

**P56: Benchmark Commercial AI product for Medical Imaging Services Provider**

**Client: Simon Poon**

**ISYS3888\_TU15\_02\_Group2**

**Tutor: Yoki**

**Group members details**

* Bronte Lee 490420056
* Cailin Feldman 480439062
* Kiran Papalkar 490411205
* Zichen Li 500174041 (Second group: SOFT3888\_W16\_02, project: P25 - Travel plan and diary application)
* Jingbang Men 500401903
* Zheyuan Zhang 500023392
* Baoxi Liu 480022439 (Second group: SOFT3888\_TU12\_01, project: P08 - WeBike)

**Submission Date**

**5th September 2022**



**School of Information Technologies**

**Faculty of Engineering & IT**

**ASSIGNMENT/PROJECT COVERSHEET - GROUP ASSESSMENT**

**Unit of Study: ISYS3888**

**Assignment name: First Group Report**

**Tutorial time: Tuesday 15:00 Tutor name: Yoki**

**DECLARATION**

**We the undersigned declare that we have read and understood the *University of Sydney Student Plagiarism: Coursework Policy and Procedure*, and except where specifically acknowledged, the work contained in this assignment/project is our own work, and has not been copied from other sources or been previously submitted for award or assessment.**

**We understand that failure to comply with the *Student Plagiarism: Coursework Policy and Procedure* can lead to severe penalties as outlined under Chapter 8 of the *University of Sydney By-Law 1999* (as amended). These penalties may be imposed in cases where any significant portion of my submitted work has been copied without proper acknowledgement from other sources, including published works, the internet, existing programs, the work of other students, or work previously submitted for other awards or assessments.**

**We realise that we may be asked to identify those portions of the work contributed by each of us and required to demonstrate our individual knowledge of the relevant material by answering oral questions or by undertaking supplementary work, either written or in the laboratory, in order to arrive at the final assessment mark.**

| **Project team members** | | | | |
| --- | --- | --- | --- | --- |
| **Student name** | **Student ID** | **Participated** | **Agree to share** | **Signature** |
| **Bronte Lee** | **490420056** | **Yes** | **Yes** | **Bronte Lee** |
| **Zichen Li** | **500174041** | **Yes** | **Yes** | **Zichen Li** |
| **Cailin Feldman** | **480439062** | **Yes** | **Yes** | **Cailin** |
| **Jingbang Men** | **500401903** | **Yes** | **Yes** | **Jingbang Men** |
| **Kiran Papalkar** | **490411205** | **Yes** | **Yes** |  |
| **Baoxi Liu** | **480022439** | **Yes** | **Yes** |  |
| **Zheyuan Zhang** | **500023392** | **Yes** | **Yes** | **Zheyuan Zhang** |

**Level 2, SIT Building, J12 T +61 2 9351 3423 ABN 15 211 513 464**

**The University of Sydney F +61 2 9351 3838 CRICOS 00026A**

**NSW 2006 Australia E** [**sit.info@sydney.edu.au**](mailto:sit.info@sydney.edu.au)

## Executive Summary

P56: Benchmark Commercial AI product for Medical Imaging Services Provider

**Group Members**

* Cailin Feldman
* Bronte Lee
* Kiran Papalkar
* Baoxi Liu
* Zichen Li
* Zheyuan Zhang
* Jingbang Men

**Project Aims/Goals**

The aim of this project is to evaluate the product qXR using the ECLAIR guidelines, with a focus on diagnostic test accuracy and user experience. The goal is for our clients to use the evaluation with high confidence to support their decision about qXR.

qXR is a highly reputable technology, with the ability to detect multiple abnormal findings on a chest X-Ray in less than one minute. However, our project will evaluate where this product stands compared to other AI TB detection products. Our team will use qXR as a benchmark against other technologies and compare accuracy and usability to determine if qXR is technology that should be suggested or would an alternative be more advisable.

**Project Overview**

The project is divided into two primary sub-parts. The first part evaluates the diagnostic test accuracy of different AI products which diagnose tuberculosis in comparison to qXR that will also include an interactive dashboard for our client to easily compare accuracy metrics. The second part evaluates factors affecting user acceptance of qXR products for diagnosing tuberculosis. A systematic review using PRISMA will be conducted to collect papers for the DTA and user experience meta-analyses.

In addition, a high-level overview of the other ECLAIR guidelines questions will be provided.

A recommendation will be made to the client which will be whether or not to purchase or trial qXR, and suggest an alternative if needed.

**Acknowledgements**

Simon Poon who is an Associate Professor in the School of Computer Science who is providing guidance throughout our research.

Dr Neysa Petrina who is currently leading a consulting study with a national health services provider in medical imaging who has provided our team with deeper context in our ISYS3888 project.

**Table of Contents**

[Executive Summary](#_heading=h.bq44qy9r3z7e)

[1 Introduction](#_heading=h.29iub5abssiy)

[1.1 Problem Statement](#_heading=h.smb0kwj703qx)

[1.2 Purpose](#_heading=h.bsfkyvpsiabt)

[1.3 Project Vision Objective](#_heading=h.e2zt0957hued)

[1.4 What will the Project Achieve](#_heading=h.rnod3gr3wsix)

[1.5 The key stakeholders, what do they do, and how they interact](#_heading=h.oar4gf350c1a)

[1.6 Identification of resources and risks involved in the project](#_heading=h.qav5g5od0exv)

[2 Overview of system from user view](#_heading=h.xix9dn1c6l6n)

[2.1 Overview](#_heading=h.leghe36c926s)

[2.1.1 User](#_heading=h.fr8yikrrj97b)

[2.1.2 Administrator](#_heading=h.v11coam1z20b)

[2.2 User Stories](#_heading=h.9ndbnbvrq3kj)

[2.2.1 Functional Requirements](#_heading=h.uj5kdlfunmsf)

[2.2.2 Non-Functional Requirements](#_heading=h.piq5wab9hdl1)

[2.3 Constraints](#_heading=h.z8z70j254vin)

[3 Evaluation](#_heading=h.pd60i6w4xjr)

[3.1 Overview](#_heading=h.au57fw96fg0q)

[3.2 Details of Tests](#_heading=h.bmwqx9bbv7v8)

[3.2.1 Article Quality](#_heading=h.yrhcdinoz7pd)

[3.2.2 Systematic Review Quality](#_heading=h.y74p444vleeb)

[3.2.3 Meta-Analysis Quality](#_heading=h.8sdn87xy2v22)

[3.2.4 Evaluation Quality](#_heading=h.l53gl78yyvwo)

[3.2.5 Report Quality](#_heading=h.3e1t84ylzga0)

[3.2.6 Presentation Quality](#_heading=h.ihhqocvwsfvu)

[3.3 Conclusions](#_heading=h.f155rw241bdu)

[4 System structure overview](#_heading=h.6rcsiyk1922t)

[5 Tools to build system](#_heading=h.60cv5ftuww06)

[6 Information search/research and discipline](#_heading=h.ykxqmv92mvnh)

[6.1 Research Discipline: Sensitivity and Specificity](#_heading=h.r75g2ihsdpmc)

[6.2 Research Discipline: ROC and AUC](#_heading=h.bkob69nxqpa4)

[6.3 Research Discipline: User experience](#_heading=h.qcbhin75he9k)

[6.4 Knowledge use](#_heading=h.u8u2evoqissw)

[6.5 Implementation](#_heading=h.tzh9zapxybbx)

[7 Group processes, reflections and conclusions](#_heading=h.vfh51lwhmyd7)

[7.1 Team Structure](#_heading=h.77obe56hea4b)

[7.2 Risks and Challenges](#_heading=h.hh15t66n0bby)

[7.3 Limitations: User experience](#_heading=h.zc101frcr9dz)

[7.4 Limitations: Tech team](#_heading=h.x55wy1g4srbc)

[8 Individual contributions](#_heading=h.ejs5t5duicxr)

[9 Appendices](#_heading=h.sp0w56psw0jy)

[A1. User stories](#_heading=h.u7fjzcawobu9)

[A2. Systematic Review Quality Example](#_heading=h.ks85y9pjvpv2)

[A3. DTA Example Table](#_heading=h.8ur1wtj21lfw)

[A4. Draft Final Report Structure](#_heading=h.rf6hegt43se6)

[A5. Example of Dashboard](#_heading=h.fjxtlkzeps22)

[10 References](#_heading=h.1oqctdcuvq0d)

## 

## 1 Introduction

### 1.1 Problem Statement

At the present moment, there is a large range of certified CAD products that may be used for TB detection and to make professional recommendations. However, an issue arises: what product is most fit for purpose? Ideally, you want a product that is the most accurate at detection and offers the best user experience.

However, due to the amount of research and the difficulty of keeping up with the fast-paced medical technology field, finding all the needed information for analysis is difficult.

### 1.2 Purpose

Tuberculosis infects 10 million people a year and kills 1.5 million. If TB is diagnosed and treated early, most people are cured within six months of diagnosis. The WHO End TB Strategy aims to reduce TB deaths by 90% and to cut new cases by 80% between 2015 and 2030. To achieve this goal, we need to find the missing cases. We do this through early screening of people at greater epidemiological risk than the general population (who are assumed to be disease-free). The first line of prevention is detecting people who may be infected and who require further investigation and treatment. The major techniques for screening include screening for symptoms (such as persistent cough and fever) by clinicians, low-cost rapid molecular assays (mWRD) and examination for radiological abnormalities indicative of TB. The latter has been performed mostly by expert human readers, such as radiographers, radiologists, and trained health care practitioners.

To achieve these goals, there is a need to implement computer-aided detection (CAD) commercial products that will be able to diagnose TB. This technology will improve workflow efficiency.

The ongoing development of AI products for medical diagnosis reveals encouraging results and these products have begun being integrated into the clinical environment. These products can be used to improve workflow, support radiologists and improve patient care. There are multiple certified AI products which diagnose tuberculosis, the performance of each differs in sensitivity, accuracy, effectiveness and other aspects.

The client is seeking more information about the product qXR, particularly its diagnostic test accuracy compared to its competitors and user experience.

### 1.3 Project Vision Objective

The purpose of this project is to evaluate the product qXR using diagnostic test accuracy and user experience. The diagnostic test accuracy of qXR will be benchmarked against its competitors through a systematic review and meta-analysis. User experience studies will be synthesised to identify factors affecting user acceptance.

qXR is a highly reputable technology, with the ability to detect multiple abnormal findings on a chest X-Ray in less than one minute. However, to see where this product stands compared to other AI TB detection products, our team will use qXR as a benchmark against other technologies and compare accuracy and usability to determine if qXR is technology that should be suggested or would an alternative be more advisable.

### 1.4 What will the Project Achieve

The project’s aim is to evaluate the product qXR using the ECLAIR guidelines. Primarily, this involves identifying the best computer-aided detection (CAD) commercial products that will be able to diagnose TB that is on the market for our clients to use with high confidence and research to support the decision. In addition, the team also plans to have an evaluation on user experience by identifying factors affecting user acceptance and the product’s impact on the clinical workflow.

A final report will analyse the various AI technologies using the qXR product from the company Qure.ai as a benchmark. At this stage we will evaluate if qXR is the best choice, or suggest an alternative product instead. Factors we will consider are its accuracy and an identification of factors affecting user acceptance. All of these must be taken into consideration by the client when deciding which AI technology to implement.

To ensure the quality of the content, we are using industry-accepted protocols and guidelines including PRISMA (for our source finding) and ÉCLAIR (for the structure of our evaluation) to ensure high quality systematic reviews and evaluation.

### 1.5 The key stakeholders, what do they do, and how they interact

The key stakeholders are:

* Information System researchers, specifically our client, Simon Poon
* Medical health care professionals who do the screenings for TB, as they are the ones who will be using the technology.
* Program directors who authorise the use of technology and clinical trials
* Software and technical managers who ensure the technology operates as expected.
* The funders of the project to see what the ROI would be.
* Regulators who authorise the use of the technology.
* Patients who are screened using the technology.
* TB screeners who use legacy systems or alternative techniques such as molecular diagnosis.

### 1.6 Identification of resources and risks involved in the project

Publicly available research will be compiled using meta-analysis to determine the validity of each AI.

To ensure the quality of the content, we are using industry-accepted protocols and guidelines including PRISMA (for our source finding) and ECLAIR.

Project risks:

* An identified problem is that our UX evaluation cannot be specific to qXR, rather it is generic with the assumption that the UX experience of qXR is similar. But there is a risk that qXR’s user experience is different to the AI’s we study.
* Poor communication of information remains a risk.
* Quality issues with data collected.
* Differing time zones of members.
* Scope changing during the week.

## 2 Overview of system from user view

### 2.1 Overview

The users for this project include the client that requested the report as well as the radiologists who will use the AI product that is chosen by ABC.

#### 2.1.1 User

The end user of the report is considered to be the x-ray technicians and radiologists that will be most impacted by the choice of AI product for medical imaging purposes, thus, the choice should reflect the most appropriate technology for their usage. These users will access the report in order to understand the choices made by the client in their choice of AI product

#### 2.1.2 Administrator

The administrator in this case is the company that has commissioned the benchmarking report, they will have access to the end report to inform their decision making in the most desirable AI product for their use, in addition to this they will also be provided with an interactive dashboard for the technical viability of these products in order to better understand the capabilities of individual AI products and the ways in which they compare to similar products. [Appendix A4](#_heading=h.fjxtlkzeps22) shows an example of the view the client will have of the dashboard.

### 2.2 User Stories

Due to the nature of this project as a benchmarking report the user stories are limited to what the client and radiologists require from the report, that is, the knowledge they expect to gain through seeking consultants to benchmark these products as well as their general requirements for quality of the report, the numbering of the user stories is an indication of importance to the client. The user stories from this project are also limited because they can only be “completed” when the report itself is completed, therefore, the progress of these user stories can not be appropriately conveyed in this report. See [Appendix A1](#_heading=h.u7fjzcawobu9) for the list of user stories.

#### 2.2.1 Functional Requirements

The functional requirements for this project have been defined as:

* Prove that AI is technically viable
* Benchmark qXR against other AI products
* Identify AI accuracy
* Identify AI efficiency
* Identify AI impact on clinical workflow
* Prove AI is more effective than human interpretation

#### 2.2.2 Non-Functional Requirements

The functional requirements for this project have been defined as:

* Report covers all requirements
* Meets client’s time demands
* Report uses reliable research

### 2.3 Constraints

The major constraints in regards to this project include the gathering of reliable results, due to the use of meta-analysis in the research we require a wide array of research papers that may not be available due to the recent nature of some of these technologies. The project is also limited by the client’s interpretation of the report, if the report is not readable and organised the overall results may not be correctly viewed leading to a failure in the project, further, the dashboard that is constructed from the data is required to be usable and simple.

## 3 Evaluation

### 3.1 Overview

The client will use the final report and presentation of this project to inform their decision about whether to move forward with the product qXR. Therefore, it is essential that the project is of a high quality so the client is confident they can rely upon the evaluation of qXR and recommendation of next steps.

The project has **[number]** user stories involving two different users and each story has a different acceptance criteria which can be viewed in [Appendix A1](#_heading=h.u7fjzcawobu9). However, they share a consistent theme, a detailed and reliable evaluation of qXR.

To assess the quality of the project, a testing plan was developed which has six areas. They are article quality, systematic review quality, meta-analysis quality, evaluation quality, report quality, and presentation quality. A criteria for each area was developed to evaluate its quality.

### 3.2 Details of Tests

#### 3.2.1 Article Quality

The articles collected are the foundation of this project, they are the evidence used for evaluation and if they are of a poor quality the evaluation would be unreliable. The criteria to ensure article quality is as follows:

* Use databases and Google Scholar to search for articles.
* The articles must be limited to journal articles.
* The company who makes qXR, Qure.ai, must not have funded the study undertaken in the article. Excluding Qure.ai funded studies reduces bias, as companies typically make their own products look better.

The quality of the articles collected will be assessed by reviewing each article to check that it satisfies the above criteria. This process occurs during the article collection stage of the project and any articles which do not satisfy this objective criteria are excluded.

#### 3.2.2 Systematic Review Quality

It is important to ensure the systematic review is completed thoroughly because the quality of it can significantly drop if an irrelevant article is included or an important article is left out. The PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) process is followed to ensure that the systematic review is of a high quality. PRISMA was designed to improve transparency and reporting of systematic reviews (Page et al., 2021) and forms the basis for the measures listed below to ensure the systematic review quality:

* Design search queries with keywords relevant to the research question.
* Include all results of search queries to ensure nothing is missed.
* Define a clear inclusion criteria relevant to the question.
* Screen the title and abstract of articles using the inclusion criteria and provide justification for the decision of each article.
* Read the full text of the articles and use the inclusion criteria to select articles to be included and provide justification for the decision of each article.

The quality of the systematic reviews will be assessed by following the measures above and completing a PRISMA flow diagram. A complete example of using the PRISMA method and applying the above quality measures is available in [Appendix A2](#_heading=h.ks85y9pjvpv2).

#### 3.2.3 Meta-Analysis Quality

Meta-analysis is a crucial part of evaluating the diagnostic test accuracy and user experience of qXR. The criteria to assess the quality of the meta-analyses is a methodical process developed with the intent of providing a transparent and reliable analysis.

**Performance and Validation**

* Spreadsheet design: spreadsheets create a uniform and systematic way of recording the data extracted from the articles. An example spreadsheet is located in [Appendix A3](#_heading=h.8ur1wtj21lfw).
* Data extraction: data will be extracted from the articles manually and entered into a spreadsheet. The data will be checked by a second person to minimise the risk of data errors.
* Analysis: compare sensitivity and specificity and use ROC and AUC to visualise the comparison.
* Transparency: all analysis conducted will be recorded and any exclusion of data will be clearly justified.

**User Experience**

* Spreadsheet design: spreadsheets create a uniform and systematic way of recording the data extracted from the articles. A spreadsheet will be designed to record data from user experience articles and data from all the articles will be recorded in the same spreadsheet.
* Data extraction: data will be extracted from the articles manually and entered into a spreadsheet. The data will be checked by a second person to minimise the risk of data errors.
* Analysis: TAM (Technology Acceptance Model) and UTAUT (Unified Theory of Acceptance and Use of Technology) model will be considered as the primary basis for identifying factors affecting user acceptance. If the article does not use one of the models, the data will be fitted to the relevant factor. From these models conclusions about user experience and the effect of the product on the clinical workflow can be drawn.
* Transparency: all analysis conducted will be recorded and any exclusion of data will be clearly justified.

#### 3.2.4 Evaluation Quality

The evaluation of qXR will follow the ECLAIR (Evaluating Commercial AI solutions in Radiology) guidelines because the guidelines are specific to the type of product qXR is. The guidelines are a set of questions which provide a practical guide to evaluate commercial AI solutions (Omoumi et al., 2021). The five areas of evaluation are:

* Relevance
* Performance and validation
* Usability and integration
* Regulatory and legal aspects
* Financial and support services considerations

These areas contain important questions which must be answered when evaluating a commercial AI solution. The client expects a high-level overview of the relevance, regulatory and legal aspects, and financial support services considerations questions. To provide this, high-quality articles will be located and a brief answer containing all key points will be written for each question.

For performance and validation, a much deeper level of analysis is expected by the client. All questions will require a more detailed response, however, specifically for question 2.3, “How has performance been evaluated?”, a systematic review and meta-analysis is required to compare the diagnostic test accuracy of qXR to its competitors and a human to provide benchmarking.

The client also expects a deeper level of analysis for usability and integration. All questions will require a more detailed response, however, specifically for question 3.2, “How exactly does the application impact the workflow?”, requires a systematic review and meta-analysis to assess user experience to identify the product’s impact on the clinical workflow and the factors affecting user acceptance.

The quality of the evaluation completed will be assessed by checking that every question in the ECLAIR guidelines has been answered to the extent that the client expects and the expectations are outlined above.

#### 3.2.5 Report Quality

The report is the document provided to the client and it must be of a high quality so the client can confidently rely on it to inform their decision about the product qXR. In addition to ensuring the content of the report is of a high quality, which is described above, the report itself has the following criteria:

* Structure the report in a clear and logical manner. Initial discussions have been held with the client about the structure of the final report and a draft of this is available in [Appendix A4](#_heading=h.rf6hegt43se6) and discussed further in the [System structure overview](#_heading=h.6rcsiyk1922t). This will be assessed by having the report structure approved by the client and following the structure.
* Use academic writing. This will be assessed by proofreading the document to ensure there are no spelling mistakes, grammatical errors, or informal language. Use clear and concise language, have logically structured paragraphs, and correctly use a consistent referencing style.
* Report is provided to the client by the date determined by them.

#### 3.2.6 Presentation Quality

The purpose of the presentation is to present a summary of the final report. Areas to assess quality are:

* Content: the presentation will summarise the ECLAIR guidelines sections, with the majority of time spent on performance and validation and usability and integration. It will also have a strong focus on the outcome of the evaluation, recommendation, and limitations.
* Slides: the slides are informative without being text heavy and have a visually pleasing and professional look.
* Demonstration: visual aids, such as a dashboard or set of graphs and diagrams, are clear and support understanding the evaluation of qXR.
* Presenters: all team members present in a professional manner, including projecting their voice clearly and making eye contact.
* Q&A: presenters are knowledgeable in the sections of the report they completed so they are prepared to answer any questions the client has.
* Demonstration:
* The presentation is presented to the client by the date determined by them.

### 3.3 Conclusions

The testing plan has been methodically developed with a strong focus on ensuring academic rigour. However, due to the nature of the project, much of the criteria developed is qualitative which means that the quality of the project will likely have some level of subjectivity. To mitigate this, accepted processes, guidelines, and models have been incorporated where possible to raise the quality of the project.

## 

## 4 System structure overview

The final report will first make an introduction to our project, then we will explain the relevance of the project. For example, What our products intended to solve. After that, we will show the performance of the product and provide validation. We will explain how clear the algorithm specification has been designed, how it is trained, how the performance be evaluated, and etc.

We also will show to the client the usability of the product and how it is integrated. We will show how the application can be integrated into the clinic workflow, how it impacts the clinic workflow, the interoperability, and also the requirements of the IT infrastructure. Regulatory and legal aspects will also be explained, and we will also explain to the client how to consider finance and support aspects. The licensing model will be considered, how the user training and follow-up will be handled and etc.

Finally, we will give a conclusion. We will make a recommendation of the products to our client and we will also explain the limitations of the product.

## 5 Tools to build system

Due to the nature of our project, we don’t need any programming skills or tools. We are focusing more on analysis by following some professional guidelines and methods. The relevant methods and guidelines we are using include systematic review (using PRISMA), meta-analysis, ECLAIR guidelines, Excel and BI tool Tableau.

* **Systematic Review (PRISMA):** This is a technique to systematically search and appraise articles. It synthesizes the results of all selected articles to do the comparison between different products. Specifically, we are using PRISMA to compare qXR to all other AI products such as CAD4TB, RADIFY and inferRead DR. In addition, PRISMA also helps us to select appropriate articles and filter out any irrelevant article.
* **Meta-Analysis:** This is a statistical process to analyse the results from several similar articles which are selected in the previous step using PRISMA. It’s also part of systematic review which focuses on the quantitative study to systematically assess the results of all previous articles and make conclusions where necessary.
* **ECLAIR Guidelines:** Our group uses these guidelines as a general structure for our written report. These are the leading guidelines for assessing commercial AI solutions in radiology. It has 5 parts in total which are Relevance, Performance and Validation, Usability and Integration, Regulatory and Legal Aspects and Financial and Support Services Considerations. We’ll mainly focus on the performance and validation as well as usability and integration parts which are mostly relevant to our project.
* **Excel:** This is the main tool we use to collect all the relevant data from each article and it will be used to help make visualization diagrams using BI tools based on the collected data.
* **Tableau:** This is one of the BI tools we will use to build our project prototype. It can create some visualization diagrams based on the import data which is collected using Excel. We may include other BI tools as well depending on the future requirements of our project.

Other tools including Bitbucket and Google Drive are convenient for the communication and collaboration of each team member. Although we don’t have any source code, we use Bitbucket to record all the relevant articles and probably some analysis as well. In addition, Google Drive is used to store all the necessary documents since our project requires a lot of research and analysis.

## 6 Information search/research and discipline

The target audience of our report is a health services company (ABC) who is considering purchasing qXR because they can choose the most suitable AI product for them according to the analysis results in the report. Considering that the decision-making level does not necessarily have a background in medicine and statistics, we need to explain some terms, principles, etc.

### 

### 6.1 Research Discipline: Sensitivity and Specificity

For product comparison, we use sensitivity and specificity. Sensitivity is the ability of a test to identify patients with a disease correctly, specificity is the ability of a test to identify people without the disease correctly. However, the project team cannot directly compare these two values, and high-specificity experiments are likely to cause over-medicalization. Because highly specific experiments tend to capture some people who do not have Disease, it also creates additional anxiety for patients and wastes unnecessary resources. Referring to these two abilities simultaneously, the ability of Ai products to identify patient status can be obtained and avoid false positives and negatives. At the same time, when comparing AI products, to determine the significance of the results, we usually compare another value under specific sensitivity/specificity conditions.

### 

### 6.2 Research Discipline: ROC and AUC

In the comparison process, in order to visualize our results, we still use ROC and AUC. ROC is a graph showing the performance of a classification model at all classification thresholds. The ROC curve is composed of True Positive Rate (the proportion of true positive in true positive + false negative) and False Positive Rate (false positive in false positive + proportion of true negatives). And AUC is the area of this curve to the x-axis (from point(0,0) to point(1,1)). And the larger the area of AUC, the better the model is at distinguishing between patients with the disease and no disease.

### 

### 6.3 Research Discipline: User experience

As for the user experience part of this evaluation, the plan our team designed was to implement the PRISMA process procedurally for both systematic review and meta-analysis. As mentioned in the project scope section and tool section, the aim of this project is to conduct a systematic review and meta-analysis. Based on research from the professional medical research journal BMJ (Liberati et al., 2009), the PRISMA process is the standard process to follow for meta-analysis on technology acceptance. Our client, Simon had also instructed the team on using the PRISMA process to achieve the desired outcome for meta-analysis.

### 

### 6.4 Knowledge use

To summarize what the PRISMA process is, referencing a guideline from the University of North Carolina (The University of North Carolina, Libguides: Creating a PRISMA flow diagram: Prisma 2020), it is a five-step procedure including collecting all the relevant articles can be found in multiple databases, remove duplicates, remove unrelated articles base on abstract, assess the quality of the full text of the remaining researches and extract the valuable data contained. After this process, if followed strictly, it is highly likely that our team is able to form valid benchmarking on different aspects of the user experience of AI models used in medical imaging.

### 

### 6.5 Implementation

Currently, the user experience team has successfully carried out the third step which is the evaluation based on the title and abstract. Before this, we searched all the related articles on user experience evaluations not only toward commercial AIs which function as tuberculosis detectors, but also AI models which are used in medical imaging as a whole. The main reason is that after preliminary research, the UX team was concerned that there were not enough mature studies done on the user experience for AIs detecting TB specifically. Following a discussion and confirmation with our client, our team expanded the targeted AI to the whole medical imaging field. There are approximately eighty articles collected from the three of us on the UX team. The three major databases our team chose the articles from were Google Scholar, PubMed, and Proquest, these databases were selected because of their already proven quality and their positive reputation among the academic community. Apart from that, the google search engine was also used for searching articles, resulting in the inclusion of

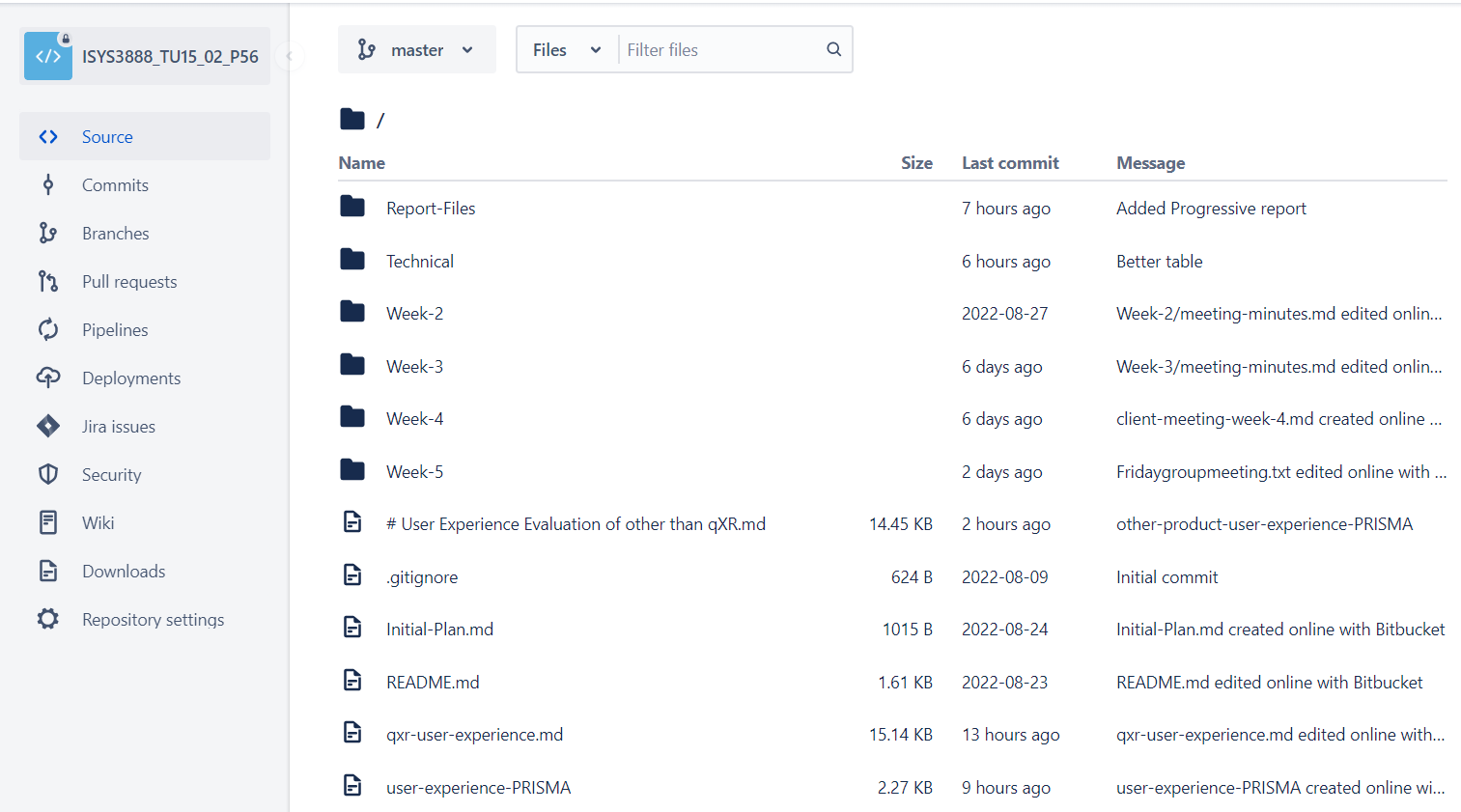
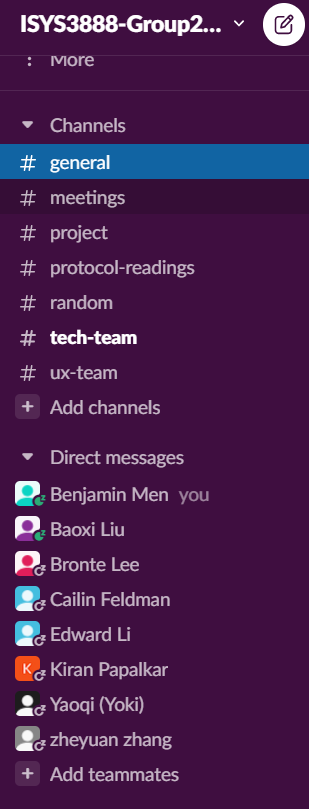
some articles from the well-known medical journal ‘the Lancet' and some other medical journals. Furthermore, our clients Simon and Neysa offered us tremendous help. Multiple articles came from their recommendations, from the comprehension of systematic review and meta-analysis to actual evaluations on both the technology and the user experience aspects.

## 7 Group processes, reflections and conclusions

### 

### 7.1 Team Structure

Unlike the other teams which have tasks focusing on software engineering and front-end design, the task for our team is mainly evaluating products that are already on the market. Hence, it is important for our team to develop a mindset of analysts instead of programmers, and this unorthodox project objective has left the team structure from extreme programming little use. After a serious consultation with the tutor and client, our team decided to adopt a different structure, the structure simply dividing the team into two separate teams according to the two objectives of our project. There are two aims of our project: one is the technical evaluation and the other is the user experience evaluation, and four and three members are assigned to the technical group and user experience group correspondingly. In order for this structure to function effectively, there is one manager for the whole group who is Cailin, she manages the general group matters like distributing tasks and communicating with tutors and clients, especially clients. More importantly, there is one leader for each group to lead the systematic review and meta-analysis process, the leader for the technical team is Kiran and the leader for the UX team is Bronte. Under this structure we are able to set up one client meeting and two group meetings weekly through Zoom, the meeting insured unobstructed communication with the clients and within the team. The tool used for team communication is slack, it is effective in the perspective of timing, and the tools utilized for uploading and editing team documents are bitbucket and google drive.



### 

### 7.2 Risks and Challenges

The team members all hold a positive attitude toward taking responsibility and all demonstrate a certain level of academic comprehension of the information system, this makes the task distribution an easy process. However, there are some potential limitations due to the different nature of our project scope. Although the communication of our team is effective in general, the separation of groups has made communication between the two groups difficult. The data and information shared between the two groups appear to be way less compared to within the two groups, the technical team, and user experience team seldomly exchange ideas and present results inside their own analysis disciplines. Another limitation is that this team consists of both remote students and on-campus students, two of the members are studying off-shore, this combination limits face-to-face interactions and can cause performance issues when presenting to clients and tutors. Sometimes, the internet connection is unstable and results in poor communication and presentation quality. These limitations made the preliminary stage of the systematic review process a slow start but the team was able to adjust and fully committed to the real evaluation later on.

The evaluation process is on track, but there are still some improvements to be made. The most significant risk for tech teams is that no one team tests all AI products (which we are comparing), which leads to different results for different articles (even for the same AI product). This restricts us from using different results from other articles for comparison (but in some cases, we still make cross-article comparisons because there is little relevant data).

### 

### 7.3 Limitations: User experience

The major limitation for the user experience team is the range of the research to be included, at first the user experience team only aimed to find user experience evaluations on commercial AIso which help detect TB, but the team struggled to find any research regarding this aspect. After serious consultation with the tutor and clients, the team decided to expand the scope of this evaluation of the user experience of all the medical imaging AIs. Since the adjustment of scope was made, the team had been finding success with the PRISMA process.

### 

### 7.4 Limitations: Tech team

The major limitation for the tech team is insufficient articles of some AI products. Some AI products' start time is too late, so there are not many comparative experiments and related information, such as JLD-02K. The main reason for lacking relevant information on some AI products is language barriers, such as VUNO. As a Korean product, VUNO is hard to find in an English-language online environment (in fact, I did find some related content on Korean-language websites). Another reason why we have to drop some data is because the data is unbelievably good. Some products can only find some evidence on its official website, and the accuracy obtained far exceeds all products currently on the market. So we will assume the data is biased and therefore it's excluded.

## 8 Individual contributions

## 

Table of Technical Contributions

| Name of Task | **Bronte Lee** | **Zichen Li** | **Cailin Feldman** | **Jingbang Men** | **Kiran Papalkar** | **Baoxi Liu** | **Zheyuan Zhang** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| qXR performance and validation |  |  |  |  |  | <https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/commits/5b36bc077748e60d8e27071921be9d7b97f37585> |  |
| Systematic review of user experience of qXR | [✔](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/qxr-user-experience.md) |  |  |  |  |  |  |
| Systematic review of user experience of other medical imaging product |  |  |  | [✔](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/%23%20User%20Experience%20Evaluation%20of%20other%20than%20qXR.md) |  |  |  |
| Systematic review of user experience for products similar to qXR |  |  | [✔](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/commits/e18397d46cf06d95686a67eb20e0e336fb9628e9) |  |  |  |  |
| Systematic review of technical data for qXR vs. CAD4TB and RADIFY |  |  |  |  | [✔](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Technical/Technical_Articles.docx)  (red) |  |  |
| Systematic review of technical data for qXR vs. Lunit and Vuno |  |  |  |  |  | [✔](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Technical/Technical_Articles.docx)  (blue) |  |
| Conduct a comparison between qXR and Lunit/VUNO |  |  |  |  |  | Commit code:  [8eb7663](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/commits/8eb7663d1246a70b1f5f6830a6dc76b3bb423dd5),  [03dcf75](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/commits/03dcf7572c8dbaad3d4202f88ac6f3d550f36fbe) |  |
| Systematic review of technical data for qXR vs. InferRead and JLD-02K |  | [✔](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Technical/Technical_Articles.docx) |  |  |  |  |  |
| Systematic review of technical data for qXR vs. ChestEye and AXIR |  |  |  |  |  |  | [✔](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Technical/Technical_Articles.docx)  (yellow) |
| Conduct a comparison between qXR and ChestEye, QXIR |  |  |  |  |  | Commit code:  [b48e7d3](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/commits/b48e7d3f6582483fbc702de7b13929a9b66213d2),  [03dcf75](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/commits/03dcf7572c8dbaad3d4202f88ac6f3d550f36fbe) |  |

Table of Non-Technical Contributions

| **Name of Task** | **Bronte Lee** | **Zichen Li** | **Cailin Feldman** | **Jingbang Men** | **Kiran Papalkar** | **Baoxi Liu** | **Zheyuan Zhang** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| XP Summary | [Completed XP roles section](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/wiki/XP%20Summary) | [Completed 12 core practices of XP](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/wiki/XP%20Summary) |  |  | [Iteration Summary](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/wiki/XP%20Summary) |  |  |
| Scope | [For version 1, I edited and elaborated on most sections, particularly deliverables and scope. For the revised version I re-wrote the background and aim sections](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Report-Files/Scope%20Statement.pdf) | [Contribute to the sections including objectives and success criteria](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Report-Files/Scope%20Statement.pdf) |  | [Wrote the original version of the project scope statement](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Report-Files/Scope%20Statement.pdf) | [Contributed to most sections, wrote human resources and other resources. Helped edit whole doc for version 2](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Report-Files/Scope%20Statement.pdf) |  |  |
| 1st Presentation | [Expected outcomes and deliverables slide](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Report-Files/Week%203%20Group%202%20Slide%20Deck.pdf) | [Member responsibilities slide](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Report-Files/Week%203%20Group%202%20Slide%20Deck.pdf) | [Created slide deck](https://docs.google.com/presentation/u/0/d/1mK0g73wGf2sa7chJEEwdXjU4YzEnP2NzN3lbsPK59XM/edit) | [Constructed the project timeline.](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Report-Files/Week%203%20Group%202%20Slide%20Deck.pdf) | [Intro Slide](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Report-Files/Week%203%20Group%202%20Slide%20Deck.pdf) | [project scope](https://docs.google.com/presentation/d/1mK0g73wGf2sa7chJEEwdXjU4YzEnP2NzN3lbsPK59XM/edit#slide=id.g1456fc6c867_0_210) | [project scope](https://docs.google.com/presentation/d/1mK0g73wGf2sa7chJEEwdXjU4YzEnP2NzN3lbsPK59XM/edit#slide=id.g1456fc6c867_0_210) |
| Initial Plan | [✔](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Initial-Plan.md) |  |  |  |  |  |  |
| README File | [✔](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/README.md) |  |  |  |  |  |  |
| Progressive Report | [Description and scope sections](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Report-Files/Progressive%20Report.pdf) | [Achievements and major deliverables](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Report-Files/Progressive%20Report.pdf) | [Quality](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Report-Files/Progressive%20Report.pdf) | [Planned activities and major issues.](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Report-Files/Progressive%20Report.pdf) | [Risks and external dependencies](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Report-Files/Progressive%20Report.pdf) | [effort section](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Report-Files/Progressive%20Report.pdf) | [Planned Activities Section](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Report-Files/Progressive%20Report.pdf) |
| First Group Report | [3 Evaluation](#_heading=h.pd60i6w4xjr) | [5 Tools to build the system](https://docs.google.com/document/d/1q0sRp2ZpBMlMjl53FUaUgykhQ0brlNBl/edit) | [Executive Summary](#_heading=h.bq44qy9r3z7e)  [1.1](#_heading=h.smb0kwj703qx)  [1.2](#_heading=h.bsfkyvpsiabt) [1.3](#_heading=h.e2zt0957hued) [1.4](#_heading=h.rnod3gr3wsix) [1.5](#_heading=h.oar4gf350c1a)  [1.6](#_heading=h.qav5g5od0exv) | [6.3](#_heading=h.qcbhin75he9k)  [6.4](#_heading=h.u8u2evoqissw)  [6.5](#_heading=h.tzh9zapxybbx)  [7.1](#_heading=h.77obe56hea4b)  [7.2](#_heading=h.hh15t66n0bby)  [7.3](#_heading=h.zc101frcr9dz) | [2 Effort section Overview of system from user view](#_heading=h.xix9dn1c6l6n) | section 6 and section 7 (with Benjamin) | [4 System Structure Overview](https://docs.google.com/document/d/1q0sRp2ZpBMlMjl53FUaUgykhQ0brlNBl/edit#) |
| User Stories |  | [✔](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Report-Files/User%20Stories.md) |  |  | [Wrote 6/8 user stories and acceptance criteria. Added whether each user story defined a functional or non-functional feature. Ranked user stories by priority.](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Report-Files/User%20Stories.md) |  |  |
| Tasks for project Document | [Task breakdown in orange highlighting](https://docs.google.com/document/d/1AVxG6s9NkYbB-WqgTa1ZU8FMHCglE6tz_XlD5Oj9pKs/edit?usp=sharing) |  | [Task list](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/commits/751961088f7ac2a91198ed376ba5d00a8090097f) |  |  |  |  |
| Meeting Minutes | [Week 2](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Week-2/), [week 4 client meeting](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Week-4/client-meeting-week-4.md), [week 5 tutor meeting](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Week-5/tutor-meeting-5.md), [week 5 tutorial meeting](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Week-5/tutorial-meeting-5.md) | [Week3 group meeting, tutor meeting and client meeting](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Week-3/) | [Week 5 Client](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Week-5/Client%20Meeting) | [Week4](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Week-4/week4groupmeetingFriday.txt)  [group meeting](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Week-4/week4groupmeetingFriday.txt)  [Week4 tutorial meeting](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Week-4/week4tutorialmeeting.txt)  [Week5 group meeting](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Week-5/Fridaygroupmeeting.txt) |  |  |  |
| Second Presentation | [Quality of work slide](https://docs.google.com/presentation/d/1QKzF1JPhqQd9GyI5PBYnrRoD06RLYM57Rc0aRnZK8KM/edit?usp=sharing) | [Demonstration](https://docs.google.com/presentation/d/1QKzF1JPhqQd9GyI5PBYnrRoD06RLYM57Rc0aRnZK8KM/edit#slide=id.g14b5138ff0d_0_60) | [Slide deck design](https://docs.google.com/presentation/d/1QKzF1JPhqQd9GyI5PBYnrRoD06RLYM57Rc0aRnZK8KM/edit?usp=sharing)  [intro](https://docs.google.com/presentation/d/1QKzF1JPhqQd9GyI5PBYnrRoD06RLYM57Rc0aRnZK8KM/edit?usp=sharing) | [Group process](https://docs.google.com/presentation/d/1CaUGvRq_3NjSAG6SgAUOFc-M1hGC-uDNIAhPYtBrBGw/edit#slide=id.g14b5028479a_2_0) | [System Specification Slide](https://docs.google.com/presentation/d/1CaUGvRq_3NjSAG6SgAUOFc-M1hGC-uDNIAhPYtBrBGw/edit#slide=id.g14b5028479a_2_0) | [discipline](https://docs.google.com/presentation/d/1CaUGvRq_3NjSAG6SgAUOFc-M1hGC-uDNIAhPYtBrBGw/edit#slide=id.g14b5028479a_2_0) | [Plan for Remaining Requirements/User Stories Potential Project Risks](https://docs.google.com/presentation/d/1CaUGvRq_3NjSAG6SgAUOFc-M1hGC-uDNIAhPYtBrBGw/edit#slide=id.g14b5028479a_2_0) |

## 

## 9 Appendices

### A1. User stories

The final version of the user stories is available in the [BitBucket](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/Report-Files/User%20Stories.md)

1. As a medical imaging provider, I want a comprehensive report so that I can justify my use of qXR.
   * Acceptance Criteria: Report covers all topics that client has requested
   * Non-Functional
2. As a radiologist, I want to have an accurate AI product to detect TB so that I can give the correct diagnosis.
   * Acceptance Criteria: Most accurate AI product identified in comparison to others
   * Functional
3. As a radiologist, I want to have an efficient AI product to detect TB so that I can identify issues and assist my patients in recovering as soon as possible.
   * Acceptance Criteria: Efficiency of various AI products is compared.
   * Functional
4. As a radiologist, I want to ensure that qXR is the most appropriate AI tool for my needs so that I can treat my patients accordingly.
   * Acceptance Criteria: qXR successfully benchmarked against competitive AI tools that are certified
   * Functional
5. As a medical imaging provider, I want a cheap alternative to medical imaging so that I can save money.
   * Acceptance Criteria: Proof that AI is a technically viable and superior to human interpretation
   * Functional
6. As a medical imaging provider, I want a reliable report so that my decision is well informed.
   * Acceptance Criteria: Report is well researched and thorough
   * Non-functional
7. As a medical imaging provider, I want to have the report by my required date so that I can make decisions in a timely manner
   * Acceptance Criteria: Report is submitted by the due date
   * Non-functional
8. As a radiologist, I want a product that will not interrupt my workflow so that I can maintain my focus on treating patients.
   * Acceptance Criteria: Show that AI products will be an asset to radiologists
   * Functional

### A2. Systematic Review Quality Example

An example of applying PRISMA and the systematic review quality measures is the qXR user experience systematic review which is located in [BitBucket](https://bitbucket.org/andersen_liu/isys3888_tu15_02_p56/src/master/qxr-user-experience.md).

### A3. DTA Example Table

| Study Title | Product Name | Number of CXRs used in training | Number of CXRs used for testing | Sensitivity (95% CI) | Specificity (95% CI) | AUC (95% CI) | Reading time (seconds) |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

### 

### A4. Draft Final Report Structure

1. Abstract
2. Introduction
3. Relevance

3.1. What problem is the application intended to solve, and who is the application designed for?

3.2. What are the potential benefits, and for whom?

3.3. What are the risks associated with the use of the AI system?

1. Performance and Validation

4.1. Are the algorithm’s design specifications clear?

4.2. How was the algorithm trained?

4.3. How has performance been evaluated?

4.4. Have the developers identified and accounted for potential sources of bias in their algorithm?

4.5 Is the algorithm fixed or adapting as new data comes in?

1. Usability and Integration

5.1. How can the application be integrated into your clinical workflow?

5.2. How exactly does the application impact the workflow?

5.3. What are the requirements in terms of information technology (IT) infrastructure?

5.4. Interoperability - How can the data be exported for research and other purposes?

5.5. Will the data be accessible to non-radiologists (referring physicians, patients)?

5.6. Are the AI model’s results interpretable?

1. Regulatory and Legal Aspects

6.1. Does the AI application comply with the local medical device regulations?

6.2. Does the AI application comply with the data protection regulations?

1. Financial and Support Services Considerations

7.1. What is the licensing model?

7.2. How are user training and follow-up handled?

7.3. How is the maintenance of the product ensured?

7.4. How will potential malfunctions or erroneous results be handled?

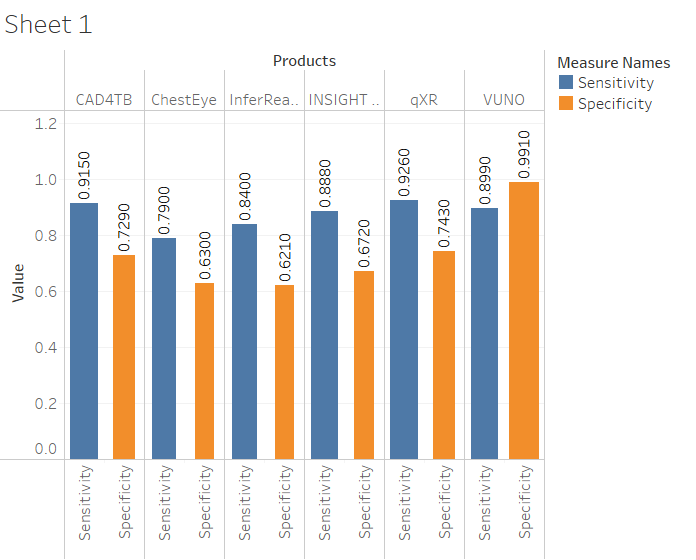
1. Conclusion

8.1. Recommendation

8.2. Limitations

1. Appendix
2. References

### A5. Example of Dashboard



## 10 References

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009, July 21). *The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: Explanation and elaboration*. The BMJ. Retrieved September 4, 2022, from <https://www.bmj.com/content/339/bmj.b2700>

*Libguides: Creating a prisma flow diagram: Prisma 2020*. PRISMA 2020 - Creating a PRISMA flow diagram - LibGuides at University of North Carolina at Chapel Hill. (n.d.). Retrieved September 4, 2022, from https://guides.lib.unc.edu/prisma

Omoumi, P., Ducarouge, A., Tournier, A., Harvey, H., Kahn, C., & Louvet-de Verchère, F. et al. (2021). *To buy or not to buy—evaluating commercial AI solutions in radiology (the ECLAIR guidelines).* European Radiology, 31(6), 3786-3796. doi: <https://doi.org/10.1007/s00330-020-07684-x>

Page, M., McKenzie, J., Bossuyt, P., Boutron, I., Hoffmann, T., & Mulrow, C. et al. (2021). *The PRISMA 2020 statement: an updated guideline for reporting systematic reviews*. BMJ, n71. doi: <https://doi.org/10.1136/bmj.n71>

Swift, A., Heale, R., & Twycross, A. (2019). What are sensitivity and specificity? Evidence Based Nursing, 23(1), 2–4. https://doi.org/10.1136/ebnurs-2019-103225

‌