RESEARCH PROTOCOL

IMPROVING THE ACCURACY OF ARTIFICIAL INTELLIGENCE TRIAGE IN PRIMARY CARE

(AI TRIAGE ACCURACY)

Version history

Version	Date	Summary of changes
1.0	2024-03-01	Initial draft
1.1	2025-03-28	Minor clarifications of: participant age range and ethnicity coding, university contact details, section numbering, data sharing and storage arrangements, total sample size
1.2	2025-07-03	Removal of sharing online consultation text and messages with the research team

Contents

1) RESEARCH TEAM & KEY CONTACTS

Chief Investigator: Name: Dr Benjamin Brown, Clinical Senior Lecturer Address: Centre for Primary Care and Health Services Research Williamson Building University of Manchester M13 9PL Email: benjamin.brown@manchester.ac.uk	***Co-investigators:*** Name: Prof Evan Kontopantelis, Professor in Data Science and Health Services Research Address: Centre for Primary Care and Health Services Research Williamson Building University of Manchester M13 9PL Email: evan.kontopantelis@manchester.ac.uk Name: Prof Niels Peek, Professor of Data Science and Healthcare Improvement Address: THIS Institute University of Cambridge Strangeways Research Laboratory 2 Worts' Causeway Cambridge CB1 8RN Email: niels.peek@thisinstitute.cam.ac.uk
Sponsor: Name: The University of Manchester Contact: Ms Lynne Macrae, Faculty Research Practice Governance Coordinator Address: Faculty of Biology, Medicine and Health Room 4.64 Simon Building University of Manchester M13 9PS Email: fbmhethics@manchester.ac.uk Telephone: 0161 275 5436	***Lead R&D Trust contact:*** Name: NIHR Clinical Research Network Greater Manchester Contact: Mr Dayle Roberts, Study Support Service Senior Manager Address: Citylabs 1.0 Nelson Street Manchester M13 9NQ Email: researchsupport.crngm@nihr.ac.uk

2) PLAIN ENGLISH SUMMARY

Background

When patients contact their GP practice, the first step is to work out what kind of help they need, who can provide it, and how quickly it's needed. This is called 'triage' and is important for safety and making the best use of resources. Since 2020, patients have been able to contact GP practices online, but this can lead to a lot of requests at once, making triage harder and causing delays in care.

Artificial Intelligence (AI) can help with triage by processing requests faster and working around the clock, which may reduce these delays. While AI is already being used in the NHS, we don't know how accurate it is or if it treats all patients fairly. Our study will focus on a system called Patchs, which helps patients contact their GP online and uses AI for triage.

What will we do?

We will conduct our research in four steps:

1. Look at data from GP practices using Patchs without AI to see how they currently triage patients and what problems they face.
2. Use data from GP practices using Patchs (both with AI on and off) to make the AI more accurate.
3. Check data from GP practices using Patchs with AI off to measure how well the updated AI system works.
4. Give the improved AI system to GP practices already using AI.

At each step, we will test whether patients from different backgrounds are treated fairly.

What difference will we make?

Our research will show the problems with online triage without AI and explain how an improved AI system could help patients get the care they need more quickly.

3) SCIENTIFIC SUMMARY

Background

GP practice staff triage patients contacting them to make the best use of resources and maintain patient safety. Online consultation systems are used by most GP practices and allow patients to contact their GP practice using an online form. They can be submitted without talking to a member of staff, thereby circumventing the usual triage process. Online consultation systems can triage patients using 'Artificial Intelligence' (AI), though there is a lack of research on their performance. We (The University of Manchester; UoM) propose to fill this gap by collaborating with an online consultation system provider with optional AI triage functionality (Patchs).

Research questions

Overall research question: is it possible to develop AI models that can replicate clinicians' triage decisions?

1. What challenges do patients and GP practices face when triaging patients in primary care, and what are their drivers?
2. What is the best performing AI model for triaging patients in primary care?
3. Is AI triage performance maintained across different geographical regions?
4. Is AI triage performance maintained over time?
5. How does AI triage performance compare to current clinical practice?

6. Does AI triage performance change when deployed into clinical practice?
7. Does AI triage work fairly for all patients?

Methods

Workstream 1: Triage problem quantification. We will analyse anonymised historic data from GP practices using Patchs with AI triage disabled. Where publicly available, we will compare this to practice-level data from GP practices not using Patchs (control practices). We will undertake descriptive and inferential analyses to understand potential triage problems and factors that influence them, such as delays in providing patient care.

Workstream 2: AI development. We will use anonymised historic data from GP practices using Patchs to build new versions of the AI triage models currently in use with four different approaches: logistic regression, XGBoost, long short-term memory (LSTM), and large language model (LLM). We will use internal-external cross-validation by geographical region and compare their performance using random-effects meta-analysis and sub-group analyses to assess fairness (e.g. across ethnicities). We will compare their performance to the current AI triage models in use. The final version of the best-performing AI models will be developed using the entire dataset.

Workstream 3: Prospective background evaluation. We will obtain predictions from the best-performing AI models on prospectively collected data from GP practices using Patchs without AI triage by running the models in the 'background'. We will undertake sub-group analyses to assess fairness as described above.

Workstream 4: Prospective implementation evaluation. In accordance with the normal Patchs software updates, we will update the AI models in GP practices already using AI triage with the best-performing versions. We will prospectively measure how often GP practice staff and patients agree with the new versions' triage predictions to test whether its performance translates to real patient care. We will undertake sub-group analyses to assess fairness as described above.

Anticipated benefits

We will help understand the problems currently faced by GP practices during online consultation triage. If we developed improved AI models, there may be improved patient safety (e.g. by helping patients receive help sooner) and reduced GP practice workload (e.g. by automating the triage process). GP practices and their patients in Workstream 4 would benefit immediately. We will provide evidence for GP practices not currently using AI triage whether to adopt it.

4) SUMMARY OF MAIN ETHICAL, LEGAL, OR MANAGEMENT ISSUES

We believe this study:

- Is research (not service evaluation) because it is designed to produce generalisable findings by sampling data and conducting analyses to make the results transferable to populations outside the study (1).
- Requires NHS Research Ethics Review because we will prospectively collect data from users of NHS GP services in Workstreams 3 and 4 (2).
- Is a clinical study of a UKCA marked device for a labelled indication, involving no change to standard care or randomisation between groups in Workstream 4 (3).

Issue 1: Anonymised data sharing

Description: When patients, their carers, or GP practice staff use Patchs they are informed their anonymised data may be shared with UoM for research purposes.

Mitigations: Patients or their carers can opt out of sharing their data with UoM at any point using a toggle button in the Patchs system without affecting their ability to continue to use Patchs to access GP services. They are informed how to do this when they use Patchs. A risk assessment based on guidance from the Information Commissioner's Office suggests the data are effectively anonymised.

Issue 2: Declaration of interests

Description: Dr Brown is a part-time employee of Patchs Health as Chief Medical Officer and is a shareholder in the company. Patchs Health develop the Patchs software.

Mitigations: 1) Co-investigators have no conflict of interest and will hold the Chief Investigator to account to ensure that the research is conducted rigorously. 2) We will pre-register our study protocol in advance of undertaking the research and making all details openly available online to prevent outcome switching and promote independent evaluation. Any out-of-protocol analyses will be reported as such in publications. 3) We will declare all interests in study protocols, reports, and publications. 4) We will make analysis code available when the findings are published. 5) We will disseminate our findings regardless of study outcome including submitting papers reporting negative results for publication in peer-reviewed journals.

5) BACKGROUND

Triage in primary care

GP practices in England deliver over 30 million patient appointments per month (4). A proportion are for urgent medical conditions that require treatment within 24-48 hours such as infections requiring antibiotics (18%) (5). A smaller proportion are for medical emergencies which require more immediate treatment, including from emergency services, like heart attacks. Delays in urgent or emergency treatment in primary care can lead to patient harm, including hospital admission and death: in a study of over 300,000 urinary tract infections in elderly patients, 13.4% experienced a delay in treatment, which was associated with a higher chance of hospital admission and sepsis (odds ratio 7.12) (6). At the other end of the spectrum, it is estimated that 13% of patients presenting to primary care consist of minor illness could have been treated by self-care or at community pharmacies (7). With finite capacity in the system, each time these patients are treated they are potentially at the expense of patients that can only be treated by a primary care clinician, which could lead to delays in treatment. With continued growth in demand for NHS services and capacity remaining the same (and potentially diminishing due to staff leaving the profession), a key challenge in primary care is to identify the best person, place, and speed to treat each patient that contacts them to make the best use of resources and ensure patient safety. This is usually called 'triage' (8) and can be achieved by asking questions similar to those in Table 1 for each patient.

Table 1: Triage questions

Short version	Long version
What?	What is the patient contacting about (e.g. sore throat, headache, medication query, admin question)?
Who?	Who is best placed to help the patient (e.g. clinician or non-clinician)?
How?	How should they help the patient (e.g. in-person appointment, telephone, written message)?
When?	When should they help the patient (e.g. on the same day or can it wait)?
Where?	Where is the most appropriate setting to help the patient (e.g. pharmacy, GP practice, emergency department)?

Online consultations

Online consultations allow patients to contact their GP practice about health problems using an online form (9). All English GP practices have been mandated to provide online consultations since April 2020 (10). Their adoption has been further catalysed by the COVID-19 pandemic, and they have been available in 85% of GP practices since May 2020 (11). An NHS England Freedom of Information request in 2022 suggests they are available in 96% of all GP practices. In GP practices using online consultations, it is estimated they accounted for 72% of all patient requests for appointments in 2021 (12).

Although online consultations offer many benefits including patient convenience and improved access (9), they have the potential to exacerbate delays in providing the most appropriate care for patients. Unlike traditional methods of contacting the GP practice (e.g. telephone), online consultations can be submitted at any time by patients without waiting in a queue or talking to a member of GP practice staff, therefore bypassing the triage process. GP practices can therefore receive many online consultations in short periods of time, including when they are closed, some of which may be more appropriately dealt with by other health care providers. GP practice staff must read each online consultation one-by-one to perform triage, which can lead to delays in patients getting the care they need and GP practices unnecessarily consulting patients. For example, urgent and emergency online consultations can be 'hidden' from view and may not be processed in an appropriate timeframe. Patients who should have attended the emergency department, pharmacy, or other health care provider could have attended earlier without contacting the GP practice.

Artificial intelligence (AI) triage: an evidence gap

One potential solution to reduce these delays is for the online consultation system to automatically triage patients as soon as they submit a query to their GP practice (9). This could reduce delays because it can happen instantaneously and work 24 hours a day, seven days a week. This is considered 'Artificial Intelligence' (AI) because it automates activities typically associated with human thinking such as decision-making and problem-solving (13).

We recently conducted the largest and most up-to-date systematic review of empirical research up to February 2022 on real-world use of online consultations in primary care (9). Out of 63 papers studying 31 different systems from nine countries, only four papers evaluated systems with AI triage, and only one of those papers evaluated the AI triage element (14). This AI triage system triaged patients with COVID-19 and focused on answering the 'When' and 'Where' triage questions: it classified patients as requiring either urgent, emergency, non-urgent, or self-care treatment (14). Furthermore, there are concerns that AI can create or exacerbate inequalities in health care due to algorithmic bias ('unfairness') (15), but this has not yet been studied in the context of triage. Consequently, there is an evidence gap on the wider use of AI triage for all clinical conditions encountered in primary care, whether it can be used to answer the other three triage questions described above ('What', 'Who', and 'How'), and whether it works the same for patients from different backgrounds.

Despite this lack of evidence, AI Triage is already used by 5/33 (15%) of NHS online consultation systems (16) so there is an urgent need to fill this gap. These systems are proprietary and typically do not routinely share their data with university research teams for analysis. We (The University of Manchester; UoM) propose to fill these evidence gaps by conducting a research study in collaboration with an online consultation system provider that offers AI triage (Patchs). GP practices using Patchs have the option to enable different AIs that address each of the five triage questions described above (What', 'Who', 'How', 'When', and 'Where').

AI triage under evaluation in this study: Patchs AI

Commercial collaborator Patchs Health developed Patchs (www.Patchs.ai) which has been available to GP practices since 2020. It is currently used by approximately 1000 GP practices in England and Wales

(approximately 20% of all GP practices). Patients access Patchs the system via their GP practice website and fill out an online form by describing their queries in natural language (unstructured free text) in response to standarised open-ended questions that mimic a typical primary care consultation. When the form is submitted by the patient it enters an inbox, where GP practice staff can respond. Receptionists typically review incoming Patchs online consultations first, deal with any queries they can, and assign those they cannot to other staff within Patchs. Patients are then contacted to resolve their query – either by written message or video consultation within Patchs, telephone or by arranging an in-person appointment. When GP practice staff process patients' online consultations in Patchs, they record various triage decisions prompted by questions covering the five triage questions from Table 1 (Table 2, and Appendix 1, Screenshot 1). The Patchs triage questions and options have been developed based on qualitative research and workshops conducted with 22 GP practice staff and 37 patients. Triage decisions must be added for an online consultation to be completed by the GP practice.

Based on qualitative research with GP practice staff and patients (in preparation for publication), the triage decisions made by staff have been used to develop Patchs AI models (Tables 2 and 3, and Appendix 1, Screenshot 1). The AI models analyse the text written by patients, along with details about the patient themselves (e.g. sex and age) and type of request to automatically triage patients. When an AI makes a triage prediction it pre-populates the triage decisions in Patchs (Appendix 1, Screenshot 2). Models have been optimised for the evaluation considerations in Table 3. If GP practice staff disagree with the triage predictions from Patchs AI they can change the triage decision. This provides a feedback loop to monitor and re-train the AIs. Approximately 20% of GP practices using Patchs have AI enabled.

Table 2: How Patchs triage decisions relate to triage questions and AIs

Patchs triage questions	Patchs triage options for GP practice staff	Relevant triage questions (Table 1)	AI monitoring and re-training
How clinically urgent is this message?	Emergency - Patient could be harmed if this request is not resolved on the same day. Urgent - Patient could be harmed if this request is not resolved within the next 48 hours. Routine - Patient unlikely to be harmed if this request is not resolved within the next 48 hours. Administrative or medication request	When? Where?	Urgency AI
What are the main topics of this message?	Over 160 different clinical topics covering common primary care symptom and conditions*	What?	Topic AI
Ideally, who do you think ***should have*** dealt with this message?	Thirteen options including different types of primary care staff, other services (e.g. emergency services, dentists), and 'me'	Who? Where?	Assign AI
What actions are required from you to resolve this message?	Twelve options e.g. Prescribe new medication, Telephone patient, See patient face-to-face*	How?	F2F AI

*Multiple can be selected

Table 3: Patchs AI models

AI	Function	Model evaluation considerations*
Urgency AI (17)	Predicts whether online consultations are 'urgent' or 'emergency'. Negative predictions are 'routine'. Emergency and urgent online consultations are highlighted with red and orange flags respectively and ordered to the top of the Patchs inbox (Appendix 1, Screenshot 3).	High true positive rate is most important to minimise missing urgent or emergency online consultations.
Topic AI (18)	Predicts the clinical topics related to the online consultation from over 170 different options. Each topic can be linked to further actions e.g. asking patients further questions at the point of submission to gather specific information about their request (Appendix 1, Screenshot 4).	High true positive rate for each topic is most important to ensure patients are asked all pertinent clinical questions.
Assign AI (19)	Predicts whether online consultations may require input from clinicians or not. Those that may require input from clinicians are automatically assigned to a 'Clinical' inbox (Appendix 1, Screenshot 5).	High true positive rate is most important so all online consultations that require input from a clinician receive it quickly.
Face-to-face (F2F) AI (20)	Predicts whether online consultations may require a F2F appointment (e.g. for a physical examination). Those that may require a F2F appointment are highlighted with an icon in the inbox (Appendix 1, Screenshot 6).	High true positive rate is most important so all online consultations that require an in-person appointment receive it quickly and because false positives can be dealt with in-person but false negatives cannot be dealt with remotely.
Signpost AI (21)	Uses predictions from Urgency AI to send messages to patients suggesting they could self-care, or access urgent or emergency services. Uses predictions form Topic AI present relevant supporting information from NHS.uk to support the message (Appendix 1, Screenshots 7 and 8).	See Urgency and Topic AIs

*Evidenced by our previous qualitative and PPI work

Patchs AI: regulatory compliance

Both Patchs (without AI) and Patchs AI are currently used in routine NHS clinical practice. They comply with NHS DCB0129 standards (22), and have been assured by NHS Digital to be available on the NHS Buying Catalogue since its inception in 2021 (23). Patchs AI is a Class I medical device because it offers 'triage and signposting of next steps based on filters by severity and probability of a match' without 'direct diagnosis' (24). Patchs AI received its UKCA mark in October 2021 (MHRA registration number 8387). Patchs AI is an optional feature in Patchs, which GP practices can enable themselves at any time. GP practices are recommended to familiarise themselves using Patchs without AI first and undergo specific training before they use it (25). During this onboarding, GP practices are made aware that the AI models may be periodically updated when newer versions are developed (25).

6) RESEARCH QUESTIONS

Overall research question: is it possible to develop AI models that can replicate clinicians' triage decisions?

1. What challenges do patients and GP practices face for triaging patients in primary care, and what are their drivers?

2. What is the best performing AI model when triaging patients in primary care?
3. Is AI triage performance maintained across different geographical regions?
4. Is AI triage performance maintained over time?
5. How does AI triage performance compare to current clinical practice?
6. Does AI triage performance change when deployed into clinical practice?
7. Does AI triage work fairly for all patients?

7) STUDY DESIGN & PROTOCOL

Figure 1 shows our overall approach. We structure this protocol following RECORD guidelines (26) for Workstream 1, TRIPOD+AI guidelines (27) for Workstreams 2 and 3, and DECIDE-AI guidelines for Workstream 4 (28). Given this is a protocol, we focus on the methods section of these reporting guidelines. We have attempted to minimise bias in our study design by assessing our approach using the PROBAST tool for Workstreams 2-4 (29).

7.1 Participants

We will recruit GP practices using the Patchs system both with AI enabled and disabled. Data relating to consultations from all patients and all types of consultations will be eligible for inclusion. There is no upper or lower patient age limit.

7.2 Study Intervention and/or Procedures

Participants will not receive any additional contact, intervention, or procedure outside usual care and Patchs system updates during this study.

Workstreams 1 and 2

Anonymised routinely collected historic data from GP practice staff and patients at GP practices who have used Patchs will be collected retrospectively. Where publicly available, we will also include practice-level data from GP practices not using Patchs (control practices).

Workstream 3

Anonymised routinely collected data from GP practice staff and patients at GP practices currently using Patchs with AI disabled will be collected prospectively after the best-performing model from Workstream 2 has been developed. Predictions from the model for each online consultation will be collected by running the model in the 'background' i.e. GP practice staff and patients will not be presented with the predicted triage outcome and any associated action from the AIs (Table 3). We will aim for a data collection period of maximum six months for practical reasons.

Workstream 4

We will update the AI model in GP practices with AI enabled to the best-performing one from Workstream 2. This will be done in stages to identify and address any potential issues during deployment. Anonymised routinely collected data from GP practice staff and patients at GP practices at these practices will be collected prospectively. The difference from data collected in Workstream 3 is that practice staff and patients will have been presented with the predicted triage outcome and any associated action from the AIs (Table 3). These GP practices have previously received specific training on Patchs AI, including help articles (30), eLearning modules, and have accepted a Clinical Safety Case Report as per NHS clinical risk management standards for software (DCB0129 (22) and DCB0160 (31)). GP practices are aware from the documentation that the AI models are regularly updated as part of routine improvements to the system.

This workstream will run parallel to Workstream 3 and will therefore also aim for a data collection period of maximum six months.

Participants will not be contacted by the research team during this project. GP practices will be contacted by the Patchs Health team in Workstream 4 to inform them when their AI models have been updated as part of their regular routine software update communications via email.

7.3 End of study

The study will end when the final AI has been prospectively tested in Workstream 4. If the new versions of the AIs are more accurate in the real world than the ones currently used, they will continue to be used by Patchs in routine care.

8) STUDY PARTICIPANTS AND DATA

Eligible GP practices, patients, and online consultation data will be identified by the Patchs Health team using the Patchs database.

When GP practices, patients or their carers (as verified by the GP practice (32)) use Patchs, they are informed their anonymised data may be shared with UoM for research purposes (Appendix 1, Screenshot 9). Patients and their carers can opt out of sharing their anonymised data at any point using a toggle button in the Patchs system without affecting their ability to continue to use Patchs to access GP services (Appendix 1, Screenshot 10). GP practices, staff, patients and their carers, are under no time pressure to decide whether they want to share their anonymised data with UoM for research purposes.

If information becomes available during the research that may be relevant to participants' continued participation, this will be communicated to GP practice staff and/or patients via email using the address they use to access Patchs. As this is identifiable information relating to Patchs customers, any email addresses will be accessed and information sent by Patchs Health not the research team.

8.1 Inclusion Criteria

GP practices

Workstream 1 – GP practices who have actively used the Patchs system with AI disabled for at least six months to capture those that have embedded the system in their organisational workflows. Where publicly available, we will also include practice-level data from GP practices not using Patchs (control practices).

Workstream 2 – GP practices who have used Patchs with AI enabled or disabled.

Workstream 3 – GP practices who have never had AI enabled.

Workstream 4 – GP practices with AI enabled.

Patients

All workstreams – To minimise bias and generate real-world evidence, we will include data from all patients registered at GP practices who use Patchs. Patients can only use Patchs if they are at least 16 years old, although carers can use Patchs on behalf of patients under 16 if they have guardianship or parental responsibility as manually verified by GP practice staff (32).

Data

All workstreams – Completed online consultations (triage decisions must be added for an online consultation to be completed by the GP practice), submitted by patients or on their behalf by carers and

GP practice staff. There are no minimum requirements for data quality or missingness, although there are minimum character input requirements for patients to submit an online consultation in Patchs.

Workstream 1 – Historic online consultations. If GP practices have had AI triage enabled, only online consultations from before it was first enabled will be included.

Workstream 2 – Historic online consultations written in English (no language translation usage recorded in Patchs (33)).

Workstream 3 – Prospective online consultations. If GP practices subsequently enable AI triage, only online consultations from before they first had Patch AI enabled will be included.

Workstream 4 – Prospective online consultations. If GP practices subsequently disable AI triage, only online consultations from before they first disabled AI will be included.

8.2 Exclusion Criteria

GP practices

All workstreams – GP practices who have not used the Patchs system for clinical care i.e. test or demo practices (all workstreams).

Workstreams 2-4 – GP practices eligible for evaluations of AI triage impact e.g. IRAS 331286.

Patients

All workstreams:

- Patients recorded as over 120 years old
- Patients with names indicating they are test patients (e.g. ‘test’, ‘dummy’, ‘donotuse‘, 'xxx’ or 'patchs’ in their first or last names, patients called ‘Mickey Mouse’ or ‘Minnie Mouse’)

Data

All workstreams – Online consultations deleted by GP practices and GP practice-initiated outbound messages.

Workstreams 2-4 – Online consultations regarding patients, or processed by GP practice staff, that are eligible for other evaluation datasets e.g. IRAS 264891.

8.3 Sampling

Sampling will occur sequentially: GP practices, then patients, and data last.

GP practices

Workstream 1 – Nil.

Workstream 2 – To mitigate including triage decisions where GP practice staff have accepted incorrect triage suggestions made by the AI (‘automation bias’) (34), we will sample GP practices after they have started using AI triage who are not outliers in overriding AI triage predictions. The AI triage models have been developed to achieve high true positive rates (Table 3) therefore we will visually inspect histograms to identify outliers with low false discovery rates as this represents a failure to identify false positives. We will not identify outliers with low false omission rates because false negatives are expected to be rare

(Table 3). We may also sample GP practices not using AI who are not outliers in their manual triage decisions in a similar way depending on the distribution of triage behaviour observed.

Workstreams 3 and 4 – Primary analyses will be undertaken on a random sample to match the GP practices used to develop the AI models in Workstream 2 (35) on as many characteristics as possible in Table 4 as closely as possible using the Stata package *repsample* (36) or similar approach. In Workstream 4, GP practices who are not outliers in terms of their false discovery rates will be sampled using the same approach described in Workstream 2 (35). Secondary analyses will be undertaken without sampling.

Table 4: Characteristics on which to match GP practices in Workstreams 3 and 4

Characteristic	**Matching criterion**
Patient population	
Index of Multiple Deprivation (37)	Mean or median
Rural Urban Classification (38)	Mean or median
Patient morbidity	Mean or median proportion of patients with more than one chronic condition (39)
Geography	Number of regions represented from internal-external cross-validation (IECV)
Staff triage behaviour	
Population size (39)	Mean or median
Triage decisions	Mean or median proportion of online consultations with the outcome triage decision
Volume of online consultations	Mean or median count of monthly online consultations per 1000 patients
Number of GPs	Mean or median number of whole-time equivalent GPs per 1000 patients (40)
Care quality	Mean or median Quality and Outcomes Framework performance (39)
Duration using online consultations	Mean or median number of months actively using Patchs
Language used in online consultations	
Online consultations submitted by staff on behalf of patients	Mean or median proportion of online consultations
Clinical online consultations submitted	Mean or median proportion of clinical online consultations

Patients

Workstream 1 – Nil.

Workstream 2 – Patients who have contributed at least one online consultation will be randomly sampled in an attempt to obtain a population representative of England according to the UK 2021 Census (41) on as many characteristics as possible in Table 5 as closely as possible using the Stata package *repsample* (36) or similar approach. Our aim is to mitigate against algorithmic bias ('unfairness') that could lead to unequal impacts between different groups (15,42).

Workstreams 3 and 4 – Primary analyses will be undertaken on a randomly sampled patient population representative of England using the same approach described above for Workstream 2 (35). Secondary analyses will be undertaken without sampling.

Table 5: Characteristics of patients to obtain a population representative of England according to the UK 2021 Census

Characteristic	Categories
Age	4 years and under 5 to 9 years 10 to 15 years 16 to 24 years 25 to 34 years 35 to 49 years 50 to 64 years 65 to 74 years 75 years and over
Sex	Male Female Not specified
Ethnicity	White Non-white Not specified
Deprivation (Index of Multiple Deprivation decile) (37)	Decile
Rural Urban Classification (38)	Decile

Data

Workstream 1 – Nil.

Workstream 2 – Where patients have submitted multiple online consultations, we will randomly select one for inclusion to minimise clustering effects at the patient and online consultation levels. To obtain clinically reliable triage decisions, we will only include online consultations:

- With at least one relevant triage decision applied or reviewed by a clinical member of staff (i.e. assigned to them or completed by them)
- Processed by only non-clinical staff members if triaged as 'Administrative or medication request' (Table 2) i.e. a non-clinical topic that would not normally be processed by clinicians (Urgency AI only)
- Processed by only non-clinical staff members if triaged as 'Emergency' (Table 2) indicating the patient may have been signposted to emergency services immediately by the non-clinician (Urgency AI only)

Workstreams 3 and 4 – Primary analyses will be undertaken on a sample of online consultations using the same approach described above for Workstream 2 (35). Secondary analyses will be undertaken without sampling.

8.4 Missing data

Patients must enter their age, sex, ethnicity, online consultation request type and request text to use Patchs. Ethnicity and sex may be 'not specified' by the patient and will be treated as a standalone category. Data could be missing following anonymisation of online consultation request text (43), though this cannot be reliably imputed. Where patient socioeconomic deprivation are missing, we will use the GP practice data.

8.5 Participants who withdraw consent or lose capacity to consent

This study will use anonymised data only. If participants withdraw consent to share their anonymised data or lose capacity after we have received their data, we will be unable to identify them to exclude them.

9) OUTCOME MEASURES

Workstream 1

Patient-level – counts and proportions

Triage outcomes in online consultations based on staff triage decisions (Table 2):

- Urgency of online consultations (administrative or medication, routine, urgent, emergency, urgent or emergency; denominator = number of online consultations
- Clinical topics (e.g. sore throat, headache, medication query, admin question; denominator = number of online consultations)
- Staff member roles involved in processing online consultations (e.g. clinician or non-clinician; denominator = number of online consultations)
- Mode of how online consultation was resolved (e.g. in-person appointment, telephone, written message; denominator = number of online consultations)
- Most appropriate care setting (e.g. pharmacy, GP practice, emergency department; denominator = number of online consultations)
- Delayed action and completion of emergency online consultations (denominator = number of emergency online consultations)
- Delayed action and completion of urgent online consultations (denominator = number of urgent online consultations)
- Delayed action and completion of urgent and emergency online consultations combined (denominator = number of urgent and emergency online consultations combined)

Patient-level – continuous

- Time until online consultation was actioned from submission
- Time until online consultation was completed from submission

GP practice-level – counts

Counts per 1000 patients:

- Online consultations
- Total appointments
- Emergency department attendances
- Emergency hospital admissions

GP practice-level – proportions

Patient satisfaction as measured in the GP Patient Survey (44)

GP practice-level – continuous

Appointment waiting times

Workstreams 2-4

AI model outcome definitions and evaluation metrics for Workstreams 2-4 are provided in Tables 6 and 7, respectively. Our aim is to minimise false negatives (Table 3) whilst balancing false positives to avoid

'alert fatigue' where predictions are ignored due to frequent 'false alarms' (47). Therefore, our primary outcome measure for model evaluation is the $F1$ score (micro-averaged across all classes in Topic AI).

Table 6: AI model outcome definitions

AI	Outcome	Definition (see Table 2)
Urgency AI	Binary classification: Whether an online consultation is urgent or emergency.	Whether a member of staff has chosen either the 'urgent' or 'emergency' option. If triage decision made by a clinical member of staff, use the most urgent clinician triage decision. If no triage decision made by a clinical member of staff, use most urgent non-clinician triage decision.
Topic AI	Multi-class classification: Clinical topic(s) relevant to the online consultation (from over 160 different options).	Clinical topic(s) chosen by clinical staff members.
Assign AI	Binary classification: Whether an online consultation requires input from a clinician.	Whether an online consultation has been assigned to a clinical staff member.
Face-to-face AI	Binary classification: Whether an online consultation requires a face-to-face appointment.	Whether the 'see patient face-to-face' option has been chosen by a clinical member of staff.
Signpost AI	See Urgency and Topic AI.	See Urgency and Topic AI.

Table 7: Model evaluation metrics

Purpose	Metric
Discrimination	Accuracy (overall proportion of correct predictions)
	True positive rate*
	True negative rate*
	Positive predictive value*
	Negative predictive value*
	$F1$ score**
	C-Statistic*
	AUC of Precision-Recall Curve*
Calibration	Calibration curve*
	Calibration slope*
	Calibration-in-the-large*
Clinical utility	Decision curve analysis* (45,46)

*Reported for each predicted class in Topic AI and micro- and macro-averaged across all classes

**Primary outcome measure for model evaluation

10) DATA COLLECTION, SOURCE DATA AND CONFIDENTIALITY

Patient-level data

We will use anonymised routinely collected Patchs data held by Patchs Health. Personal identifiable data will not be used. Data are collected automatically each time both patients and staff interact with Patchs. This includes data collected 'passively' when users visit areas within the system, in addition to data intentionally inputted such as triage decisions made by staff or when patients enter details. Data will be anonymised by Patchs Health before being shared with UoM. Patchs Health will assign a randomly generated identification number to each patient. Patchs Health will then delete the mapping key using

hard drive eraser software before sharing the data with the research team. The mapping key will not be shared with the UoM research team. The following anonymous patient-level data will be used:

- Patient randomly generated identification number
- Patient year of birth
- Patient sex
- Patient ethnicity according to methods mandated for NHS organisations and response codes set out in the NHS data dictionary (48) i.e. White, Mixed, Asian, Black, Other
- GP practice Organisation Data Service code (49)
- Index of Multiple Deprivation (37)
- Whether online consultation was submitted by patient or someone else (carer or staff member)
- Type of online consultation chosen e.g. health problem or administrative request
- Language the online consultation was written (33)
- GP practice staff triage decisions (Table 2)
- Staff member roles who processed the online consultation (e.g. clinician or non-clinician)
- Date-time online consultation was submitted by the patient
- Date-time of actions to process online consultation by GP practices staff
- Date-time online consultation was completed by GP practices staff, or cancelled by patient or carer
- Triage predictions made by AI triage

A risk assessment by Patchs Health based on guidance from the Information Commissioner's Office (50) concludes there is a low risk of individuals being identified from the data by a motivated intruder through 'singling out', 'data linkages', or 'inference', and the data are therefore effectively anonymised (Appendix 2).

GP practice-level data

In addition to Patchs data, we will also use the following GP practice-level data that we will link to Patchs data via the GP practice Organisation Data Service (49) code:

- Rural Urban Classification of Lower Layer Super Output Areas in England and Wales (freely available online) (38)
- Emergency Care Dataset from NHS Digital Data Access Request Service (application required) (51)
- National General Practice Profiles from Public Health England (freely available online) (52)
- GP Practice Workforce from NHS Digital (freely available online) (40)
- GP appointment data from NHS Digital (freely available online) (4)
- GP online consultation data (53)
- Index of Multiple Deprivation from Office for National Statistics (freely available online) (37)
- Quality and Outcomes Framework performance from NHS Digital (freely available online) (39)
- Counts of patients registered at a GP Practice (freely available online) (54)
- GP patient survey (freely available online) (44)

Analyses

Patchs Health will develop the AI models in Workstream 2, otherwise all other analyses will be conducted by the UoM research team. Study data and material may be looked at by individuals from the University of Manchester, from regulatory authorities or from the NHS Trust, for monitoring and auditing purposes, which may include access to personal information.

Data storage

Patchs Health – Data will be stored on secure Amazon Web Services servers in London (EU West 2). These servers meet industry-standard data security and privacy specifications including the NHS Data Security and Protection Toolkit (55) and Cyber Essentials Plus (56).

UoM – Data will be stored on secure Research Data Storage servers (57). These servers are only accessible by specified UoM users on the UoM network via secure connection Virtual Private Network (VPN) and require multi-factor authentication. Only members of the UoM research team will have user permissions to access the data.

Data transfer

All data will be transferred as password-protected comma separated value files using industry-standard Hypertext Transfer Protocol Secure over a Transport Layer Security connection. All data will be encrypted at rest and in transit. Data sharing agreements will be signed between UoM and Patchs Health, and UoM and NHS Digital (where relevant).

Data for workstreams 1 (Triage problem quantification) and 2 (AI triage development) will be transferred to UoM at the beginning of the study period. Data for workstreams 3 (Prospective background evaluation) and 4 (Prospective implementation evaluation) will not be transferred from to UoM until after the follow-up period for practices using the new AI models has finished. Only data from patients who have not opted out of sharing their data with UoM for research purposes will be transferred to UoM.

Data will not be shared with The University of Cambridge.

11) STATISTICAL CONSIDERATIONS

11.1 Statistical Analysis

Workstream 1: Triage problem quantification – retrospective data from GP practices with AI *disabled*

Descriptive analysis

GP practices will be descriptively analysed by patient population size (39), Index of Multiple Deprivation (37), monthly volume of online consultations per 1000 patients, geographic region, Rural Urban Classification (38), number of whole-time equivalent GPs per 1000 patients (40), levels of patient morbidity (39), Quality and Outcomes Framework performance (39), and length of time actively using online consultations. Patients will be descriptively analysed by age, sex, ethnicity (48), Index of Multiple Deprivation quintile (37), and non-English language usage in online consultations (33). Characteristics of GP practices and patients will be compared to those in the wider UK. Patients who have registered to use Patchs but not submitted an online consultation will be compared with those who have used the system at least once. Online consultations will be descriptively analysed in terms of type (e.g. health problem or administrative request), date-time submitted, role of staff who have processed them, and each outcome measure described above. We will plot counts of outcome measures as a monthly time series for both individual GP practices and all GP practices combined.

Modelling

To analyse count outcome measures, we will use negative binomial regression models (mixed-effects for patient-level count measures) with offset terms for the denominator. To analyse continuous and proportion outcome measures, we will use mixed-effects linear regression models. To analyse patient-level binary outcome measures, we will use mixed-effects logistic regression models with offset terms for the denominator. GP practice-level variables will include patient population size (39), Index of Multiple Deprivation quintile (37), monthly volume of online consultations per 1000 patients, geographic region, Rural Urban Classification (38), number of whole-time equivalent GPs per 1000 patients (40), levels of

patient morbidity (39), Quality and Outcomes Framework performance (39), and proportion of health request types. Patient-level variables will include age, sex, ethnicity (48), Index of Multiple Deprivation quintile (37), and non-English language usage (33). Time will be a continuous variable of the number of months GP practices have been actively using online consultations. We will also attempt to include month as a categorical variable to account for seasonality.

Where GP practice-level outcome measures are available prior to when they first started using online consultations, they will be analysed using an interrupted time series analysis (58). The main exposure of interest will be before vs after the practices first used online consultations modelled as binary (0/1). Where GP practice-level data are also available for practices not using Patchs (control practices), the main exposure of interest will be membership to the Patchs or control group modelled as binary (0/1). The main parameter of interest will be the interaction term between practice group (Patchs vs control) and study period (before vs after online consultations); post-estimation commands will be used to obtain estimates for each study period by practice group.

If multiple online consultations are submitted by the same patient, we will randomly sample one per patient for analysis. Multiple online consultations from the same patient are expected to be infrequent, though if this approach adversely affects reaching our sample size target we will instead include all online consultations with a patient-level variable in our models.

Sensitivity analyses

Limitations of our approach include that staff triage decisions could be applied by non-clinicians which could be systematically different to those applied by clinicians (59), and that the true urgency of an online consultation may only be apparent when further information has been obtained from the patient (e.g. over the telephone). We will therefore conduct sensitivity analyses where we restrict triage decisions to those only made by clinicians and use the final triage decision when the online consultation is complete.

Workstream 2: AI development – retrospective data from GP practices with AI disabled *and* enabled

Descriptive analysis

Data will be descriptively analysed using the same approach described in Workstream 1.

Model outcomes and predictors

Models will be built for each AI to predict the outcomes in Table 6. Model predictors will consist of data recorded in Patchs believed to influence triage decisions of online consultations based on our PPI work with patients and GP practice staff (Table 8).

Table 8: Predictors for all AI models

Purpose	Details
Patient age	At the time of online consultation submission. Continuous in non-LLM models. Transformed to text in LLM models (Table 9).
Sex	Categorical: male, female, not specified
Online consultation type	Categorical: Clinical (i.e. 'new health problem', 'ongoing health problem', 'other'), non-clinical (i.e. 'admin request', 'medication request')

Online consultation text	All data inputted by patients and GP practice staff and online consultation text written by patients in response to clinically relevant questions in the Patchs online consultation form will be included. Other questions (e.g. patient availability or preferred clinician) will be excluded as they are not clinically relevant and may contain personal identifiable information. Written by patients. Vectorised in non-LLM models. Tokenised in LLM models.

Modelling approaches

We will evaluate four different modelling approaches:

1. Logistic regression
2. XGBoost
3. Long short-term memory (LSTM)
4. Large language model (LLM)

The choice of model was based on testing the value of increasing complexity of models that have shown promising performance in text classification tasks (60). The current versions of the AI models used in clinical practice will also be evaluated.

Internal-external cross-validation (IECV)

To account for potential regional variations in the incidence of outcomes and predictors, we will use internal-external cross-validation (IECV) to assess the geographical transportability of the models (61). We will split GP practices into ten geographical regions with a similar total number of online consultations completed across all constituent GP practices. We will withhold data from one region and develop the model on the remaining data. We will then apply the model to data from the withheld region to obtain predictions for each online consultation. We will repeat this process so that each region is excluded once from the development data.

We will pool model evaluation metrics (Table 7) with standard errors from each withheld region and use random-effects meta-analysis to obtain a summary estimate of model performance to compare the different models. The proportion of total variability in performance due to heterogeneity between regions will be quantified by I^2 (with 95% CI) (62). Multivariable meta-regression will examine potential contributory factors to heterogeneity (e.g. mean age of patients, proportion of patient ethnicities, mean deprivation score, proportion of practices using AI triage) (63). We will also explore heterogeneity across GP practices too.

We will use pooled individual online consultation predictions to assess fairness by comparing true positive rates across relevant subgroups (e.g. patient ethnicity, 10-year age bands, sex, and socioeconomic deprivation quintiles) (64) (42). Subgroup analyses will assess how evaluation metrics vary based on GP practice characteristics (e.g. monthly online consultation volume, proportion of triage decisions, whether AI triage is enabled).

We will take the best-performing model based on the $F1$ score (our primary outcome measure for model evaluation), taking into account measures of calibration and clinical utility (Table 7), and fairness, and fit it to the entire Workstream 2 dataset (65).

Threshold selection

Each AI model will output a probability, which will be converted to an outcome if it reaches a specified threshold. To find the optimal threshold when developing each model, we will randomly split

development data into two groups: one for model development (90-95%) and another for threshold selection (5-10%) balanced by outcome (Table 6). We will select the threshold that minimises the difference between positive predictive value and true positive rate per our primary outcome measure, whilst keeping the true positive rate higher than positive predictive value to minimise clinical safety issues (Table 3).

Hyperparameter tuning

We will hyperparameter tune (where model (hyper)parameters are 'learnt' according to different settings) using K-fold cross-validation stratified based on outcome during model development. The set of hyperparameters that achieve the highest average $F1$ score across all folds (micro-averaged across all classes in Topic AI) will be selected. This model and set of hyperparameters will then be fit to the entire development data in that IECV split.

Text pre-processing

Text data will initially be anonymised at source by Patchs Health prior to extraction using a validated algorithm (43). It will then be concatenated into a single text variable using spaces between each answer to questions in the Patchs form. For non-LLM models, the text will be sanitised by removing HTML tags, numbers, non-alphanumeric characters, and stop words (e.g. 'a', 'the', 'is' etc) except negations (i.e. 'no' and 'not'); expanding contractions (e.g. 'do not' instead of 'don't'); converting to lower case; and spell-checking (using pyspellchecker library) on words longer than six characters. For LLM models, we will follow the same except removing numbers and stop words. For non-LLM models, individual words will be vectorised using Global Vector (GloVe) embeddings with 100 parameters (66). For LLM models, we will use text tokenisation inherent to the model.

Logistic regression

We will fit a binary (or multinomial for Topic AI) logistic regression model that applies a logistic function on a linear combination of the input variables (67). Sex and online consultation type predictors will be converted to binary variables ('one hot encoding'). Age (years) will be a continuous standardised predictor with mean centred at 0 and standard deviation scaled to 1. Online consultation text written by patients will be aggregated and sanitised through the text processing pipeline. We will use L2 regularisation in the estimation of linear weights. Hyperparameter tuning will focus on the inverse of the regularisation strength.

XG Boost

XGBoost (extreme gradient boosting) is an ensemble of gradient-boosted decision trees (68). We will use a log loss based on a binary (or multinomial for Topic AI) logistic function to develop the model. As with logistic regression, sex and online consultation type input predictors will be converted to binary variables ('one hot encoding'). Age (years) will be a continuous predictor with mean centred at 0 and standard deviation scaled to 1. Online consultation text written by patients will be aggregated and sanitised through the text processing pipeline. Hyperparameter tuning will focus on the maximum depth of each tree and the minimum number of samples required to create a new node in a tree.

Long short-term memory (LSTM)

We will fit a long short-term memory (LSTM). Online consultation text written by patients will be passed through two parallel bi-directional LSTM layers, one with a longer sequence length and one with a shorter sequence length. Sex and online consultation type predictors will be converted to binary variables ('one hot encoding') and passed through a single deep layer with age (years) as a continuous variable (standardised with mean centred at 0 and standard deviation scaled to 1). Outputs of the LSTM and deep

layer will be concatenated and passed through a final output layer. We will use dropout regularisation to control for overfitting (69). Hyperparameter tuning will focus on the numbers of neurons in the layers, sequence length, weights for loss functions, and degree of dropout regularisation.

Large language model (LLM)

We will fit a BERT-based LLM model (70) with an added classification layer because this has shown promise in both general (71) and medical text classification tasks (72). The BERT LLM and its extensions are based on an encoder-only transformer architecture and use bidirectional information in the text which makes them better suited to text classification tasks than decoder-only models like Generative Pre-trained Transformers (GPTs) (73). We will explore other LLM architectures, including GPTs, where appropriate. BERT models usually only take text data as their sole input (70). Therefore, in addition to the concatenated initial online consultation text written by patients, we will also concatenate the other input variables separated by separator tokens before online consultation text answers. Patient age is the only continuous input variable that requires transforming into text. This will be achieved by mapping patient age ranges to text descriptions, which has performed better than alternative methods of including numerical data for BERT models in text classification tasks (74). We will use accepted terms from the UK NHS Digital Service Manual (Table 9) (75). Hyperparameter tuning will focus on the learning rate, number of epochs, and batch size.

Table 9: Transformation of patient ages to accepted UK terms for Large Language Model (LLM) development

Age range	**UK NHS term** (75)
Up to 1 year	Baby
1 to 3 years	Toddler
4 to 12 years	Child
13 to 18 years	Teenager
19 to 39 years	Young adult
40 to 64 years	Middle aged adult
65 to 84 years	Elderly adult
85 and above	Very elderly adult

Current clinical practice

In GP practices not using AI, online consultations are usually triaged by non-clinicians first (typically receptionists) before being assigned to clinicians when appropriate (9). We will sample online consultations that have been first triaged by non-clinicians and subsequently reviewed by a clinician to estimate the triage performance of non-clinicians. We will evaluate their performance, in addition to the current AI model version used in clinical practice, using the IECV approach described above.

Workstream 3: Background evaluation – prospective data from GP practices with AI *disabled*

Descriptive analysis

Data will be descriptively analysed using the same approach described in Workstream 1.

Model evaluation

The best-performing model from Workstream 2 will be used to obtain predictions for each online consultation in the dataset and model evaluation metrics calculated (Table 7) (45). As prospective data, this will assess the temporal transportability of the model (64)(65). We will assess fairness by comparing true positive rates across relevant subgroups (e.g. patient ethnicity, 10-year age bands, sex,

socioeconomic deprivation quintiles, non-English language usage) (64)(42). Subgroup analyses will also assess how evaluation metrics vary based on GP practice characteristics (e.g. monthly online consultation volume, proportion of triage decision outcomes at baseline).

The predictions will be compared to usual current clinical practice where online consultations are triaged by non-clinicians first (typically receptionists) using the same sampling method described in Workstream 2. We will populate a contingency table and compare the AI vs non-clinical staff triage decision using McNemar's test for paired nominal data (76) for binary models and Stuart-Maxwell's test (77) for Topic AI. Comparisons will made for overall accuracy, true positive rate, true negative rate, positive predictive value, and negative predictive value.

We will undertake a failure case analysis to explore factors why the model may have predicted incorrectly (78) by quantitatively comparing the characteristics of patients who have submitted online consultations classified as false positive and negatives to those of true positive and negative predictions ('error auditing') (79). Additional factors tested will include those related to the online consultation (including type of online consultation, time and day of submission), GP practice staff who applied the triage decision (including role, experience using Patchs), and GP practice (including size, geographic location, experience using Patchs). The Patchs Health Clinical Safety team will qualitatively review a sample of anonymised misclassified online consultations as per their internal quality assurance processes to comply with MHRA and NHS DCB0129 standards (22) to understand how they occurred. We will review their summary findings to understand if there are patient groups or online consultation topics that are at higher risk of misclassification by the AI models and test those findings in quantitative analyses (80).

Sensitivity analyses

As per our sampling approach described in section 6.3, primary analyses will be undertaken on a random sample of GP practices, patients, and online consultations to match those used to develop the AI models in Workstream 2 (35). Sensitivity analyses will be undertaken without sampling and by using the same approach described in Workstream 1.

Workstream 4: Implementation evaluation – prospective data from GP practices with AI *enabled*

Descriptive analysis

Data will be descriptively analysed using the same approach described in Workstream 1.

Model evaluation

Like Workstream 3, the best-performing model from Workstream 2 will be deployed in GP practices using AI triage and model evaluation metrics calculated (Table 7). Metrics will differ to those in Workstream 3, because they will account for how GP practices make triage decisions after being presented with an AI triage prediction. Where an AI prediction is unchanged by staff, we will consider it a true positive or negative. Where an AI prediction is changed by staff, we will consider it a false positive or negative. Model evaluation metrics will be compared to results from Workstream 3 to quantify 'automation bias' (34). We will assess fairness by comparing true positive rates across relevant subgroups (e.g. patient ethnicity, 10-year age bands, sex, socioeconomic deprivation quintiles, non-English language usage) (64)(42). Subgroup analyses will also assess how evaluation metrics vary based on GP practice characteristics (e.g. monthly online consultation volume, proportion of triage decision outcomes at baseline). We will also undertake error auditing (79) and review of Patchs Health Clinical Safety team's qualitative findings described in Workstream 3. The latter will include any reported patient safety incidents from GP practices.

We will analyse the evaluation metrics of the new model in a monthly interrupted time series analysis (58). This will include usual current clinical practice where online consultations are triaged by non-clinicians first (typically receptionists) using the same sampling method described in Workstream 2, and previous version of the AI model they have used. We will use mixed-effects logistic regression models with appropriate offset terms (number of online consultations). The main exposure of interest will be before / after the updated AI model is deployed as binary (0/1). Models will be adjusted for GP practice characteristics including patient population size, Index of Multiple Deprivation (37), Rural Urban Classification (38), baseline proportion of triage outcomes, online consultation volume, length of time using Patchs, and other features enabled in Patchs. Time will be modelled as continuous to account for trends in the pre-intervention period. We will also attempt to include month as a categorical variable to account for seasonality.

Sensitivity analyses

As per our sampling approach described in section 6.3, primary analyses will be undertaken on a random sample of GP practices, patients, and online consultations to match those used to develop the AI models in Workstream 2 (35). Sensitivity analyses will be undertaken without sampling and by using the same approach described in Workstream 1. We will conduct a further sensitivity analysis where we include online consultations cancelled after receiving a signpost message from Signpost AI.

Statistical software

Workstreams 1, 3, and 4 will be carried out in Stata and Python using the following libraries: *pandas* and *scikit-learn*. Workstream 2 will be carried in Python using the following libraries: *kedro*, *pandas*, *nltk*, *gensim*, *contractions*, *pyspellchecker*, *scikit-learn*, *tensorflow*, *xgboost* and *transformers*.

11.2 Sample Size:

We have calculated minimum required sample sizes of online consultations for all workstreams. The impact of between-GP practice variability ('clustering') is difficult to predict; to minimise these effects we aim to recruit as many GP practices as possible. For prospective Workstreams 3 and 4, we have estimated minimum number of GP practices required to achieve our online consultations sample size during our target data collection period of maximum six months (section 5.2). Our total minimum required sample size of online consultations across all workstreams is 226821.

Workstream 1

Primary outcomes in this workstream are proportions of triage outcomes in online consultations based on staff triage decisions (Table 2). Using the Agresti-Coull approach for estimating a binomial proportion (81) in the R package *presize* (82) and a worst case scenario of estimating a proportion of 50% within 1% with 95% confidence, we estimate we require a minimum sample size of 38411 online consultations.

Workstream 2

There is no agreed approach for estimating sample sizes to develop XGBoost, LSTM, or LLM to predict clinical outcomes. Modern modelling techniques like these may require over 10 times as many events per variable to minimise overfitting than classical techniques such as regression (83). We therefore used Riley et al.'s approach (84) for estimating sample sizes required to develop regression-based clinical prediction models implemented in the R package *pmsampsize* (85) and multiplied its outputs by 10 (Table 10) (83). Further details are provided in Appendix 3.

Table 10: Sample sizes required to develop each AI

AI	Minimum required number of online consultations to develop the AI
Urgency	50320
Assign	28380
Face-to-face	89610

Workstreams 3 and 4

There is no agreed approach for calculating sample sizes to estimate the $F1$ score (our primary outcome measure) other than bootstrapping, which is too computationally expensive for modern modelling techniques like LSTMs and LLMs (64). We therefore used Riley et al.'s approach (86) for estimating sample sizes required to evaluate clinical prediction models implemented in the R package *pmvalsampsize* (87) (Table 11). Further details are provided in Appendix 3.

Table 11: Sample sizes required to evaluate each AI

AI	Minimum required number of online consultations to evaluate the AI	Minimum number of GP practices required to collect the online consultations
Workstream 3		
Urgency	3806	13
Assign	1568	5
Face-to-face	96323	321
Workstream 4		
Urgency	2477	8
Assign	1785	6
Face-to-face	1568	5

12) DATA AND INTERVENTION ACCESS POST-STUDY

Personal data will not be stored as part of this study. Anonymised data will be stored on secure UoM Research Data Storage servers for a minimum of 5 years after publication of our results per UoM's Record Retention Schedule (88). Only the UoM research team will have access to these data during this time. After this period, following consideration of all legal and ethical perspectives, interests and contractual stipulations of third-party funders and other stakeholders, as well as aspects of confidentiality and security, the data will be deleted using hard drive eraser software. We will document deletion and destruction of data, and make it accessible for possible future audit.

13) MONITORING AND QUALITY ASSURANCE

The study will be subject to the audit and monitoring regime of The University of Manchester. As a UKCA-marked medical device used in routine clinical practice, any safety issues associated with the AI Triage intervention will be reported and managed in the usual way by Patchs Health as per UKCA and NHS DCB0129 standards.

14) PEER REVIEW

Members of the UoM project team who are experts in data science (Peek and Kontopantelis) have reviewed this protocol. A statistician and expert in clinical prediction modelling at UoM, and the lead data scientist at Patchs Health, both of whom are independent of the UoM project team, have also reviewed the protocol.

15) PATIENT AND PUBLIC INVOLVEMENT (PPI) AND STAKEHOLDER ENGAGEMENT

Input to this plan

We conducted interviews and focus groups with GP practice staff (n=16) and patients (n=37) to gather feedback on our study plans. They provided input into our approach including: AI functions to develop, predictors to use in the models, selection of primary model evaluation metrics, and which benchmarks to use when comparing the models to existing clinical practice.

Input during this study

We have a PPI group of six members of the public who use GP services and a stakeholder group of six primary care members of staff (GPs, receptionists, and practice managers). Both groups will contribute to data interpretation, analysis (where possible), and project outputs. The groups will meet separately up to 10 times during the project every 2-3 months. Meetings will be face-to-face or via video conference, though members will also be able to contribute by other means (e.g. email, phone, or post), dependant on individual circumstances.

16) ETHICAL AND REGULATORY CONSIDERATIONS

16.1 Approvals

NHS Research Ethics Committee and Health Research Authority approval will be obtained before commencing research. Submission to the MHRA is not required because the intervention is a UKCA-marked medical device (Class I) already used in routine clinical practice that will continue to be used within scope of its intended purpose. The study will be conducted in full conformance with all relevant legal requirements and the principles of the Declaration of Helsinki, Good Clinical Practice, and the UK Policy Framework for Health and Social Care Research 2017.

16.2 Risks

Risks to participants

There is a theoretical risk of identifying participants from the research data. A risk assessment (Appendix 2) based on guidance from the Information Commissioner's Office (50) has been undertaken and described above. It concludes the data are effectively anonymised and the risk of a motivated intruder identifying a participant is low. Security arrangements for protecting the data include using industry-standard practices for transferring data as password-protected files using Hypertext Transfer Protocol Secure over a Transport Layer Security connection encrypted both at rest and in transit, and storing the data on secure servers.

Risks to researchers

Researchers will analyse routinely collected data only; we have identified no additional risks to them in conducting the study.

16.3Benefits

This research will help understand the problems currently faced by GP practices during online consultation triage and identify potential actions to address them. If we build better performing AIs than those currently in use, there may be benefits by improving patient safety by helping patients receive help sooner and reducing GP practice workload by automating the triage process. GP practices and their patients in Workstream 4 would realise these benefits immediately; those outside the study would benefit as the updated AI models are rolled out more widely. The study will also provide evidence for GP practices not currently using AI triage whether to adopt it, including whether AI triage is fair.

17) STATEMENT OF INDEMNITY

The University has insurance available regarding research involving human subjects that provides cover for legal liabilities arising from its actions or those of its staff or supervised students. The University also

has insurance available that provides compensation for non-negligent harm to research subjects occasioned in circumstances that are under the control of the University.

18) FUNDING AND RESOURCES

There are no costs for participant recruitment, data collection, or data access. The only costs associated with this project are for staff.

19) PUBLICATION POLICY

This protocol will be registered on the publicly accessible Open Science Framework database (https://osf.io). Findings will be published in open-access peer-reviewed scientific journals. We will produce short evidence summaries communicating key findings in an accessible way, which will be hosted on publicly available websites (e.g. www.patchs.ai) and disseminated to participating GP practices by Patchs Health via email. We will encourage GP practices to share these findings with their patients, for example by publishing them on their website. The final AI models will be described regarding relevant parameters. Commands and software packages used will be reported in manuscripts submitted for publication. We will report our findings as per RECORD guidelines (26) for Workstream 1, TRIPOD+AI guidelines (27) for Workstreams 2 and 3, and DECIDE-AI guidelines for Workstream 4 (28). Where tables are published, cells with values less than 5 will be censored to prevent combining them with other information that could potentially identify participants. If example online consultations are published, no information will be published that could be combined with other data sources to identify participants such as age or ethnicity.

20) DECLARATION OF INTERESTS

Dr Brown is a part-time employee of Patchs Health as Chief Medical Officer and shareholder in the company. Co-Investigators have no relevant interests to declare.

21) REFERENCES

1. Is my study research? [Internet]. [cited 2024 Feb 8]. Available from: https://www.hra-decisiontools.org.uk/research/index.html

2. Do I need NHS Ethics approval? [Internet]. [cited 2024 Feb 8]. Available from: https://www.hra-decisiontools.org.uk/ethics/index.html

3. IRAS Help - Reference - Collated Guidance - Project Filter [Internet]. [cited 2024 Feb 8]. Available from: https://www.myresearchproject.org.uk/help/hlpcollatedqsg-sieve.aspx#2349

4. NHS Digital [Internet]. [cited 2023 Dec 13]. Appointments in General Practice. Available from: https://digital.nhs.uk/data-and-information/publications/statistical/appointments-in-general-practice

5. Van Staa T, Li Y, Gold N, Chadborn T, Welfare W, Palin V, et al. Comparing antibiotic prescribing between clinicians in UK primary care: an analysis in a cohort study of eight different measures of antibiotic prescribing. BMJ Qual Saf. 2022 Mar 3;bmjqs-2020-012108.

6. Gharbi M, Drysdale JH, Lishman H, Goudie R, Molokhia M, Johnson AP, et al. Antibiotic management of urinary tract infection in elderly patients in primary care and its association with bloodstream infections and all cause mortality: population based cohort study. BMJ. 2019 Feb 27;364:l525.

7. Fielding S, Porteous T, Ferguson J, Maskrey V, Blyth A, Paudyal V, et al. Estimating the burden of minor ailment consultations in general practices and emergency departments through retrospective review of routine data in North East Scotland. Fam Pract. 2015 Apr;32(2):165–72.

8. Nguyen H, Meczner A, Burslam-Dawe K, Hayhoe B. Triage Errors in Primary and Pre–Primary Care. J Med Internet Res. 2022 Jun 24;24(6):e37209.

9. Darley S, Coulson T, Peek N, Moschogianis S, van der Veer SN, Wong DC, et al. Understanding how the design and implementation of online consultations impact primary care quality: Systematic review of evidence with recommendations for designers, providers, and researchers (Preprint). J Med Internet Res. 2022 Feb 22;

10. NHS England. GP Contract. 2019 [cited 2022 May 9]. GP Contract documentation 2019/20. Available from: www.england.nhs.uk/gp/investment/gp-contract/gp-contract-documentation-2019-20

11. Bakhai M. NHSX. 2020 [cited 2022 May 9]. The use of online and video consultations during the COVID-19 pandemic - delivering the best care to patients. Available from: https://www.nhsx.nhs.uk/blogs/use-online-and-video-consultations-during-covid-19-pandemic-delivering-best-care-patients/

12. Clarke GM, Dias A, Wolters A. Access to and delivery of general practice services: a study of patients at practices using digital and online tools [Internet]. London: The Health Foundation; 2022. Available from: https://www.health.org.uk/publications/access-to-and-delivery-of-general-practice-services

13. Bellman R. An Introduction to Artificial Intelligence: Can Computers Think? Boyd & Fraser Publishing Company; 1978. 168 p.

14. Judson TJ, Odisho AY, Neinstein AB, Chao J, Williams A, Miller C, et al. Rapid design and implementation of an integrated patient self-triage and self-scheduling tool for COVID-19. J Am Med Inform Assoc. 2020;27(6):860–6.

15. Paulus JK, Kent DM. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. Npj Digit Med. 2020 Dec 1;3(1).

16. NHS Digital. Buying Catalogue. 2022 [cited 2022 May 9]. Find Buying Catalogue Solutions. Available from: https://buyingcatalogue.digital.nhs.uk/catalogue-solutions

17. Brown B. PATCHS Help Centre. 2022 [cited 2022 May 9]. Urgency AI. Available from: https://help.patchs.ai/hc/en-gb/articles/360058979713-Urgency-AI

18. PATCHS Support [Internet]. 2023 [cited 2024 Feb 29]. Topic AI. Available from: https://help.patchs.ai/hc/en-gb/articles/360060879913-Topic-AI

19. PATCHS Support [Internet]. 2024 [cited 2024 Feb 29]. Assign AI. Available from: https://help.patchs.ai/hc/en-gb/articles/1500010485241-Assign-AI

20. PATCHS Support [Internet]. 2024 [cited 2024 Feb 29]. Face-to-Face (F2F) AI. Available from: https://help.patchs.ai/hc/en-gb/articles/4407650662167-Face-to-Face-F2F-AI

21. Brown B. PATCHS Help Centre. 2022 [cited 2022 May 9]. Signpost AI. Available from: https://help.patchs.ai/hc/en-gb/articles/4410509916183-Signpost-AI

22. NHS Digital [Internet]. [cited 2023 Dec 13]. DCB0129: Clinical Risk Management: its Application in the Manufacture of Health IT Systems. Available from: https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/dcb0129-clinical-risk-management-its-application-in-the-manufacture-of-health-it-systems

23. NHS Digital. Buying Catalogue. 2021 [cited 2022 May 9]. PATCHS Online Consultation. Available from: https://buyingcatalogue.digital.nhs.uk/catalogue-solutions/10046-006/features

24. MHRA. Guidance: Medical device stand-alone software including apps (including IVDMDs) v1.08. 2021.

25. PATCHS Support [Internet]. 2024 [cited 2024 May 30]. Introduction to Patchs AI. Available from: https://help.patchs.ai/hc/en-gb/articles/4411218478871-Introduction-to-Patchs-AI

26. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLOS Med. 2015 Oct 6;12(10):e1001885.

27. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024 Apr 16;e078378.

28. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. The BMJ. 2022;

29. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51–8.

30. Patchs AI (Artificial Intelligence) – PATCHS Support [Internet]. [cited 2023 Dec 13]. Available from: https://help.patchs.ai/hc/en-gb/sections/360012017013-Patchs-AI-Artificial-Intelligence-

31. NHS Digital. DCB0160: Clinical Risk Management: its Application in the Deployment and Use of Health IT Systems - NHS Digital. NHS Digit [Internet]. 2018; Available from: https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/dcb0160-clinical-risk-management-its-application-in-the-deployment-

32. PATCHS Support [Internet]. 2023 [cited 2023 Dec 15]. How to verify carer relationship. Available from: https://help.patchs.ai/hc/en-gb/articles/1500004882962-How-to-verify-carer-relationship

33. Brown B. What is PATCHS Translate? [Internet]. 2023. Available from: https://help.patchs.ai/hc/en-gb/articles/1500005595042-What-is-PATCHS-Translate-

34. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc. 2012 Jan;19(1):121–7.

35. Sperrin M, Riley RD, Collins GS, Martin GP. Targeted validation: validating clinical prediction models in their intended population and setting. Diagn Progn Res. 2022 Dec 22;6(1):24.

36. Kontopantelis E. A Greedy Algorithm for Representative Sampling: repsample in Stata. J Stat Softw [Internet]. 2013;55(Code Snippet 1). Available from: http://www.jstatsoft.org/v55/c01/

37. Office for National Statistics. Community and society. 2019 [cited 2022 Sep 30]. English indices of deprivation 2019. Available from: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019

38. Office for National Statistics. 2011 rural/urban classification - Office for National Statistics [Internet]. [cited 2024 Jul 31]. Available from: https://www.ons.gov.uk/methodology/geography/geographicalproducts/ruralurbanclassifications/2011ruralurbanclassification#

39. NHS Digital [Internet]. [cited 2023 Dec 13]. Quality and Outcomes Framework. Available from: https://digital.nhs.uk/data-and-information/publications/statistical/quality-and-outcomes-framework-achievement-prevalence-and-exceptions-data

40. NHS Digital [Internet]. [cited 2023 Dec 13]. General Practice Workforce. Available from: https://digital.nhs.uk/data-and-information/publications/statistical/general-and-personal-medical-services

41. Office for National Statistics. Census [Internet]. [cited 2024 Jul 31]. Available from: https://www.ons.gov.uk/census

42. Standard 4: consider health and care inequalities and bias mitigation [Internet]. NICE; 2018 [cited 2024 Apr 25]. Available from: https://www.nice.org.uk/corporate/ecd7/chapter/how-to-meet-the-standards#standard-4-consider-health-and-care-inequalities-and-bias-mitigation

43. Fernandes AC, Cloete D, Broadbent MTM, Hayes RD, Chang CK, Jackson RG, et al. Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records. BMC Med Inform Decis Mak. 2013 Dec 11;13(1):71.

44. GP Patient Survey [Internet]. [cited 2024 Jul 3]. Available from: https://www.gp-patient.co.uk/

45. Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ. 2024 Jan 15;e074820.

46. Vickers AJ, Van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. 2019 Dec;3(1):18.

47. Ancker JS, Edwards A, Nosal S, Hauser D, Mauer E, Kaushal R. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Med Inform Decis Mak. 2017 Dec 10;17(1):36.

48. NHS Digital [Internet]. [cited 2023 Dec 13]. Ethnicity. Available from: https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets/mental-health-services-data-set/submit-data/data-quality-of-protected-characteristics-and-other-vulnerable-groups/ethnicity

49. NHS Digital [Internet]. [cited 2023 Dec 13]. About the Organisation Data Service. Available from: https://digital.nhs.uk/services/organisation-data-service/about-the-organisation-data-service

50. Information Commissioner's Office. How do we ensure anonymisation is effective? [Internet]. Information Commissioner's Office; 2021 Oct. Available from: https://ico.org.uk/media/about-the-ico/documents/4018606/chapter-2-anonymisation-draft.pdf

51. NHS Digital [Internet]. [cited 2023 Dec 13]. Emergency Care Data Set (ECDS). Available from: https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets/emergency-care-data-set-ecds

52. Public Health England. National General Practice Profiles. Crown Copyr. 2023;1–14.

53. NHS England Digital [Internet]. [cited 2024 Dec 4]. Submissions via Online Consultation Systems in General Practice. Available from: https://digital.nhs.uk/data-and-information/publications/statistical/submissions-via-online-consultation-systems-in-general-practice

54. NHS Digital. NHS England Digital. [cited 2024 Jul 31]. Patients Registered at a GP Practice. Available from: https://digital.nhs.uk/data-and-information/publications/statistical/patients-registered-at-a-gp-practice

55. Data Security and Protection Toolkit [Internet]. [cited 2024 Mar 11]. Available from: https://www.dsptoolkit.nhs.uk/

56. About Cyber Essentials [Internet]. [cited 2024 Mar 11]. Available from: https://www.ncsc.gov.uk/cyberessentials/overview

57. University of Manchester. Research Data Storage Documentation [Internet]. Available from: https://ri.itservices.manchester.ac.uk/rds/

58. Kontopantelis E, Doran T, Springate DA, Buchan I, Reeves D. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. BMJ. 2015;350(4):h2750.

59. Sexton V, Atherton H, Dale J, Abel G. Clinician-led secondary triage in England's urgent care delivery: a cross-sectional study. Br J Gen Pract. 2023 Jun;73(731):e427–34.

60. Gasparetto A, Marcuzzo M, Zangari A, Albarelli A. A Survey on Text Classification Algorithms: From Text to Predictions. Information. 2022 Feb 11;13(2):83.

61. Austin PC, Van Klaveren D, Vergouwe Y, Nieboer D, Lee DS, Steyerberg EW. Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. J Clin Epidemiol. 2016 Nov;79:76–85.

62. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 2016 Jun 22;i3140.

63. Steyerberg EW, Nieboer D, Debray TPA, Van Houwelingen HC. Assessment of heterogeneity in an individual participant data meta-analysis of prediction models: An overview and illustration. Stat Med. 2019 Sep 30;38(22):4290–309.

64. Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Van Calster B, et al. Evaluation of clinical prediction models (part 1): from development to external validation. BMJ. 2024 Jan 8;e074819.

65. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal–external, and external validation. J Clin Epidemiol. 2016 Jan;69:245–7.

66. Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation. Proc 2014 Conf Empir Methods Nat Lang Process EMNLP. 2014;1532–43.

67. McCullagh P. Generalized Linear Models. 2nd ed. New York: Routledge; 2019. 532 p.

68. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. San Francisco California USA: ACM; 2016 [cited 2024 May 9]. p. 785–94. Available from: https://dl.acm.org/doi/10.1145/2939672.2939785

69. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting.

70. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019;

71. Mohammadi S, Chapon M. Investigating the Performance of Fine-tuned Text Classification Models Based-on Bert. In: 2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS) [Internet]. Yanuca Island, Cuvu, Fiji: IEEE; 2020 [cited 2024 Jun 5]. p. 1252–7. Available from: https://ieeexplore.ieee.org/document/9408054/

72. Babu A, Boddu SB. BERT-Based Medical Chatbot: Enhancing Healthcare Communication through Natural Language Understanding. Explor Res Clin Soc Pharm. 2024 Mar 1;13:100419.

73. Benayas A, Sicilia MA, Mora-Cantallops M. A Comparative Analysis of Encoder Only and Decoder Only Models in Intent Classification and Sentiment Analysis: Navigating the Trade-Offs in Model Size and Performance [Internet]. 2024 [cited 2024 Jun 5]. Available from: https://www.researchsquare.com/article/rs-3865391/v1

74. Gu K, Budhkar A. A Package for Learning on Tabular and Text Data with Transformers. In: Proceedings of the Third Workshop on Multimodal Artificial Intelligence [Internet]. Mexico City, Mexico: Association for Computational Linguistics; 2021 [cited 2024 May 30]. p. 69–73. Available from: https://www.aclweb.org/anthology/2021.maiworkshop-1.10

75. nhs.uk [Internet]. [cited 2024 May 30]. Age - NHS digital service manual. Available from: https://service-manual.nhs.uk

76. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947 Jun;12(2):153–7.

77. Maxwell AE. Comparing the Classification of Subjects by Two Independent Judges. Br J Psychiatry. 1970 Jun 29;116(535):651–5.

78. Rivera SC, Liu X, Chan A wen, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. BMJ. 2020 Sep 9;m3210.

79. Oakden-Rayner L, Dunnmon J, Carneiro G, Re C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: ACM CHIL 2020 - Proceedings of the 2020 ACM Conference on Health, Inference, and Learning. Association for Computing Machinery, Inc; 2020. p. 151–9.

80. Moore G, Audrey S, Barker M, Bonell C, Hardeman W, Moore L, et al. Process evaluation of complex interventions: UK Medical Research Council (MRC) guidance. 2014.

81. Brown LD, Cai TT, DasGupta A. Interval Estimation for a Binomial Proportion. Stat Sci. 2001;16(2):101–17.

82. Lenz A, Haynes AG, Limacher A. presize: Precision Based Sample Size Calculation [Internet]. 2023 [cited 2024 Jul 24]. p. 0.3.7. Available from: https://CRAN.R-project.org/package=presize

83. Van Der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol. 2014 Dec;14(1):137.

84. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020 Mar 18;m441.

85. Ensor J. pmsampsize: Sample Size for Development of a Prediction Model [Internet]. 2019 [cited 2024 Jul 15]. p. 1.1.3. Available from: https://CRAN.R-project.org/package=pmsampsize

86. Riley RD, Snell KIE, Archer L, Ensor J, Debray TPA, Van Calster B, et al. Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study. BMJ. 2024 Jan 22;e074821.

87. Ensor J. pmvalsampsize: Sample Size for External Validation of a Prediction Model [Internet]. 2023 [cited 2024 Jul 15]. p. 0.1.0. Available from: https://CRAN.R-project.org/package=pmvalsampsize

88. The University of Manchester [Internet]. [cited 2023 Dec 13]. Retention of records. Available from: https://www.manchester.ac.uk/discover/privacy-information/freedom-information/record-retention/