APPENDIX 3: DETAILS OF SAMPLE SIZE CALCULATIONS FOR MODEL DEVELOPMENT AND VALIDATION

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

WORKSTREAM 2 (AI MODEL DEVELOPMENT)

There is no agreed approach for estimating sample sizes to develop multiclass regression models, so we used Riley et al.'s approach for developing clinical prediction models with binary outcomes (1). This approach is based on obtaining a large enough sample size to estimate the overall outcome proportion with sufficient precision, targeting a shrinkage factor of 0.9, and a small optimism of 0.05 in the apparent R^2 using estimates from existing models (1). We did not target a small mean absolute prediction error because the number of parameters we use is more than 30 (1).

We used the implementation in the *pmsampsize* package in R (2), which can be calculated using the following inputs:

- Number of predictor parameters for the prediction model
- Prevalence of the outcome in the model development dataset, which should usually be derived from previous studies in the same population
- The expected value of the C-statistic or Cox-Snell R-squared of the model, which is usually taken from previous similar models

We set the number of predictor parameters at 107 for all models (Table 8 in main document: intercept = 1, patient age = 1, sex = 3, online consultation type = 1, online consultation text vectorised using GloVe embeddings = 100) and estimated values for prevalence and the C-statistic from analyses of the AI models currently used in Patchs. We did not use the Cox-Snell R-squared because this was not available for all models at the time of writing.

Because our modelling approach utilises data from GP practices who have both AI enabled and disabled, we calculated sample sizes by analysing data from both cohorts (Table 1). We then took the largest figure as our final sample size and multiplied it by 10 to account for the 'data hungriness' of more modern modelling techniques (3) (Table 10 in main document).

WORKSTREAMS 3 AND 4 (PROSPECTIVE AI MODEL EVALUATION)

There is no agreed approach for estimating sample sizes to evaluate multiclass models, so we used Riley et al.'s approach for evaluating clinical prediction models with binary outcomes (4). This approach is based on obtaining a large enough sample size to precisely estimate the observed/expected statistic, calibration slope, c statistic, and standardised net benefit using estimates from existing models (4).

We used the implementation in the *pmvalsampsize* package in R (5), which can be calculated using the following inputs:

- Prevalence of the outcome in the model development dataset, which should usually be derived from previous studies in the same population
- The expected value of the C-statistic, sensitivity, and specificity of the model in the evaluation sample
- The expected mean and standard deviation of the linear predictor distribution from a normal distribution
- The expected threshold of the model

We used the same threshold in the AI models currently used in Patchs and estimated the remainder of the values from analyses of these models in clinical practice. To match the GP practices that will be recruited in each workstream, we analysed data from GP practices with AI *disabled* to calculate the sample size for Workstream 3, and data from GP practices with AI *enabled* to calculate the sample size for Workstream 4 (Table 2).

Assuming GP practices process 50 requests per month (a conservative estimate) and a target data collection period of maximum six months (section 5.2), we also calculated the number of GP practices required to achieve the minimum required number of online consultations during the study period (Table 2).

Table 1: Parameters and estimated sample sizes for AI model development

AI Model	GP practice cohort AI status	Prevalence*	C-statistic	Estimated sample size
Urgency	Disabled	0.096	0.869	5032
	Enabled	0.374	0.920	2095
Assign	Disabled	0.590	0.835	2501
	Enabled	0.418	0.817	2838
F2F	Disabled	0.101	0.787**	8961
	Enabled	0.515	0.787**	3450

* Prevalences between GP practices with AI enabled and disabled may differ due to ‘automation bias’ (6)

** C-statistic only available for data from GP practices with AI enabled, so was also used for GP practices with AI disabled

Table 2: Parameters and estimated sample sizes for AI model evaluation

AI	GP practice AI status	Prevalence*	C-statistic	Sensitivity	Specificity	Linear predictor mean (standard deviation)	Threshold	Estimated sample size (online consultations)	Estimated sample size (GP practices)
Urgency	Disabled	0.096	0.869	0.835	0.751	-3.910 (2.025)	0.060	3806	13
	Enabled	0.374	0.920	0.95	0.845	-3.11 (2.299)	0.060	2477	8
Assign	Disabled	0.590	0.835	0.920	0.750	0.545 (2.228)***	0.500	1568	5
	Enabled	0.418	0.817	0.907	0.317	0.545 (2.228)***	0.500	1785	6
F2F	Disabled	0.101	0.787**	0.839	0.832	0.545 (2.228)**	0.722	96323	321
	Enabled	0.515	0.787**	0.919	0.819	0.545 (2.228)**	0.722	1568	5

* Prevalences between GP practices with AI enabled and disabled may differ due to ‘automation bias’ (6)

** C-statistic and linear predictor mean and standard deviation were only available for data from GP practices with AI enabled therefore these values were also used for GP practices with AI disabled

*** Linear predictor mean and standard deviation unavailable for Assign AI therefore values from F2F AI were used (chosen instead of Urgency AI because the thresholds are similar)

REFERENCES

1. 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.

2. 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

3. 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.

4. 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.

5. 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

6. 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.