Challenges for Responsible AI Design and Workflow Integration in Healthcare: A Case Study of Automatic Feeding Tube Qualification in Radiology

ANJA THIEME, Microsoft Health Futures, UK
ABHIJITH RAJAMOHAN, BENJAMIN COOPER, HEATHER GROOMBRIDGE, ROBERT SIMISTER, and BARNEY WONG, UCLH NHS Foundation Trust, UK
NICHOLAS WOZNITZA, UCLH NHS Foundation Trust & Canterbury Christ Church University, UK
MARK A. PINNOCK, University College London, UK
MARIA T. WETSCHEREK, University of Cambridge & Cambridge University Hospitals NHS Foundation Trust, UK
CECILY MORRISON, Microsoft Research Cambridge, UK
HANNAH RICHARDSON, FERNANDO PÉREZ-GARCÍA, STEPHANIE L. HYLAND, SHRUTHI BANNUR, DANIEL C. CASTRO, KENZA BOUZID, ANTON SCHWAIGHOFER, MERCY P. RANJIT, and HARSHITA SHARMA, Microsoft Health Futures, UK
MATTHEW P. LUNGREN, Microsoft Health Futures, University of California & Stanford University, US
OZAN OKTAY, JAVIER ALVAREZ-VALLE, and ADITYA NORI, Microsoft Health Futures, UK
STEPHEN HARRIS and JOSEPH JACOB, University College London, UK

Nasogastric tubes (NGTs) are feeding tubes that are inserted through the nose into the stomach to deliver nutrition or medication. If not placed correctly, they can cause serious harm, even death to patients. Recent AI developments demonstrate the feasibility of robustly detecting NGT placement from Chest X-ray images to reduce risks of sub-optimally or critically placed NGTs being missed or delayed in their detection, but gaps remain in clinical practice integration. In this study, we present a human-centered approach to the problem and describe insights derived following contextual inquiry and in-depth interviews with 15 clinical stakeholders. The interviews helped understand challenges in existing workflows, and how best to align technical capabilities with user needs and expectations. We discovered the trade-offs and complexities that need consideration when choosing suitable workflow stages, target users, and design configurations for different AI proposals. We explored how to balance AI benefits and risks for healthcare staff and patients within broader organizational, technical, and medical-legal constraints. We also identified data issues related to edge cases and data biases that affect model training and evaluation; how data documentation practices influence data preparation and labelling; and how to measure relevant AI

Authors' addresses: Anja Thieme, anthie@microsoft.com, Microsoft Health Futures, Cambridge, UK; Abhijith Rajamohan; Benjamin Cooper; Heather Groombridge; Robert Simister; Barney Wong, UCLH NHS Foundation Trust, UK; Nicholas Woznitza, UCLH NHS Foundation Trust & Canterbury Christ Church University, UK; Mark A. Pinnock, University College London, UK; Maria T. Wetscherek, University of Cambridge & Cambridge University Hospitals NHS Foundation Trust, UK; Cecily Morrison, Microsoft Research Cambridge, UK; Hannah Richardson; Fernando Pérez-García; Stephanie L. Hyland; Shruthi Bannur; Daniel C. Castro; Kenza Bouzid; Anton Schwaighofer; Mercy P. Ranjit; Harshita Sharma, Microsoft Health Futures, UK; Matthew P. Lungren, Microsoft Health Futures, University of California & Stanford University, US; Ozan Oktay; Javier Alvarez-Valle; Aditya Nori, Microsoft Health Futures, UK; Stephen Harris; Joseph Jacob, University College London, UK.

ACM 1557-7325/2025/2-ART
https://doi.org/10.1145/3716500

outcomes reliably in future evaluations. We discuss how our work informs design and development of AI applications that are clinically useful, ethical, and acceptable in real-world healthcare services.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; • **Computing methodologies → Machine learning algorithms**.

Additional Key Words and Phrases: Radiology, AI, healthcare, responsible AI, socio-technical systems, feeding tubes, NGT

1 INTRODUCTION

Artificial Intelligence (AI) is increasingly gaining recognition as an important application in radiology [48, 66, 67, 111, 136]. AI has been applied to detect and diagnose key clinical findings (e.g., [68, 74, 102, 152]), or to analzse anatomical structures on medical images [55, 59]. Latest advances in foundation models (FMs) [20], which are powerful, general-purpose models that can be adapted to various healthcare and radiology-specific tasks, suggest even greater potential for AI to revolutionize clinical practice [25, 101, 131, 143, 157] as demonstrated in tasks such as medical knowledge extraction [117], clinical text modification (e.g., [76, 85, 103, 105]) and new forms of medical question-answering [89, 132]. Recent approaches towards *multimodal* FMs, that integrate medical images alongside text representations (e.g., BioVIL(-T) [11, 19], ELIXR [159], MAIRA-1 and -2 [10, 69], or Med-PaLM M [145]), further expand the scope of possible, innovative use cases. For AI-assisted radiology, this includes capabilities to: automatically generate a radiology report from a medical image (e.g., [10, 66, 69, 170]); to answer questions about a radiology image using text queries (cf. [159]); or detect errors in a radiology report text by comparing it with the image [167].

Despite all the excitement and remarkable progress in AI research and development in recent times, the successful translation of such technical innovations into clinical practice however remains challenging [13, 14, 51, 108, 110, 136, 146, 149–151, 153, 171]. In what has been described by Andersen et al. [110] as "a race for getting the technology right before exposing human end-users to new promising AI tools"; many researchers warn against the dangers of developing AI in isolation [99] – without considering the specific needs of the intended users and the downstream implications of the technology [91]. This currently leaves a significant gap between research investigations, compelling technical proofs-of-concepts, or lab experiments, and larger ambitions of integrating and deploying AI-enabled systems successfully in routine clinical care [37, 81, 149, 153, 171].

Some of the challenges for real-world healthcare integration arise from a lack of trust in the 'black-box' nature of advanced AI outputs [120, 149, 151, 153, 173, 175]; and the difficulties for designers and healthcare experts to appropriately understand and productively work with new AI capabilities [30, 162, 165, 168]. There are also perceptions AI may replace clinicians from their jobs [151]; and uncertainty about the value that AI applications bring to clinical practice. This is evident where systems are perceived to not provide useful information [12, 114]; provide information that cannot be actioned [54]; or where system interactions are too time-consuming or cumbersome to work well in practice [151] – reducing AI utility. Alongside this, there are increasing demands for the evaluation and monitoring of AI in real-world settings [90] that foreground clinical validation of AI algorithms [149] and effectiveness trials (cf. [51]).

New AI solutions should also better align with context-specific patterns of care [114] and account for the disruptions they can create on existing workflows [30, 149] – including the social relationships that often characterize collaborative healthcare delivery [46]. Interlinked with this are many broader organizational, technical, ethical and regulatory issues that affect the adoption and use of AI systems in healthcare (cf. [14, 126, 127, 136]). Therefore, designing human-AI interactions effectively and responsibly within the clinical domain presents a complex socio-technical problem that requires a holistic, multidisciplinary approach that deeply engages with the real-world problems and contexts that AI solutions aim to solve [46, 49, 110]; carefully matching technical capabilities with user needs [91, 143, 153].

The work presented in this paper builds on recent human-centred healthcare AI research (e.g., [14, 26, 33, 58, 73, 96, 109, 126, 150, 151, 153, 160, 162, 168]) that studies clinical workflows and corresponding AI integration challenges [14, 26, 33, 149–151, 160]; and that provides insights into AI design [26, 64, 162] for configuring effective human-AI interactions [58, 142, 168]. Starting as early as problem identification and ideation stages [31, 72, 86, 156, 166], our research explores the opportunities and challenges of developing and integrating AI within an end-to-end radiology imaging workflow such that it can achieve *clinical utility*. By deepening understanding of existing clinical workflows and data documentation practices, we want to derive key insights and requirements that can meaningfully guide AI development (e.g., assist with data labels, edge case understanding, or evaluation metrics). More specifically, we seek to extract more clearly the *interrelations* [16, 173] between AI data work or systems and corresponding clinical needs, workflows, and broader AI acceptance and organizational considerations. In this regard, our human-centred approach focuses on the specific use case of AI assisting the interpretation of Chest X-ray images (CXRs) to verify the correct placement of nasogastric feeding tubes in intensive care patients with the goals of improving workflow efficiency and patient safety.

1.1 Use Case: Nasogastric Feeding Tube Placement Qualification on Chest X-ray Images

A nasogastric tube (NGT) is a thin tube that is inserted via the nose and passed into the stomach. It is used for short- to medium-term nutritional support, medication administration or aspiration of stomach contents [61]. NGTs are amongst the most commonly used medical lines in critically ill patients in intensive care units (ICU) and emergency departments for life-supporting purposes [164], and high-dependency units and departments where patients require nutritional support (i.e., Stroke). Due to increases in the number of hospitalized patients, it is estimated that approximately 10 million NGTs are used annually in Europe, 1 million of which are used in the UK (approx. 1.2 million in the US) [144]. Previous research highlights a variety of complications associated with NGT placement [107, 174] – especially the accidental insertion into the patients' lungs (Figure 3B), which can cause aspiration of feeds and pneumothorax [53] that increase time spent in intensive care, treatment costs [5], and patient morbidity and mortality, highlighting the importance of accurate feeding tubes placement [164]. Yet, clinical studies demonstrate that up to 3% of NGTs are reported as misplaced within the lungs, causing complications in up to 40% of these cases [84]. Seeking to reduce risks of sub-optimally or critically placed NGTs being missed or delayed in their detection, initial AI developments demonstrate the feasibility of robustly assessing NGT placement on CXR images [43, 128]. Despite great technical advances, gaps remain in understanding how best to design; practically integrate; and responsibly use any such models as part of clinical workflows; as well as how best to evaluate the effectiveness of any prospective AI application. We examine these issues in the context of a UK hospital ICU, with some comparisons to Stroke care, which is detailed in Section 4.

1.2 Research Questions & Contributions

Our work presents a rare example of an in-depth case study that engages early in the AI design process with the end-to-end workflow and concrete use context of CXR-based NGT placement verification with key domain stakeholders. The study aim is to understand unique opportunities and challenges for creating clinically useful AI applications. Specifically, we ask: (1) what are the right types of applications; (2) how to effectively and responsibly design those from a human-centred perspective; and (3) how can insights into the specific use context meaningfully guide AI development and evaluation? Against this backdrop, our work makes three main contributions:

(1) *We surface complex interrelations between human, technical and organizational factors that determine perceived AI utility and successful adoption*; and we propose the systematic mapping of identified factors as a tool to clarify important benefit-risk and feasibility trade-offs.

(2) *We propose 'Human-Process Integration of AI' as an approach to future AI development to foster AI acceptance*. We argue that AI should not be seen as a separate entity that needs human 'verification', but framed as part of existing human (safeguarding) processes of information review, guideline adherence, and patient concern.
(3) *We extract key insights into real-world data availability and data production practices, and discuss their implications for AI development*. Specifically, we reflect on dataset curation, model training, and outcome evaluations.

2 RELATED WORK

We begin with a concise summary of: (i) existing AI approaches for automatically detecting and localizing medical lines and tubes on CXRs; and (ii) the relevant literature on human-centered AI research in healthcare and radiology.

2.1 AI & Machine Learning for Automated Detection of Medical Lines and Tubes on CXRs

In recent years, we witness a growth of AI research and development in the automatic detection and localization of medical lines and tubes on CXRs, seeking to help prioritize and shorten turn-around times especially for critical cases (e.g., [88, 121, 130, 164]) and thereby improve the effectiveness of clinician workflows and patient safety [44, 130, 169]. The majority of existing studies is focused on detecting one specific tube type, most commonly *central venous catheters (CVC)* [88, 128, 133, 137, 169], which are thin tubes inserted via the patient's veins to draw blood and give treatments [133]; and also *endotracheal tubes (ETT)* [80], which are airway tubes to assist in lung ventilation. Given that critically ill patients often have multiple lines and tubes inserted (i.e., patients are intubated for air ventilation and fed via a feeding tube), many studies explored the differentiation of multiple tube types on a CXR image [1, 5, 21, 44, 62, 82, 121].

Only few studies to date specifically address the placement of feeding tubes [43, 130]. Most notably, Drozdov et al. [43] report the development and evaluation of a deep learning (DL) approach for NGT misplacement detection. Their model achieves high performance on various NGT classification tasks (e.g., AUC of 0.98 for lung misplacement). The authors are also amongst few (e.g., [124]) who study how CXR AI can assist clinicians in critical tube findings detection and enhance clinical decision-making. Their study with five clinicians reviewing 335 CXR images with and without AI revealed an increased of overall accuracy in decision to feed from 69% (unaided) to 78% (with AI), suggesting greater clinician confidence in decision making, and the potential for AI to reduce NGT misplacement complications.

More generally, for tube placement detection, we find a variety of ML approaches applied – mostly to CXR image analysis – and in rare cases to radiology reports (cf. [128]). These approaches seek to assist in tasks such as: (i) detecting the *existence of a tube* on a CXR [62, 80, 137]; (ii) classifying the *tube type* present [1, 62, 82, 137]; (iii) identifying or classifying *tube tip position* [80, 88, 128, 169] and *relevant anatomical landmarks* (e.g., carina point) [133]; (iv) and classifying the *accuracy* (i.e., normal, borderline, abnormal [1, 5, 21, 43, 82, 121]) or *criticality of the tubes placement* (e.g., critical vs. non-critical [124, 130]). Most of the datasets used for analysis are either self-curated; or derived from much larger publicly available datasets like: MIMIC [78], NIH ChestX-ray14 [154] and RANZCR CLiP [138]. See Table 7 in the Appendix A.1 for an overview of these studies, including datasets used, and reported AI performance outcomes. In general, across these studies, tube misplacement classification performance is high, with many reporting accuracies of 90-95%, or more.

In terms of more commercially oriented developments, *Qure.ai* reported receiving FDA approval[1] for ML confirmed placement of breathing tubes. The company *annalise.ai* employed a deep learning (DL) model on a large scale of CXRs (over 800,000 images) with clinician curated labels for a wide range of clinical findings; their model

[1] https://qure.ai/news_press_coverages/qure-ais-breathing-tube-placement-ai-technology-receives-fda-clearance/

was able to robustly predict suboptimally placed NGTs with high AUC (0.984) alongside other catheter types such as central lines, ETTs and pulmonary arterial catheters [124]. Lastly, based on the research by Drozdov et al. [43] reported above, *Bering Ltd* recently received UK CA marking[2] for BraveNGT, an AI software that detects NGT misplacement on CXRs to provide effective decision support for clinicians whether feeding in patients can be safely performed.

All these works suggest growing research and commercial interest and increasing promise of utilizing AI capabilities in CXR analysis to provide useful insights to lines and tubes (mis)placement qualification. Simultaneously, it demonstrates the need to move from technical solutions and clinical proof-of-concepts to understanding AI system deployment and how design choices implicate both utility and risks.

2.2 Human-Centered AI Research & Design in Healthcare and Radiology

Responsible AI focuses on centering people and their goals in design processes, considering the benefits and potential harms of AI systems on individuals and society. Extended to the complex domain of healthcare, this requires human-centered approaches to AI research and design [14, 73, 173]. Below, we first provide a brief summary of common challenges for clinical AI development, and then extend to the specific domains of medical imaging and radiology.

2.2.1 Challenges for Responsible Clinical AI Design. A recent systematic review of human-centered approaches to clinical decision support systems (CDSS) [153] highlight that developing AI solutions for healthcare requires addressing many complex, contextual factors for successful implementation and user adoption (e.g., [26, 73, 96, 126, 151, 160]).

Amongst those factors is the need to ensure *availability of good quality and representative healthcare data* – as the building block for training, adapting or evaluating AI models (cf. [38, 143, 157]); and to address potential risks of inequality and discrimination that may arise from biased or unfair data algorithms [15, 36, 106, 143, 172]. Where healthcare data is the focus, concerns also arise about how patients can control, consent or opt out of data uses; and how their data privacy and security can be ensured [140, 156].

Furthermore, a substantial body of human-centred AI research is focused on improving AI acceptance into clinical practice by fostering *trust* and *confidence in AI*. Here, especially the field of eXplainable AI (XAI) (e.g., [4]) aims to make AI more transparent and understandable by employing various methods of explanation and feedback that enable clinicians to contest [63, 115], verify [160], negotiate with [134], and learn from AI outputs [30]; seeking to configure effective human-AI collaboration [141]. For instance, recent research by Burgess et al. [26] investigated healthcare practitioners response to prototypes designed to support medication selection for Type 2 diabetes melilitus. Studying how AI insights could find acceptance within the health practitioners workplace, the authors found – amongst others – that AI systems were judged against reputable methods of clinical knowledge generation (e.g., RCTs) and that trust in offered insights were moderated by users understanding of, and their confidence in the methods that generated the AI insight.

In recent times, studies have increasingly focused on *broader organizational, technical, social, ethical and regulatory implications of AI systems* (cf. [14, 126, 127, 136]); emphasizing the need for greater alignment of AI development with clinical workflows [151, 153] and stakeholder needs [168]. For example, Wang et al. [151] found that an AI-driven decision support tool in rural Chinese clinics was misaligned with local contexts and work practices that differed from text-book approaches. In their study, everyday time-constraints meant that it was impossible for clinicians to gather all the information required by the AI systems to make accurate and comprehensive diagnosis to aid their work, nor did 80% of daily tasks even involve the need for a diagnosis (e.g., medical refills, specialist referrals). System recommendation utility was further limited due to a lack of interoperability with other health information systems (e.g., laboratory, pharmaceutical) as well as lacking

[2]https://icaird.com/2022/icaird-partner-bering-achieves-ukca-mark-for-ai-supported-chest-x-ray-classification/

consideration of broader social-economical factors that meant algorithm-suggested medicine was often not affordable to lower-income patients.

Moreover, for AI acceptance and adoption, researchers have studied health professionals understanding [30], attitudes and perspectives of AI [167]; and how AI may come to be perceived as a threat to their *professional autonomy* and sense of control over clinical decision-making processes [151, 168]; alongside broader questions of medical-legal *accountability* and responsibility – especially where AI-assisted decisions in healthcare span individual users, healthcare institutions, and insurance providers (e.g., [52, 113, 118, 149]).

As a diverse and heterogeneous domain that encompasses different specialties, settings, workflows, and stakeholders, AI solutions for healthcare therefore require rigorous definition and tailoring to the specific needs of the intended users and beneficiaries; and within its anticipated use context (cf. [14, 126, 151]). Against this backdrop, our work seeks to develop a deeper understanding of existing work practices, stakeholder values, and the clinical problem space that surrounding NGT placement verification within an ICU setting, with the aim to better understand the specific design requirements, and potential impacts of any prospective AI intervention.

2.2.2 Human-centered AI in Radiology. Human-centered AI research in medical imaging spans investigations in the fields of ophthalmology [8, 14], pathology [29, 57, 58, 93], and radiology [6, 17, 32–34, 108, 150, 158]. In pathology, for instance, Cai et al. [29, 30] presented their user research and development process for the *SMILY* prototype, a prediction tool for prostate cancer diagnosis. They showed how enabling users to interactively refine the predictions improved the clinical utility of their tool and user trust in the algorithm. Lindvall et al. [93] designed *Rapid Assisted Visual Search*, a human-AI interface to help pathologists assess colorectal cancer. Conducting an evaluation with six pathologists, they demonstrated how their interface reduced pathologists' search time for regional lymph nodes with signs of metastasis.

In radiology more specifically, AI research so far mainly focused on making AI outputs more understandable to domain experts [6, 33, 34, 108]. For example, Atad et al. [6] used counterfactuals to explain the AI diagnosis of CXR findings (e.g., cardiomegaly) by highlighting what feature changes in the image would lead the model to give a different outcome. Ontika et al. [108] proposed a hybrid AI system for multiparametric MRI to help prostate cancer diagnosis; using visualizations to facilitate human comprehension of AI generated outputs. The authors conducted contextual inquiries and interviewed five radiologists to better understand how to create "an impactful human-AI collaborative environment in radiology" (p. 395). Their work revealed the variability of medical imaging interpretation, workflow differences across radiology centers, and potential for automating tedious and repetitive tasks.

Another growing area of research is the evaluation of AI radiology models and studies of their clinical impact. More technically-oriented investigations focus on new approaches to systematically evaluate the accuracy and suitability of model outputs, and identif relevant radiology-specific metrics (e.g., assessing factual completeness and consistency of AI generated radiology text [100]). Often these involve domain experts in reviewing and categorizing AI errors [10, 69, 170]. Other studies examine the effects of AI use on clinicians, demonstrating improved radiologists classification accuracy with AI [124] as well as reduced diagnosis time and error rates (cf. AI assisted *BreastScreening* for breast cancer diagnosis [33, 34]). Some studies also explored how different ways of presenting AI outputs to clinicians (e.g., via an assertive or non-assertive communication style) can reduce medical errors [32]. For instance, Bernstein et al. [17] showed how incorrect AI outputs can bias radiologists to make incorrect follow-up decisions when they were correct without AI. However, adding a bounding box on the region of interest (e.g., to verify AI results) reduced human errors.

Lastly, few studies investigate radiology workflows or focus on understanding current needs of radiologists in their daily practice [109, 149, 150, 158, 167]. An exception is work by Xie et al. [158], who conducted an early phase need-finding and design study that included a survey, low-fidelity prototype design, and high-fidelity evaluation to identify opportunities for AI-assistance in radiology X-ray work. Exploring the barriers to AI

adoption in radiology imaging, Verma et al. [149, 150] reported an interview study with seven imaging experts in oncology. Their investigations revealed rich insights into clinicians' concerns about black-box models, small training datasets, and quality control in training data [150]; as well as demonstrating a divergence in perceptions, intentions and scope of AI between *clinical* and *research* workflows [149]. Finally, a recent interview study by Yildirim et al. [167] sought to achieve better alignment between clinician/ radiologist needs and new AI capabilities of vision-language models (VLMs) to identify clinically useful radiology applications. Amongst others the authors examined participants perspectives on AI generating a draft report from CXR images. Findings revealed: design preferences for short-from texts and easy content review and edits; nuance about what report parts radiologists wishes to remain in control over (as opposed to have it AI-generated); and considered the impact of such an AI assistance feature on radiology trainee education and learning experiences.

Crucially, these works (e.g., [17, 32]) evidence that AI implementation choices affect human-AI performance; emphasizing the need to further investigate AI design and integration challenges and opportunities. We extend this line of work by investigating the end-to-end workflow in verifying feeding tube placement via CXRs to learn about challenges and identify design requirements for clinically relevant, responsible AI development.

3 STUDY METHOD

Aiming to clarify opportunities and challenges for AI assisted NGT placement verification, our user research involved a combination of ICU ward observations and semi-structured staff interviews. To conduct observational work and recruit staff, the lead researcher (AT) partnered with a Junior Clinical Fellow (AR) for this project, who she accompanied and 'shadowed' on four of their regular work shifts. This included one short (8 hours, 8am-4pm), and three long shifts (13 hours): split into two day-shifts (8am-9pm) and one night-shift (8pm-9am). The observation days spanned two different ICU settings (the main hospital ICU and a private unit) to observe existing catheter placement procedures, ward dynamics, and data documentation practices. Through our presence on the wards and additional hospital contacts of the broader research team, we were able to recruit 15 hospital staff to our interviews. These were either held in a vacant hospital room during observations days, or conducted as remote calls using Microsoft Teams software.

3.1 Ethics

The research study was carefully reviewed and monitored for compliance and privacy regulations; and approved by the NHS Health Research Authority (REC reference: 22/HRA/4824). Informed consent was sought by all participants in writing prior to the study.

3.2 Participants

Our interview participants reflect a range of professions and included predominately junior and more senior clinicians and nurses from ICU care, as well as two Stroke clinicians and three radiographers, who were reporting X-rays across the hospital more generally. The sample also presented a mix of staff who had been in their professional role for 0-2 years (n = 8), with the remaining describing 2-5 (n = 2), 5-10 (n = 2) or more than 10 (n = 3) years of experience. Gender was balanced across the study cohort with 8 self-reporting as *female* (7 as *male*). Each participant has been given a unique identification number to protect their anonymity, reported as: D1-D8 for doctors in ICU or Stroke care, N1-N4 for ICU nurses, and RR1-RR3 for reporting radiographers. See Table 1 for participant details.

3.3 Interview Procedure

Our staff interviews were aimed at 1 hour in duration with some variation depending on staff availability, especially for on-ward meetings, which we needed to accommodate more flexibly (MD = 60 mins, M = 51 mins,

Table 1. Participants professional role, care setting, amount of years they had been working in this specific role, and their gender; alongside details on the location and duration of the interview.

Code	Professional Role	Setting	Time in Role	Gender	Location	Duration
D1	Junior Doctor	Intensive Care (ICU)	0-2 years	male	Teams Call	60 mins
D2	Junior Doctor	Intensive Care (ICU)	0-2 years	male	Teams Call	60 mins
D3	Junior Doctor	Intensive Care (ICU)	0-2 years	female	On-Ward	60 mins
D4	Junior Doctor	Intensive Care (ICU)	0-2 years	male	Teams Call	60 mins
D5	Registrar	Intensive Care (ICU)	0-2 years	female	On-Ward	45 mins
D6	Consultant	Intensive Care (ICU)	2-5 years	female	On-Ward	20 mins
N1	Staff Nurse	Intensive Care (ICU)	>10 years	female	On-Ward	40 mins
N2	Senior Staff Nurse	Intensive Care (ICU)	>10 years	male	Teams Call	35 mins
N3	Senior Staff Nurse	Intensive Care (ICU)	5-10 years	female	On-Ward	45 mins
N4	Charge Nurse	Intensive Care (ICU)	0-2 years	female	On-Ward	35 mins
D7	Registrar	Stroke Care	0-2 years	male	Teams Call	65 mins
D8	Consultant	Stroke Care	>10 years	male	Teams Call	60 mins
RR1	Consultant Reporting Radiographer	Imaging Dept. (ID)	2-5 years	male	Teams Call	65 mins
RR2	A&E Superintendent & Reporting Radiographer	Imaging Dept. (ID)	5-10 years	male	Teams Call	60 mins
RR3	Consultant Reporting Radiographer	Imaging Dept. (ID)	0-2 years	female	Teams Call	55 mins

min = 20 mins, *max* = 65 mins). The interviews were semi-structured and involved three main parts that sought to better understand: (1) NGT end-to-end workflow and data documentation practices; (2) challenges with existing processes; and (3) AI opportunities to assist NGT verification via CXRs.

Following the capture of basic demographic information, the first part asked participants to describe the current NGT placement and CXR verification process step-by-step; about their specific role and responsibilities within that workflow; and how various steps are being documented (where and by whom). The second part of the interview explored the difficulties or problems that participants faced at any stage of this process. We asked them about their reasons for why NGTs might be misplaced and not detected; and how they could prevent or minimize delays in identifying misplaced NGTs. For staff who checked CXRs to confirm NGT placement, we also inquired about situations of doubt or ambiguity in examining the image to understand what kind of help could assist their assessments. Whenever relevant, we asked staff to give specific examples to better illustrate their experiences.

Lastly, we tried to spark participants imagination about prospective AI by asking ”what-if” questions to probe into different ways in which AI could provide insights to the NGT CXR verification process. Specifically, we explored five AI proposals (Table 2) that would span three main categories of intended AI uses in clinical practice: decision support; prioritization; and task automation [173]. The proposals were further informed by radiology AI research describing functionality to: detect critical image findings [124]; prioritize radiologist reading lists [9]; produce image segmentations for lines and tubes [50, 155]; and generate radiology reports from medical images [10, 69, 170]. Translating these to the use context of NGTs, we thus proposed for AI to either (1) provide an (early) alert for a misplaced tube that could be presented to clinicians; or (2) prioritize the image in the radiology reporter worklist. We also imagined AI could (3) act as a cross-checker of image assessments made by an image reporter to flag up any potential errors (e.g., if it identifies discrepancies between human text report and what it itself detects from the image). The AI may also (4) offer a visual overlay or segmentations that highlight the tube line and tip on the CXR image, or it can trace key anatomical landmarks to assist image interpretation. Lastly, the AI may (5) auto-generate a (preliminary) report of the NGT CXR to speed up clinician review or radiology sign-off. To engage in deeper conversation about the prospective use of such AI, we asked: How would you use this AI information if it was available to you? In what scenarios do you think having this AI functionality

Table 2. Overview of five different proposals for how AI could come into assisting NGT CXR verification practices.

AI proposals for assisted NGT CXR verification
(1) (Early) NGT misplacement detection to alert clinicians to speed up correction
(2) (Early) NGT misplacement detection to prioritize for radiology reporting
(3) AI cross-checker for detecting errors in human assessment of the NGT CXR
(4) Segmentation overlay of NG tube + tip/ key anatomical landmarks to guide visual assessment
(5) Auto-generate (preliminary) report of NGT for clinician review or faster radiology sign-off

would be beneficial? What are the advantages and disadvantages? We also inquired about any concerns that the participants might have regarding any of these suggestions.

We acknowledge that different AI model type vary in their capabilities and limitations, which can profoundly influence their success in the world; and are mindful that each proposal comes with different AI requirements (e.g., classification of NGT placement is likely an easier task to realize technically than auto-report generation). However, at this early problem investigation and formulation stage, we deliberately chose to remain more agnostic to any specific model configuration or performance. We only assume that the AI models can handle multi-modal data such as radiology images, clinical texts, or other electronic patient record (EHR) data (e.g., [10, 69, 145]). The AI proposals serve as examples to explore different system goals and factors that may enhance or reduce clinical utility.

3.4 Data collection & analysis

The main aim of this research was to better understand the opportunities and challenges for developing clinically meaningful AI for NGT CXR placement verification within real-world hospital workflows. Our research is based on a constructionist epistemology, taking the perspective that knowledge and *meaning* are actively constructed by individuals and become socially produced via an interplay of subjective and inter-subjective construction [28]. In this regard, *reflexive* Thematic Analysis (TA) [23] was chosen as it offers a systematic, qualitative approach for analyzing the unique perspectives of different hospital staff – including their aspirations and concerns for, and broader complexities shared about AI use within this context – whilst embracing reflexivity and valuing the researcher's subjectivity as an analytical *resource* [23, 24]. Reflexive TA is also commonly used in HCI healthcare research [22]. Reflexive TA acknowledges the *researchers active, interpretive role* in creating meaning through deep interaction with, and identification of patterns, in the data [24, 28]. The lead researcher (AT) conducting the analysis is a Westernized, female researcher with a background in HCI; who is employed at a highly resourced AI industry lab. She came into this project with over ten years of experience working as a human-centered researcher in healthcare, including prior studies in UK hospitals (albeit not in ICU). This provided a base understanding of how the NHS is organized and the pragmatic challenges of facilitating such research. As someone actively working at the intersection of AI innovation, HCI and healthcare, she understands the prospective benefits of AI use in healthcare whilst being aware of the many complex, socio-technical challenges surrounding its successful application. Her own commitment to creating *responsible AI applications* that seek to *improve societal outcomes* mean that she takes a *critical view on techno-solutionism*. Going into the research, she was also aware that clinical stakeholders may not have a good understanding or misaligned expectations of AI (e.g., due to AI hypes or public failings) that can make it more difficult to invite their imagination about novel technology capabilities that were not previously possible. This meant, she made extra efforts to explain AI workings and possible failure cases to participants to facilitate discussion and considerations of potential risks.

Our data analysis is predominantly based on the interviews with hospital staff, which were audio-recorded. The lead researcher fully anonymized the recordings; transcribed them with Microsoft Teams software, and checked and edited these for correctness. Following the six-phase process for data engagement, coding and theme development by Braun & Clarke [24], the lead researcher familiarized herself with the interview data; systematically coded it regarding the main research question; and iteratively developed, reviewed and refined themes. The experiential-oriented data analysis was part-deductive: guided by existing AI research; concrete research goals documented in a study protocol with clearly defined topics, research activities and desired outcomes; and assumptions about AI use (Table 2) that provided a lens through which the data was analyzed. Predominantly, however, the lead researcher followed an inductive, open-coding approach based on the data itself. High-level themes such as "Existing workflow challenges, safeguarding mechanisms & their limitations" (Section 5.1) comprise of multiple sub-themes such as "Delays to timely NGT placement verification" (Section 5.1.1) that includes facets of: resource constraints & emergency-led care; communication inefficiencies; and broader ward and workflow dynamics. For instance, the facet "resource constraints & dependencies on emergency-led care" is developed from semantic codes like "NGT assessment delayed if ward/ staff busy" or "Emergency-only resources at night (NGT not acute)".

Given the richness of insights in each theme, the finding presentation is divided into two parts: *Part A* focuses on themes related to the identification of clinically meaningful applications of AI to assist NGT CXR verification practices. It entails two main themes: existing NGT workflow challenges (Section 5.1), and perceived clinical utility of AI (Section 5.2); and derived from these, includes an example for "mapping out" the interrelations that the analysis surfaced (Section 5.3). *Part B* centers on the theme of opportunities and challenges for data preparation, labelling & interpretation (Section 5.4), and from this extrapolates insights into measuring AI intervention success (Section 5.5). Throughout our analysis we seek to derive practical insights for the development of AI for this use case.

Lastly, there were additional data sources and discussions that helped contextualize the interviews and supported this interpretive work. Firstly, the lead researcher kept a study diary, taking notes of her observations as well as capturing room set-ups, relevant machinery, and data artefacts with photography[3]. Secondly, to better understand how the NGT verification process was captured in data, the lead researcher reviewed during her ward visits (under supervision) selected ICU patient records to better understand the timings, data location, and format of key events. This review served to better understand what types of data is commonly generated as part of the overall NGT position qualification process; to check for consistency in entries; and assess the viability of existing data entries for use as outcome metrics (e.g., reduction in delays to feeding) in any prospective AI pilot study. Thirdly, data results were reviewed independently with clinical research collaborators (AR, JJ, NW, SH, TW) that helped confirm and adjust interpretations of Part A of the findings (Sections 5.1f); whilst a focused group discussion with AI researchers (DC, FP, HS, KB, SH, SB) provided additional analytical depth to the data and ML development implications of the Part B findings (Sections 5.4f).

4 BACKGROUND: STUDY CONTEXT & PROCESS OF FEEDING TUBE QUALIFICATION VIA CXR

This section summarizes the everyday work practices shared by hospital staff during the interviews, detailing the NGT placement verification workflow end-to-end. Where relevant, this context summary is contextualized by on-ward observations and literature references; providing an overview of the specific hospital context and the research motivation before presenting the main findings on AI opportunities and challenges in subsequent Section 5.

[3]Any hospital or person identifiable information were either cropped from the image or blocked out as a black bar on any imagery to protect patient and staff anonymity.

4.1 Hospital context

The research was conducted at the University College London (UCL) Hospitals NHS Foundation Trust that serves as a major teaching hospital, presents a world-wide centre for medical research, and is well-known for its provision of first-class acute and specialist services. Our investigation focused particularly on AI integration within intensive care units (ICU). Through ward visits and conversations with the Junior Clinical Fellow (AR), we learned that the hospitals' ICU has 48 inpatient beds (including individual patient rooms for those with infectious diseases). It encompasses a main ICU, which is divided into two separate parts: North and South (23 beds). Three additional ICUs are located in the same building and other hospital sites, which includes a private ICU ward for specialist oncology patients (10 beds). During the day, each ICU, or its parts, has a care team comprising of: a consultant, a registrar and a team of junior doctors. Overnight, one consultant is on-call at home and a registrar looks after the unit (or its two parts); alongside a team of junior doctors. Each patient also has a bedside nurse 24/7 dedicated to them. Despite our studies primary focus on ICU care, we also included the review of existing practices in the hospitals' Stroke department as another inpatient area that frequently places NGTs, to explore how gathered insights and proposed avenues for AI may translate and generalize across different care settings. Here, the Stroke staff we interviewed described how the approximately 40-inpatient Stroke cohort (excluding patients with Stroke on ICU) is split between a hyper acute unit, and longer-stay patients, who are not in the hyper acute stage of their care. Nursing staff ratios differ depending on patient acuity and range between one nurse to four patients, to one nurse to eight patients.

4.2 End-to-end ICU Workflow of Nasogastric Feeding Tube (NGT) Verification via Chest X-ray (CXR)

This section gives an overview of the ICU end-to-end workflow of qualifying the placement of a nasogastric feeding tube (NGT) via chest radiography. Starting at NGT insertion, hospital staff described that – unless the patient arrives with an NGT already inserted (e.g., by theater surgeons) – the decision to place such a tube is a complex, multi-faceted process that considers the patients: (i) nutritional needs; (ii) ability to swallow that is needed for oral feeding; (iii) acuity (e.g., if a patient is very unstable, feeding is often not a priority), and (iv) other risk and comfort factors (e.g., insertion could cause delirium). Given other, often more acute patient conditions, patient feeding is commonly sought of as less 'urgent' and NGT insertion tasks or placement checks often delayed within a targeted 24-48 hour period from devising an NGT feeding plan to the patient being fed (see Figure 1 for a workflow overview).

ICU nurses described how they are usually responsible for placing the NGT and noted that any difficulties during the tubes insertion or adverse patient response (e.g., oxygen saturation drop) can already signal a potential misplacement. To verify correct placement, the nurses apply various methods. This includes the rare use of a laryngoscope during placement, which directly visualizes the tube location. More commonly, and in keeping with the hospitals detailed NGT insertion and position verification policy, the NGT verification process involves as first line test: the *obtaining of gastric aspirate*. This means a syringe is used to acquire gastric content from the stomach via the placed NG tube. This content is then checked for its pH value; a pH value of 5.5 or less confirms the NGTs location in the stomach, deeming it as safe to use (cf., guidelines by [116]). At this hospital, various staff described how aspirate-based pH tests however were only successful in 8-10% of ICU cases, since many patients receive proton pump inhibitors like pantoprazole – a type of antacid medication that contributes to generally higher pH values (cf. also [139]). As a result, for most NGT placement checks, and especially upon any newly placed NGT, Chest X-rays are requested as a more definite second line verification test.

It is ICU doctors who request CXRs for NGT confirmation via the hospital's EPIC[4] electronic patient record (EHR) system. The doctors then use a pager to contact the on-call radiographer, awaiting call back to discuss the case and schedule the CXR based on urgency triage with other imaging requests. Since the majority of ICU

[4]https://www.epic.com/

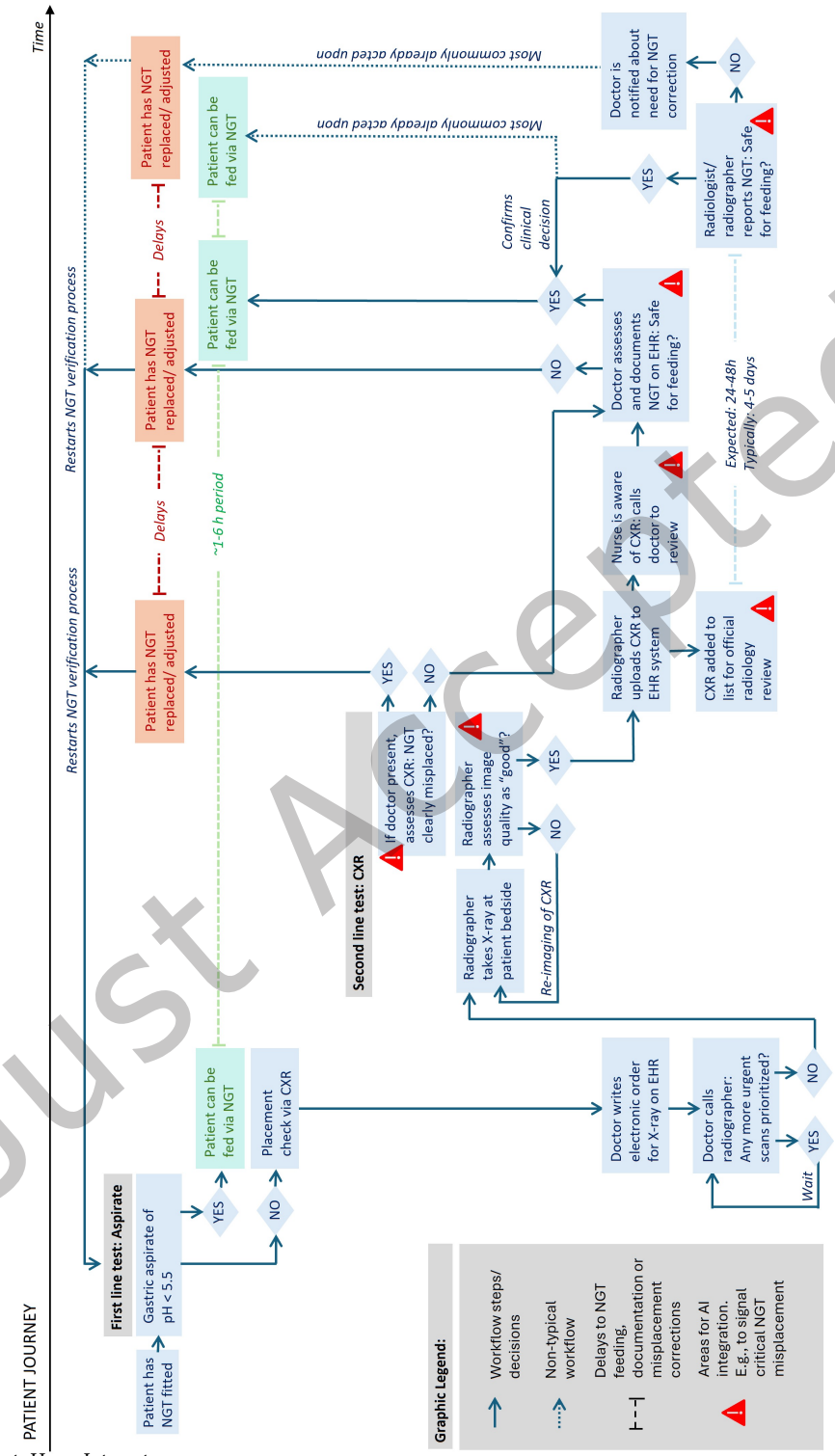


Fig. 1. Overview of a typical ICU end-to-end workflow of CXR based NGT verification from initial insertion through to decisions to use NGT for feeding or correct it, including temporal delays. The red exclamation mark symbol illustrates areas of potential integration of AI assisted critical NGT misplacement detection at different time-points and for different user groups. Note that this workflow differs from Stroke care where clinicians are not present at CXR capture, nor often review the image themselves, and rely on official radiology reporting. In ICU, NGTs are most commonly already acted upon by the time of the official radiology report.

inpatients are too unstable to be transported to the imaging department, CXRs are mostly carried out via a mobile X-ray machine that is brought to the patients bedside by a radiographer, who captures the CXR as an anterior-posterior[5] (AP) image. Through in-person observation of a mobile X-ray capture procedure on the ICU, it is apparent that the image taken immediately appears on a preview monitor of the X-ray machine. When asked about the relevance of this preview image in subsequent interviews, doctors and radiographers explained this image enables them to assess if image quality of the performed CXR is acceptable, and if not, to immediately initiate any repeate imaging. Whilst not intended as formal assessment, ICU doctors and radiographers also described how they may already identify misplaced NGTs at this image preview stage. However, they explained how no verbal assessment of the image is accepted as grounds for the actual feeding decision. Instead, decisions that the NGT is safe-to-feed need to be formally documented in writing within EPIC once the image is uploaded to the hospitals' Picture Archiving Communications System: PACS [47]. Rather than being automatic, this image upload however requires the radiographer to physically move the mobile X-ray machine and connect it to the hospital computer system, a process that can be delayed if there's a sequence of scheduled X-rays that the radiographer needs to perform.

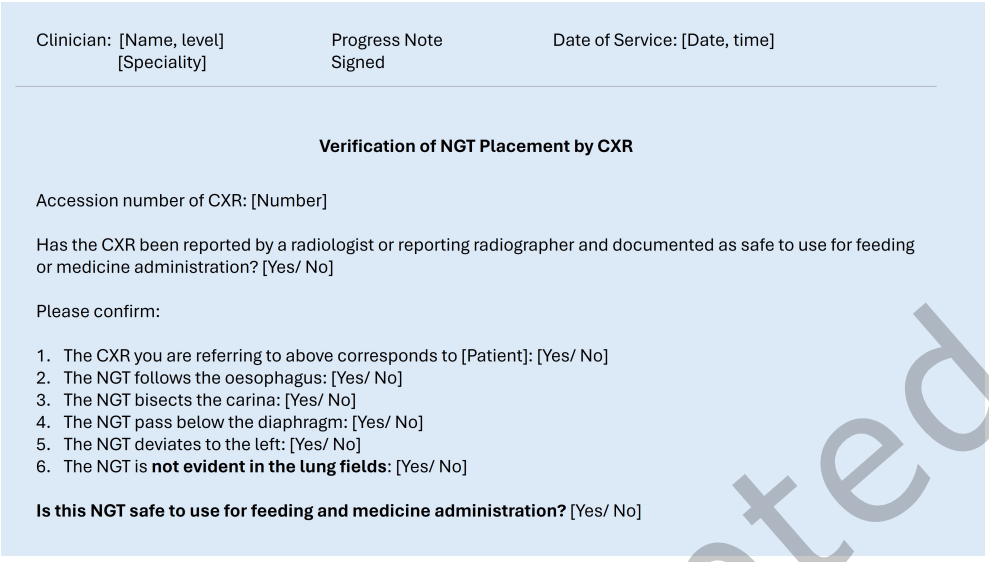
The ICU NGT CXR capture process differs from Stroke care, where patients are rarely imaged at bedside and are instead taken by a Porter to and from the imaging department (ID) – a process that requires additional resources and organization. It also means that in this alternative workflow, the patients' clinical care team (e.g., Stroke physician) is not present during image capture. Yet, since the imaging is integrated into the hospital system, image upload is instant.

Once the patient CXR is uploaded to PACS, it is added to a queue of images awaiting official radiology reporting by a radiologist or reporting radiographer. As also reported in the literature [164], the large number of CXRs obtained each day, especially in intensive and emergency care, means that image interpretation can be substantially delayed. Consequently, it is common practice for ICU doctors and more senior (Stroke) doctors to check the CXR and verify the NGT's correct positioning and suitability for use [137] prior to the radiology report being issued [138]. In our study context, ICU bedside nurses therefore reported how they would frequently review all incoming patient results and notify doctors if a CXR image became available on the system to ask for its documentation. Doctors then log into PACS to review the image, using zoom and image contrast enhancement tooling to aid their image assessment. To document NGT placement, ICU and senior Stroke doctors described utilizing a smart text template, called '.NGT', that provides them with a set of binary questions that they complete in the absence of an official radiology report, and that gives clinical permission to commence patient feeding if deemed as safe. Figure 2 shows the .NGT template, which was re-drawn from a photo taken on-site. The template questions serve as visual check points to ensure correct placement of the NGT, which in most normal patient cases is defined as a tube that follows down the oesophagus, bisects the carina, passes below the diaphragm and then deviates to the left such that the tube tip is located in the stomach.

Importantly, an NGT should not be placed into the patient's lungs. Feeding a patient through an NGT misplaced into the lungs would be a severe incident that can have critical implications – including patient death – and is classified by the UK National Health Services (NHS) as a Never Event[6]. Less critical, yet sub-optimally placed NGTs may not extend far enough into the stomach; need withdrawing; or may be kinked or coiled along the path – thereby inhibiting proper use and intended functionality. Figure 3 illustrates cases of normal, sub-optimal and critically (mis)placed NGTs as informed by the literature (e.g., [138]) and individual patient case reviews with

[5]In this imaging position and view, the patient has a metal plate placed behind their back and the radiation beam traverses through their front chest to the back of their body

[6]Never events are "serious incidents that are entirely preventable because guidance or safety recommendations providing strong systemic protective barriers are available at a national level, and should have been implemented by all healthcare providers": https://www.england.nhs.uk/patient-safety/revised-never-events-policy-and-framework/"



Clinician: [Name, level] [Speciality] | Progress Note Signed | Date of Service: [Date, time]

Verification of NGT Placement by CXR

Accession number of CXR: [Number]

Has the CXR been reported by a radiologist or reporting radiographer and documented as safe to use for feeding or medicine administration? [Yes/ No]

Please confirm:

1. The CXR you are referring to above corresponds to [Patient]: [Yes/ No]
2. The NGT follows the oesophagus: [Yes/ No]
3. The NGT bisects the carina: [Yes/ No]
4. The NGT pass below the diaphragm: [Yes/ No]
5. The NGT deviates to the left: [Yes/ No]
6. The NGT is **not evident in the lung fields**: [Yes/ No]

Is this NGT safe to use for feeding and medicine administration? [Yes/ No]

Fig. 2. Schematic representation of the '.NGT' template used to verify NGT placement outside an official radiology report.

clinicians that were also captured in photos. Any intervention to correct or replace the NGT then restarts the entire NGT verification process (Figure 1).

Once an NGT is correctly placed, ICU nurses described starting patient feeding either by following a standard protocol or a more bespoke plan devised by ICU dietitians based on the patient's weight and caloric needs. Although the hospital aims to report inpatient X-rays within 12 hours for acutely unwell or emergency care patients (ideally under 4 hours during normal work-hours)[7], actual turn-around times are much longer (e.g., more than 7 days), eliminating the report's practical utility and relevance in clinical decision making. The Stroke clinicians we spoke to further explained how reporting times differ in their care setting, where junior doctors are not permitted to access CXRs and rely on reports by radiographers, radiologists or a senior clinician. Turnaround times, while faster than for ICU-requested scans, can still be delayed for non-acute CXRs or requests made outside main work hours (e.g., at night, weekends).

Where CXRs are assessed by non-radiology staff, and when working in stressful care environments with staff and resources stretched to capacity, previous research highlighted that image interpretation is prone to human errors [44]. This suggests potential utility of leveraging AI image analysis capabilities, for example, to *alert clinicians to critical placement* or *have NGT CXRs prioritized for urgent radiology reporting*; enabling earlier detection of misplaced NGTs to speed-up their correction and prevent any additional complications [138]. AI may also assist in *detecting human errors* in image assessment, *provide visual guidance* to human interpretation, or help *automate NGT assessment* as captured in our AI proposals (Table 2).

5 FINDINGS PART A: CLINICALLY MEANINGFUL AI APPLICATIONS FOR NGT CXR VERIFICATION

Our investigation into meaningful AI NGT applications begins with: insights into existing workflow challenges, safeguarding mechanisms and their limitations (Section 5.1). We then detail how perceived clinical utility of

[7] https://www.england.nhs.uk/long-read/diagnostic-imaging-reporting-turnaround-times/

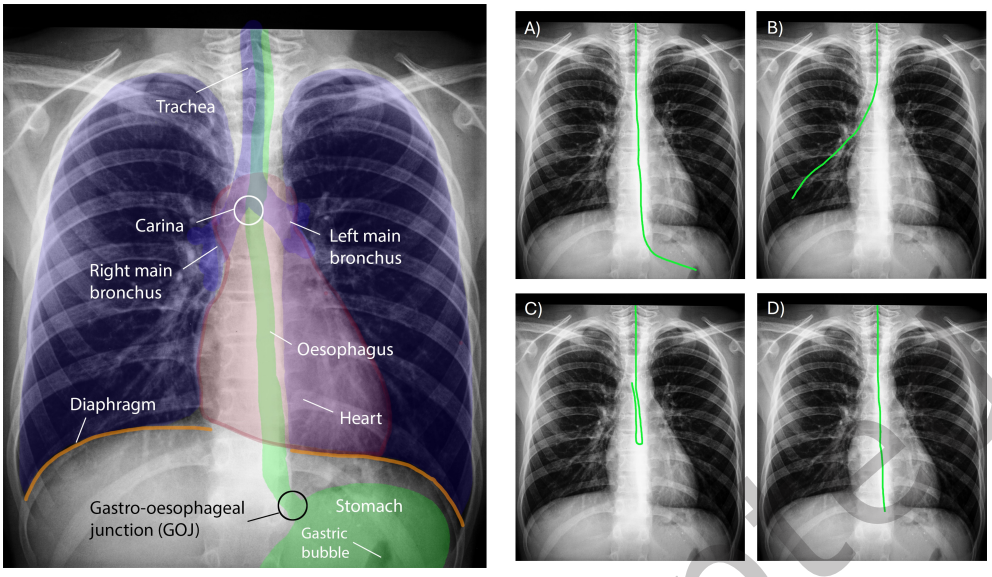


Fig. 3. Left: Simplified illustration of a patient's key anatomical structures of relevance to NGT placement assessment (the heart is drawn for clarity only). Right: different NGT placements: A) correct placement of NGT with tip projecting in the stomach; B) critical misplacement of the NGT into the patient's right lung (via the trachea and the right main bronchus); C) misplaced NGT which is coiled in patient's mid oesophagus; and D) misplaced NGT with tip in distal oesophagus requiring further advancement to reach the stomach.

proposed AI functionality is bound-up by a complex interplay and decisions surrounding: the target user group; desired AI outcome prioritization; workflow integration and AI design choices; as well as broader AI acceptance and adoption challenges (Section 5.2). Based on these two themes, we conclude this Part A with an example of how identified factors of application type, workflow stage, user group, design choice and technology performance requirements can be mapped to clarify and trade-off the potential clinical impact and feasibility of different AI proposals (Section 5.3).

5.1 Existing Workflow Challenges, Safeguarding Mechanisms & their Limitations

This section first outlines broader workflow challenges surrounding the end-to-end NGT CXR verification process that expand the scope for AI opportunities as well as give important context for data work. We then specifically describe the prevalence of errors, their reasons, and existing safeguarding mechanisms in NGT image assessment and subsequent safe-to-feed decisions, which presents the main focus of our AI investigation.

5.1.1 Delays to Timely NGT Placement Verification as Most Prevalent Workflow Problem.

When asked about key challenges or pain points in the NGT verification process, rather than describing difficulties in CXR image interpretation, staff most prominently mentioned the length of 'time' that it takes, and the various types of 'delays' that span the entire process from initial NGT placement, through to the CXR being scheduled, captured, uploaded, reviewed and documented, and the patient eventually being fed. We learned that timeliness of NGT placement verification is bound up by a complex interplay of: (i) staff resourcing and emergency-led care requirements; (ii) communication inefficiencies; and (iii) broader ward dynamics.

Resource Constraints & Emergency-led Care Affecting Staff Ability to Perform NGT Tasks. With hospitals operating on "bare minimum of staff for everything, particularly at night", staff availability (and level of expertise) required to execute NGT-related tasks is especially limited at weekends, and out of main working hours. This affects ICU staff, who can be busy or caught up in a crash call[8]. Similarly, image capturing or reporting staff can be in high-demand. For example, one radiographer alone may need to cover both CT and X-ray imaging, which then often leaves NGT checks last due to a need to prioritize more urgent patient cases, as well as its balancing with needs of other busy hospital departments:

> "(...) it's all [based] on this understanding of emergency need. And so X-rays to check NG positioning sit awkwardly in the way that most hospitals plan their overnight resources, or all of their acute flow, because they're not quite acute, and yet at the same time, they need to be done soon." (D8, Stroke)

Low staff resources and emergency-centric care mean that X-rays may not get timely reported unless acute. These delays, often extending over several days, can result in the loss of the reports' clinical relevance, particularly for assessing a misplaced NGT or guiding clinical decisions. As a result, ICU staff tend to "not wait" (D5), "rely on" (D4), nor "check" (D2, D3) the radiology report for NGTs:

> "Yeah, I guess the main thing is if you're not immediate reporting them, are they really of any use to anyone? That's the thing, isn't it? Unless you're getting a definite result right there, is there any point of reporting something that's three days old and that they're already using? I don't know. I guess that's one of the main challenges." (RR3, ID)

Communication Inefficiencies & Constantly Chasing-up Tasks: Along the NGT process, staff described human checks (e.g., keeping taps on patient records to identify if a report has been issued) or in-person verbal exchanges whereby nurses 'chase-up' doctors about the need to order an X-ray or document it, alongside electronic messages or system notifications. In their accounts, they describe communication inefficiencies such as risks of missing digital messages and notifications on EPIC in good time; incomplete handover information across staff teams that delay actions; and radiographers forgetting to carry their bleep or to call back (D1); as well as doctors missing call-backs, forgetting to order, or being late to review the CXRs. All this suggests the need for a more reliable communication system and better assistance with administrative burden:

> "The frustrating parts of being a doctor is how much time you spend doing kind of admin tasks such as being on the phone all the time, waiting for bleeps back from people, or waiting to bleep people, waiting for reports of things to come, waiting for emails. All of these types of things which AI might make our lives quicker and easier in, and free us up to do the decision-making part of the job or the assessment part of the job would be really beneficial." (D1, ICU)

Ward dynamics of Task Offloading based on Staff Confidence, Flow & Shift Handover Times: Timeliness of NGT task completion and their prioritization are further moderated by (i) staff's ability to safely execute a task (e.g., their competence, time availability to place the NGT, or do checks). More implicitly, staff described their awareness of strain on under-staffed services, which they responded to by offloading, for example, tasks from night shifts to the day team. Describing considerations of heightened risks in conducting certain, less urgent procedures at night due to increased tiredness and awareness of a pending shift hand-over, a junior doctor stated:

> "So, so this all happened overnight and I remember being very busy nightshift where I think there were only a few hours left of the night shift. And we were like: actually, this is probably safer left to the morning team. Just don't use it, don't touch the NG at the moment (...)" (D3, ICU).

Staffs considerations of shift patterns in decision to prioritize, hold or post-pone tasks is further evident in an account whereby an NGT is placed on ICU at night or early morning, but does not require urgent confirmation. In such cases, the doctor may suggest to the radiographer on-call that image capture could wait for the day

[8]In a crash call on ICU, all doctors are required to be at hand for a deteriorating patient who experiences a cardiac arrest and needs resuscitation.

team, being mindful that requesting an image scan close to shift hand-over (usually 8am or 8pm) can mean for a radiographer having to stay at work longer:

> "But I also know that, if they [the radiographers] hand over at 7:30 AM to their colleagues and I'm bleeping them at 7:15 and they go, it's kind of day team come. If I say no, we need it now, then they're not going home until late. So then, human nature, you're kind of like, yeah, it's fine." (D1, ICU)

Task prioritization is further balanced with staffs desires and needs to protect the "flow" of clinicians and nurses in completing tasks. The below quote illustrates how a senior ICU staff nurse (N1) is conscious to not want to interrupt doctors too frequently about various tasks. Instead, she keeps a "rolling list" of things that are less urgent that she then informs them about upon their next exchange:

> "I mean, the doctors are in and out anyway, so I tend to have a little rolling list of things I want to tell them. So I don't forever be interrupting them in their flow as well. If some thing's can wait till the evening ward round, I just kind of wait and then just say this needs doing, you know, and then we can get it all sorted in one go. Otherwise, it's a bit fragmented for them. They're kind of jumping between patients all the time, and it's quite…They lose their flow of thought as well. I don't think all nurses think that way, though I don't think more junior ones would think I need to get this sorted, I'm gonna talk to the doctor and they are forever talking to the doctor about things, but I can kind of prioritise what they need to know straight away, yeah." (N1, ICU)

All of this draws attention to complex human factors and work rhythms that determine the prioritization and timeliness of NGT task completion and data documentation. These broader workflow dynamics have two main implications: Firstly, suggested delays and inefficiencies in existing NGT verification broaden the scope for how AI could assist current radiology practices above and beyond image interpretation (e.g., focused on optimizing communication/ triage management processes). See further Appendix A.2 for additional ideas that surfaced throughout this study for how AI could assist NGT workflows. Secondly, interdependencies caused by variations in staff resourcing; emergencies; or shift pattern present key variables that may hinder staff's ability to act upon specific AI insights (e.g., an AI alert to a misplaced NGT) and therefore require considerations in any evaluative studies that seek to assess AI impact (see further Section 5.5).

5.1.2 Prevalence of, Reasons for Errors and Existing Safeguards in NGT CXR Image Interpretation & Safe-to-Feed Decisions.

Outside process delays, and given the severe implications of missing and potentially using an NGT that has been critically misplaced into the patients' lung, we next report staffs' accounts on: the prevalence of NGT misplacement; reasons for errors in NGT assessment; and current hospital risk mitigation strategies. These serve to better understand where AI could add value; and may sit alongside, or replace existing safeguarding practices.

While staff confirmed that feeding tube placement occurs frequently on ICU and Stroke wards, NGTs were described as rarely misplaced. On occasion, NGTs were found to be sub-optimally placed, whereby the tube is either coiled, needs advancing or pulling back; most commonly as a consequence of difficulty placing the tube or the patient dislodging a tube that is already in place. Its critical misplacement into the lungs, however, was described as very rare, with an estimate of 2 cases within a 6-week period on ICU. Almost all of these tend to be spotted prior to any feeding, meaning that so called 'never events' rarely ever occur. In fact, the majority of participants reported to have only ever encountered such a case once or twice – most often with no direct involvement (e.g., reported by a colleague, occurrence on the ward/ department). Asked to speculate about reasons for why a critically or sub-optimally placed tube was not spotted in time was attributed most often to (i) *human error* whereby either the wrong image was reviewed or the image was not assessed correctly; as well as (ii) *additional (technical) challenges* pertaining to poor image quality/ tube visibility and also specific patient factors and edge cases. Next, we detail these challenges, their contributing factors, and describe existing safety procedures and their limitations for ensuring safe image assessment.

Human Errors in Image Assessment. Errors in image review were commonly attributed to the person reviewing the CXR being unable to think clearly due to a busy, understaffed shift, or tiredness during the night. Errors are also bound up with lower image reader expertise and confidence, and lack of proper adherence to hospital policy and safeguarding processes (e.g., documentation protocols). Whilst more senior clinicians (e.g., registrars, consultants) generally described checks of NGT position as one of the easiest, most straightforward CXR assessment tasks, more uncertainty was expressed by more junior, less-experienced staff (D2, D3, D5, RR2), who would commonly ask for peer review. A reporting radiographer reflects:

> "(...) when I was first starting, even though I felt like I was sure it was in the right place, I was never confident it was always in the right place, so I'd always ask for someone else to check, but I think over time, once you've done enough of them and you've unfortunately seen ones that go into the lung, you can discern between ones that are in the right place and aren't in the right place a bit more." (RR2, ID)

To support confident image review practices, our analysis surfaced three main risk mitigations: (i) cultivation of a mindset of caution to prevent patient harm and frequent engagements in human peer review; (ii) mandatory, standardized reporting on EPIC via a template that enforces key visual checkpoints; and (iii) requirements for nurses to check the safe positioning of the NGT on the patients' nose and conduct regular aspiration checks at the beginning of their shift, and at 4-hourly intervals.

Mindset of Caution & Peer Review: Entrusted with their patients' care, staff described being cautious and vigilant, and to ask for a second opinion or conduct extra checks if they had any doubt about the NGT's position to avoid potential mistakes. An ICU nurse reflects on this error-preventing mindset:

> "So we just need to give them [the patients] topmost care, so I think each and everything, we don't need to go drastically. We need to take second opinion or some others opinion. Because if I have any doubt then I just need to stop there. I just need to escalate and I just need to take a second opinion on something like that. And we have a lot of ways to check. Because if anything wrong happened, at that time, we can't do anything. So there is a word that prevention is better than cure. So we just need to prevent everything." (N2, ICU)

Outside of specific reasons for uncertainty, it was generally also regarded as "good practice" (D4), especially for junior doctors, to consult ideally a senior colleague to review their assessment. In fact, some of the junior doctors stated they engaged in peer review "every time" they needed to document an NGT (e.g., "I just want to be 100% on this type of thing" (D1)). Yet, the ability to connect with a more senior staff as peer review can be more difficult at night or on a busy ward, which may potentially force a more junior, tired doctor to make a decision without additional human safeguards:

> "There could be emergencies in the unit, so the registrar and the rest of the team are busy with that, but your patient needs that NG tube confirmed. So that's when a potential never event could happen because the pressure of you need to confirm it but you don't have a person to second ask. So you might confirm it just for confirming sake." (D2, ICU)

Mandatory, Standardized Documentation: In the ICU, clinicians typically perform their own readings of CXR images to verify NGT placement and initiate necessary clinical actions before an official report is issued – often days later – by a radiologist or radiographer. Hospital policy mandates that doctors document on EPIC that the NGT is safe for feeding, ensuring they review the image and providing a legal trace. To minimize interpretation errors, clinicians must use the NGT-specific smart text reporting template (see Figure 2), which mandates responses to six image assessment checkpoints. Despite the existence of the NGT template, and its predominant use over free text reporting (which also exists), a recent audit conducted by one of the junior doctors (D2) revealed that only 83% of NGT templates had all questions completed. Describing the implications of negligence in properly following the review protocol, the doctor described a case where a senior doctor did not trace the tube path to check if it bisected the carina and went down the oesophagus; while the tube tip looked

like it was below the diaphragm in the area of the stomach, the clinician had missed that the tube instead had pierced side-ways through the lungs (D2).

Although mandatory requirements of written NGT documentation and template adherence can increase safety, we found that such safeguards also invite friction. For one, three clinicians remarked on the irritating design of the final visual assessment question that – through its enforced negation – could cause confusion then therefore lead to false answer. Secondly, mandatory documentations was also perceived to compete with desires to speed-up clinical decisions and patient feeding. In this regard, one ICU registrar describes the documentation and training requirements for NGT confirmation as "a lot of over-doing" and unnecessary for clear cases:

> "In my personal opinion only, it's a lot of over-doing. I've worked in other places and it wasn't done like that. We would put the NG feed in, we would use what is known as the whoosh[9], which is basically blowing air inside and listening to see if it, or feeling that it's in the stomach, if it's there and if you can aspirate anything out of it, then it's fine. We wouldn't document it. It's like a routine procedure that we do, we wouldn't document it. (...) I'm saying if there's a doubt and we would also order a Chest X-ray and we will always document what we saw in the X-ray. But we don't have like a special... I never used a special entry template. I never had to go through like a test to make sure I can do that. It's taken for granted that if you are a doctor that you are supposed to be able to look at a Chest X-ray, you don't need to be tested for it." (D5, ICU)

Continuous NGT Position & Aspiration Checks: Following NGT placement confirmation, ICU bedside nurses described how they would check the NGT daily, at every shift, and at 4-hour intervals, to look for any signs of displacement. Changes to measures of the tube length and checking if it is still securely attached to the nose can be key indicators that an NGT may have moved and be misplaced. Yet, such NGT measurements are not error-proof as a tube can, i.e., coil in the mouth or oesophagus:

> "The concern to me is that if it's sub-optimally placed, advancing and then I often ask them to re-advance it to check that it's not coiling. And that's often people will try to advance it and actually, the coiling applies in the mouth, so we've reached a sort of problematic, anatomically, down there, in that placement. So coiling above the X-ray position. (...) You generally can see that if you put 58 centimeters in [of tube] and it has just gone into the mouth. You can generally see that, but there are occasions where you advance it, and instead of advancing at that point, all that happens is you create a kink point within the mouth. I mean this is infrequent, this is but that's often why I would want to advance and check rather than just advance. So if its too short you would want to check that it has actually moved rather than. An awake patient would tell you there's something weird in my mouth, but it's the intubated ones." (D6, ICU)

Furthermore, ICU nurses responsibilities include '4-hourly aspiration' checks of the NGT as continued confirmation of its correct placement where possible. If no aspirate could be obtained, the result will be counter-checked by a second person, and, should results not match, a third checker. If in doubt, any feed that had been started would be stopped and the patient send for X-ray. Across the ICU and Stroke setting, CXR was perceived across hospital staff as a more definite, reliable test for assessing feeding tube placement rather than aspirate (e.g., "Everybody would rather do an extra Chest X-ray to check absolutely that tubes in the right position" (N1)). Especially in Stroke care, aspirate checks and required 4-hourly wait periods in-between unsuccessful tests were criticized for holding up X-rays and causing overall delays. Expressing frustration about the process and how aspirates, if not obtained successfully first time, most likely won't be obtained second time, the Stroke registrar shared:

> "And I also think, this is perhaps a bit more philosophical, but I also think that in introducing like little barriers and delays into something that is very essential, both to get right and also essential not to get wrong. It's really dumb to have delays. It's really dumb to frustrate people with the process. In other places, the guidelines have been able where it's like if there's any uncertainty, get a chest X-ray. And I actually really

[9]The whoosh test is a method whereby air is rapidly injected down an NGT while auscultating over the epigastrium. Listening to the resulting sound, a gurgling indicates air entering the stomach, whilst its absence suggests the tip of the NGT is elsewhere (e.g., lung, oesophagus) [39]

> don't understand quite how much of a barrier it is in this organization. I mean I can understand why it's evolved, but it wouldn't be my chosen thing. But it is what it is." (D7, Stroke)

To conclude, while the above processes and staff training are sought to *increase safety*, engagements in peer review; requirements for written NGT documentation; and (multiple) aspiration checks prior to CXR orders contribute to delays. In a hospital culture already perceived as very safety-oriented, this surfaced friction where the desire for patient safety competes with the need to expedite clinical decisions and patient feeding. This raises questions about how AI can meaningfully fit within, may compete with, or serve as a better alternative to existing safeguarding practices.

Technical Challenges in Image Assessment (Unclear NGT Visibility & Moderating Patient Factors): Finally, based on the interviews, we learned that image assessment difficulties and subsequent mistakes can be rooted in technical challenges pertaining to: *poor image quality* or *tube visibility* due to insufficient image penetration; difficulties to fully trace the tube path and its tip on the CXR; or external artifacts obscuring its view. Other patient factors, such as *obesity* and other *opacities at the lung base* (e.g., consolidation) that show as white on the image – alike the NGT – further hinder a clear view of the tube. Where the *patient is (hugely) rotated*, it can also: "give the impression that [the] NG tube is in a different place, so you're not getting the true position of it" (RR2). To trouble shoot those image quality and assessment difficulties, clinicians predominantly described improvements to image capture or viewing via: (i) the use of better X-ray machines, (ii) improve image capture settings or imaging modality (e.g., via CT), and (iii) review on higher-resolution monitors.

Lastly, we learned that patients can have an *unusual anatomy*, for example due to lung pathology, esophageal or stomach surgery. Where their anatomy has changed (e.g., due to a gastric pull-up or a gastroesophageal resection), correct NGT placement verification can be more difficult and usually requires a more individual assessment, oftentimes involving consultations with other clinicians (e.g., gastrologists) and reviews of the patients (image) history to get necessary context information to aid NGT assessment. In patients over the age of 50, a condition called *hiatus hernia*[10] is also very common, whereby their stomach moves up into their chest and can show above the diaphragm. Whilst deviating in their position from the standard protocol (Figure 2), the NGT can still be correctly placed into the stomach and be safe for feeding. Similarly, in very rare cases, a patients' stomach can be to the right, rather than left, caused by a congenital abnormality called *situs inversus*[11] whereby the person's major visceral organs are in a reverse position.

Building on these insights, Section 5.4 will expand descriptions as well as reflections on how understanding of image quality constraints; patient edge cases; and relevance of important patient history information present critical insights to data work in this space – e.g., pointing to potential data biases.

5.2 Perceived Clinical Utility of AI: A Complex Interplay of Multiple, Interwoven Goals and Constraints

This section details how perceived clinical utility of proposed AI functionality is bound up with (i) the target user group; (ii) AI outcomes that are being prioritized; (iii) workflow integration and AI design choices; as well as (iv) broader AI acceptance and adoption challenges.

5.2.1 Determining the Right Target User Group for the AI. Asking hospital staff in the interview to describe the end-to-end NGT verification workflow, we learned about the involvement of different stakeholders. This included ICU or senior Stroke clinicians; ICU nurses; imaging and reporting radiographers; and radiologists or radiology registrars. Considering each group as a potential recipient of AI outputs, we next describe their involvement

[10] Hiatus Hernia https://www.nhs.uk/conditions/hiatus-hernia/
[11] Situs Inversus https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8901252/

in the ICU CXR workflow, including additional reflections on the role that AI could play within their clinical practice.

Lower Involvement of Radiologists in NGT CXR Workflow: Contrary to the research team's initial assumptions that placed radiologists as the most likely user group (e.g., to prioritize clinical findings in their reading list), we learned through the interviews that interactions with radiologists or their text reports rarely featured within the ICU NGT CXR verification workflow. ICU clinicians and reporting radiographers would only reach out and call a radiologist if they had a very specific, urgent clinical question, which rarely occurs for any NGT-related queries except for very complex or edge cases (e.g., gastric surgery or very obese patient). In general, there was a sense that X-ray reporting, especially for NGTs "distracts the radiologist, who's reporting, from other things" (D8), such as more complex imaging (e.g., CT, MRI) or diagnostically more challenging tasks that require their specific expertise; suggesting CXR reviews for NGTs could be offloaded to other professions:

> "[About prioritizing critically misplaced feeding tubes in the reading order of radiologists] "I think that's not a bad idea. Only tricky thing about that is how many could they get and a lot of the time when it's not, I think obviously speaking from my experience, but I think a lot of the time when it is misplaced, it's quite clearly misplaced. So I'm thinking is it worth taking a radiologist time to check or can a clinician, trained looking at NG tubes, like for any of the SHOs[12] on critical care, just have a quick glance at it before escalating it to a radiologists." (D4, ICU)

Offloading to ICU Clinicians or Senior Stroke Staff: As alluded to in the above quote, doctors with permission to verify NGTs may present a suitable target user group of an AI alert to critical misplacement to speed up its correction especially where official radiology reporting is otherwise delayed. While ICU doctors are all specifically trained and usually the first people to respond in assessing CXR images to verify NGTs, this is not common practice across all hospital care settings, and may not easily generalize to other hospitals. For example, in Stroke care, where junior doctors are not allowed to assess the X-ray to make a safety-critical feeding decision. Instead, Stroke doctors describe how junior clinicians have to refer to a more senior staff, most commonly the registrar, to confirm and document the NGT. In the absence of senior colleagues, Stroke doctors – in contrast to ICU doctors – therefore rely on the official radiology report. This official reporting is either provided by few (~8-9) reporting radiographers, or by radiology registrars or radiologists, who report the NGT CXR as part: of acute reporting; radiology registrars training; and to increase reporting capacity (there are ~120 consultant radiologists employed at the hospital).

Upskilling ICU Nurses: In the interviews with ICU nurses, as the ones who place the NGT, we also discussed the possibility of them becoming the recipient of an AI alert. ICU nurses described to have a good understanding of any difficulties surrounding the NGT placement, and often had already build-up some understanding of recognizing NGTs on CXRs. While this suggests that (ICU) nurses could be upskilled to take on NGT CXR reviews, they are currently *not authorized* to access radiology images (and cannot access PACS), nor would they have the clinical expertise to potentially detect other medical conditions of the patient that may show on the X-ray and fall into the diagnostic remit of doctors; thereby blurring diagnostic boundaries and medical-legal responsibilities.

Radiographers' Role in Early Intervention & for Generalization Across Care Settings: As a user group, we distinguish the roles of imaging radiographers, who operate imaging equipment to capture X-rays (and other medical images); from reporting radiographers, who interpret and report on those captured images. It was common in the reports of different staff that imaging radiographers often spot NGT misplacements straight way, at point of acquisition (see workflow diagram, Figure 1), when it is clinically most relevant as it can be acted upon immediately – whereas reporting radiographers may not see the image until days later. Imaging radiographers can also alert nurses to the need for immediate review or re-insertion of a misplaced tube:

[12]SHO stands for Senior House Officers and represents an umbrella term for multiple doctor grades in the UK.

> "(…) sometimes it's caught by the radiographers, the radiographers are very good at this and they can spot this straight away and then they would contact the clinician straight away to inform it, because that's why, actually probably the most important thing is that it's caught straight away, because by the time we've got to us as reporting radiographers. As a reporting clinician, it could be 2 days down the line. Patient may have been fed by that point. So the crucial point, the crucial point, is at that point of acquisition." (RR2, ID)

When considering that feeding tubes can be placed anywhere in hospital and by many different staff, upskilling all those professions to NGT CXR assessment was described as unlikely to scale. Instead, imaging radiographers, who take the X-ray image, were suggested as the "only constant" (RR1), involved in NGT CXRs at day and night, and across any ward and hospital in the world; suggesting their upskilling to this task:

> "Also we're looking at radiographer-led as well. So we can look at doing some training for the radiographers and if they're on the ward, at night time or something like that and then they will assess it [the NGT] and make their judgment themselves which is not quite off the ground yet, but we're looking at it." (RR3, ID)

Currently, imaging radiographers without postgraduate education however are not authorized to provide a radiology report. Nonetheless, it is expected of their professional competence at the point that they appraise the image – should they identify critical findings on an examination – to take direct and timely action by alerting the referrer[13]. Furthermore, one of the consultant radiographers described an initiative that plans to standardize NGT reporting and build competencies for all imaging radiographers in England in the near future:

> "(…) And that intervention is aimed at standardizing training and education with uniform competency assessment for all radiographers who take chest X-rays where a nasogastric tube maybe present. So the diagnostic radiographers you would have met in ICU. The ambition is to train everyone of those, in England, to be able to identify a nasogastric tube on a chest X-ray and make a safe to feed, not safe to feed decision basis. (…) So the Royal College of Radiologists and the College of Radiographers facilitated by HEE [Health Education England] have commissioned a piece of work that is active at the moment to do that stepped process." (RR1, ID)

In contrast, CXR-based NGT checks feature high in the workload of reporting radiographers, who reported that approx. 10% of their inpatient reports include requests to confirm NGT placement. Despite this prevalence, reporting radiographers tend to report predominantly within main-working hours, and are currently not part of the main pathway for clinicians to reach out to for any urgent NGT CXR reviews. Instead, reporting radiographers described to come too late into the image review process, stating that any misplaced NGT would already been identified by the ward clinicians. In few instances where radiographers were first to spot a misplaced tube, it was because clinicians hadn't yet looked at the images themselves.

Furthermore, reporting radiographers explained how they tend to report more narrowly – for example only chest and skeletal X-rays – compared to radiologists, who also report on other modalities (e.g., MRI, CTs) and subspecialities (e.g., brain, abdomen). This narrower reporting focus was suggested to be more suited to prioritizing critical findings in their reporting worklist given the smaller range of possible acute findings that could surface and otherwise compete with each other for urgency:

> "If you are a consultant radiologist working on the acute desk and you are reporting a MRI spine for cord compression. You are reporting CT chest for pulmonary embolism. You are reporting a nasogastric tube for placement confirmation, and you are reporting a CT abdomen for perforation. How many prioritization algorithms do you run? And what trumps what in a worklist? So is your bleed on a CT brain more or less urgent then your misplaced nasogastric tube, which is more or less important than your pneumothorax, which is more or less important… Where one, I'm [a] simple creature. I can report a handful of things, and therefore I'm easy to prioritize, because I do this few things." (RR1, ID)

To conclude, this section highlighted the prospective benefits and feasibility constraints of different AI target user groups. While an imaging radiographer could help identify a critically misplaced NGT early on the X-ray

[13]https://www.hcpc-uk.org/standards/standards-of-proficiency/radiographers/

machine, integrating AI functionality with vendor software is likely more complex. Current hospital policies and reporting protocols also do not *yet* permit their involvement in this verification and reporting task (in writing). Therefore, proposals to alert the clinical care team early about a likely misplacement, allowing them to act or mobilize a radiology review, and prioritizing CXR reporter work lists seem more aligned with existing care pathways. In Section 5.3 below, we continue to map out the implications of these choices on staff practices and patient risks and benefits.

5.2.2 Balancing Across Goals of: Patient Safety, Assessment Confidence and Process Efficiency.
Next we detail how participants perceived the potential utility of our five different AI proposals that describe desires for: (i) improving patient safety; (ii) staff confidence in image assessment; and (iii) overall process efficiency; and we outline some of the tensions that can exist between those goals.

Patient Safety as the Most Important Requirement. When asked what type of AI application participants would find most beneficial; they recognized value across the various AI proposals, with some even expressing a preference for a combination of different AI capabilities to provide more effective safeguarding support in the NGT verification process. In particular, staff emphasized *patient benefit and safety* as the most prominent driver for clinical care. As a result, AI functionality that (i) assists the detection of sub-optimal or critically placed NGTs – either as an alert directed at (ICU) clinicians or for image review prioritization (especially in Stroke care) – or that (ii) flags up potential reporting errors was ascribed most clinical relevance to speed up immediate problem detection.

Early Detection of NGT Misplacement (for ICU Clinicians or in Prioritizing Reporting Lists): Having an AI automatically alert or flag-up an abnormal NGT early in the verification process was described as "definitely helpful" (D7) as it (i) enables *immediate action*: ICU doctors would "go and look at that image straightaway" (D4). An early misplacement alert would also (ii) invite extra "caution" (D2) and closer inspection (e.g., via peer review), which was regarded to improve patient safety by *reducing risks of human or contextual factors* that can cause for a misplaced tube to be missed and the patient being misfed. All of this also *reduces overall delays to patient feeding*:

> "I think it could only be a good thing, right? Because sometimes patients are waiting days before they get fed. So if this is highlighted straight away and it's cutting down a risk of misfeeding, I think it would be great if you, even if it had to wait 5-10 minutes for somebody to put in the official report. I think it could only be a benefit really." (RR3, ID)

In Stroke care, where junior clinicians cannot verify CXR NGTs themselves but rely on the official radiology report, the proposal to re-prioritize the reading list of image reporters to speed up assessment of critical NGTs was also regarded as "extremely helpful" (D7). Reporting radiographers too described prioritizing critical examinations that needed their attention as one of the most beneficial and "appropriate" uses of AI (RR1), suggesting that amongst a list of 200 patients: "(…) if there was something that could hopefully highlight the ones that needed more urgent attention, that would be great." (RR3). Alike early alerts to ICU clinicians, sooner reporting due to AI prioritization of critical NGTs was proposed to *reduce delays in taking corrective actions*, and in *subsequent feeding or medicine administration*. It was also considered to *reduce risks of inappropriate self-reporting* due to formal reporting delays that otherwise could present a safety concern:

> "In practice, reasonable wards would not be feeding until they've got the report, and really the problem is delays. And in practice, whoever is reading the report, in practice good systems wouldn't take matters into their own hand and attempt to report the scan themselves, even if not competent. But I completely understand if you have a system immediately for flagging that the tube is in the wrong place, you are minimizing the chance of either feeding happening, assuming it's alright, delays to feeding and delays to medicine, and the last which would be inappropriate self-reporting because of delays to formal report and therefore wrong conclusion of safety. All of those sound straightforward potential benefits of the technology." (D8, Stroke)

Upon reflection, despite its potential, an open question remains to what extend an early NGT misplacement alert could indeed speed-up clinical review or report generation – and consequentially correction times. On ICU, nurses tend to chase doctors to review the image as soon as is feasible, as they too have an urgency to work through their tasks; whilst doctors, attending to a busy ward, may not be able to review the image any earlier than they currently do. Similarly, in Stroke care, even if a critically misplaced NGT was brought higher up in the reading list of a radiographer, there may still be a long reporting gap, if a review was requested out-of-reporting-hours.

AI for Error Checking: Similar to critical placement detection, having the AI flag a potential reporting error prior to the report or clinical note submission was described by clinicians to invite caution and a closer look, and to serve as a useful "little backup" (RR3) and "extra layer of security" (D4) to *prevent human errors* of misreading the image or accidentally clicking the wrong answer option on the .NGT template caused by rushing, distractions, or inexperience:

> "(…) I think for us doctors the most important thing is patient safety as opposed to efficiency and trying to get things done fast and trying to identify things quicker, and workflows, and all these kind of things. I think for us it's safety and the only one that really safeguards, as a safeguard, is that one option [error-checking], because it makes you take a step back and think again." (D4, ICU)

Benefits to have AI serve as a second checker of their X-ray assessment was especially valued by ICU doctors with lesser experience in assessing NGT placement, who may feel pressured to decide or may not have access or time to consult a (senior) colleague. Here, having the AI take on the role of a second checker was suggested to reduce overall NGT confirmation time as it: "would probably cut down the time that you might check with a senior colleague when you didn't have the confidence to know" (D3).

Upon reflection, while alerting to a human error in image assessment could catch mistakes that otherwise can have significant implications on patient safety, uncertainty remained in staffs accounts about the prevalence of actual reporting errors for NGTs. If at all detected, reporting errors are not commonly captured outside of any *official critical incident reporting* and documentation practices, leaving this as an open challenge for data work.

Assisting Image Assessment Confidence as 'Nice to Have': Having an AI produce a visual overlay on the CXR image, where its "literally highlighting where it thinks the tube is" (D3) including its path and tip, and key relevant landmarks (e.g., the carina bifurcation point), were described as helpful for *improving image assessment confidence and speed* – especially in more junior clinicians; and for simplifying the completion of the .NGT template.

> "It would make the workflow much easier. Because, some of the difficult times, like, especially if it's junior colleagues, they would be delaying the documentation because they are not complete 100 percentage sure. So if AI can provide a certain amount of assurity to it and they can also visualize OK, this bisects the carina, and it's below the diaphragm then it definitely reduces the time in which it is confirmed." (D2, ICU)

Despite these proposed advantages in assisting reporter confidence and speedy reporting; there was general agreement across hospital staff that feeding tubes are commonly "quite easy" to identify (D3, D4); especially since they present the only tube that extends that far down into the body, and when compared to other, more complex neck and PICC lines. Similarly, key landmarks like the stomach, which tends to show as a gas bubble on the CXR, or the patients' diaphragm (Figure 3), are regarded as simple anatomy for a doctor to locate; all of which questioned the extend of the usefulness of an AI overlay, outside more complex patient cases: The Stroke registrar draws a parallel to other AI software:

> "Yeah, I mean, I personally I think that [the AI overlay] would be really good. Number one, I'd find it interesting, I guess another question about whether I'd find it like helpful, but I would find it interesting. I would definitely, if the image was there, I would definitely look at it. To draw an analogy with the CT scans

> we do in Stroke, I'm always intrigued by what the brainomix[14] – I don't know if you've ever seen it? I always scroll through. Basically you have the scan that we look through and then there's another like basically an extra study which is overlaid on the scan, its the AI's interpretation of the scan. And I do look through that, cause I do think it is interesting and once in a while, what it flags as an abnormality, I would have a second look at." (D7, Stroke)

Consequently, when weighing-up our various AI proposals, staff ascribed more clinical utility towards proposals that more directly and immediately aided the detection or critical findings or human errors, as opposed to boosting staff confidence or reporting speed:

> "If all that's gonna happen is you're going to make me as a reporter 1% better. No one's perfect, right? I'm not perfect. AI is not perfect. No one's perfect. We accept that people get things wrong. (…) If you are just making experts a very, very, very tiny bit better, that's good, but that's not where the money is. The money is identifying or prioritizing those examinations, the critical finding, who need your attention first. So that's useful for me, because I just report Chest X-rays and Skeletal X-rays. So if you identify critical findings, of either of those two things that needs to be the next thing that I report." (RR1, ID)

Increasing Reporting Capacity (or Speed), but not at the Expense of Safety: While the above proposals of AI assistance in critical findings prioritization and AI error checking serve as safeguards to clinical assessment and to increase patient safety, they do not increase image review capacity or reduce image backlogs that contribute to official reporting delays. However, proposals of having an AI auto-generate a (preliminary) NGT report or have it pre-fill the .NGT template received mixed responses. In terms of clinical utility, participant feedback varied depending on how auto-report functionality was implemented to either: (i) *assist* or (ii) *fully replace* existing NGT image review practices.

Weighing-off Benefits and Risks of AI-Assistance in NGT Reporting: Where AI image interpretation and reporting functionality is intended to assist clinicians review practices (e.g., ICU doctor, reporting radiographers) doubt was expressed about it providing much utility in terms of time-savings or to guide clinical decision making since clinicians "would still look at the x-ray [them]self" (D7) and spend time to check, verify, and if necessary, correct the AI-generated report. Furthermore, staff described that they would still consult with a more senior colleague if there was uncertainty about their assessment of the image – despite AI input. Consequently, there was agreement that, as long as it remains the responsibility of the clinician, radiologist or radiographer to sign off on the report, that they would do their "own checks" even if, ultimately, they state at the bottom to "agree with AI interpretation" (D7):

> " (…) let's say the algorithm said it's in the right place like how much confidence would I have to agree with that algorithm, or would I still going to have a look at the X-ray myself? And then, if I'm looking at the x-ray myself then how much help is the algorithm, because essentially if I'm putting my license on the line for this NG placement, then how much do I trust the algorithm to get it right? And I think that's kind of the issue. (…) personally for me, I would still always look, even if there was some sort of system in place to say that it is in the right place." (D4, ICU)

Simultaneously, if NGT auto-report functionality existed, doctors expressed concerns they might become "lazy" (D2, D3) and either might not open the X-ray at all, if the AI for instance directly auto-populated the clinical note (D2); or might not "always always always check everything on the protocol" (D3).

> "So the only problem then is there's a chance that they might not open the X-ray, we are lazy people. Maybe you need to at least open the X-ray to get it auto-populated into the thing, because if it auto populates into the note directly then we might not check the X-ray. It is a possibility, especially on busy nights and stuff like that. There's a chance that if AI does that automatically, so at least I think." (D2, ICU)

[14]https://www.brainomix.com/

Furthermore, even if humans check and always have the "final say" on assessment (RR2), risks remain where the AI errs, but clinicians, especially more junior ones (D7, RR3), may "not [be] confident in overruling" what has been flagged by the AI (RR1), or be "brave" enough to go with their "own convictions" (RR2).

> "I think I broadly like it. If someone had asked me to look at that, the chest X-ray, and an AI report was already there. I would read the AI report. I would definitely read; I wouldn't just ignore it. I would. And I would still look at the image itself, look at the image with my own eyes. I guess the question comes in, if there is a more junior or less confident colleague who looks at the AI report, is less confident about disagreeing with the AI report. If they look at the picture themselves, where does the responsibility lie?" (D7, Stroke)

Accounting for Medical-legal Responsibility of Full Automation in NGT Reporting: Where AI use is envisioned as providing good enough performance to automate NGT assessment or reporting, some staff regarded this as useful to offload work and save ICU clinicians time:

> "I can tell you that if the machine was doing it [the NGT assessment], then it would offload us from doing it and then nurses will just read the machine thing and start using the NG. So the machine would write that the NG is safe to use and it will spare me 50 minutes of my day." (D5, ICU).

However, for this proposal to come to its fruition, it would require for the AI to also take over "full responsibility" for the medical assessment. One of the junior ICU doctors explained:

> "(…) if we're giving the sole responsibility of a patient to a computer program, then fine, but if we're not, then I don't really see where, apart from being a helping hand, I can't see it being that useful, if that makes sense, and that's kind of my take on it." (D4, ICU)

Simultaneously, AI automation use cases like automated image assessment and reporting also face regulatory and legal challenges that limit its practical realization, as highlighted by one of the radiographers:

> "At the moment that is illegal, it's not compliant with radiation legislation in the UK. So anything will only ever be of clinical decision support or a triage tool. Triage and decision support doesn't create capacity, it just reallocates existing capacity because you still have to report everything." (RR1, ID).

5.2.3 Workflow Integration & AI Interaction Design Considerations.

This section illustrates how the utility of different NGT AI proposals is bound up not only with its integration within existing workflows, and also human-AI interaction configurations and design choices that surface tensions between desires for seamless, unobstrusive AI integration and needs to check and verify the AI.

Enabling Seamless AI Workflow Integration. In our conversations about AI integration within current practice, staff described the need for efficiency and seamlessness in workflow integration to not impede reporting speed or flow. For instance, in the context of reading list prioritization, reporting radiographers described the desire that high priority cases, like a misplaced NGT, would be flushed "straight up to the top of our list, so it comes up as our next patient" (RR3) to not "interrupt my workflow" (RR1) and instead to auto-load the next case, once review is finished and signed:

> "My personal preference, again this is not science, is that having a report prioritization to top of work list is useful and so it doesn't interrupt my workflow. So I'm reporting a case and finished my report. It then auto-prioritizes into the reading worklist that when I hit sign, and then it auto loads the next case for me, that case is the next one that it loads. So seamless prioritization in the worklist is really important. I don't wanna have to keep going in and out of things." (RR1, ID)

Reflecting on the placement of an AI alert to a critical findings, staff radiographer's proposals varied from notifications upon opening the patient page within EPIC; through to its integration as an AI tab within the image viewer PACS; and annotations directly to the radiology image to bring extra caution to image review practices. By way of example, one of the registrars drew on historical radiographer practices whereby they would put a "red dot" on a physical X-ray image to indicate that a scan was *obviously abnormal*; which made doctors second check. He described how in some hospitals this practice still existed digitally, through a white text box stating

"Red dot" or an "asterix" that is put on the image, indicating: "(...) This study contains a critical result. I find that helpful, so it's not intrusive, but it's unavoidable if you're looking at the scan." (D7, Stroke).

While direct image annotations can bring key AI outputs to the forefront, in the context of AI producing a visual overlay of the NG tube, reporting radiographers also remarked on the importance for AI outputs to not interfere with their visual search strategy through the image that might prohibit the finding of other injuries or pathologies. Instead, they expressed preferences for the AI to show as a toggle or second image within a separate AI image tab – as is commonly seen with other AI software (D7, RR2).

Furthermore, design considerations such as (i) the *timing or ordering of AI results* within human workflows; or (ii) the *level of clinical AI interpretation* of the image are bound up with risk and concerns about AI over-reliance, and human skill acquisition:

Designing Human-Led AI Interactions. For instance, discussing how AI functionality, like an error checker would become integrated, staff expressed a preference for doctors having to make their assessment first, before receiving an AI response to not interfere with their own clinical reading, nor with the formation of key medical skills and competencies in CXR image assessment:

> "So if we say it's a very good system, but it's not perfect. And it wasn't generating huge amounts of work, as in, there weren't masses of discrepancy in general between what it [the AI] was finding out and what you were finding, then having an AI safeguard would be helpful, yes! I think if you've done it first. ..and I think what would also be useful in that scenario is that as a clinician you wouldn't deskill, you would still be forced to look at the image. You would still have to make a decision. You wouldn't rely on technology to make that decision for you. So you still are competent in assessment, which I think is a skill." (D3, ICU)

However, reflecting on earlier insights: any requirements of clinicians to make their own assessment at first, or having to double-check the AI generated (report) output, limit proposed AI efficiency and utility gains.

Focusing on Lower-Level AI Interpretation. In the context of how best to articulate a critical AI finding, most clinical value was ascribed to an understanding whether the tube is unsafe for feeding. In this regard, one of the radiographers expressed a preference to communicate the AI output as a binary finding (Is it safe to feed: Yes/ No) without any context why the AI made that decision such that he still needs to critically appraise the X-ray. This is sought to reduce automation bias and over-reliance on the AI (RR1).

> "The most important question that AI would be able to answer is whether it's unsafe for feeding. So for example, if it's not bisecting the carina, the AI would be able to pick it up and it could potentially prevent a never event from happening." (D2, ICU)

Others preferred for the AI to only flag-up and alert towards an abnormally placed NGT rather than for it to interpret the image – by stating about the NGT that "it's safe" to use (D7) – to reduce risks of any negative implications in instances where the AI was incorrect. Given that the role of image reporters is to interpret the image and to "only say what they see (...) [not] decide what should be done with the patient" (D5). This suggests the role of AI to be better placed and to be lower-risk if focused on *assisting* image assessment; thereby leaving any more interpretative medical conclusions – whether feeding is safe or not – to the clinical team.

Emphasizing Human Skills Acquisition. Many of the staff also described the potential of using AI, especially the visual overlay, to assist in NGT training or skill development scenarios. For example, they proposed for AI to: (i) show the "optimum position for the tip" (RR3) on the image; (ii) learn where tube tip needs to be (how far advanced) to provide guidance to junior doctors; or (iii) link to demonstrations of the correct way to textually describe the NGT. Simultaneously, they raised concerns about how AI use might complicate skills acquisition. Reflecting on the question of how novices of today become experts of tomorrow, staff emphasized the importance of developing appropriate internal feedback loops that ensure clinicians won't repeat certain mistakes. They also cautioned not to train people to report *only with AI*, which risks AI over-reliance. To better illustrate how AI use

could risk de-skilling staff, one ICU registrar drew a parallel to ECG auto-impression, which may lead staff to not take the time to think, and can remove opportunity for developing important image interpretation skills:

> "I can tell you on the ECG today, the machine is calculating all kinds of stuff and coming up with diagnostics, like writing 'patient has this'. I never read it. Because it's very often wrong, and because it makes you not think. The same like when you download an article. When you make ChatGPT write your assignment, then you didn't learn anything. Maybe I'm old-fashioned. (…) if all my life I looked at what the machine was interpreting I would never know how to interpret an ECG. (…)" (D5, ICU)

5.2.4 AI Acceptance more Broadly: Setting and Managing Appropriate AI Expectations.
Lastly, clinical utility of our AI proposals for assisting NGT placement verification is bound-up with AI performance requirements as well as overall AI acceptance and adoption by medical professionals and health organisations.

Need for High AI Performance: Surpassing Human Capabilities for AI Utility. Across AI proposals, clinicians described chances of the AI being incorrect or to miss important instances as the "main risk" (D3) and "worry" (D4); even if an AI system was "really, really good" (D7). Especially for more ambitious AI concepts such as auto-generation (or auto-completion) of the NGT report (template), staff expressed "doubts" (N2) and a lack of trust in AI's capabilities, suggesting it being useful: "if we can show that it is 100 percentage accurate" (D2), which is likely an unrealistic expectation of AI to fulfil. Probing deeper into what would be considered as *good enough performance for AI to take full responsibility in NGT assessment*, one of the ICU clinicians suggested it needed to be "better than a human interpreting" (D4) – alike computers that beat human chess players (D5): "It may mean that we may trusting something that may fail, but humans fail. So who is more likely to fail?" (N4). As described above, in cases where technology performance at least surpassed human capability, full-automation in NGT reporting was perceived to potentially offload work and save ICU clinicians time (if medical-legally permitted).

Costs & Disruptions Incurred by Imperfect AI. Even for less risky AI uses, such as an AI error checker application, clinicians emphasized the need to achieve good enough performance to not overly constrain human processes in cases where it is incorrect. Technically, cases where the AI may flag an otherwise unnoticed human error, should only occur very rarely. Thus, flagging-up disagreements in cases where clinicians were competent to make the assessment, or the AI was incorrect, can cause additional overhead and reporting delays (e.g., requiring additional peer review); and may mean AI alerts are becoming ignored going forward:

> "What I don't know is, how many times do qualified professionals say an NG tube is appropriately placed and then the AI disagrees. How often does that come up? Because it might be quite annoying if I look at 10 X-rays, and all of them, the AI disagrees with me, but it all ends up being like they're fine. In which case people just gonna click through that alert and like, never bother with it." (D7, Stroke)

Desired AI Performance: Theoretical Ideal vs. Reality. When discussing areas of opportunities for AI, we noticed tensions in staffs' accounts between clinicians suggesting its use to support instances that were difficult for humans to assess – for example, have AI "add an extra layer" of insight (D4) that could save clinicians having to call a radiologists, thereby reducing demands on their time – especially for more complex patient or poorer image quality cases – whilst simultaneously expressing doubts in AI capabilities to perform accurately and reliably in these cases.

For example, most clinical utility for a visual AI overlay was ascribed to recognizing the NGT path in *more complicated cases* where patients either: (i) have had a poor posture during image capture (e.g., ICU patients are often slumped badly, N4); (ii) have multiple tubes or lines – internally or externally – overlying their chest, complicating their differentiation; (iii) have other pathologies (e.g., big effusions) that can complicate the identification of key landmarks such as the diaphragm; or (iv) present with an unusual anatomy. Reflecting upon this, there is an open question however, how well AI would perform in those more complicated, often edge cases. Given that even expert clinicians and official radiology image reporters can be unsure about the NGT placement

in these instances (e.g., asking for better image quality), this may also limit their ability to verify and contest the AI output, posing further risks in cases where AI results could (more likely) be false.

In general, our findings indicate how it can be difficult for clinicians to frame realistic expectations of AI. For example, while one ICU clinician could see the prospect of AI making their "life easier" (D5) and described new AI capabilities (like ChatGPT) as "extraordinary" (D5) she was mindful this technology is still new, thus regarding such AI capabilities as more of a *theoretical proposition rather than a practical reality*:

> "(…) and it's still just starting, so it's not perfect, I mean. Right now, AI is also making up a lot of nonsense. It's just not true. Yeah, but here we are discussing [a] theoretical thing. If it works, good, I would like to have it. If it works 50% and the other times it's lying completely, then I wouldn't wanna use it." (D5, ICU).

Participants poorer performance expectations and lack of trust in AI capabilities are partially grounded in experiences with other prior technical innovations. Making reference particularly to patient electrocardiogram (ECG) readings that have an auto-generated impression of the ECG signal printed at the top of paper strip (e.g., "This ECG shows ST elevation across the anterior leads"), ICU doctors described how they would still make up their "own opinion" of the ECG, even "recalculating" the machine outputs (D4), or not reading its impression at all as they found it many times to be "wrong or out of context" (D5).

> "So I think that goes back to what I was saying about ECG, because that's essentially what the ECG machine does. It gives you a preliminary report, but despite that, every time I'd still look at the strip myself and even some of the calculations, I recalculate them by hand. Because, I mean, at the end of the day, obviously the ECG machine works more on plan recognition as opposed to obviously using big data or what, so it's quite hard to trust exactly what's being presented to you. (…) It does get it wrong sometimes. So you see things that are reported or preliminary reports on the ECG strip and you're going: well, that's clearly not what's going on with this strip. So I think that's where the double check is needed unfortunately." (D4, ICU)

Broader Considerations for Managing AI Expectations & Acceptance. Lastly, our participants also touched on broader considerations for how AI acceptance and adoption would need to be achieved, including: (i) need for AI education for staff to develop appropriate AI literacy; (ii) requirements for effectiveness and practice integration studies to better understand the (negative) effects of AI use on people's reporting practices; (iii) desires for carefully staged approaches to AI performance assessments and its continuous monitoring; and (iv) clarity on Trust policy on AI responsibility and how much certainty clinicians can give to the AI outputs:

> "I think it really depends on what the Trust decides is like a level of certainty that we give to the AI report. Because I think if junior doctors were told the AI report is just a guide, but it's not sufficient, they would have a very different attitude to it, for if they were told we have faith in the AI guide. If you're not sure, speak to a senior. I think those like there's, like, very different responsibility burdens of proof, depending on how the AI is framed." (D7, ICU)

All these considerations shape appropriate expectations of AI and how it becomes adopted within clinical practice.

5.3 Mapping Intended AI & its Potential Implications to Distill Clarity and Facilitate Discourse

The above research insights paint a complex picture of the design space for AI-assisted NGT CXR support within an ICU hospital workflow. Extrapolating from this, we next exemplify how these insights can be mapped out in more systematic ways – as illustrated in Table 3 – to help research and development teams create more clarity across different intended AI uses; design configurations; and their implications on direct and indirect stakeholders to better weigh-off prospective AI benefits and risks.

For the use case of AI assisted NGT misplacement detection, our example mapping shows how seemingly the same functionality enabled by AI – *the detection of a (critical) NGT misplacement* – could become incorporated at different workflow stages. For each of three example instances depicted in Table 3, the various implementations would involve a different target user group (imaging radiographer vs. ICU clinician vs. radiologist or reporting

radiographer), different software integrations (e.g., X-ray machine vs. EPIC alert vs PACS viewer) and design choices (e.g., alert notification vs. worklist reordering). To help think through the potential benefits and risks of each instantiation, we chose to map desired AI performance of correctly identified (critical) misplacement events against likely errors (e.g., false classification of correct NGTs as misplaced), and reflected on its direct impact on staff (practice) and indirectly, on patient outcomes. This clarified that the use case likely offers patient benefits, with comparatively lower risks as long as AI performance is optimized to avoid the misclassification of correctly placed NGTs.

The break-down also shows that clinical relevance and patient benefits of a misplacement alert are likely highest at the point of image acquisition and may more easily scale across care settings (e.g., beyond ICU) given the consistent role and presence of radiographers at CXR capture. However, current technical and medical legal-constraints suggest lesser practical feasibility for this instantiation. Whilst reporting radiographers and radiologists are most qualified to appropriately review CXRs and critically appraise AI outputs (to avoid risks of AI over-reliance), their often late reporting of ICU NGT CXRs suggests that ICU doctors may be the most beneficial target user group to act upon a timely alert for greater clinical utility. We may conclude that alerting ICU doctors early to potential NGT misplacement with high AI confidence is perhaps the most promising application for improving patient safety. This approach fits well within current workflows and medical-legal constraints. As discussed further in Sections 6.2.2 and 6.3, decisions on next steps however should also consider broader stakeholder needs, economic impact assessments, and real-world data production practices.

It is worth noting that our mapping example suggests a simple, binary AI misplacement classification with a restricted scope of possible errors. For other use cases, such mapping would benefit from extending. For instance, an AI-generated NGT visual overlay may surface a range of AI errors such as: missed, false or broken path tracing; no or false indication of the tube tip; or NGT misclassified as a different tube type. For NGT report generation, error types likely extend even further such as: errors in distance quantification when specifying tube location. Likely those different error types can have more or less severe implications if remained undetected. For instance, while the misclassification of an NGT as placed correctly, when it is however leading into the lung can be fatal to the patient's life; a technical mistake in misclassifying a *naso*gastric tube as an *oro*gastric tube (a feeding tube that is inserted via the mouth, not nose) however – that nonetheless is correctly identified in the area of the stomach where it should be – is unlikely to pose any significant risks to patient safety. This suggests the need to bring closer attention to developing and deploying more human-centered AI evaluation methods that help formalize and systematically assess the (likely) clinical implications of different system errors above and beyond technical performance metrics.

FINDINGS PART B: DATA OPPORTUNITIES & CHALLENGES FOR AI DEVELOPMENT AND EVALUATION

This finding section details learnings about existing NGT hospital data – it's production, availability and characteristics – for NGT-specific AI development (Section 5.4) as well as insights from reviewing data records for choosing or defining relevant outcome metrics (Section 5.5), if a prospective NGT AI application was to be deployed and studied within healthcare practice.

5.4 Opportunities and Challenges for Data Preparation, Labelling & Interpretation

This section illustrates how insights into real-world practices surrounding NGT image capture, review and reporting can surface challenges for, as well as assist in processes of, effective data preparation, labelling and image interpretation.

Table 3. Example of mapping a proposed AI functionality across different workflow stages; illustrating variances in user groups, design choices, potential benefits and risks of desired versus erroneous AI outputs on staff and patients; and broader context considerations.

AI Capability	Workflow Stage	User	Design Choice	AI Performance	Potential Implications on Staff/ Practices	Potential Implications on Patients	Broader considerations
Detection of (critical) NGT misplacement	CXR preview image	Imaging radiographer	Alert on (mobile) X-ray machine	Correct critical or sub-optimal misplacement detection	*Benefits* • Earliest possible review of CXR to detect NGT misplacement (when its clinically most critical) • Informing clinical team to review for sooner NGT documentation and correction • Most Consistent user group across care settings (all CXRs beyond ICU)	*Benefits* • Highest impact on reducing risk of missed NGT critical (or sub-optimal) displacement + related complications • Reduced delays to patient feeding via more timely correction	*Technically feasibility:* requires AI implementation within radiology equipment software *Medical-legal constraints:* imaging radiographers often not permitted to report NGT CXRs
				False alert: NG tube is correctly placed	*Risks* • Alert fatigue • Unnecessary disruption to doctor/ nursing workflows • Increase in time spent in image review or peer review causing delays	*Risks* • Other (potentially urgent) patient imaging / care becomes delayed	*AI performance:* avoid misclassification of correct NGTs in model optimization
	CXR image review	ICU doctor	Critical findings alert in EPIC	Correct critical or sub-optimal misplacement detection	*Benefits* • More immediate review of CXR image on PACS • Earlier detection + action to correct critical NGT misplacement • Reduced risks of critical NGT misplacement remaining undetected (due to human or contextual factors)	*Benefits* • Reduced risk of missed NGT critical displacement + related complications • Reduced delays to patient feeding via more timely correction	
				False alert: NG tube is correctly placed	*Risks* • Increase in time spent in image review or peer review causing delays • Risk of misinterpretation of tube location • Confusion about AI output/ loss in AI trust	*Risks* • Other (potentially urgent) patient imaging / care becomes delayed	
	CXR reporting	Reporting radiographer/ Radiologist	CXR image with (critically) misplaced NGT is prioritized at top of PACS reading list	Correct critical or sub-optimal misplacement detection	*Benefits* • Earlier official reporting of (critically) misplaced tube that reduce risks of inappropriate self-reporting (e.g., by more junior doctors) • Reduced risks of critical NGT misplacement remaining undetected (due to human or contextual factors)	*Benefits* • If detected prior to clinical team: reduced risk of missed NGT critical displacement + related complications • Reduced delays to patient feeding via more timely correction	*Workflow:* Often (in ICU care), clinical team will have already acted upon NGT CXR prior to official report especially if reporting only happens within-hours
				False alert: NG tube is correctly placed	*Risks* • Competes with other critical urgent findings that can become deprioritized	*Risks* • Other (potentially urgent) patient imaging / care becomes delayed	*Prioritization:* Difficulty managing different, potentially competing 'critical findings' in urgency triage

Table 4. Summary of patient factors and potential confounders in image or report data that can implicate AI analysis.

Edge-cases & Confounders for AI Data Analysis	
Image quality (acquisition)	Poor image penetration/ exposure (e.g., image capture settings) Lower image resolution (e.g., type of imaging system) Image annotations (e.g., AP ERECT SITU, SPINE ITU) Patient rotation
Patient factors	Impaired consciousness (e.g., Patient Glasgow Coma Scale) Patient obesity (e.g., BMI) Chest abnormalities that show as opacities (e.g., lower lung consolidation) Other patient history/ conditions (e.g., gastric surgery, hiatus hernia, situs inversus)
Radiology report quality	Differences between in-house vs. external reporting

Understanding NGT Data Requirements & Confounders: Extending on previous descriptions (Section 5.1.2) on how poor image quality can implicate CXR interpretation, Table 4 summarizes key factors that influence image quality and may introduce AI model biases.

Amongst others we learned from clinicians that portable X-rays machines used on ICU produce *lower resolution images* compared to radiology imaging department machines used i.e., in Stroke care. Radiographers further described practices of adding *image annotations* like "AP ERECT SITU" or "SUPINE ITU" and other hospital or patient-related information directly onto the X-ray, which has additional implications for data anonymization. We also learned that some radiology reporting gets outsourced to external companies, which may introduce *variances in reporting style or quality* compared to in-house reports. Discussing these insights with AI researchers on the team, they brought forward the need in model development to consider factors such as: systematic differences in image resolution, image annotations/ masks, patient rotation (e.g., if a proxy for other patient pathology), or any particularities in 'outsourced' reporting (e.g., high prevalence of this occurring at night) – as these can potentially risk introducing spurious correlations.

Lastly, clinicians explained how poor image quality can also be a consequence of *patient factors* such as their 'level of impaired consciousness'[15] whereby the inability of unconscious patients to proactively hold-in their breath at the time of image capture correlates with poorer image quality. As previously described, other factors like patient obesity and chest abnormalities that show as white opacities on the image, or specific patient conditions (e.g., history of gastric surgery, hiatus hernia or situs inversus) also need consideration as they can either complicate image interpretation (by human or machine), and can mean that correct placement for an NGT likely varies from more standard cases.

Challenges in Effectively Linking Radiology Image(s) with Reference Text: Much AI development in radiology imaging involves AI training on carefully curated datasets. While a radiology study may include multiple X-ray images (e.g., various frontal and/or lateral views) and their corresponding radiology report text (see prominent datasets like MIMIC [78]), model training often requires a simpler mapping, for example in the form of 'one image - one report' or 'one image - (multiple) label' pairs. To achieve such formats by processing real-world data – whose production may be less standardized and messy – however can be more challenging. To illustrate this, we next describe three examples of current CXR data review and documentation practices that were reported by radiographers or became apparent through data case reviews; and then extrapolate their relevance to data preparation efforts:

[15]In types of acute medical and trauma patients level of impaired consciousness is recorded via the Patient Glasgow Coma Scale)

(1) *Multiple (potentially sub-optimal) radiology images can be taken to complete one CXR order.* For instance, if one CXR images the upper part of the chest well, but misses the apices (at the bottom) of the lungs and another CXR captures better the lower abdomen (crucial for NGT assessment), both images would be sent to PACS to provide "a better diagnostic picture" (RR2) to the image reader, who would mentally assembly them.
(2) *Multiple images being documented within one radiology report.* Radiographers remarked on the non-ideal, yet common reporting of image series, especially for inpatients where X-rays are taken daily. In those cases, radiographers load 3-4 images into the same accession and reports them chronologically within the same report.
(3) *Multiple lines and tubes being reported within one report sentence.* Our review of NGT data records showed how multiple patient lines and tubes that would show on the CXR would be referenced together within a single radiology report sentence (e.g., "Appropriately sited left CVC and NG on most recent radiograph"; "On the most recent image appropriately sited NG and ET tube approximately 4.5 to 5 cm above the carina").

Discussing those real-world data practices with AI researchers, we concluded their relevance to data preparation efforts as follows: (i) example 1 contributes to decisions to include or exclude, i.e., incompletely imaged studies from AI analysis; (ii) example 2 suggests greater complexity and higher chances of errors for applying, i.e., automated approaches to precisely map radiology images with report findings; and lastly, (iii) example 3 has implications for extracting key entities (e.g., a sentence mentioning the NGT and its position qualifier) for data label generation purposes.

Ambiguity in Defining & Labelling NGT (Mis)Placement: Furthermore, we found that, although NGT review was perceived as generally a straight-forward assessment task, there were differences in how participants defined if a tube is *correctly* or *sub-optimally* placed. While there was agreement across clinicians that a *critically misplaced* NGT was one that led into the lung – the condition that most existing ML models also predict for best (e.g., [43]); assessment of correct or sub-optimal NGTs could vary in practice from text-book definitions. Clinicians described how their review was influenced by: (i) the *likely implications of a sub-optimal NGT placement on the patient* (e.g., a tube that may be too far down poses lower safety risks than one not advanced enough); and (ii) *reporter experience and judgement on whether the NGT is likely in a good-enough position* -– even without clear visibility of the tube tip (e.g., to not have to repeat the X-ray). Furthermore, inter-personal assessor variances in exact location/ NGT definitions as well as in their reporting styles (e.g., level of precision, interpretation, recommendations for action) can *create ambiguity and variance in establishing robust NGT-positioning labels*. We illustrate this futher below.

Correctly Placed NGTs: Generally, a correctly placed NGT is "in the right place in the stomach" (D3). To make assessments of correct placement, ICU doctors referred to the NGT template or similar criteria, protocols or medical guidelines/ rules that enable their assessment of whether a tube is correctly placed. However, to conform with those visual checks requires the ability to (i) *fully trace the path of the NGT*; and, ideally, to (ii) see the *tube tip*. While one ICU consultant expressed lesser need to see the actual tip ("I know that's in the GI tract probably rather than necessarily see the tip", D6), most described how difficulties to locate the tip would cause hesitation in decision making given the high-stakes implications of any mistakes. A Stroke registrar explains the judgement call that is then made:

> "It's quite common for either the problem to be it's in the GI tract, but the tip is so far that it's actually off the bottom of the X-ray. (...). And in that situation there's a bit of a judgment call to be made of: Do you think the tube is in the right place, but the X-ray didn't go far enough? Or do you think the tube is too far in? And in those situations either you might say, ask the nurse to pull the tube back a few centimetres, up to five centimetres. Try and re-aspirate. And if they can't re-aspirate repeat the Chest X-ray. Or I just call the radiographer saying the tube looks like it is definitely in the right place. I think that X-ray just didn't go

quite far enough, could you repeat that X-ray just going down a little tiny bit further? That's quite common." (D7, Stroke)

Sub-optimally Placed NGTs: For sub-optimally placed NGTs, clinicians and radiographers distinguished between three types: (1) the NGT is 'folded up', 'bend' or 'coiled' within the patient's mouth, oesophagus or stomach, which can obstruct feed to pass through and increase risks of lung aspiration; (2) the NGT does *not reach far enough into the stomach*; or, more rarely, (3) is *inserted too far in*. In those two latter cases, all reporting radiographers and some clinicians described difficulties and judgement calls to determine exactly how far the NGT would need to be advanced for it to overall be considered still safe for feeding, and when to alert clinicians that its exact position is not within the stomach:

> "(...) one of the things that is hard to bottom out is: at what point in the stomach is it safe to feed? Because if it's at the gastroesophageal junction, do you run the risk that any patient head movement will then dislodge the tube that becomes oesophageal? And how do you say, you know, it's 5 centimeters past the GOJ[16]? Part of the problem about standardizing the interpretation; expectation and assessment for radiographers is, is it 5 or 10 centimeters passed the GOJ, because once you've set that standard, that's it. For me, as long as it appears radiographically clear of the region of the GOJ, obviously the Chest X-ray is 2D flat. You know about where the GOJ is, if it is over the stomach. It's safe to feed. It's OK." (RR1, ID)

Some ambiguity how individual reporters exactly define or distinguish what counts as safe NGT placement is also reflected in reporting practices. One radiographer remarks that some reporters "consider sort of 5cm past the gastroesophageal junction as adequate, other people say you need to advance it" (RR3); commenting on report variations:

> "I guess the other challenge is you know the variability to people's opinions and I guess variability of what people would say is an optimal position. What are the guidance they're gonna give, what advice they're gonna give in their report as well, cause a lot of people would just say satisfactory, something like that, or advanced, but they won't go into specifics, so there's that variation of reporting." (RR3, ID)

5.5 Measuring AI-Intervention Success

In this final section, we reflect on the study's insights regarding relevant quantifiable outcome metrics for evaluating the effectiveness of an AI NGT application in practice. We also consider the practical feasibility and reliability of assessing these metrics based on existing data documentation. Table 5 summarizes potential outcomes of interest with regards to AI-assisted feeding tube (mis)placement verification.

Improved Patient Safety. Our interview findings surfaced clinicians key motivation for applications of AI is to help improve patient safety by reducing risks of sub-optimal or critically placed NGTs being used for feeding. This may be achieved: (i) by reducing occurrences of NGT misplacement remaining undetected; or (ii) by speeding up their detection.

Reducing undetected misplacements: Perhaps the most obvious metric for patient safety is to evidence a reduction in number of Never Events, which are officially reported via DATIX – a risk management information system designed to collect and manage data on adverse events [42]. Whilst highly significant for patient safety, hospital staff reported extremely low prevalence of such events can mean it is too sparse to serve as a reliable metric – at least for any smaller-scale, short-term pilot study. NGT misplacement into the lungs without feeding, will likely occur more frequently, and requires as early detection as possible. For those cases we learned through our review and discussion of EHR patient data that ICU nurses documented a patient's physiological responses via so

[16]The GOJ (gastro-oesophageal junction) is the part of the gastrointestinal tract where esophagus and stomach are joined. It's a key landmark to assess if the NGT extends far enough and reaches into the stomach. However, the GOJ is not directly visible on a Chest X-ray

Table 5. Summary of potential outcome metrics for assessing improvements to feeding tube placement assessment, and timelier verification or use with goals to improve patient safety or workflow efficiency.

NGT AI: Outcomes of Interest	
Improved patient safety	Reduced number of Never Events (e.g., as indicated in DATIX records) Reduced number of misplacement-induced complications Reduced number of missed NGT misplacements Reduced time to identification of misplaced NGT Reduced time to documentation of misplaced NGT (e.g., in radiology report/ clinical note) Reduced time to correction of misplaced NGT (e.g., NGT removal/ replacement) Reduced errors in image assessment
Improved workflow efficiency	Increased reporter confidence Reduced requests for human peer review Reduced time to patient feeding

called CEASE sings[17]. We also learned that detected increases in heart rate or decreases in oxygen saturation levels during NGT insertion can be indicative of placement problems, which are monitored and recorded by ICU machinery. The Stroke registrar we interviewed further suggested evaluating any subsequent patient X-rays (e.g., using AI) for signs of lung aspiration. This could indicate that an NGT misplacement had been missed in earlier verification steps, for example, due to a deceptive aspirate test or CXR image interpretation error. Technically, reducing image assessment errors serves as a valuable metric for evidencing AI success. However, as stated previously, the prevalence of NGT CXR image interpretation errors remains unclear, as these errors may go unnoticed or are not formally documented – except in the case of extremely rare Never Events.

Speeding up misplacement detection: At the outset of the study, the research team assumed that an early alert to a misplaced NGT to clinicians, or to re-prioritize reading lists could speed up the time of its detection. Reviewing EHR records of patients with fitted and verified NGTs revealed that tracing the exact *time that a misplaced NGTs has been detected* can be complex. For instance, a radiographer might identify a misplacement at the time of image acquisition, but the official documentation may only be carried out hours later by an ICU clinician. This suggest that earlier documentation of misplacement through faster radiology reporting turnaround for misplacement tubes, or quicker clinical documentation might be a more feasible and reliable metric. Ideally, earlier misplacement detection should lead to *reduced time to its correction*. This could be evidenced by data entries documenting the removal and reinsertion of an NGT post-CXR (e.g., on a patient avatar within EPIC documentations); and time-stamped clinical notes describing corrective actions.

Improved Workflow Efficiency: In Section 5.1.1, we described the complexities of ICU work dynamics, indicating that improvements in workflow efficiency might be challenging to capture or evidence. However, our interviews discussing the potential benefits of AI-assisted NGT verification suggest the following potential outcome metrics: (i) increases to staff confidence in image interpretation and therefore reduced need for peer review due to AI assistance; and (ii) speedier NGT correction and patient feeding.

More confident, speedier image review: When discussing AI proposals that notify about potential human reporting errors or enhance visual image review (e.g., via an overlay) especially more junior ICU clinicians described their expectations of AI-assistance to increase their confidence in image interpretation. This improvement could

[17]CEASE signs stand for: Coughing, Extreme agitation, Abdominal distension, Stoma site leakage and Elevated temperature. These can all indicate complications that arise from feeding tubes and can be recorded in the patient record.

be qualitatively assessed and indicated through proxies such as a reduced need for peer review and faster review times.

Reduced delays to patient feeding: As a consequence of more effective misplacement detection and correction, hospital staff described assumptions of AI assistance to help reduce delays to patient feeding. To achieve and assess this reliably however requires: (i) accurate data recording; and (ii) staff ability to act upon earlier alerts and review requests. Reviewing data records, we learned that machine generated data, image order requests, their upload and signed documentations as well as feeding records provide more accurate timing information than, for example, nursing notes. Nursing notes are composed and added to throughout the shift and are only signed and time-stamped at the end of that shift. Furthermore, it could be argued that (ICU) doctors already look at the NGT CXR at their earliest convenience. As described earlier, we also need to account for time dependencies based on other ward dynamics (e.g., shift rhythms, emergencies, staff resourcing, time of day), or changes in plans to feed the patient (e.g., patient may undergo surgery first). This suggests speeding-up the detection of an NGT misplacement as best captured by the timing of its documentation, which likely presents a more reliable metric than subsequent clinical action of patient feeding.

In summary, this section highlighted the need to include additional data in evaluations for key outcomes of interest (e.g., DATIX data, patient complications at time of NGT placement, or follow-up CXRs to assess for missed sub-optimal NGTs). For more quantified, automated assessments, our findings suggest focusing on robust and reliable measures of AI effectiveness using data items that require mandatory documentation; are set-up to be documented at specific times or time-intervals (e.g., 4-hourly aspirates or feeding logs); and less time-dependent on broader hospital dynamics.

6 DISCUSSION: CONTEXT-SPECIFICITY & COMPLEXITIES FOR REALIZING AI UTILITY

Much of current AI development has been criticized for being technology-driven, decontextualized from concrete use scenarios and the many human and organizational factors that affect their adoption and integration within healthcare practice (e.g., [91, 99, 143]). To better address the disconnect between technical AI capabilities and real-world stakeholder needs [37, 110, 173], this paper contributes an in-depth case study describing our approach and the learnings of bringing a human-centered process to early stage AI innovation work that seeks to understand current clinical practice and identify the right problems for AI to solve. Practical insights gained from this study are synthesized in Table 6.

Next, we discuss the implications of this work that: surfaced complex interrelations between human, technical and organizational factors that determine perceived AI utility (Section 6.1); discuss challenges surrounding human expectations of AI and configurations of human-AI interactions for fostering AI acceptance (Section 6.2); and reflect how key insights into real-world data production and its characteristics can usefully guide AI development processes (Section 6.3). Across these areas, we draw out directions for future work in healthcare AI, and for radiology more specifically. We conclude with some of the imitations of our work (Section 6.4).

6.1 Interrelations in Perceived AI Utility

By grounding our AI work within the specific use context of an ICU hospital and concrete NGT CXR verification workflow, we learned how perceived clinical utility of AI capabilities and their potential realization within healthcare are bound-up by a complex interplay of multiple factors. In this section, we draw out how identified factors of AI goals and implications; workflow design and use integration; real-world data production and technical realization of AI are interlinked. Additionally, broader considerations such as medical-legal requirements, IT infrastructure, and resource constraints determine trade-offs in benefit-cost relations that affect perceived AI utility. We suggest systematic mapping as a tool to clarify those relations and facilitate cross-disciplinary discussions and decisions on directions forward.

Table 6. Summary of findings and implications for developing and evaluating clinically relevant healthcare AI.

Clinically Meaningful AI Applications	
Scope	• Given broader design goals (increase radiologist effectiveness, improve patient safety), remain open to identifying alternative, potentially more important opportunities for AI applications (e.g., beyond a focus on image analysis). *Ask: What are the most important problems to focus on? How could AI be best placed to assist those?* • Consider how clinical value and risks of different AI proposals depend on their realization in practice. *Ask: How well will AI need to perform to realize its potential? How well can we assist AI verification/ error detection? What are the implications of undetected AI errors? For what types of use cases would an imperfectly performing AI still provide utility?* • Clarify medical accountability of AI assisted clinical practice. *Ask: Is the purpose of the AI to "augment" or "automate" (new/ existing) human practices? What is the added value proposition for use cases that require AI use "only with human oversight"? What are medical-legal and broader organisational requirements?*
Target users	• When clarifying target users for AI, consider opportunities & implications of (re)defining AI-assisted role responsibilities including training requirements, overall workload burden, and clinical responsibilities. *Ask: How generalizable may an AI application be across care professions and care settings (e.g., different wards, hospital environments)? Are there any (future) changes anticipated for certain roles/ professions?*
Workflow integration/ design choices	• Be mindful of frictions introduced by new (safety) practices or substantial changes to existing workflows that AI might introduce (e.g., requirements of time, training, changes to routine). *Ask: How does the new, AI approach sit alongside, or presents an improvement to, other current (safeguarding) practices? What would be the simplest implementation of AI that causes least disruption to existing work, but would still be insightful?* • Map out where within an end-to-end workflow AI could be incorporated and its implications to more systematically guide choices (see Table 3 as an example). *Ask: Where are areas of opportunity to assist the workflow and what would be the benefits and risks of different AI implementations to direct and indirect stakeholders?* • Balance requirements for seamless workflow integration with needs for AI transparency and verification. *Ask: How to introduce the right types of friction (e.g., through careful alerts/ notification frequency based on good-enough AI performance) to ensure greater safety balanced with avoiding overburdening or delays? What types of AI applications or their design allow for easy, fast verification and correction by humans?*
Data Constraints & Opportunities	
Data labelling	• Investigate how data generation and documentations may differ from previous data(sets). *Ask: What additional data processing stages may need to be performed (e.g., to anonymize data, extract valid labels)? What are existing, standardized data capturing processes that could be leveraged (e.g., to assist data labelling)?* • Consider differentiating between data items that are clear (ground truth) vs. more ambiguous to label. *Ask: What are more factual labels that could be introduced that avoid interpretation variances?*
Model training/ evaluation	• Clarify relevant patient factors (e.g., BMI, consciousness) and patient edge cases (e.g., hiatus hernia, situs inversus, gastro/ oesophageal surgery), and other confounders (e.g., image quality) for the AI task. *Ask: What other relevant (meta/ EHR) data should be included in model development or evaluation?* • Map out different AI error types with regards to the use case to test model performance against. *Ask: What error types might occur & how can these be (automatically) detected in data evaluations?* What are the implications of different error types for users assessment of the image (or other clinical data)?
Evaluating AI application success	• Consider real-world data generation practices and their limitations to clarify clinically relevant success criteria. *Ask: What can be evidenced reliably with the data that is available? What additional data sources need including?* • Define realistic measures for evidencing AI success by considering the broader use context. *Ask: What moderator variables (e.g., hospital dynamics related to resourcing, shift patterns, emergencies) may implicate staffs ability to timely act upon AI insights and may need including in evaluation study protocols?*
Broader Implementation/ AI Adoption Considerations	
AI expectations, training & organisational framing	• Help clinical stakeholders develop realistic expectations of AI. *Ask: What are current expectations of AI capabilities and how well do they map to technical performances? For what use cases is AI more and less likely to assist with?* • Devise plans for staff training & onboarding (see [30] for guidance), and clarify AI (pilot) deployment or broader roll-out strategy with hospital leadership.

6.1.1 Trading-Off Intended Use with Actionability, Data Availability and Broader Organizational Constraints. When reviewing our five proposals for how AI could assist NGT CXR practices, we identified desires for patient safety as a predominant driver of clinical relevance. While goals for patient safety can be assisted by efforts to increase staff confidence in image assessment, and by speeding up overall processes (e.g., reducing delays to misplaced NGT detection and correction); AI proposals that more concretely pronounced *patient benefit and safety* received most support. This aligns with prior research emphasizing the importance of considering the patient as the primary beneficiary where applications are intended for clinical use [49, 149]; and suggests a focus on AI that facilitates *more immediate detection of critical findings*, or *human errors*. When we investigated where and how within existing workflows such AI functionality could be clinically most relevant and impactful (e.g., to speed-up and inform actual care decisions) our research surfaced interdependencies with factors such as: (i) *timing* and *staff's ability to act* upon AI insights; (ii) broader *clinical care pathway and organizational set-up* considerations; and (iii) *data availability* constraints, which are expanded on next.

Ability to Act within the Context of Existing Workflows and other Hospital Dynamics. Investigating different workflow stages (illustrated in Figure 1), we learned about the roles of various stakeholder groups involved (as potential AI users) and how the timing of an AI intervention interlinks with clinical utility. For example, having an AI potentially alert to any "critical' or 'sub-optimal' NGT placement detection ideally *as early as the image acquisition stage* would be most beneficial. This workflow stage involves imaging radiographers, who were also identified as the only constant across care settings, suggesting greater potential for an AI application to *scale* and expand reach of benefits. Yet, imaging radiographers current *professional qualifications* and *medical responsibilities*, mean that they can only inform referring clinicians or nurses about any noteworthy NGT observation, but often do not have permission to report the NGT CXR. This workflow integration proposal may further be complicated by *technical constraints* such as requirements to have AI run on X-ray machinery, which has its own software separate to EPIC or PACS. In contrast, an AI that would support 'human error detection' in the reports of radiology reporters may be technically easier to integrate within reporting software like PACS, but is likely have a lesser impact on clinical practice where ICU doctors have already acted upon a CXR images days prior to its official reporting.

The latter example hints at the importance of *staff being able to act upon AI insights*, ideally in ways that improves current processes or patient outcomes, for AI utility to become realized; as is also discussed in other research [92, 173]. *Staff role responsibilities and competencies* aside, our study further showed how *resource availability, shift patterns* and *broader ward dynamics (like emergencies)* influence staff's ability and decisions to complete, or hold-off on NGT specific tasks. Given these constraints and that ICU clinicians described to already try and review NGT CXRs as early as possible, or that reporting radiographers may not be available out-of-hours; this surfaces the question, to what extent, realistically, staff could 'better action' any early alerts or worklist prioritization in response to a critical findings detection. At a minimum, our findings suggest that key context variables such as: *staffing levels*; *shift and work hours*; or *prevalence of emergencies* may need careful consideration in any evaluative study protocols that seek to assess AI effectiveness in practice. In addition to these *temporal work rhythms* and broader clinical workflow considerations (e.g., clinicians being constantly on the move; logging in and out of different computer systems as described by [163]), research by Beede et al. [14] also draws attention to considerations of the *built environment*. For example, their work showed, how poor clinic light conditions at a local clinic negatively implicated the performance of their AI system that otherwise showed high accuracy in lab tests.

Broader Clinical Care-Pathway & Organizational Set-up Considerations. Furthermore, we need to consider the development of any one AI application in the context of the broader organizational set-up. For one, our work surfaced how *care pathways varied* between ICU and Stroke services, which was evident in differences in: (i) image capture technology that can mean different software integration requirements, or image quality outputs; and (ii) role responsibilities that suggest a greater reliance on official radiology reporting in Stroke care.

These suggest potentially the *need for tailoring solutions and configurations of AI to different pathways*. Further, we need to account that sub-optimally placed NGTs, of course, present only one of many potentially 'urgent' findings that could be detected and prioritized on a CXR. For example, Seah et al. [124] list 34 'critical' findings to detect on CXRs, and that could compete for clinical attention. Furthermore, research by See et al. [125] shows that – even outside of any AI use – and despite the availability of a well-established electronic notification system and mechanisms to notify referring clinicians about an abnormal radiology report, urgent findings can still be missed or their communication be delayed; arguing for the importance to have *the right organizational set-up to be able to translate key clinical findings into prompt action* (cf. [46] – above and beyond mechanisms to detect or alert to a critical finding).

Balancing Data Availability Constraints with Broader Opportunities for AI Innovation & Impact. Data availability limitations can affect the feasibility and evaluation of some AI solutions. For instance, for AI to detect reporting errors in image assessment, it is unclear how frequent these errors are, as they may go unnoticed or unreported in historical data, or are only documented in critical incident reports. However, not all initial errors lead to critical incidents. This makes it hard to obtain and process such data for AI purposes. While data, its quality and scale, necessarily presents the key building block for the practical realization of any AI, it is important that this does not necessarily limit explorations of other potentially more viable use cases and areas of AI opportunity (cf. [140]). While our research investigation and AI proposals specifically focused on addressing NGT CXR image assessment or report generation challenges, our study surfaced 'delays' that spanned across the entire NGT verification process as the most prevalent workflow problem. This suggest *broader opportunities* to assist in clinical process optimizations. We list alternative opportunities where AI could assist NGT workflows as identified through the user research in Appendix A.2. Furthermore, our work more generally highlighted: (i) *communication inefficiencies* whereby staff described the constant checking and chasing-up of tasks; as well as (ii) *tedious efforts to extract relevant patient information from EHR records* (cf. [122] for a report on costs and need to simplify administrative burden in healthcare). This also foregrounds how definitions of desired AI functionality may best evolve as an iterative dialogue between user need and data availability.

6.1.2 Systematic Mapping as a Tool for Achieving Clarity about Key Factors and their Interrelations. Given the complexity of the various human, data and organizational factors that can implicate perceived AI utility, we found it helpful to systematically map out key factors and their interrelations across different AI proposals. We illustrated an excerpt of such mapping in Table 3, which depicts links between: workflow stages, users, design choices, AI error type breakdowns, direct and indirect stakeholder implications, and broader technical or medical-legal constraints. We believe that this – or similar activities (cf., Yang et al.'s [161] AI design complexity map; or design resources for scaffolding AI concept ideation by Yildirim et al. [165]) – can serve as a useful exercise to trade-off perceived AI benefits and risks; thereby guiding AI research and development teams in make more informed choices on AI use cases and their configuration. Insights revealed may further provide valuable inputs to current responsible AI practices such as *impact assessments* [70, 98], and serve as *boundary objects* [77, 87] – as common frames of reference and 'shared vocabulary' for facilitating inter-disciplinary team collaboration (cf. [7, 31, 168])when discussing i.e., clinical priorities; realistic AI performance goals; or necessary risk mitigations.

6.2 Human-Process Integration of AI

This section discusses how perception of AI utility and acceptance of future AI applications are bound-up with often high-expectations of AI performance, and how AI functionality becomes positioned and integrated within human work.

6.2.1 Setting & Managing Appropriate AI Expectations. Our study findings surfaced how participants evaluations of perceived AI utility was based on how well it met their expectations of its performance. For AI proposals like NGT report generation, to realize its full potential, they suggested the AI needed to be high-performing, ideally 100% accurate. More value was also ascribed to AI that would surpass human capabilities – particularly for assistance with very complex and difficult-to-interpret patient cases. However, it is unclear how well AI would handle unusual edge-cases; and how realistic an almost perfect AI performance would be. Other research [96, 173] has also warned that unrealistic and overly high expectations of AI can negatively affect its perceived usefulness. We also noticed a tension between aspirations and excitement for how AI could be transformative to staff practices or patient care (e.g., references to ChatGPT capabilities), and staffs previous experiences with other AI, decision-support or automated systems (e.g., auto-generated ECG impression) that made them doubt AI capabilities. This tension made more ambitious AI proposals appear as more of a theoretical proposition rather than practical reality. This sentiment is echoed in findings by Verma et al. [149] who found that oncologists in their interview study had a positive disposition towards AI utility and its potential in transforming healthcare provision in the future, yet perceived existing AI as being in its *infancy* with a clear disparity between the positioning of AI potential and its actual adoption into clinical workflows. Zając et al. [173] further found that there is more familiarity with AI use for decision support (including quality assurance) and prioritization, than with AI for automation scenarios, which are still very rare in healthcare (cf. [14] for an exception). All this complicates the *development of realistic expectations of (new) AI capabilities by healthcare professionals* – a well-recognized challenge in human-centred AI design (e.g., [30, 161, 165]). It implies a need for better understanding of how healthcare professionals currently perceive AI (cf. [114]), as well as more support of cross-disciplinary and education initiatives that enhance healthcare stakeholders' knowledge of AI [149]. In this regard, Cai et al. [30] provide useful guidance on how AI capabilities and limitations; as well as their functionality, underlying configurations and design decisions should be communicated to medical professionals both in on-boarding and interface design. All this will enable a more informed perspective and better participation in discussions of feasible goals and strategies for AI (system) innovation – both short-, and long-term.

6.2.2 Appropriately Configuring Human-AI Relations. Whilst participants described the potential benefits (e.g., time-savings) of an AI to automatically assesses the CXR image and generate the NGT report; to unlock most utility, they also described for the AI needing to take 'full responsibility' of its outputs, alongside requirements for high performance and current medical-legal barriers. Where, instead of full automation, AI functionality was positioned to only assist existing human practices (e.g., by showing an image overlay to help visual analysis, or creating a 'preliminary' report) it was harder for staff to identify the value added by AI. This aligns with recent findings by Burgess et al. [26] who reported how physician's perceive little value in an AI tool that offered insights they already knew, and other research [14, 75, 151, 167] that describes clinicians concerns about AI potentially increasing clinician workload, contributing to lesser technology adoption, or risking of human burnout.

Interlinked with the above is the fact that, if clinicians would still have to do their own checks according to current practice, *they would remain fully accountable and legally responsible for any medical judgments* (cf. [149]). This not only emphasizes the current difference in accountability between AI and clinicians (cf. [18]); it adds to their existing workload expectations that: (i) clinicians are able to understand AI workings, and (ii) can spot and correct any potential AI errors. Consequently, any suggested benefits of, and potential trust in, the AI are weighed-off with AI risks. This includes potential negative effects on staff workflows and patient outcomes due to AI errors (e.g., increasing peer review when false outputs cause confusion), and risks of AI over-reliance; or conversely, of AI alert fatigue, if its outputs were incorrect often. Proposed AI benefit also needs to be high enough for staff (and hospital organizations) to be willing to invest time and effort (e.g., staff training) to potentially adjust work practices to accommodate new AI functionality; along with required IT infrastructure and other financial resources and information governance processes involved in developing healthcare AI services [119].

All this sits alongside considerations of opportunity cost whereby resources are expended in one area (e.g., AI), but not another [92]. Thus, given the substantial resources that are needed to successfully build and deploy AI systems in healthcare, it is important to understand the conditions under which AI can be effectively leveraged to maximize the benefits that these investments can bring to healthcare delivery [114, 127]. Reviewing the perceived value across different clinical decision support systems, Wang et al. [153] also emphasized the need for AI to make clinical assessment demonstrably *easier* or *faster* than conventional approaches to unlock real benefit. Even where research and development demonstrate clear clinical benefits, balanced with *costs of disruption or change*, there still needs to be willingness to pay for new (AI-enabled) services. Consequently, this calls for a much broader *economic impact and cost effectiveness assessment* that would focus on wider stakeholder engagements and context variables than any more immediate, narrower study outcomes (cf. [35]).

6.2.3 Shifting from 'Human-AI-Verification' to 'Human Process Integration of AI'. Where effective, responsible realization of (full) AI automation may present a longer-term ambition, we speculate that greater utility may come from: a better integration of AI within human healthcare work; and shifts in clinical role responsibilities.

Limitations of Current Strategies to De-Bias and De-Risk Potential AI-Errors. Our research findings echo previous studies that surface tensions in *human-assisted* or *human-in-the-loop AI* systems to balance desires for a seamless, unobtrusive AI integration [142, 162] and to protect and respect clinicians professional autonomy [126, 149, 151, 153, 168] with approaches to ensure humans appropriately interact with AI outputs by taking time to check its correctness. Especially in healthcare, where clinicians work under time-constraints, they may not have the interest, ability, nor technical expertise to engage more deeply, or more critically with AI outputs [26, 73, 126, 142, 151]. When extra time is spent on extra data entry, data review, waiting for AI outputs, or sorting through (false) AI alerts, this can mean AI systems are not used [151]; disrupt work practices; or distract clinicians from their focus on patient care [127, 173]. Similar concerns about preservation of professional autonomy and fit within workflows are also reported in recent research by Bach et al. [8], who specifically study bias mitigations in clinical AI support by asking humans: to make their assessment first; to provide decision justification; or to explicitly consider opposing AI outputs. Their study showed that while such strategies can reduce bias and improve diagnostic accuracy, the *additional burden required to engage with AI in this way*, significantly decreased work efficiency. Clinicians also did not appreciate for AI to correct rather than support them in their tasks – describing the experience as 'condescending'; emphasizing desires for humans to remain in *control* over healthcare decisions [149]. All this suggests a more complex picture when trying to appropriately configure human-AI experiences that are *seamless, effective* and that *responsibly* consider the nuances of different AI use and error scenarios.

Taking Inspiration from Existing Safeguarding Practices. There the goal is to create AI applications that feel empowering to healthcare providers, we suggests a departure from positioning humans as 'error-checker for (imperfect) AI' that not only adds extra burden of time, but also positions them as the 'blame-takers' for any AI failings - when clinicians primary concern is patient wellbeing. Here, we can draw parallels to prominent work in the security community by Adams & Sasse [2], where humans are often framed as the 'weakest link' within security systems (rather than focusing on improved user experiences); and also Elish's [45] notion of humans as 'moral crumpet zones' that bear the burnt of the moral and legal responsibility when the overall (AI) system fails. Such a positioning of the humans risks conflicting with desires for creating technology that feels *empowering*.

Thus, for a safe, effective integration of AI outputs within clinical work, we therefore wonder if instead more inspiration could be taken from existing, accepted hospital safeguarding practices designed to exercise appropriate caution and prevent human errors. Amongst the range of risk mitigation strategies we discovered were: (i) a clear hospital NGT policy, staff training and skills test; (ii) the cultivation of a cautionary mindset; (iii) human peer review; (iii) mandatory, templated reporting of the NGT; and (iv) continuous (4-hourly) position and aspiration checks. These strategies and their combination, whilst not in themselves perfect, are designed to catch human

errors. In this regard, we suggest that future work could extend investigations how AI insights can become part of familiar steps and processes of clinical information review, peer checks, guideline adherence, and other practices to clinically correlate and validate medical insights. Where AI is conceived as an 'instrument' or 'tool' used to support healthcare practices – and not as a 'constituent' or 'expert-type agent' offering *independent* advice (see Verma et al. [149] for this differentiation) – how might a framing of AI insights within such practices and information ecosystems aid its utility and acceptance? Treating AI insights pragmatically alike other data 'tools' – *that offer unique insights and have their limitations* – the goal for users should be less on identifying whether a specific AI output is correct or not (or how it arrived at a particular output/ prediction). Instead, we should aid clinicians to effectively triangulate AI insights– as "a source of additional evidence" [168] – with other clinical evidence and information they have about the patient such that it brings about caution or confidence in taking next steps. By taking a stronger focus on empowerment, we should seek to identify ways in which health professionals want to take the responsibility for working with AI. For this, it may be more acceptable to have this responsibility shared with other (human) agents and safeguarding processes; bringing greater acknowledgement to how Wang et al.'s [151] participants described the diagnosis process as: "a highly interactive, communicative, and social event".

In the context of our study, this suggests to not de-contextualize how AI will help radiologists in interpreting images [56] and instead to connect AI outputs to other criteria 'external' to the model [129, 142] and its bringing into context with other, trustworthy information agents or resources (cf. [26, 160]). Such other (human) agents and resource can *facilitate meaningful clinical correlations that enable clinicians to (part) accept, reject or ignore AI outputs similarly to how they would treat other 'imperfect' clinical assessments.* Future work therefore should explore more deeply the integration of AI insights within existing (safeguarding) processes (i.e., multi-disciplinary team meetings); alongside continued advances in technical solutions such as: self-consistency prompting [131], LLM-generated explanations [104], or correctness predictions [79]) – that serve to reduce AI errors and associated risks.

Adapting AI to Different Users & Supporting Shifting Role Responsibilities. To expand reach of radiology services and reduce the implications of delayed official radiology reports, we learned how all ICU clinicians are trained and permitted to assess NGT CXRs; and about the role of reporting radiographers – including near-term plans to upskill imaging radiographers to be able to make all-important safe-to-feed decisions. Such shifts in roles and responsibilities expands the workforce that takes on more radiology-specific tasks, which may indeed be necessary to increase reporting capacity required to address ever growing imaging backlogs [95]. On the one hand, this suggests *AI may need to be adapted for different user types and their needs.* On the other hand, this surfaces the question: *how AI could play a useful role in building-up image assessment confidence without getting in the way of important human skills-acquisition*? Here, tasks like CXR NGT verification may be particularly suited to AI support as it has a clear question (e.g., is the tube in the correct place?) and involves an observation-based assessment rather than a complex clinical interpretation. Another example may be: is there a bone fracture? For these types of tasks, visual inspection following standard, protocolled procedures is sufficient and *does not demand more advanced, expert-level radiology evaluation*, as is often the case for higher-risk AI tasks that involve: diagnosis, treatment suggestions, and other decision-support functions (cf. [167]). Whilst this suggests opportunities to leverage a wider workforce – at least for certain radiology tasks – this is not without challenges, as a CXR may reveal other medical conditions that require clinical intervention; raising questions about diagnostic boundaries and medical-legal responsibilities that need to be addressed in future work.

6.3 Relevance of Real-World Insights into Data Production and its Characteristics for AI Development

The last section of our findings details learnings about existing NGT hospital data – it's production, availability, and characteristics – and discusses their relevance in guiding AI development specific to NGTs on CXRs, and beyond.

6.3.1 Implications for Data Preparation & Model Training/ Evaluation.

Accounting for Potential Data Biases & Including Relevant Context Data for Image Interpretation. Through our contextual inquiry and interview research, we identified key factors that can influence image quality; introduce model biases; or otherwise complicate image interpretation. For example, we described how key patient factors such as obesity, conscious impairedness, and obscuring structures or pathologies of their chest can hinder the visibility of the NGT path or tip, whilst edge cases such as an unusual or changed patient anatomy can mean deviations in image interpretation (e.g., what's considered 'correct' placement) from more standard cases. This suggests the inclusion of key *image meta* and *EHR patient data* (e.g., patient BMI, Patient Glasgow Scale, patient medical history) as important context information to image assessment and AI model training/ testing.

Our findings also identified specific data characteristics that can invite spurious correlations that may need controlling for. For instance, specific image makers that we observed in CXR NGT data (e.g., 'AP ERECT SITU' annotations), whilst not entailing any patient-identifiable information that otherwise would be removed, can be indicative of the patient being in 'ICU' care. Even where those annotations are masked (e.g., using a black box), the visual remains and specific annotation locations can become spurious correlations – these are 'shortcuts' an AI model might learn instead of desired image features (cf. [40]). Here, new methods such as RadEdit [112], a generative image editing approach may be particularly promising, as it allows to systematically mask image parts to diagnose potential spurious correlations and other biases.

As another example on the risks of spurious correlations, Drozdov et al. [43], who developed a deep learning model to classify NGT position, found their model performance to be adversely impacted, and the model to be very sensitive to images that showed sudden changes in 'system manufacturer' and 'institutional department', which were linked to differences in the imaging machinery used, and resulting image quality. This foregrounds the importance to include details such as hospital department or imaging system used within the data strategy; alongside more commonly considered attributes like: patient age, sex, or ethnicity, which are often collected to ensure a diverse, representative dataset composition [147], and to assess AI fairness [3, 97].

Understanding Data Production for Effective Data Curation.

In Section 5.4, we described how real-world NGT reporting practices can implicate effective dataset curation. For example, in much AI analysis that uses prominent Chest X-ray datasets (e.g., MIMIC [78], NIH ChestX-ray14 [154], PadChest [27], Indiana CXR collection [41]); and also more lines and tubes specific ones (e.g., RANZCR CLiP [138]), data is often represented as "image – report" or "image – (multi) label" pairs. Achieving a simple mapping between data inputs (e.g., 1 image - 1 report sentence) may be harder, where such information needs to be disambiguated from text reports that address multiple images; or where multiple images are linked to one report. These are important considerations where label generation as part of imaging dataset creation increasingly uses more automated methods (cf. CheXpert [71], PadChest [27]) that utilize advanced natural-language processing capabilities to identify and extract relevant text entities from radiology reports at scale, and in-lieu of requiring expensive 'human' label annotations. Fortunately, recent advances in large-language models (LLMs), and their combination with careful 'prompt' strategies, demonstrate that *entity extraction from radiology report texts* can be very effective for a range of tasks, even if only few examples are given to the LLM (cf. uses of GPT in processing radiology text as reported in [94]).

Complex Data Inclusion/ Exclusion Decisions. Our study findings also surfaced common challenges of CXRs potentially omitting key parts of the relevant patient anatomy, e.g., cropped the lower abdomen, or the

apices of the lungs are missing. With current trends in healthcare AI that focus on improving data quality in training (e.g., [60, 83, 94]), this might suggest excluding such studies from analysis. To a certain extent this *data exclusion may be necessary to ensure robust AI training* – given that the relevant area for the target AI task might not (fully) show on the image. Moreover, when deploying the AI system, *it is important to define what kind of inputs are acceptable for the model to produce reliable and safe outcomes*. Simultaneously, however, such data exclusion practices can mean that a substantial amount of imaging studies could be rejected for analysis, limiting the potential reach and perceived utility of a resulting AI application. For example, in the context of the RANZCR CLiP dataset [138], 33% of NGT images (2748 out of 8344) were labelled as "incompletely imaged", which constitutes a considerable amount of studies that could risk not being predicted for alongside other *data exclusions* (e.g., restrictions to adult populations).

Defining (Ground Truth) Data Labels & Distinct Outcomes. Interlinked, our study surfaced how there can be ambiguity and variance in cases where assessment of a correct or sub-optimal NGT presents a more borderline case. Here, we identified reporters making judgement calls that are guided by (i) their experience and considerations of the (ii) likely implications of errors in those judgements (e.g., it is less problematic if a tube is slightly too far advanced vs. not far enough extended into the stomach). The clinical reasoning they apply may differ from common text-book definitions that often suggest for the NG tube tip to be "at least 10 cm below the gastroesophageal junction", a definition often used in NGT data label generation efforts [43, 138]. This echoes previous research [151, 168]. For instance, Yoo et al.'s [168] study on AI treatment recommendations for sepsis clearly foregrounded how treatment was a very dynamic process, characterized by a *range* of plausible next steps rather than a clearly defined optimal solution. Therefore *the 'accuracy' of real-world clinical decisions can be unknowable and contentious*. Consequently, they described how their study participants often chose which aspects of an AI recommendations they would accept into their decision making or not (e.g., fluid dosage suggestion); thereby balancing or 'negotiation' AI insights with their own intuition rather than a more simplistic, dichotomous 'accepting' or 'rejecting of the AI.

While the ability to 'negotiate' which AI insights to consider might address ambiguity in clinical decision making by humans in clinical practice; ambiguity may still need to be addressed where data is being 'labelled' into distinct categories to indicate specific observations or outcomes to effectively train an algorithm. While some datasets include labels generated or confirmed by a human expert (e.g., [11, 65]); others employ multiple experts to generate annotations and assess agreement across them to achieve more certainty about the accuracy and overall quality of the resulting labels (e.g., referred to as 'gold' labels [19]). However, such labelling efforts often require substantial resources as well as access to domain experts, which limits possibilities of their development. Another route forward to addressing the challenge of identified variations in clinical assessments is to take inspiration of from existing hospital practices and *leverage more standardize clinical data where available*.

Leveraging Existing, More Standardize Clinical Data. To manage uncertainty in image interpretation, ICU clinicians and reporting radiographers described referring back to the .NGT template or similar protocols, and to revert to describing the tube's location rather than making any safe-to-feed decisions or other recommendations. Aligning with the provision of clear visual descriptions that the clinical team can then interpret in the context of an individual patients' circumstance may also present a less risky, potentially clinically more useful AI output than, for instance, more binary, high-level NGT classifications (e.g., predicted 'correct' placement). In this regard, the .NGT template (Figure 2) offers a unique opportunity for framing AI tasks more specifically around visual assessments. Not only does it provide a focus on visual checks of the feeding tube with regards to the CXR (e.g., "The NGT bisects the carina"; "The NGT pass below the diaphragm"); its standardized format that records yes/ no responses to specific visual questions also makes it particularly suitable for auto-label extraction; and offers concrete criteria to evaluate AI models against (see for example the design of decision rubrics as part of visual-question-answering (VQA) approaches). In other words, by discovering the '.NGT' reporting template, we identified not only a useful short-cut for label extraction; as a clear, standardized checklist of what the AI needs

to achieve to provide utility specific to the NGT verification task, the template itself may be regarded as *defining the capabilities that the AI models needs to be trained for*, and to provide *clear indications of assessment or success criteria to evaluate a model* against.

6.3.2 Evaluating AI Effectiveness (in Practice). Lastly, we clarified different outcomes that could be utilized to assess the effectiveness of an AI NGT application within clinical practice. Previous work by Wang et al. [153] surfaced common evaluation metrics of: improvements in patient assessment or management; time spent on patient care; frequency of patient interactions; precision and recall in information retrieval; and appropriateness of a suggested clinical intervention or procedure order. For our specific AI use case and context, and guided by clinical goals to improve patient safety and workflow efficiency, our data investigations surfaced how choices of suitable metrics depend on: *ease of data availability and access* (e.g., routinely collected EHR data vs. DATIX data vs. data that is not directly captured); the *reliability of (timely) data capture*; especially where the goal is to assess temporal relations (e.g., to speed-up detection of critical findings); and the *possibility for AI outcomes to affect change* in the context of other hospital dynamics. Furthermore, our data investigations surfaced new opportunities for analysis not considered in our study such as: detecting patient's physiological responses during NGT placement (e.g., CEASE signs, detected increases in heart rate, or reductions in oxygen levels) that are recorded in EHR data as potential inputs to NGT misplacement detection; and an analysis of follow-up CXRs for feed-induced 'lung aspiration' to detect a potentially missed NGT misplacement or un-intentional changes to its position. Furthermore, our mapping example in Section 5.3 brought attention to the need for future work to develop more established methods to *qualitatively explore AI failure cases* (e.g., with domain experts); and *help formalize and systematically assess the (likely) clinical implications of different AI error types*. This will nuance aspects of AI models that perform better or need improving; clarify more or less permissible AI error types (based on likely implications); and help concretize definitions of what constitutes successful AI performance or outcomes within development teams. In addition, where AI development moves towards deployment, *evaluations of the user experience* such as assessments of: clinicians willingness to use the AI system; system understandability or learnability; workflow integration fit; and trustworthiness in its outputs warrant consideration (cf. [153]). Likley, as expressed by Yoo et al. [168], the more realistic the AI integration within actual clinical decision-making, the harder it may become to assess the impact of AI insights on overall outcomes – especially with 'standard' validation techniques – and when considering how AI acceptance or reliance may vary across cases, and can change over time.

6.4 Study Limitations

Our study is limited by its specific research context, chosen study method, and selected sample:

We conducted our research in the UK, which is a high-income country, and within a hospital well-known for its medical research and excellence in acute and specialist services. It is one of the few places that employs EPIC as an EHR system, which usually entails a substantial financial commitment for larger hospitals. The hospital also has advanced digital radiology workflows and funding for this research, indicating the availability of resources and technological infrastructure to facilitate AI innovation. We recognize that this represents a more advantaged healthcare situation. Moreover, we also acknowledge that our specific focus on ICU workflows limits the applicability of our findings to other departments, and other hospitals within and outside the UK context.

Our work focused on early-stage AI use case exploration and development requirements elicitation that prioritize an understanding of existing work practices and feedback on prospective ideas for how AI could assist NGT processes. As a result, we did not yet create or utilize concrete visual user interface sketches (cf. [26, 142, 162]) nor any functional AI prototypes (cf. [148, 158]). Instead, our research insights seek to inform future human-centered AI work (e.g., key questions to ask early within research and development) and guides design concepts within the context of NGT CXRs, and healthcare AI more broadly (e.g., learnings about AI acceptance barriers,

interaction design choices). As such, this work presents the starting point of a learning journey how existing healthcare data, models, users, workflows and organizations inter-relate and would have to come together to meaningfully identify, shape and realize AI innovation use cases. Future work will need to build on these initial insights through situated, iterative AI (prototype) design and development cycles.

Finally, our study sample reflects a breadth of participants rather than a more representative sample of one specific clinical user group. Jointly, with a research team also comprising experts in compliance, research governance and IT – alongside AI, HCI and healthcare, this inclusion of a broader range of professional expertise is more common for innovation studies that seek lab-to-clinic transition [173]. Even though our research team included two consultant radiologists (MTW + JJ), who assisted in designing the study protocol and interpreting the research findings, our main focus and participant selection on the ICU *excluded radiologists* as a user group in our study. Lastly, we also recognize the potential value of incorporating *patient perspectives*, for instance through more extensive public and patient involvement (PPI) beyond ethical approval, into this and future work.

7 CONCLUSION

Seeking to pave the way forward in closing the gap between innovative AI research and its translation into healthcare practice, we presented a detailed case study of how we adapted a human-centered approach to early stage AI innovation in radiology. An in-depth contextual inquiry and interviews with 15 clinical stakeholders revealed rich insights into current clinical practices, highlighting existing workflow challenges, safeguarding mechanisms and their limitations. Evaluating different AI proposals in this context, we discussed how multiple factors and complexities, such as the timing and ability to act upon AI insights, the broader clinical care pathway and organizational set-up, real-world data production and technical realization of AI capabilities as well as human expectations and interactions with AI, affect the perceived AI utility and acceptance. We proposed using systematic mapping as a tool to clarify the trade-offs and interrelations among these factors. Moreover, we argued that in configuring human-AI relations to shift from a focus on 'humans needing to verify AI' towards a closer 'positioning of AI as integrated within existing human (safeguarding) processes' of: clinical information review, guideline adherence and concerns for patient safety. This brings AI as a data insight to patient assessments that is clinically correlated with other information and insights external to the AI model(s) at work. Further, we discussed challenges and opportunities of existing image reporting practices and data production, and their implication for AI data curation. Specifically, we: (i) drew attention to image quality and patient factors that can bias model training and affect image interpretation for NGT placement; (ii) described potential trade-offs between efforts to achieve high-quality training data for producing reliable AI outputs and consequential data exclusion for the potential reach of a resulting AI application; and (iii) discussed difficulties in achieving robust 'ground truth' data labels for more borderline NGT cases. We suggested that standardized reporting templates may help to define the AI model's capabilities as well as in systematic evaluations for the NGT verification task.

Whilst our research centred on the specific use case of NGT verification, the insights provided translate well to other medical lines and tubes investigations (e.g., CVCs, ETTs), as well as the detection of clinically relevant or critical findings within radiology imaging. More broadly, our leanings – further summarized in Table 6 – guide future research team that seek to design and develop AI applications that are clinically useful, ethical, and acceptable in real-world healthcare services.

8 ACKNOWLEDGMENTS

Special thanks go to all our study participants for their time and invaluable input to the research. JJ was supported by the Wellcome Trust [209553/Z/17/Z] and the NIHR UCLH Biomedical Research Centre, UK.

In Memoriam of Dr. Barney Wong, who sadly passed away in 2023. Barney was a dedicated Stroke doctor at

UCLH, whose passion for medicine and commitment to his patients, colleagues and advancement of knowledge were truly inspiring. His expertise and hard work were greatly appreciated, particularly by the Elderly Care team at Homerton Hospital. He is deeply missed by all who knew him.

REFERENCES

[1] Moneeb Abbas, Muhammad Usman Akram, and Anum Abdul Salam. 2022. Automatic detection and classification of correct placement of medical tubes on chest x-rays using auxiliary head and test time augmentation. (2022).

[2] Anne Adams and Martina Angela Sasse. 1999. Users are not the enemy. *Commun. ACM* 42, 12 (1999), 40–46.

[3] Muhammad Aurangzeb Ahmad, Arpit Patel, Carly Eckert, Vikas Kumar, and Ankur Teredesai. 2020. Fairness in machine learning for healthcare. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 3529–3530.

[4] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I Madai, and Precise4Q Consortium. 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making* 20 (2020), 1–9.

[5] Milan Aryal and Nasim Yahyasoltani. 2021. Identifying Catheter and Line Position in Chest X-Rays Using GANs. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 122–127.

[6] Matan Atad, Vitalii Dmytrenko, Yitong Li, Xinyue Zhang, Matthias Keicher, Jan Kirschke, Bene Wiestler, Ashkan Khakzar, and Nassir Navab. 2022. Chexplaining in style: Counterfactual explanations for chest x-rays using stylegan. *arXiv preprint arXiv:2207.07553* (2022).

[7] Amid Ayobi, Jacob Hughes, Christopher J Duckworth, Jakub J Dylag, Sam James, Paul Marshall, Matthew Guy, Anitha Kumaran, Adriane Chapman, Michael Boniface, et al. 2023. Computational Notebooks as Co-Design Tools: Engaging Young Adults Living with Diabetes, Family Carers, and Clinicians with Machine Learning Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[8] Anne Kathrine Petersen Bach, Trine Munch Nørgaard, Jens Christian Brok, and Niels van Berkel. 2023. "If I Had All the Time in the World": Ophthalmologists' Perceptions of Anchoring Bias Mitigation in Clinical AI Support. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.

[9] Ivo Baltruschat, Leonhard Steinmeister, Hannes Nickisch, Axel Saalbach, Michael Grass, Gerhard Adam, Tobias Knopp, and Harald Ittrich. 2021. Smart chest X-ray worklist prioritization using artificial intelligence: a clinical workflow simulation. *European radiology* 31 (2021), 3837–3845.

[10] Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, et al. 2024. MAIRA-2: Grounded Radiology Report Generation. *arXiv preprint arXiv:2406.04449* (2024).

[11] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. 2023. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15016–15027.

[12] Amie J Barda, Christopher M Horvat, and Harry Hochheiser. 2020. A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. *BMC medical informatics and decision making* 20, 1 (2020), 1–16.

[13] Sally L Baxter, Jeremy S Bass, and Amy M Sitapati. 2020. Barriers to implementing an artificial intelligence model for unplanned readmissions. *ACI open* 4, 02 (2020), e108–e113.

[14] 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*. 1–12.

[15] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623.

[16] Marc Berg, Jos Aarts, and Johan van der Lei. 2003. ICT in health care: sociotechnical approaches. *Methods of information in medicine* 42, 04 (2003), 297–301.

[17] Michael H Bernstein, Michael K Atalay, Elizabeth H Dibble, Aaron WP Maxwell, Adib R Karam, Saurabh Agarwal, Robert C Ward, Terrance T Healey, and Grayson L Baird. 2023. Can incorrect artificial intelligence (AI) results impact radiologists, and if so, what can we do about it? A multi-reader pilot study of lung cancer detection with chest radiography. *European Radiology* (2023), 1–7.

[18] Hannah Bleher and Matthias Braun. 2022. Diffused responsibility: attributions of responsibility in the use of AI-driven clinical decision support systems. *AI and Ethics* 2, 4 (2022), 747–761.

[19] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*. Springer, 1–21.

[20] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).
[21] Thuvanan Borvornvitchotikarn and Thongchai Yooyativong. 2022. Pre-Activation-Spatial Attention Module for Multiple Catheters and Tubes Classification. In *2022 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)*. IEEE, 238–241.
[22] Robert Bowman, Camille Nadal, Kellie Morrissey, Anja Thieme, and Gavin Doherty. 2023. Using thematic analysis in healthcare HCI at CHI: A scoping review. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.
[23] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.
[24] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative research in psychology* 18, 3 (2021), 328–352.
[25] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
[26] Eleanor R Burgess, Ivana Jankovic, Melissa Austin, Nancy Cai, Adela Kapuścińska, Suzanne Currie, J Marc Overhage, Erika S Poole, and Jofish Kaye. 2023. Healthcare AI Treatment Decision Support: Design Principles to Enhance Clinician Adoption and Trust. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.
[27] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis* 66 (2020), 101797.
[28] David Byrne. 2022. A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & quantity* 56, 3 (2022), 1391–1412.
[29] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–14.
[30] 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. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.
[31] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2021. Onboarding Materials as Cross-functional Boundary Objects for Developing AI Assistants. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.
[32] Francisco Maria Calisto, João Fernandes, Margarida Morais, Carlos Santiago, João Maria Abrantes, Nuno Nunes, and Jacinto C Nascimento. 2023. Assertiveness-based Agent Communication for a Personalized Medicine on Medical Imaging Diagnosis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.
[33] Francisco Maria Calisto, Carlos Santiago, Nuno Nunes, and Jacinto C Nascimento. 2021. Introduction of human-centric AI assistant to aid radiologists for multimodal breast image classification. *International Journal of Human-Computer Studies* 150 (2021), 102607.
[34] Francisco Maria Calisto, Carlos Santiago, Nuno Nunes, and Jacinto C Nascimento. 2022. BreastScreening-AI: Evaluating medical intelligent agents for human-AI interactions. *Artificial Intelligence in Medicine* 127 (2022), 102285.
[35] Winnie Chen, Kirsten Howard, Gillian Gorham, Claire Maree O'Bryan, Patrick Coffey, Bhavya Balasubramanya, Asanga Abeyaratne, and Alan Cass. 2022. Design, effectiveness, and economic outcomes of contemporary chronic disease clinical decision support systems: a systematic review and meta-analysis. *Journal of the American Medical Informatics Association* 29, 10 (2022), 1757–1772.
[36] Isabel Chien, Nina Deliu, Richard Turner, Adrian Weller, Sofia Villar, and Niki Kilbertus. 2022. Multi-disciplinary fairness considerations in machine learning for clinical trials. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 906–924.
[37] Enrico Coiera. 2019. The last mile: where artificial intelligence meets reality. *Journal of medical Internet research* 21, 11 (2019), e16323.
[38] Kate Crawford. 2021. The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.
[39] James Dawson. 2007. Nasogastric tube incidents and the use of the'whoosh test'. *Critical Care* 11 (2007), 1–1.
[40] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. 2021. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence* 3, 7 (2021), 610–619.
[41] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association* 23, 2 (2016), 304–310.
[42] The Medical Education Directorate. [n. d.]. What happens when I submit a Datix? https://www.med.scot.nhs.uk/trainee-doctors/learning-from-datixes/what-happens-when-i-submit-a-datix#:~:text=What%20is%20DATIX%3F,identify%20learning%20and%20implement%20improvement.

[43] Ignat Drozdov, Rachael Dixon, Benjamin Szubert, Jessica Dunn, Darren Green, Nicola Hall, Arman Shirandami, Sofia Rosas, Ryan Grech, Srikanth Puttagunta, et al. 2023. An Artificial Neural Network for Nasogastric Tube Position Decision Support. *Radiology: Artificial Intelligence* 5, 2 (2023), e220165.

[44] Abdelfettah Elaanba, Mohammed Ridouani, and Larbi Hassouni. 2021. Automatic detection using deep convolutional neural networks for 11 abnormal positioning of tubes and catheters in chest X-ray Images. In *2021 IEEE World AI IoT Congress (AIIoT)*. IEEE, 0007–0012.

[45] Madeleine Clare Elish. 2019. Moral crumple zones: Cautionary tales in human-robot interaction (pre-print). *Engaging Science, Technology, and Society (pre-print)* (2019).

[46] Madeleine Clare Elish and Elizabeth Anne Watkins. 2020. Repairing Innovation: A Study of Integrating AI in Clinical Care. *Data & Society* (2020), 1–62.

[47] Nikki Fennell, Matthew D Ralston, and Robert M Coleman. 2021. PACS and Other Image Management Systems. *Practical Imaging Informatics: Foundations and Applications for Medical Imaging* (2021), 131–142.

[48] Ross W Filice and Raj M Ratwani. 2020. The case for user-centered artificial intelligence in radiology. , e190095 pages.

[49] Geraldine Fitzpatrick and Gunnar Ellingsen. 2013. A review of 25 years of CSCW research in healthcare: contributions, challenges and future agendas. *Computer Supported Cooperative Work (CSCW)* 22 (2013), 609–665.

[50] Maayan Frid-Adar, Rula Amer, and Hayit Greenspan. 2019. Endotracheal tube detection and segmentation in chest radiographs using synthetic data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 784–792.

[51] Astrid Galsgaard, Tom Doorschodt, Ann-Louise Holten, Felix Christoph Müller, Mikael Ploug Boesen, and Mario Maas. 2022. Artificial intelligence and multidisciplinary team meetings; a communication challenge for radiologists' sense of agency and position as spider in a web? *European Journal of Radiology* 155 (2022), 110231.

[52] Stephen Gilbert, Hugh Harvey, Tom Melvin, Erik Vollebregt, and Paul Wicks. 2023. Large language model AI chatbots require approval as medical devices. *Nature Medicine* (2023), 1–3.

[53] Fernanda Raphael Escobar Gimenes, Marta Cristiane Alves Pereira, Patricia Rezende do Prado, Rhanna Emanuela Fontenele Lima de Carvalho, Janine Koepp, Ligia Menezes de Freitas, Thalyta Cardoso Alux Teixeira, and Adriana Inocenti Miasso. 2019. Nasogastric/Nasoenteric tube-related incidents in hospitalised patients: a study protocol of a multicentre prospective cohort study. *BMJ open* 9, 7 (2019), e027967.

[54] Jennifer C Ginestra, Heather M Giannini, William D Schweickert, Laurie Meadows, Michael J Lynch, Kimberly Pavan, Corey J Chivers, Michael Draugelis, Patrick J Donnelly, Barry D Fuchs, et al. 2019. Clinician perception of a machine learning-based early warning system designed to predict severe sepsis and septic shock. *Critical care medicine* 47, 11 (2019), 1477.

[55] Yu Gordienko, Peng Gang, Jiang Hui, Wei Zeng, Yu Kochura, Oleg Alienin, Oleksandr Rokovyi, and Sergii Stirenko. 2019. Deep learning with lung segmentation and bone shadow exclusion techniques for chest X-ray analysis of lung cancer. In *Advances in Computer Science for Engineering and Education 13*. Springer, 638–647.

[56] L Gorospe-Sarasúa, JM Muñoz-Olmedo, F Sendra-Portero, and R de Luis-García. 2022. Challenges of Radiology education in the era of artificial intelligence. *Radiología (English Edition)* 64, 1 (2022), 54–59.

[57] Hongyan Gu, Yuan Liang, Yifan Xu, Christopher Kazu Williams, Shino Magaki, Negar Khanlou, Harry Vinters, Zesheng Chen, Shuo Ni, Chunxu Yang, et al. 2023. Improving workflow integration with XPath: Design and evaluation of a human-AI diagnosis system in pathology. *ACM Transactions on Computer-Human Interaction* 30, 2 (2023), 1–37.

[58] Hongyan Gu, Chunxu Yang, Mohammad Haeri, Jing Wang, Shirley Tang, Wenzhong Yan, Shujin He, Christopher Kazu Williams, Shino Magaki, and Xiang'Anthony' Chen. 2023. Augmenting Pathologists with NaviPath: Design and Evaluation of a Human-AI Collaborative Navigation System. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[59] K Harrison, H Pullen, C Welsh, O Oktay, J Alvarez-Valle, and R Jena. 2022. Machine learning for auto-segmentation in radiotherapy planning. *Clinical Oncology* 34, 2 (2022), 74–88.

[60] Tianxing He, Shengcheng Yu, Ziyuan Wang, Jieqiong Li, and Zhenyu Chen. 2019. From data quality to model quality: An exploratory study on deep learning. In *Proceedings of the 11th Asia-Pacific Symposium on Internetware*. 1–6.

[61] NHS Kent Community Health. 2022. Nasogastric feeding tube. Leaflet. https://www.kentcht.nhs.uk/leaflet/nasogastric-feeding-tube/#:~:text=When%20feeding%2C%20please%20sit%20or,risk%20of%20the%20tube%20blocking.

[62] Robert DE Henderson, Xin Yi, Scott J Adams, and Paul Babyn. 2021. Automatic detection and classification of multiple catheters in neonatal radiographs with deep learning. *Journal of digital imaging* 34, 4 (2021), 888–897.

[63] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E Imel, and David C Atkins. 2017. Designing contestability: Interaction design, machine learning, and mental health. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. 95–99.

[64] Tad Hirsch, Christina Soma, Kritzia Merced, Patty Kuo, Aaron Dembe, Derek D Caperton, David C Atkins, and Zac E Imel. 2018. " It's hard to argue with a computer" Investigating Psychotherapists' Attitudes towards Automated Evaluation. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 559–571.

[65] Xinyue Hu et al. [n. d.]. Medical-Diff-VQA: A Large-Scale Medical Dataset for Difference Visual Question Answering on Chest X-Ray Images. ([n. d.]).

[66] Jonathan Huang, Luke Neill, Matthew Wittbrodt, David Melnick, Matthew Klug, Michael Thompson, John Bailitz, Timothy Loftus, Sanjeev Malik, Amit Phull, et al. 2023. Generative Artificial Intelligence for Chest Radiograph Interpretation in the Emergency Department. *JAMA network open* 6, 10 (2023), e2336100–e2336100.

[67] Shigao Huang, Jie Yang, Na Shen, Qingsong Xu, and Qi Zhao. 2023. Artificial intelligence in lung cancer diagnosis and prognosis: Current application and future perspective. In *Seminars in Cancer Biology*. Elsevier.

[68] Brian Hurt, Andrew Yen, Seth Kligerman, and Albert Hsiao. 2020. Augmenting interpretation of chest radiographs with deep learning probability maps. *Journal of thoracic imaging* 35, 5 (2020), 285.

[69] Stephanie Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, et al. 2023. MAIRA-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv: 2311.13668* (2023).

[70] Ada Lovelace Institute. 2022. Algorithmic Impact Assessment: A Case Study in Healthcare. https://www.adalovelaceinstitute.org/report/algorithmic-impact-asssessment-case-study-healthcare

[71] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 590–597.

[72] Azra Ismail, Naveena Karusala, and Neha Kumar. 2018. Bridging disconnected knowledges for community health. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–27.

[73] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C Ahn, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. Designing AI for trust and collaboration in time-constrained medical decisions: a sociotechnical lens. In *Proceedings of the 2021 chi conference on human factors in computing systems*. 1–14.

[74] Sowon Jang, Hwayoung Song, Yoon Joo Shin, Junghoon Kim, Jihang Kim, Kyung Won Lee, Sung Soo Lee, Woojoo Lee, Seungjae Lee, and Kyung Hee Lee. 2020. Deep learning–based automatic detection algorithm for reducing overlooked lung cancers on chest radiographs. *Radiology* 296, 3 (2020), 652–661.

[75] Stefanie Jauk, Diether Kramer, Alexander Avian, Andrea Berghold, Werner Leodolter, and Stefan Schulz. 2021. Technology acceptance of a machine learning algorithm predicting delirium in a clinical setting: a mixed-methods study. *Journal of medical systems* 45, 4 (2021), 48.

[76] Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Sabel, Jens Ricke, et al. 2022. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *arXiv preprint arXiv:2212.14882* (2022).

[77] Bonnie E John, Len Bass, Rick Kazman, and Eugene Chen. 2004. Identifying gaps between HCI, software engineering, and design, and boundary objects to bridge them. In *CHI'04 extended abstracts on Human factors in computing systems*. 1723–1724.

[78] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data* 6, 1 (2019), 317.

[79] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221* (2022).

[80] E-Fong Kao, Twei-Shiun Jaw, Chun-Wei Li, Ming-Chung Chou, and Gin-Chung Liu. 2015. Automated detection of endotracheal tubes in paediatric chest radiographs. *Computer methods and programs in biomedicine* 118, 1 (2015), 1–10.

[81] Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. 2019. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine* 17 (2019), 1–9.

[82] Abdul Baseer Mohammed Khan and Syed Mahboob Abrar Ali. 2021. Early detection of malpositioned catheters and lines on chest X-rays using deep learning. In *2021 International Conference on Artificial Intelligence and Computer Science Technology (ICAICST)*. IEEE, 51–55.

[83] Elin Kjelle and Catherine Chilanga. 2022. The assessment of image quality and diagnostic value in X-ray images: a survey on radiographers' reasons for rejecting images. *Insights into Imaging* 13, 1 (2022), 1–6.

[84] Matthew C Koopmann, Kenneth A Kudsk, Molly J Szotkowski, and Susan M Rees. 2011. A team-based protocol and electromagnetic technology eliminate feeding tube placement complications. *Annals of surgery* 253, 2 (2011), 297–302.

[85] Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. 2020. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. *arXiv preprint arXiv:2005.01795* (2020).

[86] Sean Kross and Philip Guo. 2021. Orienting, framing, bridging, magic, and counseling: How data scientists navigate the outer loop of client collaborations in industry and academia. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–28.

[87] Charlotte P Lee. 2007. Boundary negotiating artifacts: Unbinding the routine of boundary objects and embracing chaos in collaborative work. *Computer Supported Cooperative Work (CSCW)* 16 (2007), 307–339.

[88] Hyunkwang Lee, Mohammad Mansouri, Shahein Tajmir, Michael H Lev, and Synho Do. 2018. A deep-learning system for fully-automated peripherally inserted central catheter (PICC) tip detection. *Journal of digital imaging* 31 (2018), 393–402.

[89] Peter Lee, Sebastien Bubeck, and Joseph Petro. 2023. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *New England Journal of Medicine* 388, 13 (2023), 1233–1239.
[90] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110* (2022).
[91] Q Vera Liao, Yunfeng Zhang, Ronny Luss, et al. 2022. Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 10. 147–159.
[92] Christopher J Lindsell, William W Stead, and Kevin B Johnson. 2020. Action-informed artificial intelligence—matching the algorithm to the problem. *Jama* 323, 21 (2020), 2141–2142.
[93] Martin Lindvall, Claes Lundström, and Jonas Löwgren. 2021. Rapid assisted visual search: Supporting digital pathologists with imperfect AI. In *26th International Conference on Intelligent User Interfaces*. 504–513.
[94] Qianchu Liu, Stephanie Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Maria Teodora Wetscherek, Robert Tinn, Harshita Sharma, Fernando Pérez-García, Anton Schwaighofer, et al. 2023. Exploring the Boundaries of GPT-4 in Radiology. *arXiv preprint arXiv:2310.14573* (2023).
[95] Giles Maskell. 2022. Why does demand for medical imaging keep rising?
[96] Stina Matthiesen, Søren Zöga Diederichsen, Mikkel Klitzing Hartmann Hansen, Christina Villumsen, Mats Christian Højbjerg Lassen, Peter Karl Jacobsen, Niels Risum, Bo Gregers Winkel, Berit T Philbert, Jesper Hastrup Svendsen, et al. 2021. Clinician preimplementation perspectives of a decision-support tool for the prediction of cardiac arrhythmia based on machine learning: near-live feasibility and qualitative study. *JMIR human factors* 8, 4 (2021), e26964.
[97] Amarachi B Mbakwe, Ismini Lourentzou, Leo Anthony Celi, and Joy T Wu. 2023. Fairness metrics for health AI: we have a long way to go. *Ebiomedicine* 90 (2023).
[98] Microsoft. [n. d.]. https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Guide.pdf
[99] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547* (2017).
[100] Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P Langlotz, and Dan Jurafsky. 2020. Improving factual completeness and consistency of image-to-text radiology report generation. *arXiv preprint arXiv:2010.10042* (2020).
[101] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature* 616, 7956 (2023), 259–265.
[102] Caroline M Moore, Elena Frangou, Neil McCartan, Aida Santaolalla, Douglas Kopcke, Giorgio Brembilla, Joanna Hadley, Francesco Giganti, Teresa Marsden, Mieke Van Hemelrijck, et al. 2023. Prevalence of MRI lesions in men responding to a GP-led invitation for a prostate health check: a prospective cohort study. *BMJ Oncology* 2, 1 (2023).
[103] Nabla. 2023. Nabla Copilot · Enjoy care again. https://www.nabla.com/ [Accessed 11-08-2023].
[104] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375* (2023).
[105] Nuance-Microsoft. 2023. Nuance and Microsoft Announce the First Fully AI-Automated Clinical Documentation Application for Healthcare — news.nuance.com. https://news.nuance.com/2023-03-20-Nuance-and-Microsoft-Announce-the-First-Fully-AI-Automated-Clinical-Documentation-Application-for-Healthcare. [Accessed 11-08-2023].
[106] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.
[107] Francis O'Connell, Justin Ong, Crystal Donelan, and Ali Pourmand. 2021. Emergency department approach to gastric tube complications and review of the literature. *The American Journal of Emergency Medicine* 39 (2021), 259–e5.
[108] Nazmun Nisat Ontika, Sheree May Sassmannshausen, Aparecido Fabiano Pinatti De Carvalho, and Volkmar Pipek. 2023. PAIRADS: Hybrid Interaction Between Humans and AI in Radiology. In *HHAI 2023: Augmenting Human Intellect*. IOS Press, 395–397.
[109] Nazmun Nisat Ontika, Hussain Abid Syed, Sheree May Saßmannshausen, Richard HR Harper, Yunan Chen, Sun Young Park, Miria Grisot, Astrid Chow, Nils Blaumer, Aparecido Fabiano Pinatti de Carvalho, et al. 2022. Exploring human-centered AI in healthcare: diagnosis, explainability, and trust. (2022).
[110] Tariq Osman Andersen, Francisco Nunes, Lauren Wilcox, Elizabeth Kaziunas, Stina Matthiesen, and Farah Magrabi. 2021. Realizing AI in healthcare: challenges appearing in the wild. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–5.
[111] Bhavik N Patel, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, et al. 2019. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ digital medicine* 2, 1 (2019), 111.
[112] Fernando Pérez-García, Sam Bond-Taylor, Pedro P Sanchez, Boris van Breugel, Daniel C Castro, Harshita Sharma, Valentina Salvatelli, Maria TA Wetscherek, Hannah Richardson, Matthew P Lungren, et al. 2023. RadEdit: stress-testing biomedical vision models via diffusion image editing. *arXiv preprint arXiv:2312.12865* (2023).

[113] Eike Petersen, Yannik Potdevin, Esfandiar Mohammadi, Stephan Zidowitz, Sabrina Breyer, Dirk Nowotka, Sandra Henn, Ludwig Pechmann, Martin Leucker, Philipp Rostalski, et al. 2022. Responsible and regulatory conform machine learning for medicine: a survey of challenges and solutions. *IEEE Access* 10 (2022), 58375–58418.

[114] Cécile Petitgand, Aude Motulsky, Jean-Louis Denis, and Catherine Régis. 2020. Investigating the barriers to physician adoption of an artificial intelligence-based decision support system in emergency care: an interpretative qualitative study. In *Digital Personalized Health and Medicine*. IOS Press, 1001–1005.

[115] Thomas Ploug and Søren Holm. 2020. The four dimensions of contestable AI diagnostics-A patient-centric approach to explainable AI. *Artificial Intelligence in Medicine* 107 (2020), 101901.

[116] Policy and Guideline Committee. 2023. . https://secure.library.leicestershospitals.nhs.uk/PAGL/Shared%20Documents/Nasogastric%20and%20Orogastric%20Tubes%20in%20Adults%20UHL%20Policy.pdf

[117] Sam Preston, Mu Wei, Rajesh Rao, Robert Tinn, Naoto Usuyama, Michael Lucas, Yu Gu, Roshanthi Weerasinghe, Soohee Lee, Brian Piening, et al. 2023. Toward structuring real-world data: Deep learning for extracting oncology information from clinical text with patient-level supervision. *Patterns* 4, 4 (2023).

[118] Rob Procter, Peter Tolmie, and Mark Rouncefield. 2023. Holding AI to account: Challenges for the delivery of trustworthy AI in healthcare. *ACM Transactions on Computer-Human Interaction* 30, 2 (2023), 1–34.

[119] Nigel Rees, Kelly Holding, and Mark Sujan. 2023. Information governance as a socio-technical process in the development of trustworthy healthcare AI. *Frontiers in Computer Science* 5 (2023), 1134818.

[120] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.

[121] Ankit Rungta. 2021. *Detection of the Malpositioned Catheters and Endotracheal Tubes on Radiographs using Deep Learning Methods*. Ph. D. Dissertation. Dublin, National College of Ireland.

[122] NR Sahni, P Mishra, B Carrus, and DM Cutler. [n. d.]. Administrative Simplification: How to Save a Quarter-Trillion Dollars in US Healthcare. McKinsey & Company. October 20, 2021.

[123] Public Health Scotland. [n. d.]. Use of the National Safe Haven. https://www.isdscotland.org/products-and-services/edris/use-of-the-national-safe-haven/

[124] Jarrel CY Seah, Cyril HM Tang, Quinlan D Buchlak, Xavier G Holt, Jeffrey B Wardman, Anuar Aimoldin, Nazanin Esmaili, Hassan Ahmad, Hung Pham, John F Lambert, et al. 2021. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *The Lancet Digital Health* 3, 8 (2021), e496–e506.

[125] Teik Choon See, Mark Callaway, Raman Uberoi, Amanda Martin, Alexandra Lipton, Sue Johnson, Alastair Henderson, Katherine Henderson, James France, Richard Roope, et al. 2023. Alerts and notification of imaging reports recommendations. *Clinical Radiology* 78, 3 (2023), e227–e236.

[126] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. " The human body is a black box" supporting clinical decision-making with deep learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 99–109.

[127] Mark P Sendak, William Ratliff, Dina Sarro, Elizabeth Alderton, Joseph Futoma, Michael Gao, Marshall Nichols, Mike Revoir, Faraz Yashar, Corinne Miller, et al. 2020. Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. *JMIR medical informatics* 8, 7 (2020), e15182.

[128] Manan Shah, Derek Shu, VB Surya Prasath, Yizhao Ni, Andrew H Schapiro, and Kevin R Dufendach. 2021. Machine learning for detection of correct peripherally inserted central catheter tip position from radiology reports in infants. *Applied Clinical Informatics* 12, 04 (2021), 856–863.

[129] Murray Shanahan. 2022. Talking About Large Language Models. *arXiv preprint arXiv:2212.03551* (2022).

[130] Varun Singh, Varun Danda, Richard Gorniak, Adam Flanders, and Paras Lakhani. 2019. Assessment of critical feeding tube malpositions on radiographs using deep learning. *Journal of digital imaging* 32 (2019), 651–655.

[131] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature* (2023), 1–9.

[132] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617* (2023).

[133] Ilyas Sirazitdinov, Matthias Lenga, Ivo M Baltruschat, Dmitry V Dylov, and Axel Saalbach. 2021. Landmark constellation models for central venous catheter malposition detection. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 1132–1136.

[134] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M Kahn, and Adam Perer. 2023. Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[135] Ruchika Sreedhar, Jeena Thomas, and Ebin Deni Raj. 2021. Detection Of Chest Catheters Using Mask R-CNN. In *2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*. IEEE, 1–6.

[136] Lea Strohm, Charisma Hehakaya, Erik R Ranschaert, Wouter PC Boon, and Ellen HM Moors. 2020. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. *European radiology* 30 (2020), 5525–5532.

[137] Vaishnavi Subramanian, Hongzhi Wang, Joy T Wu, Ken CL Wong, Arjun Sharma, and Tanveer Syeda-Mahmood. 2019. Automated detection and type classification of central venous catheters in chest x-rays. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*. Springer, 522–530.

[138] Jennifer SN Tang, Jarrel CY Seah, Adil Zia, Jay Gajera, Richard N Schlegel, Aaron JN Wong, Dayu Gai, Shu Su, Tony Bose, Marcus L Kok, et al. 2021. CLiP, catheter and line position dataset. *Scientific Data* 8, 1 (2021), 285.

[139] P Temblett and S George. 2013. pH testing to confirm nasogastric tube position on the ICU: are we wasting our time? *Critical Care* 17, Suppl 2 (2013), P245.

[140] Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 5 (2020), 1–53.

[141] Anja Thieme, Ed Cutrell, Cecily Morrison, Alex Taylor, and Abigail Sellen. 2020. Interpretability as a dynamic of human-AI interaction. *Interactions* 27, 5 (2020), 40–45.

[142] Anja Thieme, Maryann Hanratty, Maria Lyons, Jorge Palacios, Rita Faia Marques, Cecily Morrison, and Gavin Doherty. 2023. Designing human-centered AI for mental health: Developing clinically relevant applications for online CBT treatment. *ACM Transactions on Computer-Human Interaction* 30, 2 (2023), 1–50.

[143] Anja Thieme, Aditya Nori, Marzyeh Ghassemi, Rishi Bommasani, Tariq Osman Andersen, and Ewa Luger. 2023. Foundation Models in Healthcare: Opportunities, Risks & Strategies Forward. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–4.

[144] Tim Torsy, Renée Saman, Kurt Boeykens, Ivo Duysburgh, Mats Eriksson, Sofie Verhaeghe, and Dimitri Beeckman. 2020. Accuracy of the corrected nose-earlobe-xiphoid distance formula for determining nasogastric feeding tube insertion length in intensive care unit patients: A prospective observational study. *International Journal of Nursing Studies* 110 (2020), 103614.

[145] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, et al. 2023. Towards generalist biomedical ai. *arXiv preprint arXiv:2307.14334* (2023).

[146] Stephanie Tulk Jesso, Aisling Kelliher, Harsh Sanghavi, Thomas Martin, and Sarah Henrickson Parker. 2022. Inclusion of clinicians in the development and evaluation of clinical artificial intelligence tools: a systematic literature review. *Frontiers in Psychology* 13 (2022), 830345.

[147] Daiju Ueda, Taichi Kakinuma, Shohei Fujita, Koji Kamagata, Yasutaka Fushimi, Rintaro Ito, Yusuke Matsui, Taiki Nozaki, Takeshi Nakaura, Noriyuki Fujima, et al. 2023. Fairness of artificial intelligence in healthcare: review and recommendations. *Japanese Journal of Radiology* (2023), 1–13.

[148] Stephanie Valencia, Richard Cave, Krystal Kallarackal, Katie Seaver, Michael Terry, and Shaun K Kane. 2023. "The less I type, the better": How AI Language Models can Enhance or Impede Communication for AAC Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.

[149] Himanshu Verma, Jakub Mlynar, Roger Schaer, Julien Reichenbach, Mario Jreige, John Prior, Florian Evéquoz, and Adrien Depeursinge. 2023. Rethinking the role of AI with physicians in oncology: revealing perspectives from clinical and research workflows. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[150] Himanshu Verma, Roger Schaer, Julien Reichenbach, Mario Jreige, John O Prior, Florian Evéquoz, and Adrien Depeursinge. 2021. On improving physicians' trust in AI: Qualitative inquiry with imaging experts in the oncological domain. *BMC Medical Imaging, in review* (2021).

[151] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI doctor" in rural clinics: Challenges in AI-powered clinical decision support system deployment. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–18.

[152] Jiangkun Wang, Miyuka Nakamura, and Abderazek Ben Abdallah. 2022. Efficient AI-enabled pneumonia detection in chest X-ray images. In *2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech)*. IEEE, 470–474.

[153] Liuping Wang, Zhan Zhang, Dakuo Wang, Weidan Cao, Xiaomu Zhou, Ping Zhang, Jianxing Liu, Xiangmin Fan, and Feng Tian. 2023. Human-centered design and evaluation of AI-empowered clinical decision support systems: a systematic review. *Frontiers in Computer Science* 5 (2023), 1187299.

[154] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2097–2106.

[155] Siyuan Wei, Youngwon Choi, M Wasil Wahi-Anwar, Liza Shrestha, Koon-Pong Wong, and Matthew S Brown. 2023. Catheter segmentation in chest x-ray: improving imbalanced segmentation with a class frequency weighted loss function. In *Medical Imaging 2023: Computer-Aided Diagnosis*, Vol. 12465. SPIE, 433–439.

[156] Lauren Wilcox, Robin Brewer, and Fernando Diaz. 2023. AI Consent Futures: A Case Study on Voice Data Collection with Clinicians. (2023).

[157] Malwina Anna Wójcik. 2022. Foundation Models in Healthcare: Opportunities, Biases and Regulatory Prospects in Europe. In *International Conference on Electronic Government and the Information Systems Perspective*. Springer, 32–46.

[158] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang'Anthony' Chen. 2020. CheXplain: enabling physicians to explore and understand data-driven, AI-enabled medical imaging analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[159] Shawn Xu, Lin Yang, Christopher Kelly, Marcin Sieniek, Timo Kohlberger, Martin Ma, Wei-Hung Weng, Attila Kiraly, Sahar Kazemzadeh, Zakkai Melamed, et al. 2023. ELIXR: Towards a general purpose X-ray artificial intelligence system through alignment of large language models and radiology vision encoders. *arXiv preprint arXiv:2308.01317* (2023).

[160] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. 2023. Harnessing biomedical literature to calibrate clinicians' trust in AI decision support systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.

[161] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.

[162] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–11.

[163] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F Antaki. 2016. Investigating the heart pump implant decision process: opportunities for decision support tools to help. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4477–4488.

[164] Xin Yi, Scott J Adams, Robert DE Henderson, and Paul Babyn. 2020. Computer-aided assessment of catheters and tubes on radiographs: How good is artificial intelligence for assessment? *Radiology: Artificial Intelligence* 2, 1 (2020), e190082.

[165] Nur Yildirim, Changhoon Oh, Deniz Sayar, Kayla Brand, Supritha Challa, Violet Turri, Nina Crosby Walton, Anna Elise Wong, Jodi Forlizzi, James McCann, et al. 2023. Creating Design Resources to Scaffold the Ideation of AI Concepts. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 2326–2346.

[166] Nur Yildirim, Mahima Pushkarna, Nitesh Goyal, Martin Wattenberg, and Fernanda Viégas. 2023. Investigating how practitioners use human-ai guidelines: A case study on the people+ ai guidebook. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.

[167] Nur Yildirim, Hannah Richardson, Maria Teodora Wetscherek, Junaid Bajwa, Joseph Jacob, Mark Ames Pinnock, Stephen Harris, Daniel Coelho De Castro, Shruthi Bannur, Stephanie Hyland, et al. 2024. Multimodal healthcare AI: identifying and designing clinically relevant vision-language applications for radiology. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–22.

[168] Dong Whi Yoo, Hayoung Woo, Sachin R Pendse, Nathaniel Young Lu, Michael L Birnbaum, Gregory D Abowd, and Munmun De Choudhury. 2024. Missed Opportunities for Human-Centered AI Research: Understanding Stakeholder Collaboration in Mental Health AI Research. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–24.

[169] Dingding Yu, Kaijie Zhang, Lingyan Huang, Bonan Zhao, Xiaoshan Zhang, Xin Guo, Miaomiao Li, Zheng Gu, Guosheng Fu, Minchun Hu, et al. 2020. Detection of peripherally inserted central catheter (PICC) in chest X-ray images: A multi-task deep learning model. *Computer Methods and Programs in Biomedicine* 197 (2020), 105674.

[170] Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. 2023. Evaluating progress in automatic chest X-ray radiology report generation. *Patterns* 4, 9 (2023). https://doi.org/10.1016/j.patter.2023.100802

[171] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. 2018. Artificial intelligence in healthcare. *Nature biomedical engineering* 2, 10 (2018), 719–731.

[172] Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, et al. 2023. Coding Inequity: Assessing GPT-4's Potential for Perpetuating Racial and Gender Biases in Healthcare. *medRxiv* (2023), 2023–07.

[173] Hubert D Zając, Dana Li, Xiang Dai, Jonathan F Carlsen, Finn Kensing, and Tariq O Andersen. 2023. Clinician-facing AI in the Wild: Taking Stock of the Sociotechnical Challenges and Opportunities for HCI. *ACM Transactions on Computer-Human Interaction* 30, 2 (2023), 1–39.

[174] Meng Zhang, Hong Zhu, Zheng Liu, and Xuexue Deng. 2021. Malposition of a Nasogastric Feeding Tube Into the Right Pleural Cavity of a Nasopharyngeal Carcinoma Patient After Radiotherapy and Chemotherapy: A Case Report. (2021).

[175] Esra Zihni, Vince Istvan Madai, Michelle Livne, Ivana Galinovic, Ahmed A Khalil, Jochen B Fiebach, and Dietmar Frey. 2020. Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome. *Plos one* 15, 4 (2020), e0231166.

A APPENDIX

A.1 Existing ML/AI Work for Lines & Tubes Detection

Table 7 provides an overview of existing ML/ AI work for lines and tube detection from Chest X-ray images.

A.2 Additional Areas of Opportunity for AI Assistance in NGT workflows

Table 8 summarizes additional ideas for AI utilization that surfaced throughout our user research spanning AI use to: auto-extract relevant patient information; provide useful search functionality; aid patient triage optimization; as well as other means to further improve the detection of NGT-related complications.

Received 31 January 2024; revised 27 September 2024; accepted 28 December 2024

Table 7. Overview of existing ML/ AI work for lines and tube detection from Chest X-ray images.(*) Indicates studies with neonatal [62] or pediatric [80] populations.

Reference	Research Aim	Dataset	AI/ ML Outcome
Abbas et al. [1]	Detect presence of tube type & Classify position of ETTs, CVCs, NGTs and Swan Ganz as normal, borderline or abnormal.	RANZCR CLiP [138]	Use of transfer learning via EfficientNet (B7 with Auxiliary connection) achieved average AUC of 0.963; the authors also experimented with Quantization as a technique to increase inference speed and downsize model weights.
Aryal & Yahyasoltani [5]	Classify position of ETTs, CVCs, NGTs and Swan Ganz as normal, borderline or abnormal.	RANZCR CLiP [138]: 9085 (out of 30083) annotated for catheter	Use of GANs to expand dataset catheter annotations significantly improves classification accuracy from AUC of 0.87 (CNN without synthetic annotations) to 0.96.
Borvornvitchotikarn & Yooyativong [21]	Classify position of ETTs, CVCs, NGTs and Swan Ganz as normal, borderline or abnormal.	RANZCR CLiP [138]: 24,062 for training; 6,021 for validation; 3,255 for testing	Use of novel spatial attention module (called Pac-SA) based on an attention mechanism to enhance multi-label image classification. Best performing model achieves 94.65% Accuracy and 65.18% Precision.
Drodov et al. [43]	Detect NGT malposition on CXRs & evaluate model impact as clinical decision support tool.	CXRs from 14 acute sites in NHS Greater Glasgow and Clyde [123]: 1,132,142 CXRs pre-training (ImageNet); 7,081 CXRs fine-tuning; 335 CXRs evaluation	Their model ensemble achieved classification AUCs of 0.82 for satisfactory; 0.77 for malpositioned; and 0.98 for mispositioned into the lungs.
Elaanba et al. [44]	Detect (binary) position labels for ETTs, CVCs, NGTs and Swan Ganz.	RANZCR CLiP [138]	Their best performing model in this multilabel classification tasks achieved an AUC of 80%.
Hendersen et al. [62] (*)	Detect presence and tube type: ETTs, NGTs, umbilical arterial and venous catheters (UACs, UVCs).	777 neonatal AP chest + abdominal radiographs (49 no catheter; 167 with 1 catheter; 561 with 2+ catheters); labelled by medical student, reviewed by a resident and attending paediatric radiologist; Obtained from NICU of Royal University Hospital, Saskatoon, Saskatchewan, CA (2014-2015).	Average precision achieved in detecting the presence of NGTs was 97-99%; 98-99% for ETTs, 93-98% for UACs, and 93% for UVCs; authors also analysed performance based upon number of catheters in each image.
Kao et al. [80] (*)	Detect presence of ET tube and tip location in paediatric CXRs.	528 CXRs with ETTs + 816 CXRs without ETTs from 412 patients in NICU of Kaohsiung Medical University Hospital, Taiwan (Jan-July 2013).	AUC of 94.3% in detecting existence of an ETT with a tip location detection error of 1.89 ± 2.01 mm.
Khan & Ali [82]	Classify ETTs, CVCs, NGTs and Swan Ganz as present, normal, borderline or abnormal.	RANZCR CLiP [138]: 9083 segmented CXRs NIH ChestX-ray14 [154].	Use of UNet for segmentation and transfer learning via EfficientNet for classification; achieving an AUC score of 0.972.
Lee et al. [88]	Detect PICC line tip location.	600 de-identified, HIPPA compliant DICOM AP CXRs images from 600 patients with visible PICCs; obtained from Massachusetts General Hospital, USA (Jan 2015-Jan 2016).	Best model obtained absolute distances from ground truth with a mean of 3.10 mm, a standard deviation of 2.03 mm, and a root mean squares error (RMSE) of 3.71 mm.
Rungta [121]	Classify position of ETTs, CVCs, NGTs and Swan Ganz as normal, borderline or abnormal.	RANZCR CLiP [138]	Report EfficientNet accuracy of 0.89 for validation and 0.91 for test datasets.
Seah et al. [124]	Classify 127 clinical findings from chest X-ray, which included 34 crucial clinical findings (e.g., suboptimal tube placement) & evaluate model impact as clinical decision support tool.	5 Datasets: MIMIC [78], I-MED, NIH ChestX-ray14 [154], CheXpert [71], PadChest [27]: 821,681 CXR images for model training; 2,589 enriched CXRs for testing.	Report AUCs of their DL model for suboptimal: NGT (0.984), ETT (0.995), Central line (0.969), and pulmonary arterial catheter (0.992).
Singh et al. [130]	Detect critical vs. non-critical placement of enteric feeding tubes.	5475 de-identified HIPAA compliant chest + abdominal X-rays (5301 non-critical); 2 expert labellers; data source unknown.	Best performing models achieved AUCs of 0.82-0.85 in differentiating critical vs. non-critical placement.
Sirazitdinov et al. [133]	Detect malpositioned CVC tip utilizing 13 distinct anatomical landmarks.	NIH ChestX-ray14 [154]: 300 manually selected CXRs that had CVCs present; annotated 13 landmarks + CVC tip position.	Model achieved an AUC of 0.96 for identifying malpositioned CVC tips.
Sreedhar et al. [135]	Detect the presence and correct placement of multiple catheters.	RANZCR CLiP [138]	Proposal of using Mask R-CNN technique.
Subramanian et al. [137]	Detect presence and differentiate between four CVC types (PICCs, Swan-Ganz, IJ lines, Subclavian lines).	NIH ChestX-ray14 [154]: 1500 AP CXRs (608 pixel-level CVC type annotations); 3000 CXR (2381 with external medical device, image-level); 10,746 CXRs with at least 1 CVC (Image-level).	Best performing models achieved 85.2% accuracy in CVC detection (91.6% precision) and CVC type classification (95.2% precision).
Yu et al. [169]	Segment PICC line and detect its tip.	348 AP CXRs from 326 patients with visible PICCs for segmentation tasks; a subset (174 CXR) labelled for tip detection task; obtained from Quhua Hospital of Zhejiang Province and Hospital of Zhejiang University.	Their model on catheter segmentation achieved an F1 score of 0.58 on both the test and validation set; and an F1 score of 0.74 on the test and 0.69 on the validation set for tip detection.

Table 8. List of additional areas of opportunity where AI could assist NGT workflows identified through the user research.

User Need	**Application: Additional areas of opportunity**
Assist with NGT image review/ reduce administrative burden	• Auto-extract relevant patient history information from electronic health record to guide reporter review (e.g., if patient had stomach surgery, presence of hiatus hernia or situs inversus). • Auto-extract/ generate X-ray request information that includes relevant process/ triage descriptions (e.g., if methods of obtaining aspirates were followed and what pH checks revealed) to improve decision making/ reduce efforts in soliciting the information. • Provide "similar image" search capability to aid NGT interpretation (e.g., for patients with a similar, unusual anatomy or surgery). Enabling a review of how NGT placement was assessed and deemed as safe in those cases could assist clinical decision making. • Enable advanced AI search within EPIC/ auto-generate NGT specific patient trajectory summary to help reporters understand if a misplaced NGT has already been acted upon by clinical team (as documented in their records).
Improve workflow efficiency	• Optimize what type of image reporter (NGT trained clinicians vs. reporting radiographer vs. radiology registrar/ senior radiologist) should assess/ report (sub-optimal) NGT placement based on criticality, required expertise and staff availability.
Detect NGT-related complications	• Improve NGT misplacement/ complication detection by including CEASE data or patterns in patient vital changes at time of insertion (e.g., changes in heartrate, lung oxygen saturation).
Analyse additional factors for lung aspiration	• Assess implications of feeding position (45 degrees, sloped, slumped down on bed); patient state (fully awake, half drowsy); feeding regime; and feeding time or interval rates (e.g., night, breaks, 4-hour intervals) on NGT-related lung inflammation outcomes.