and J. L. Baudouin. 1988. The radiological report: what is useful for the referring physician? 39, 2 (1988), 140–143.

[49] Jennifer Lai and John Vergo. 1997-03-27. MedSpeak: report creation with continuous speech recognition. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems* (Atlanta Georgia USA). ACM, 431–438. doi:10.1145/258549.258829

[50] Luke E. Lassiter. 2005. *The Chicago guide to collaborative ethnography.* Univ. of Chicago Press.

[51] Yun Liu, Lu Wang, William R. Kearns, Linda Wagner, John Raiti, Yuntao Wang, and Weichao Yuwen. 2021. Integrating a Voice User Interface into a Virtual Therapy Platform. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2021-05-08) *(CHI EA '21)*. Association for Computing Machinery, 1–6. doi:10.1145/3411763.3451595

[52] D MacVicar. 2005. Are you sitting comfortably? 78, 931 (2005), 581. doi:10.1259/bjr/37038966

[53] Philippe Marrast, Pascale Zaraté, and Anne Mayère. 2013. How to support coordination through annotations? A longitudinal case study of nurses' work in an oncology hospital. 207–212. doi:10.1145/2441955.2442007

[54] S McGURK, K Brauer, T V Macfarlane, and K A Duncan. 2008. The effect of voice recognition software on comparative error rates in radiology reports. 81, 970 (2008), 767–770. doi:10.1259/bjr/20698753

[55] Amit Mehta, Keith J. Dreyer, Alan Schweitzer, John Couris, and Daniel Rosenthal. 1998. Voice recognition—An emerging necessity within radiology: Experiences of the Massachusetts General Hospital. *Journal of Digital Imaging* 11, 2 (1998), 20–23. doi:10.1007/BF03168173

[56] Tilo Mentler, Philippe Palanque, Susanne Boll, Chris Johnson, and Kristof Van Laerhoven. 2021. Control Rooms in Safety Critical Contexts: Design, Engineering and Evaluation Issues. In *Human-Computer Interaction – INTERACT 2021* (Cham, 2021), Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen (Eds.). Springer International Publishing, 530–535. doi:10.1007/978-3-030-85607-6_72

[57] Tilo Mentler, Philippe Palanque, Kristof Van Laerhoven, Margareta Holtensdotter Lützhöft, and Nadine Flegel. 2023. On Land, at Sea, and in the Air: Human-Computer Interaction in Safety-Critical Spaces of Control. In *Human-Computer Interaction – INTERACT 2023* (Cham, 2023), José Abdelnour Nocera, Marta Kristín Lárusdóttir, Helen Petrie, Antonio Piccinno, and Marco Winckler (Eds.). Springer Nature Switzerland, 657–661. doi:10.1007/978-3-031-42293-5_89

[58] R. E. Motyer, S. Liddy, W. C. Torreggiani, and O. Buckley. 2016. Frequency and analysis of non-clinical errors made in radiology reports using the National Integrated Medical Imaging System voice recognition dictation software. *Irish Journal of Medical Science (1971 -)* 185, 4 (2016), 921–927. doi:10.1007/s11845-016-1507-6

[59] NA. [n. d.]. REDACTED FOR ANONYMITY. ([n. d.]).

[60] Royal College of Radiologists. [n. d.]. *What does a clinical radiologist do? | The Royal College of Radiologists.* https://www.rcr.ac.uk/our-specialties/clinical-radiology/discover-clinical-radiology/thinking-about-a-career-in-clinical-radiology/what-does-a-clinical-radiologist-do/

[61] John A. Pezzullo, Glenn A. Tung, Jeffrey M. Rogg, Lawrence M. Davis, Jeffrey M. Brody, and William W. Mayo-Smith. 2008-12. Voice Recognition Dictation: Radiologist as Transcriptionist. *Journal of Digital Imaging* 21, 4 (2008-12), 384–389. doi:10.1007/s10278-007-9039-2

[62] A. A. O. Plumb, F. M. Grieve, and S. H. Khan. 2009. Survey of hospital clinicians' preferences regarding the format of radiology reports. 64, 4 (2009), 386–394. doi:10.1016/j.crad.2008.11.009

[63] Martin Porcheron, Leigh Clark, Stuart Alan Nicholson, and Matt Jones. 2023. Cyclists' Use of Technology While on Their Bike. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2023-04-19) *(CHI '23)*. Association for Computing Machinery, 1–15. doi:10.1145/3544548.3580971

[64] Sayeeda Rahman, Md Anwarul Azim Majumder, Sami F Shaban, Nuzhat Rahman, Moslehuddin Ahmed, Khalid Bin Abdulrahman, and Urban JA D'Souza. 2011-03-07. Physician participation in clinical research and trials: issues and approaches. *Advances in Medical Education and Practice* 2 (2011-03-07), 85–93. doi:10.2147/AMEP.S14103

[65] James Reason. 1990. *Human error.* Cambridge university press.

[66] Madhu Reddy and Paul Dourish. 2002. A finger on the pulse: temporal rhythms and information seeking in medical work. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work* (New York, NY, USA, 2002-11-16) *(CSCW '02)*. Association for Computing Machinery, 344–353. doi:10.1145/587078.587126

[67] Madhu C. Reddy, Paul Dourish, and Wanda Pratt. 2006. Temporality in Medical Work: Time also Matters. 15, 1 (2006), 29–53. doi:10.1007/s10606-005-9010-z

[68] Michael D. Ringler, Brian C. Goss, and Brian J. Bartholmai. 2015. Syntactic and Semantic Errors in Radiology Reports Associated With Speech Recognition Software. *Studies in Health Technology and Informatics* 216 (2015), 922.
[69] Sherry Ruan, Jacob O. Wobbrock, Kenny Liou, Andrew Ng, and James A. Landay. 2018-01-08. Comparing Speech and Keyboard Text Entry for Short Messages in Two Languages on Touchscreen Phones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018-01-08), 1–23. doi:10.1145/3161187
[70] Rimvydas Rukšėnas, Paul Curzon, Ann Blandford, and Jonathan Back. 2013. Combining human error verification and timing analysis: A case study on an infusion pump. 26 (2013), 1033–1076. doi:10.1007/s00165-013-0288-1
[71] Valéria Farinazzo Martins Salvador and Lincoln de Assis Moura. 2010. Heuristic evaluation for automatic radiology reporting transcription systems. 292–295. doi:10.1109/ISSPA.2010.5605467
[72] Kjeld Schmidt. 1994. Field Studies and CSCW [COMIC Deliverable D2.2]. (1994). https://www.academia.edu/2041422/Field_Studies_and_CSCW_COMIC_Deliverable_D2_2_
[73] David Smith and Kenneth Simpson. 2004. *Functional safety.* Routledge.
[74] Allan Stisen, Nervo Verdezoto, Henrik Blunck, Mikkel Baun Kjærgaard, and Kaj Grønbæk. 2016. Accounting for the Invisible Work of Hospital Orderlies: Designing for Local and Global Coordination. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (New York, NY, USA, 2016-02-27) *(CSCW '16)*. Association for Computing Machinery, 980–992. doi:10.1145/2818048.2820006
[75] Johanna Viitanen. 2009-09-01. Redesigning digital dictation for physicians: A user-centred approach. *Health Informatics Journal* 15, 3 (2009-09-01), 179–190. doi:10.1177/1460458209337429 Publisher: SAGE Publications Ltd.
[76] A. Wallis and P. McCoubrie. 2011. The radiology report — Are we getting the message across? 66, 11 (2011), 1015–1022. doi:10.1016/j.crad.2011.05.013 Publisher: Elsevier.

Received July 2024; revised December 2024; accepted March 2025