Accepted for publication in Medical Decision Making (14/7/2025)

Title: Determinants of physicians' referrals for suspected cancer given a risk-prediction algorithm: Linking signal detection and fuzzy-trace theory

Authors:

Olga Kostopoulou, PhD*

Bence Pálfi, PhD*

Kavleen Arora, MBChB

Valerie Rayna, PhD

Email and affiliations:

Olga Kostopoulou (o.kostopoulou@imperial.ac.uk): Imperial College London

Bence Pálfi (b.palfi@gold.ac.uk): Goldsmiths University of London

Kavleen Arora (drkavleenarora@gmail.com): Imperial College London

Valerie Rayna (vr53@cornell.edu): Cornell University

*Equal contribution

Corresponding author: Olga Kostopoulou (o.kostopoulou@imperial.ac.uk)

ABSTRACT

Background. Previous research suggests that physicians' inclination to refer patients for suspected cancer is a relatively stable characteristic of their decision-making. We aimed to identify its psychological determinants in the presence of a risk-prediction algorithm.

Methods. We presented 200 UK General Practitioners with online vignettes describing patients with possible colorectal cancer. Per vignette, GPs indicated likelihood of referral (from "highly unlikely" to "highly likely") and level of cancer risk (negligible/low/medium/high), received an algorithmic risk estimate, and could then revise their responses. After completing the vignettes, GPs responded to questions about their values with regards to harms and benefits of cancer referral for different stakeholders; perceived severity of errors; acceptance of false alarms; and attitudes to uncertainty. We tested whether these values and attitudes predicted their earlier referral decisions.

Results. The algorithm significantly reduced both referral likelihood (b=-0.06 [-0.10, -0.007], p=0.025) and risk level (b=-0.14 [-0.17, -0.11] p<0.001). The strongest predictor of referral was the value GPs attached to patient benefits (b=0.30 [0.23, 0.36] p<0.001), followed by benefits (b=0.18 [0.11, 0.24] p<0.001) and harms (b=-0.14 [-0.21, -0.08] p<0.001) to the health system/society. Perceived severity of missing a cancer vis-à-vis over-referring also predicted referral (b=0.004 [0.001, 0.007] p=0.009). The algorithm did not significantly reduce the impact of these variables on referral decisions.

Conclusions. The decision to refer patients who might have cancer can be influenced by how physicians perceive and value the potential benefits and harms of referral primarily for patients, and the moral seriousness of missing a cancer vis-à-vis over-referring. These values contribute to an internal threshold for action and are important even when an algorithm informs risk judgements.

HIGHLIGHTS

Physicians' inclination to refer patients for suspected cancer is determined by their assessment of cancer risk but also their core values; specifically, their values in relation to the perceived benefits and harms of referrals and the seriousness of missing a cancer compared to over-referring.

We observed a moral prioritisation of referral decision making, where considerations about benefits to the patient were foremost, considerations about benefits but also harms to the health system or the society were second, while considerations about oneself carried little or no weight.

Having an algorithm informing assessments of risk influences referral decisions but does not remove or significantly reduce the influence of physicians' core values.

1. INTRODUCTION

General Practitioners (GPs) are on the front line of referral decisions for suspected cancer. To expedite the earlier detection of cancer, a national priority in the UK,[1] GPs can refer patients on the two-week-wait (2WW) pathway, so that patients with suspected cancer are seen by a specialist or for specialist investigations within two weeks from referral.[2] In this paper, we used the term "referral" to signify 2WW referrals. Signal detection theory suggests that such decisions are determined by two factors: "discrimination" and "response bias".[3,4] Discrimination refers to our ability to distinguish between situations that necessitate a specific response or action (e.g., referral for suspected cancer) and those that do not. Response bias refers to our inclination to take action, and, in theory, is independent of discrimination. Two people may be equally good at discrimination but produce different responses because they differ in their response bias. For example, three social workers may review the same evidence and make the same risk assessment regarding a child in possible need. Yet, one decides to refer the child; the other decides to repeat the assessment; and the third decides to dismiss the case. These three different responses are produced by differences in an internal threshold for acting. The second and third social workers need more evidence (a stronger signal) before deciding to act.

Discrimination and response bias can be measured by recording decisions over multiple trials with known outcomes and estimating the number of correct and incorrect responses. "Hits" are the number of trials where action should be taken, and action was indeed taken; "false alarms" are the number of trials where action should not be taken but action was taken. In previous research, we measured the discrimination and response bias of large samples of UK GPs who made referral decisions about hypothetical patients with possible cancer. One study involved clinical vignettes of patients with possible colorectal cancer (N=216),[5] the other with possible upper GI cancer (N=252).[6] A subset of 165 GPs took part in both studies with approximately a 1-year interval in-between studies. Although average discrimination of these GPs was uncorrelated between studies, there was substantial correlation of average response bias: GPs who were inclined to refer patients for one type of cancer were also inclined to refer patients for the other type of cancer. This suggests that, in addition to – and independent from – risk assessment, some GPs are more inclined than others to refer patients for suspected cancer.

Decision thresholds can determine differences in healthcare professionals' decision making more than discrimination or accuracy in risk assessment.[7,8] Furthermore, there is evidence that referral thresholds at the level of clinics (GP practices), representing the collective response bias across individual GPs within each clinic, are responsible for missed cancers to a greater extent than discrimination.[9] In the present study, we investigate psychological determinants of individual GPs' referral decisions, separating risk assessment from bias to take action under uncertainty. Furthermore, given the increasing emphasis on using algorithmic models to reduce clinical uncertainty, we explored whether an algorithm informing GPs of the probability that the patient has cancer influenced both risk assessment and referral responses.

In theory, GPs could base their decisions on Swets, Dawes, and Monahan's "optimal threshold" equation.[10] The optimal threshold (i.e., the one that maximizes benefits relative to costs) depends on the base rates of signal and non-signal events (e.g., cancer present vs. absent) and how the decision maker evaluates the expected outcomes of action and inaction, i.e., the benefits of a correct response (hits and correct rejections) and the costs of an incorrect response (misses and false alarms). However, it is unlikely that GPs perform a deliberative cost-benefit analysis when deciding whether to refer a patient for suspected cancer.

Instead, psychological research indicates that perceived harms and benefits exert their influence via an intuitive mode of thinking (for a review, see Reyna and colleagues).[11] According to Fuzzy Trace Theory, decision makers encode mental representations of the gist of information (e.g., *this risk is high!*) in parallel with verbatim details (e.g., *the risk is 15%*). Gist captures the bottom-line meaning, beyond specific words, pictures or numbers.[12] The predominant gist can be identified with well-structured questionnaires and it has been found to predict healthcare decision making, e.g., antibiotics prescribing.[13] Across studies, the gist of risk often boils down psychologically to fuzzy ordinal distinctions, such as none/negligible, low, moderate, and high. To reach a decision, similarly gist-based values, a kind of personal aphorisms, are applied to the mental representations, such as *"better safe than sorry"* or *"it is bad to refer low-risk patients"* or *"missing a cancer is the worst possible thing"*. Therefore, in this study, we elicited GPs' assessments of risk using not numerical but gist-type response scale with categories ranging from negligible to high. We also elicited their values and attitudes with regards to a host of variables that we expected to predict referral responses.

2. METHODS

2.1 Approach, aim and hypotheses

We ran the study online and recruited GPs to respond to clinical vignettes, which had been used in a previous study that demonstrated impact of algorithmic risk advice on GPs' numerical risk assessments and referral responses.[14] Per vignette, GPs made categorical judgements of risk and referral likelihood both before and after seeing the output of an algorithm. Based on our previous study, we expected that both referral and risk responses would change significantly following algorithmic advice (*H*1) and that referral responses would become more appropriate with regards to the 3% NICE referral threshold[†] (*H*2).[15] After all the vignettes were completed, we elicited 1) perceptions of global harms and benefits of referral for patients, the health system/society, and GPs themselves/their clinic at different levels of cancer risk; 2) how GPs viewed missing a cancer compared to over-referring; 3) their acceptance of over-referring for detecting a single cancer; and 4) their attitudes towards clinical uncertainty. We aimed to explore how these global values and attitudes predicted GPs' earlier responses on the clinical vignettes. Specifically, we expected the following associations:

1) positive associations between the likelihood of referring the patients in the vignettes and
    a. perceived global benefits of referrals (*H*3),
    b. perceived severity of missing a cancer vis-à-vis over-referring (*H*4), and
    c. willingness to trade false-positive referrals for one cancer detection (*H*5); and
2) negative associations between the likelihood of referring the patients in the vignettes and
    a. perceived global harms of referrals (*H*6), and
    b. tolerance of clinical uncertainty (*H*7).

2.2 Sample Size Estimation and Participant Recruitment

We powered the study for a multilevel linear regression to test whether perceived global harms and benefits of referrals can predict referral of the vignettes. Using G*Power 3.1.9.7, we estimated that a minimum of 652 independent responses would be required to detect a small effect ($f^2 = 0.02$) using two-tailed tests with alpha of .05 and 95% power in a multiple linear regression. We adjusted this number by the Design Effect (DE) for data clustering.[16]

[†] The NICE referral threshold of 3% (0.03) is the Positive Predictive Value, i.e., the probability that a patient with a specific symptom or presentation has cancer. This means that, on average, (1/0.03=) 33 people need to be referred for one cancer to be diagnosed.

The DE was calculated using the formula of DE=1+(n-1)*ICC, where *n* is the cluster size, i.e., the number of vignettes to be completed by each GP (9 vignettes), and ICC is the intra-class correlation, which was estimated to be 0.116 based on the previous study that used the same dependent variable.[14] Hence, DE equals 1.928. To calculate the minimum number of GPs we would need, we multiplied the required number of independent responses by the DE and divided it by the cluster size (652*1.928/9=140). Therefore, we estimated that we would need to recruit at least 140 GPs.

Participation in this study was limited to fully qualified GPs and GP trainees at their final stage of specialty training (i.e., ST3 and above) currently practising in England. We recruited from our database of GPs who had participated in previous decision-making studies by our research group and had indicated an interest in taking part in similar, future studies. We invited 596 eligible GPs who had *not* participated in our previous study on risk algorithms in colorectal cancer referral.[14] The invitation e-mail briefly introduced the study, the benefits of participation (Amazon voucher of £30) and included a link to an expression-of-interest form, where GPs could sign up for the study by providing their NHS e-mail address. 255 GPs completed this expression-of-interest form (43% response rate). Data collection was undertaken between 02/12/2022 and 10/04/2023 (dates of first and last completion).

2.3 Materials

We slightly adapted 12 clinical vignettes that we had used in the previous study of referral decision making.[14] Each vignette described a hypothetical patient presenting to the GP with a combination of risk factors and symptoms that could indicate colorectal cancer. Three of the vignettes were used for practice purposes. These vignettes had risk scores of 1.04% (low), 6.33% (moderate), and 39.58% (high) – as calculated using a publicly available cancer risk calculator (https://www.qcancer.org). The remaining 9 vignettes were used for data collection. Three of these were of low risk (1% to 2%), three were of moderate risk (5% to 9%), and three were of high risk (21% to 40%).

All vignettes started with the patient demographics and risk factors presented in a list format: patient name and gender, age, body mass index, smoking status (never smoked/ex-smoker/ number of cigarettes per day), alcohol intake (units/week) and age of menopause for female patients under 60. Each vignette included symptoms and some non-clinical, filler information in a narrative format. The latter was intended to make the vignettes more realistic and engaging for the participants. Each vignette finished with a statement that there were no other symptoms and that examination findings were normal. The vignettes with their estimated risk and appropriate referral responses are presented in Supplement 1.

2.4 Design and Procedure

The study followed a pre-post design, with response timing (pre- vs. post-algorithm) as the within-participant factor. We used the Qualtrics XM platform (Qualtrics, Provo, UT) to create an online survey that the GPs could access via a link that they received after they signed up to the study. Initially, GPs read detailed information about the study and provided online consent. They then completed basic demographic information (gender, age, GP status [fully qualified vs. trainee] and year of qualification. Before seeing the vignettes, GPs were presented with information about the algorithm's derivation, validation, and accuracy (see Kostopoulou et al., 2022, p. 3, Box 1, for the exact wording).[14] Next, they were provided with the three practice vignettes in a random order to familiarise them with the interface and task. The procedure was identical for all vignettes.

After the practice session was over, GPs were presented with the 9 remaining vignettes in a random order. At the end of each vignette, they were asked to respond to two questions in the following order:

"How likely is it that you would refer this patient for specialist investigations within 2 weeks and/or refer on the 2WW suspected colorectal cancer pathway ***at this consultation****?"* (highly unlikely, unlikely, uncertain, likely, highly likely)

"In your clinical judgement, which of the following best describes the risk of colorectal cancer for this patient?" (negligible, low, medium, high)

Although we would expect risk assessment to precede a referral decision, we chose to ask about risk only after clinicians responded to the referral question. In this way, we felt that we would obtain more gist-like referral responses, not diluted by explicit elicitations of risk.

After responding to these two questions, the same vignette was presented again with the corresponding algorithmic score in frequency and percentage formats. GPs were reminded of their own responses and were invited to revise them if they wished or re-enter them (Figure 1).

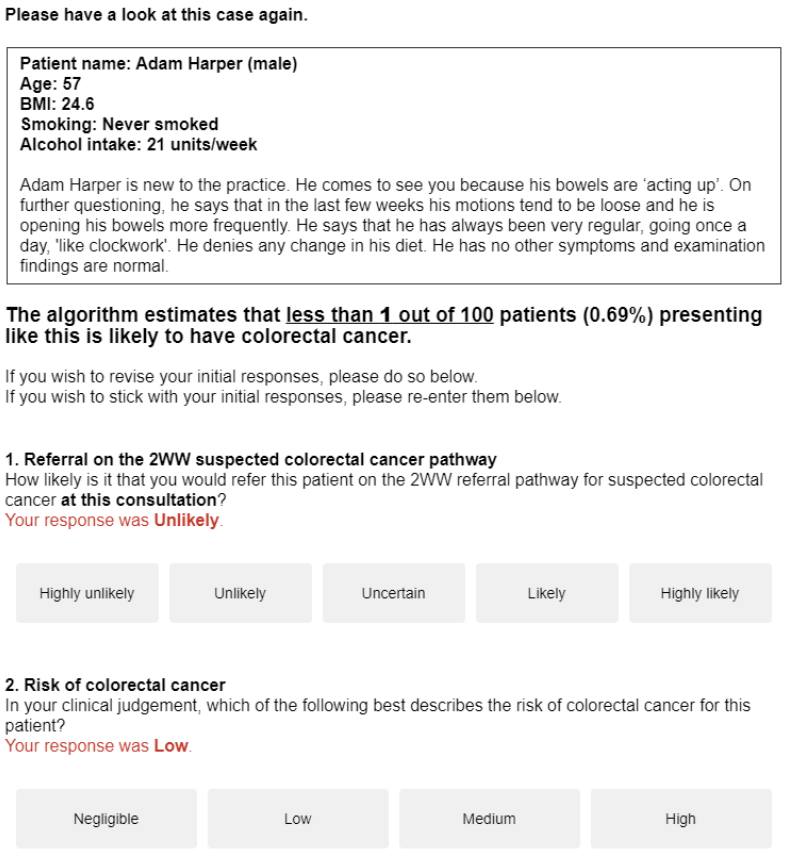


Figure 1. Screenshot of a vignette presented again to a respondent, this time with the algorithmic risk estimate and a reminder of the initial responses.

After all 9 vignettes were completed, we asked GPs a series of general questions not referring specifically to the vignettes seen earlier, and in the following sequence:

Step 1. First, we asked GPs to rate potential global harms and benefits of 2WW referrals at each of four gist levels of cancer risk corresponding to the scale used to assess risk in the vignettes: negligible, low, medium, and high. For each risk level, we asked them to consider three stakeholders: the patient, the NHS or the society, and the GP or the practice (representing patient-centred, public-centred, and self-centred values respectively). Thus, we elicited 12 ratings for potential benefits and 12 ratings for potential harms. Ratings were given on 4-point gist scales (negligible, low, medium, high – see Figure 2).

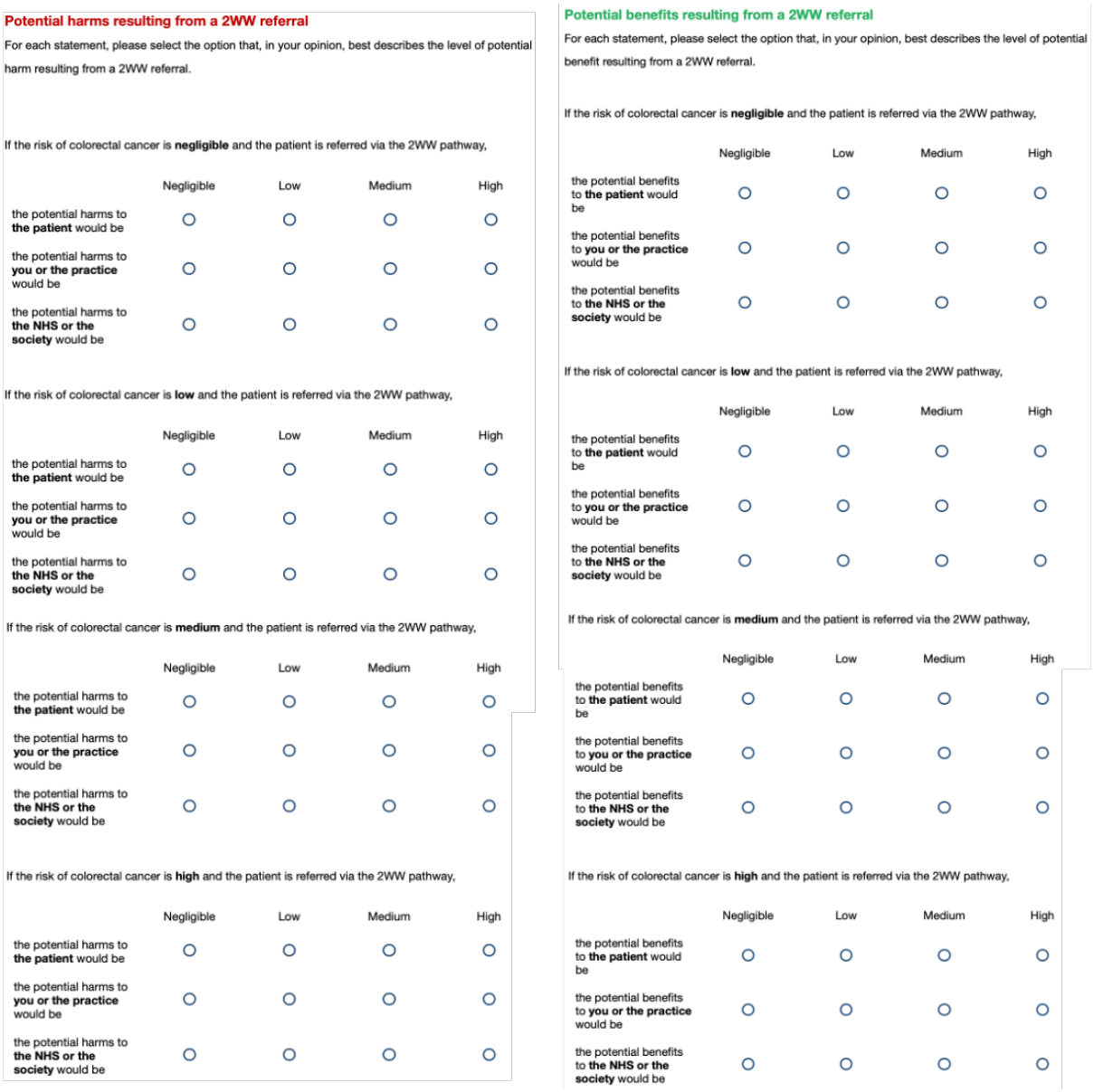


Figure 2. Rating scales for perceived harms (left-hand side) and benefits (right-hand side) for 4 gist levels of cancer risk and 3 stakeholders. The harm questions always preceded the benefits questions.

Step 2. Next, we measured perceived severity of two types of errors (i.e., how bad GPs thought they were) in relation to 2WW referrals: false alarms (*"referring a patient who should not have been referred"*) and misses *("not referring a patient who should have been referred"*). Responses were given on separate 0-100 scales (Figure 3). Note that we asked about "outcomes" rather than "errors", to avoid the negative loading of the term "error" and because referring a patient who should not be referred may not be considered an error. Reyna and Lloyd used a similar question about admission to hospital of patients with unstable angina but asked physicians to make a direct comparison between false alarms

and misses ("which of the two errors is worse?") on a 0-100 scale from "no difference at all" to "the maximum possible difference".[17]

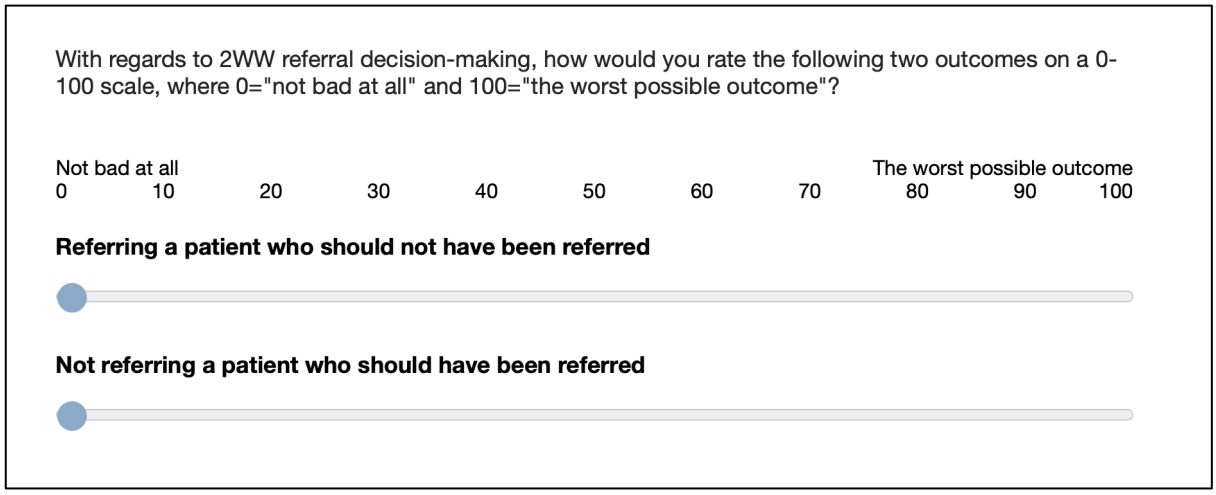


Figure 3. Rating scales for measuring perceived severity of false alarms and misses.

Step 3. Then, we measured willingness to trade false alarms for a hit by asking GPs to state how many 2WW referrals, where the patient turned out not to have cancer, they would deem acceptable for one cancer diagnosis to be made via this pathway (Figure 4).

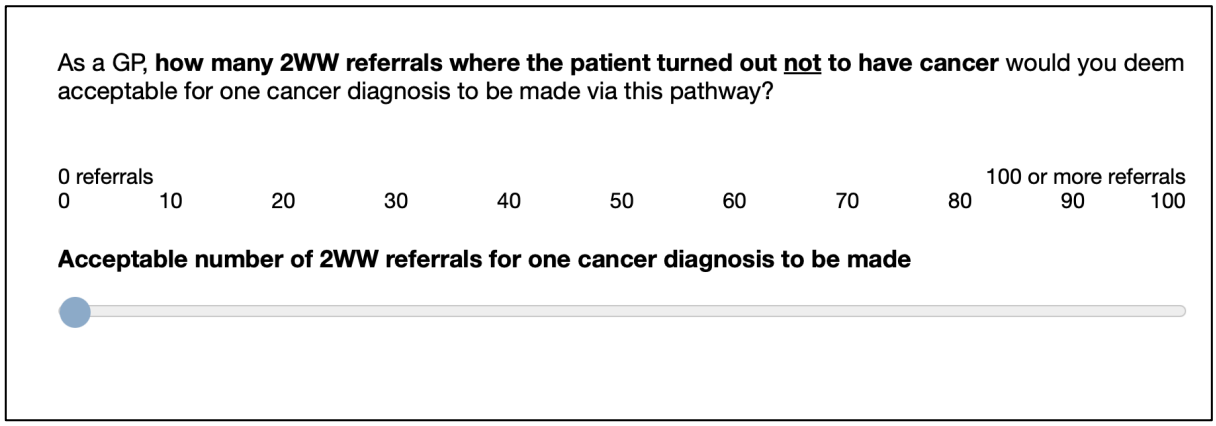


Figure 4. Rating scale for measuring willingness to trade false alarms for a hit.

Step 4. Finally, we measured attitudes toward risk and uncertainty in clinical practice using a validated questionnaire consisting of five items measured on 5-point Likert scales ranging from 1 [strongly disagree] to 5 [strongly agree].[18] Items concerned not taking any risks with physical complaints, seeking certainty about patients' diagnoses, referring to a specialist rather than wait and see, doing everything one can to establish the cause of a complaint,

and awareness that the complaint can be the beginning of a serious disease. At the end, participants had the opportunity to provide written feedback on any aspect of the study.

2.5 Statistical analyses

2.5.1 Creation of variables.

Referral responses on the nine vignettes were scored from -2 (highly unlikely) to +2 (highly likely). Judgements of risk were scored from 0 (negligible) to 3 (high). To test *H*1, we created two binary variables, one for referral responses and the other for risk judgements, which indicated whether GPs changed their responses post algorithm (0 to indicate no change and 1 to indicate change).

To measure appropriateness of referral responses (*H*2), we used the NICE risk threshold of 3%: in vignettes with QCancer risk of over 3%, we categorised “likely” and “highly likely” referral decisions as appropriate, and “unlikely” and “highly unlikely” ones as inappropriate; in vignettes with Qcancer risk of below 3%, we categorised “unlikely” and “highly unlikely” referral decisions as appropriate, and “likely” and “highly likely” ones as inappropriate.

To measure perceived error severity, we subtracted ratings of false alarms from those of potential misses. Scores could potentially vary from +100 to -100. However, we expected misses to be considered more severe than false alarms and hence most scores to be positive. For the willingness-to-trade question, we used the raw 0-100 values, indicating the number of false alarms GPs would accept for one cancer to be detected. To create a single attitude-to-uncertainty score for each participant, we averaged responses across the 5 questions of the Grol et al. questionnaire[18] and reversed the final score so that high values denoted willingness to tolerate uncertainty.

2.5.2 Regression analyses. All regression analyses were multilevel with random intercepts by GP and either vignette or risk level, depending on the model. To test the regression slopes, we used significance testing with the traditional *p*-value threshold of 0.05, and report regression coefficients and 95% CIs. The reported regression models of referral responses and risk judgements are all linear, but, since the response variable is ordinal, we also ran ordinal models to assess the robustness of our findings. Notably, results of the ordinal models are in harmony with those reported here (see Supplement 5). We conducted the analyses in R (version 4.3.1) and confirmed them in STATA 17.0. For the non-significant statistical tests, we used the Bayes factor (BF)[19,20] to distinguish between data insensitivity and evidence for the alternative vs. null hypotheses (see Supplement 2 for more details).

3. RESULTS

3.1 Study Sample

We recruited 200 GPs (193 fully qualified and 7 trainees). Their mean age was 42 years (*SD* 8.5) and 52% were female (103/200). The sample's average experience was 12 years in general practice post-qualification (*SD* 8.6, *Median* 9, range from 0 to 39). Five GPs reported one- or two-digit numbers rather than their year of qualification, so they were not included in the count or in any analyses involving the experience variable.

3.2 Algorithm impact on risk and referral responses

Both referral and risk responses changed significantly post-algorithm (referral change vs. no change OR 0.25 [0.16, 0.40] $p<0.001$; risk change vs. no change OR 0.49 [0.35, 0.59] $p<0.001$, see Table 1), thus confirming *H*1. When we regressed referral responses on Timing (pre- vs. post-algorithm), we found that the likelihood of referral reduced post-algorithm (b=-0.06 [-0.10, -0.007], p=0.025). Similarly, when we regressed risk judgements on Timing, we found that they too significantly reduced post-algorithm (b=-0.14 [-0.17, -0.11], $p<0.001$).

Examining response changes in more detail, we found that referral responses remained unchanged 75% of the time (1356/1800), but when they changed, they moved more often toward no referral (268 times) than toward referral (176 times). Similarly, risk judgements remained unchanged 65% of the time (1176/1800), but when they changed, they moved more toward lower risk (429 times) than higher risk (195 times). Although GPs were more conservative with their referrals and did not change them as frequently as risk judgements, risk and referral responses were tightly linked, such that when risk level increased by a unit, referral was more likely by almost a unit: b=0.80 [0.76, 0.84], $p<0.001$.

These results are comparable to those by Kostopoulou et al. (2022), where GPs saw 20 vignettes depicting patients with possible colorectal cancer and responded on the same 5-point referral scale, after they had provided a numeric risk estimate on a 0-100 Visual Analogue Scale (VAS). This suggests that differences in the risk response scales (5-point response scale in this study vs. 0-100 VAS in the previous study) and in the design (referral responses requested before vs. after risk assessment) did not alter the direction of the results.

Table 1. *Frequencies and within column relative frequencies (%) of Referral and Risk response categories pre- and post-algorithm*

Outcome variable	Response	Pre-algorithm	Post-algorithm	Total
Referral	Highly unlikely (-2)	63 (3.5%)	88 (4.9%)	151 (4.2%)
	Unlikely (-1)	278 (15.4%)	320 (17.8%)	598 (16.6%)
	Uncertain (0)	291 (16.2%)	236 (13.1%)	527 (14.6%)
	Likely (1)	431 (23.9%)	416 (23.1%)	847 (23.5%)
	Highly likely (2)	737 (40.9%)	740 (41.1%)	1477 (41.0%)
Risk	Negligible (0)	35 (1.94%)	79 (4.39%)	114 (3.2%)
	Low (1)	479 (26.61%)	575 (31.94%)	1054 (29.3%)
	Medium (2)	667 (37.06%)	596 (33.11%)	1263 (35.1%)
	High (3)	619 (34.39%)	550 (30.56%)	1169 (32.5%)
	TOTAL	1800	1800	3600 (100.0%)

3.3 Algorithm impact on referral appropriateness

When we excluded "uncertain" responses from the analysis, we found that 84.6% (1205/1425) of responses pre-algorithm and 86.5% (1232/1425) post-algorithm were appropriate. This increase was not significant in a multilevel logistic regression (*OR* 1.28 [0.99, 1.65], p=0.058). However, the Bayesian analysis revealed good enough evidence for the alternative hypothesis, $H2$ ($BF_{H(0,\ 1.26)}$=3.62, $RR_{BF>3}$ [1.13, 1.49]). When we included "uncertain" responses as inappropriate in the model, we detected a significant improvement from 68.7% (1236/1800) pre-algorithm to 74.7% (1345/1800, *OR* 1.52 [1.28, 1.81], p<0.001) post-algorithm. That is, the odds of a NICE-concordant referral response increased by about 50% post-algorithm, mainly due to reduced uncertainty and movement towards not referring.

3.4 Perceived harms and benefits as a function of risk level

To examine whether GPs' perception of potential harms and benefits of referrals were sensitive to the category of stipulated cancer risk (i.e., the risk level of each vignette by design), we ran 6 multilevel linear regression models with random intercept by GP, where we regressed the harms and the benefits for each stakeholder (patient, GP/practice, NHS/society) separately on the stipulated risk level. As expected, perceived harms reduced as cancer risk increased (harms to patients: b=-0.49 [-0.52, -0.45] p<0.001; harms to GP/practice: b=-0.28 [-0.31, -0.25] p<0.001; harms to NHS/society: b=-0.53 [-0.57, -0.50] p<0.001), while perceived benefits increased as cancer risk increased (benefits to patients: b=0.76 [0.73, 0.80] p<0.001; benefits to GP/practice: b=0.75 [0.71, 0.78] p<0.001; benefits to NHS/society: b=0.80 [0.77, 0.85] p<0.001) (Figure 5).

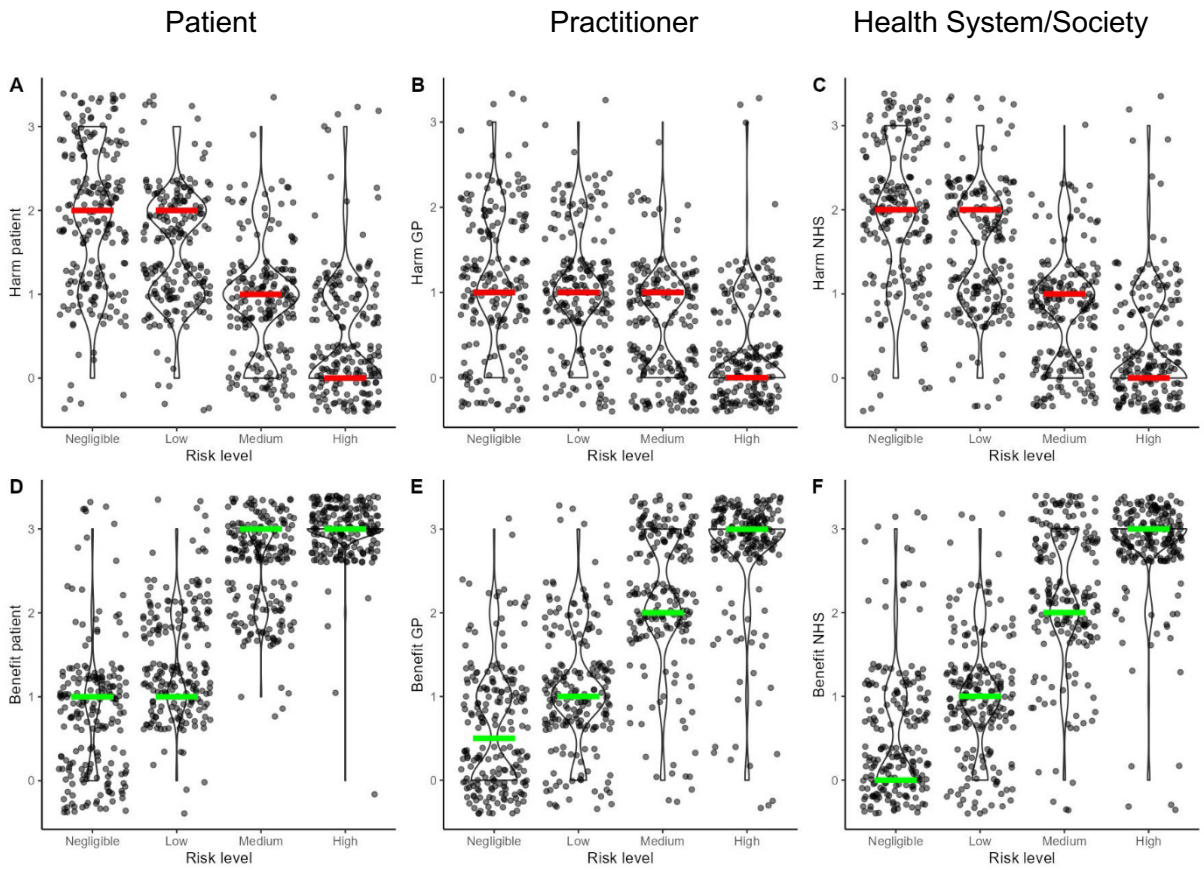


Figure 5. Violin and scatterplots depicting the relationship between the stipulated level of colorectal cancer risk and the perceived potential harms and benefits of 2WW referrals to the patient (Panels A and D), the GP or the practice (Panels B and E), and NHS or the society (Panels C and F). The coloured lines highlight the medians per level of risk (red lines for harms and green lines for benefits).

3.5 Perceived harms and benefits as predictors of referral responses

Next, we tested whether perceived global harms and benefits predicted referral responses on the vignettes, our main research question. Since the harms and benefits questions were general and not about a specific vignette, we linked the harm and benefit ratings to referral responses via the GPs' judged risk for each vignette. For example, if a GP indicated that the risk of colorectal cancer for a vignette was medium, then we matched their referral response for that vignette with their ratings of global harms and benefits for the medium level of risk. In a multilevel regression model with random intercepts for GP and vignette, we regressed all referral responses on the perceived benefits and harms for each stakeholder and included response Timing (pre- vs. post-algorithm) as a factor (Table 2). We assessed multicollinearity using the mean variance inflation factor (VIF). This was 3.48, i.e., lower than 5, and no individual VIF was larger than 5, which suggested that collinearity was not a problem. Perceived global benefits and harms for the patient and the NHS or the society, respectively, predicted referral likelihood in the expected direction, while we found no

evidence for perceived global benefits ($BF_{H(0, 0.15)}$=0.52, $RR_{3>BF>1/3}$ [0, 0.24]) and global harms for the GP/practice ($BF_{H(0, 0.15)}$=0.52, $RR_{3>BF>1/3}$ [0, 0.28]). Thus, *H*3 and *H*6 were supported for the patient and the health system/society but not for the GP. Perceived global patient benefits were the strongest predictor of referral likelihood. When we included in the model the judged level of risk associated with each referral response (i.e., the risk level that the GP assigned to a vignette after providing a referral response), global patient benefits, NHS benefits and NHS harms remained significant, though their predictive power weakened: *b*=0.15 [0.09, 0.21] *p*<0.001 for patient benefits; *b*=0.09 [0.02, 0.15] *p*=0.007 for NHS benefits; and *b*=-0.09 [-0.15, -0.02] *p*=0.007 for NHS harms. Patient harms did not pass the conventional threshold of 0.05 level of significance in that model ($BF_{H(0, 0.15)}$=0.36, $RR_{3>BF>1/3}$ [0, 0.16]). When we included interactions with Timing in the model, we found no evidence that the algorithm moderated the impact of perceived global benefits and harms on vignette referrals – even though regressing separately pre-algorithm referrals and post-algorithm referrals on global benefits and harms showed small drops in the regression coefficients post-algorithm (see Supplement 3).

Table 2. *Regression coefficients, 95% confidence intervals, and p values of a multiple regression model predicting referral responses.*

Stakeholder	Predictor	Coefficient	95% CI	p
Patient	Benefits	0.30	0.23, 0.36	<0.001
	Harms	-0.09	-0.15, -0.02	0.008
NHS/Society	Benefits	0.18	0.11, 0.24	<0.001
	Harms	-0.14	-0.21, -0.08	<0.001
GP/Practice	Benefits	0.03	-0.03, 0.09	0.358
	Harms	-0.03	-0.10, 0.04	0.413
	Timing (post-algorithm)	0.02	-0.02, 0.07	0.295

Some GPs left written comments at the end of the survey. Some comments in relation to the harm/benefit questions indicated differences in how GPs conceptualised these and some difficulties thinking about referrals in those terms (see Supplement 4). One GP did not provide any ratings of harms for any of the three stakeholders. There were also missing values for certain levels of risk for GP/practice harms (two GPs), NHS/society harms (one GP), and GP/practice benefits (one GP), suggesting that some respondents had difficulty thinking about referral benefits for GPs and referral harms in general. There were no missing values for patient benefits and harms nor for NHS/society benefits.

3.6 Perceived error severity, trade-offs, and attitudes to uncertainty as predictors of referral responses

The difference in the perceived severity between potential misses and false alarms (measured at Step 2) followed a highly negatively skewed distribution (*Median* 65, *Mean* 61.5, *SD* 24.4, range -19 to 100 – Figure 6A), suggesting that, as expected, respondents perceived the possibility of missing a cancer as more serious than over-referring. Three respondents produced negative values (over-referring more serious than misses), and three others produced 0 values (no difference), which could be the result of misinterpretation or inattentiveness. No responses were dropped from the analyses, however.

The trade-off variable (measured at Step 3) was positively skewed (*Median* 33, *Mean* 43.2, *SD* 29.7) but with a small peak at very high values (Figure 6B). That is, on average, GPs would be willing to refer 43 patients who turned out not to have cancer in order to detect one cancer but a substantial minority (17%, 34/200) indicated that they would accept 90 or more referrals for a single cancer diagnosis. This pattern of a peak close to the maximum of the response scale, with additional peaks at 50 and 20, suggests underlying categorical or ordinal gist representations rather than smooth continuous quantities.

Attitudes to clinical uncertainty and risk (measured at Step 4) followed a normal distribution, with GPs being slightly risk-avoiding on average (*Mean* 2.7, *SD* 0.6 on the 1-5 scale – Figure 6C).

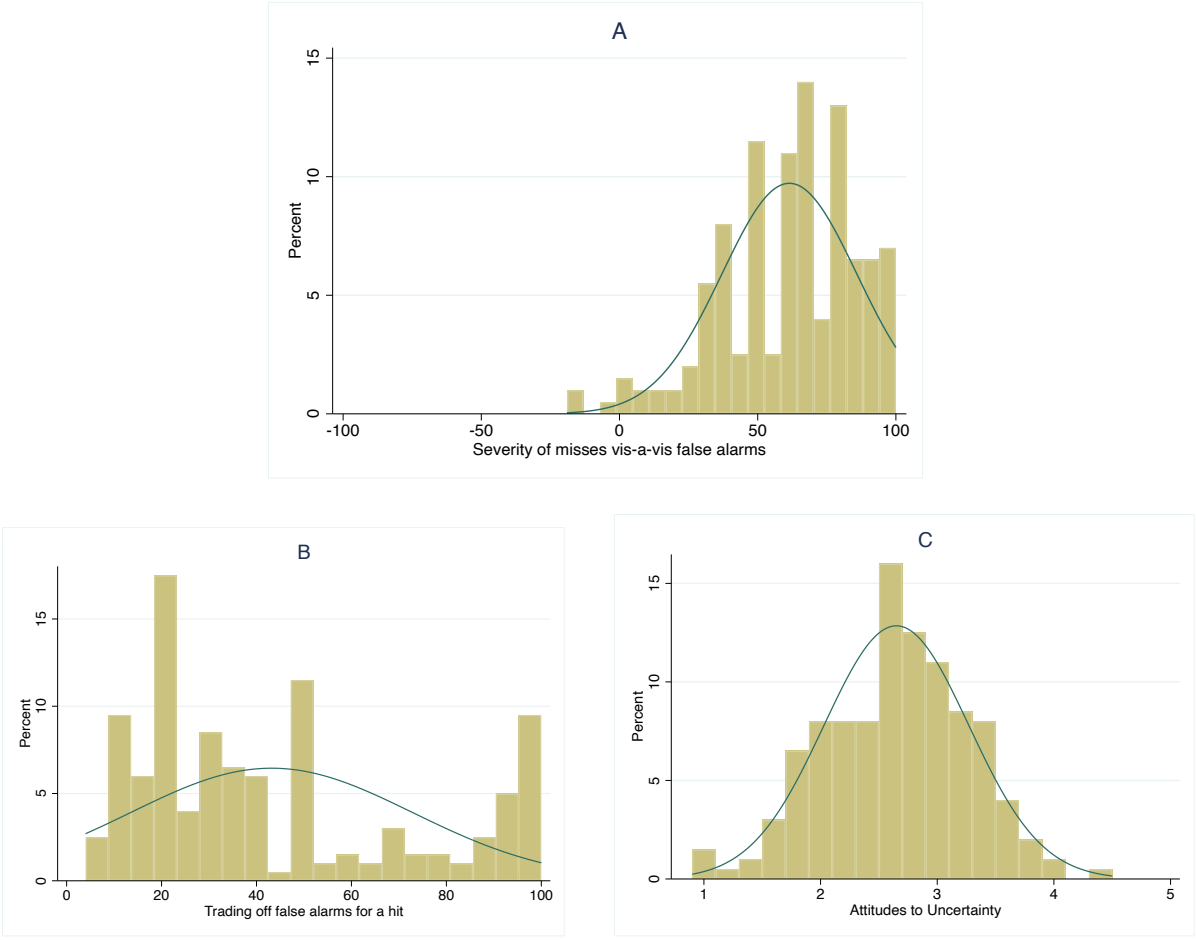


Figure 6. Histograms of the distribution for the variables of perceived error severity (A), trade-offs (B), and attitudes to uncertainty (C) with normal-density plot lines.

Difference in perceived severity of errors (misses vs. false alarms) positively correlated with trade-offs (r=0.26, p<0.001). There was also a significant correlation between trade-offs and uncertainty attitudes (r=0.092, p<0.001). When we included each of these related measures in separate regressions, only the difference in perceived error severity was a significant predictor (b=0.004 [0.001, 0.007] p=0.009), such that the greater seriousness GPs attached to missing a cancer compared to over-referring, the more likely they were to refer. Thus, we found support for *H*4.[†]

When we added perceived error severity to the large model shown on Table 2 and included the judged level of risk (risk gist), perceived error severity remained a statistically significant predictor of referrals (Table 3). Results remained unchanged, with no further significant relationships detected, when GP gender and years of experience were also added to the model.

[†] The Bayesian analysis showed data insensitivity for the tests of trade-offs ($BF_{H(0, 0.004)}$=0.56, $RR_{3>BF>1/3}$ [0, 0.007]) (*H*5) and attitudes to uncertainty ($BF_{H(0, 0.10)}$=0.79, $RR_{3>BF>1/3}$ [0, 0.11]) (*H*7).

Table 3. *Regression coefficients, 95% confidence intervals, and p values of a multiple regression model predicting referral responses.*

Stakeholder	Predictor	Coefficient	95% CI	p
Patient	Benefits	0.15	0.09, 0.21	<0.001
	Harms	-0.02	-0.08, 0.04	0.581
NHS/Society	Benefits	0.08	0.02, 0.15	0.008
	Harms	-0.09	-0.15, -0.02	0.008
GP/Practice	Benefits	-0.04	-0.10, 0.02	0.213
	Harms	-0.02	-0.08, 0.05	0.663
	Severity of misses vis-à-vis FAs	0.003	0.0009, 0.006	0.007
	Risk gist	0.56	0.50, 0.63	<0.001
	Timing (post-algorithm)	0.05	0.01, 0.10	0.011

4. DISCUSSION

Our study sheds light on psychological factors that make clinicians (dis)inclined to refer patients with suspected cancer, namely, their intuitive judgements of risk and their core values about potential errors of referring or not referring patients. Perceived benefits and harms to the patient and the health system or society all predicted likelihood of referring, with benefits for the patient exerting the strongest influence. In contrast, benefits and harms to the clinician or their clinic did not appear to impact referral likelihood, suggesting that GPs' values were patient-centred rather than self-centred. Although GPs may be subject to pressures from colleagues or the health system to refer more or fewer patients, they did not appear to give such considerations much weight relative to the well-being of the patient. Our findings are consistent with a moral prioritisation of referral decision making, where perceived benefits to the patient have the strongest impact, followed by perceived benefits and harms to the health system and society, followed by little to no weight given to self. The size of the regression coefficients clearly illustrates this moral pyramid of duty: the doctor's duty to save lives and not to consider reputational or cost implications to themselves and their practice.

In healthcare systems with salient resource constraints (i.e., where the care of individual patients can directly or indirectly impact the resources available to other patients), physicians may take both an individual-patient and a population/public-health perspective and trade these off against one another.[21] A moral dilemma occurs when physicians face a conflict

between maximising collective (or personal) interests against the interests of their individual patients. Although physicians are no doubt aware of resource constraints, in the current study, the perceived benefits to the individual patient primarily guided referral decisions. Thus, decision making in this context appears more consistent with a deontological (core values) perspective, which eschews moral trade-offs, rather than a utilitarian perspective, which focuses on maximising the greatest good for the greatest number.

When dealing with patients in everyday clinical practice, benefit and harm considerations could play a different role from those we observed, and considerations of oneself and the practice may indeed influence referrals. Anecdotally, UK GPs acknowledge conflicting pressures from health authorities to refer more patients and communications from hospitals suggesting that they refer inappropriately (i.e., false alarms). In addition, metrics of a practice's 2WW referral performance used to be in the public domain, so that practices could compare themselves with the best performers, receiving a kind of norm-nudging.[22–24] It is possible that such pressures may influence response bias of individual GPs, but our study did not detect this, as other variables exerted a larger influence. Hypothetical judgements might also have precluded or minimised consideration of external pressures. Had we made explicit those pressures before GPs responded to the vignettes (e.g., "your practice's referral rate is below average"), we might have observed an impact not only on referral responses but also on self-interest as a predictor of those responses.

Even when we included in the model the judged level of risk, a variable closely and directly associated with referral responses, physicians' values regarding patient benefits and NHS/societal benefits and harms continued to predict referral responses. The algorithm consistently lowered perceived risk and referral likelihood without overriding the impact of perceived benefits and harms. A trend to attenuate them was nevertheless observed, which could reflect the algorithm's pressure to unmoor physicians from their perceptions and core values. Had the algorithm also provided an explicit recommendation, this could have increased the pressure on physicians to conform to algorithmic advice despite possible misgivings.

We also found that the more physicians considered missing a cancer as a more serious error than referring a patient who turns out not to have cancer, the more likely they were to refer. This relationship was however not confirmed by the number of false alarms they were willing to accept for one cancer to be diagnosed. This attests to the difference the simple wording of questions can make. Fuzzy Trace Theory suggests that these exact numerical responses are incompatible with preferred modes of thinking even in numerate populations such as physicians.[11]

Decision thresholds have also been linked to emotions such as anticipated regret and anticipatory worry.[25] A future study could attempt to measure these emotions directly and assess their impact on referral decisions.

Beyond mandates and incentives, which are known to influence response bias, sometimes coercively, attempts to influence decision making could also target decision makers' cost/benefit perceptions.[26] We did not try to identify precisely which harms and benefits respondents had in mind when answering the questions and it is likely that these differed among respondents, as suggested by some comments. For example, one clinician may value the peace of mind that patients may gain from a referral that is unlikely to find cancer, while another may wish to save these patients from the pain and stress of unnecessary investigations (but see Nurek and Kostopoulou about the value patients place on invasive tests that can help to exclude a rare but serious disease).[27] Similarly, one GP may consider that a high referral rate could save the health system money in the long run by identifying and treating cancers early, while another may focus on the potential harm to other patients from increasing waiting lists. We could, of course, have given our respondents examples of harms and benefits but did not want to pre-empt what they considered important or suggest to them issues that they had not considered before. Our findings could be used to validate studies based on self-reports of harms and benefits, such as interviews. For example, we would expect that all GPs would mention benefits to patients, while very few, if any, would voice considerations about themselves or their practice.

5. CONCLUSION

Algorithmic risk estimates appear to influence physicians' perceived risks and referrals for cancer patients. However, they do not supplant core values about the benefits and harms from a referral and the moral seriousness of missing a cancer vis-à-vis over-referring. These values are thought to contribute to an internal threshold for action that prioritises benefits to the patient, while only secondarily considering benefits and harms to the health system or society and negligibly considering the self.

Statements and disclosures

Funding

The study was funded by a Cancer Research UK grant awarded to Olga Kostopoulou. Funding Scheme: Population Research Committee - Project Award, Reference A28634.

Valerie Reyna's contribution was supported by the National Institute of Standards and Technology (Grant 60NANB22D052) and the Institute for Trustworthy AI in Law and Society (supported by both National Science Foundation and National Institute of Standards and Technology Grant IIS-2229885).

Institutional Review Board Statement

Study approval was provided by the Research Governance and Integrity Team of Imperial College London (reference 23IC8172).

Informed Consent Statement

Participants provided informed consent online after reading a Participant Information Sheet. They were free to withdraw from the study at any point, by closing their browser, in which case their data would be deleted.

Data Availability Statement

The data are publicly available at https://osf.io/nydh2/.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions:

OK: conceptualisation, methodology (design and materials), formal analysis, validation, writing (original draft preparation, review and editing), project supervision, funding acquisition. **BP:** methodology (design and materials), software, project administration, formal analysis, validation, writing (original draft preparation, review and editing). **KA:** methodology (design and materials), piloting of materials, writing (review and editing). **VR:** conceptualisation, methodology, writing (review and editing). **OK** and **BP** contributed equally to the study.

REFERENCES

1. Crosby D, Lyons N, Greenwood E, et al. A roadmap for the early detection and diagnosis of cancer. *Lancet Oncol*. 2020;21(11):1397-1399. doi:10.1016/S1470-2045(20)30593-3

2. Round T, Gildea C, Ashworth M. Association between use of urgent suspected cancer referral and mortality and stage at diagnosis : a 5-year national cohort study. *Br J Gen Pr*. 2020;70(695):e389-e398. doi:10.3399/bjgp20X709433

3. Macmillan N, Creelman C. *Detection Theory: A User's Guide*. Lawrence Erlbaum Associates Inc.; 2005.

4. Stanislaw H, Todorov N. Calculation of signal detection theory measures. *Behav Res Methods Instrum Comput*. 1999;31(1):137-149.

5. Kostopoulou O, Nurek M, Cantarella S, Okoli G, Fiorentino F, Delaney BC. Referral Decision Making of General Practitioners: A Signal Detection Study. *Med Decis Mak*. 2019;39(1):21-31. doi:10.1177/0272989X18813357

6. Kostopoulou O, Nurek M, Delaney BC. Disentangling the Relationship between Physician and Organizational Performance: A Signal Detection Approach. *Med Decis Making*. 2020;40(6):746-755. doi:10.1177/0272989X20936212

7. Cheyne H, Dalgleish L, Tucker J, et al. Risk assessment and decision making about in-labour transfer from rural maternity care: a social judgment and signal detection analysis. *Bmc Med Inform Decis Mak*. 2012;12. doi:Artn 122 Doi 10.1186/1472-6947-12-122

8. Christensen-Szalanski JJ, Diehr PH, Bushyhead JB, Wood RW. Two studies of good clinical judgment. *Med Decis Making*. 1982;2(3):275.

9. Burton CD, McLernon DJ, Lee AJ, Murchie P. Distinguishing variation in referral accuracy from referral threshold: analysis of a national dataset of referrals for suspected cancer. *BMJ Open*. 2017;7(8):e016439. doi:10.1136/bmjopen-2017-016439

10. Swets JA, Dawes RM, Monahan J. Psychological science can improve diagnostic decisions. *Psychol Sci Public Interest*. 2000;1(1):1-26.

11. Reyna VF, Edelson S, Hayes B, Garavito D. Supporting Health and Medical Decision Making: Findings and Insights from Fuzzy-Trace Theory. *Med Decis Making*. 2022;42(6):741-754. doi:10.1177/0272989X221105473

12. Reyna VF. A new intuitionism: Meaning, memory, and development in Fuzzy-Trace Theory. *Judgm Decis Mak*. 2012;7(3):332-359. doi:10.1017/S1930297500002291

13. Klein EY, Martinez EM, May L, Saheed M, Reyna V, Broniatowski DA. Categorical Risk Perception Drives Variability in Antibiotic Prescribing in the Emergency Department: A Mixed Methods Observational Study. *J Gen Intern Med*. Published online 2017. doi:10.1007/s11606-017-4099-6

14. Kostopoulou O, Kavleen A, Palfi B. Using cancer risk algorithms to improve risk estimates and referral decisions. *Commun Med*. 2022;2(2). doi:10.1038/s43856-021-00069-1

15. National Institute for Health and Care Excellence (NICE). *Suspected Cancer: Recognition and Referral. Final scope.* Published 16 July 2024. https://alpha.nice.org.uk/guidance/GID-HTE10050/documents/html-to-pdf-7

16. Barratt H, Kirwan M. Clustered data - effects on sample size and approaches to analysis. 2009. http://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/clustered-data

17. Reyna VF, Lloyd FJ. Physician Decision Making and Cardiac Risk: Effects of Knowledge, Risk Perception, Risk Tolerance, and Fuzzy Processing. *J Exp Psychol Appl*. 2006;12(3):175-179.

18. Grol R, Whitfield M, De Maeseneer J, Mokkink H. Attitudes to risk taking in medical decision making among British, Dutch and Belgian general practitioners. *Br J Gen Pract J R Coll Gen Pract*. 1990;40(333):134-136.

19. Dienes Z. Using Bayes to get the most out of non-significant results. *Front Psychol*. 2014;5:781.

20. Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon Bull Rev*. 2009;16:225-237.

21. Dawes RM. Social Dilemmas. *Annu Rev Psychol*. 1980;31(1):169-193. doi:10.1146/annurev.ps.31.020180.001125

22. Bicchieri C, Dimant E. Nudging with care: the risks and benefits of social information. *Public Choice*. 2022;191(3-4):443-464. doi:10.1007/s11127-019-00684-6

23. Meeker D, Linder JA, Fox CR, et al. Effect of Behavioral Interventions on Inappropriate Antibiotic Prescribing Among Primary Care Practices. *JAMA*. 2016;315(6):562. doi:10.1001/jama.2016.0275

24. Hallsworth M, Chadborn T, Sallis A, et al. Provision of social norm feedback to high prescribers of antibiotics in general practice: a pragmatic national randomised controlled trial. *The Lancet*. 2016;387(10029):1743-1752. doi:10.1016/S0140-6736(16)00215-4

25. Robinson PJ, Wouter Botzen WJ. The impact of regret and worry on the threshold level of concern for flood insurance demand: Evidence from Dutch homeowners. *Judgm Decis Mak*. 2018;13(2):237-245.

26. Kostopoulou O. Measuring gist-based perceptions of medication benefit-to-harm ratios. *BMJ Qual Saf*. Published online July 31, 2024:bmjqs-2024-017375. doi:10.1136/bmjqs-2024-017375

27. Nurek M, Kostopoulou O. How the UK public views the use of diagnostic decision aids by physicians: a vignette-based experiment. *J Am Med Inform Assoc*. 2023;30(5):888-898. doi:10.1093/jamia/ocad019

Supplementary Materials

⇒ Supplement 1: The vignettes

PRACTICE VIGNETTES

NAME: Bryony Barnes (female)

Age: 56

BMI: 30

Smoking: Currently smokes 15 cigarettes/day

Alcohol intake: 21 units/week

Age of menopause: 51

Bryony Barnes comes to see you complaining of being more constipated in the last month. Over the last 2 months, she has also noted that she has lost about 4kg in weight and doesn't understand why. She has not been dieting and her lifestyle has not changed.

QCANCER RISK: 1.04%

Appropriate response: No referral – risk <3%

NAME: Henry Lipp (male)

Age: 75

BMI: 24.9

Smoking: Never smoked

Alcohol intake: 4 units/week

Mr Henry Lipp has come to see you concerned because he noticed some blood in his stools over the last four weeks. He has no other symptoms.

QCANCER RISK: 6.33%

Appropriate response: Referral – risk >3%

NAME: Dawn Jenkins (female)

Age: 70

BMI: 26.2

Smoking: Ex-smoker

Alcohol intake: 14 units/week

Dawn Jenkins is your next patient. She has a background of Type 2 Diabetes and no other medical problems. She has come to see you because in the last few weeks she has become increasingly aware of some abdominal pain. When you ask her about her bowels, she says she has seen some blood in her stools on and off and this has been the case for the last two weeks. She has no other symptoms. You ask her to have some blood tests done which reveal a microcytic anaemia (Hb 9.8) and a low ferritin.

QCANCER RISK: 39.58%

Appropriate response: Referral – risk >3%

MAIN STUDY VIGNETTES

NAME: Adam Harper (male)

Age: 57

BMI: 24.6

Smoking: Never smoked

Alcohol intake: 21 units/week

Adam Harper is new to the practice. He comes to see you because his bowels are ‘acting up’. On further questioning, he says that in the last few weeks his motions tend to be loose and he is opening his bowels more frequently. He says that he has always been very regular, going once a day, 'like clockwork'. He denies any change in his diet. He has no other symptoms and examination findings are normal.

QCANCER RISK: 0.69%

Appropriate response: No referral – risk <3%

NAME: Matt Crayton (male)

Age: 75

BMI: 24.4

Smoking: Never smoked

Alcohol intake: 4 units/week

Matt Crayton comes in to see you with his wife. He is seeking your advice because he has lost some weight recently. His wife intervenes and says that she and his friends have noticed this over the last 3 months and told him to see the GP. Matt says that he has not changed his diet. He has no other symptoms and examination findings are normal.

QCANCER RISK: 1.08%

Appropriate response: No referral – risk <3%

NAME: Nina Durbridge (female)

Age: 54

BMI: 27.2

Smoking: Ex-smoker

Alcohol intake: 21 units/week

Age of menopause: 51

Nina Durbridge, your next patient, works in marketing. She saw another doctor in your surgery last week and had requested to have some routine blood tests. She was called to come in for the results and has made an appointment for today. The results show microcytic anaemia (Hb 10.8) with a low ferritin. Upon enquiring about any symptoms that she may have, she tells you that she has had abdominal pain for about a month, which she has not had before. She has no other symptoms and examination findings are normal.

QCANCER RISK: 2.09%

Appropriate response: No referral – risk <3%

NAME: Antonio DiMarco (male)

Age: 78

BMI: 24.2

Smoking: Currently smokes 12 cigarettes/day

Alcohol intake: Nil

Antonio DiMarco comes to see you because he's lost some weight recently without dieting. His wife remarked upon it. Antonio also mentions that he's been passing stool more frequently than usual, which worries him a little because his father died of gastrointestinal cancer. He has no other symptoms and examination findings are normal.

QCANCER RISK: 5.16%

Appropriate response: Referral – risk >3%

NAME: Debbie Lawrence (female)

Age: 58

BMI: 21.8

Smoking: Currently smokes 3 cigarettes/day

Alcohol intake: 3 units/week

Age of menopause: 52

Debbie Lawrence comes in accompanied by her husband. She complains of abdominal pain which she has had for more than a month. She says that she cannot understand what might be causing it and that it is not getting better. She has also noticed that she is passing stool more frequently than usual (2 or 3 times a day) in the last few weeks. You order a blood test, which comes back showing microcytic anaemia (Hb 10.3) with low ferritin. She has no other symptoms and examination findings are normal.

QCANCER RISK: 4.70%

Appropriate response: Referral – risk >3%

NAME: Doris Newman

Age: 75

BMI: 29.4

Smoking: non-smoker

Alcohol intake: nil

Doris Newman has come to see you concerned about some abdominal pain that she has experienced in the last few weeks. It seems to be there most of the time. She says that she loves her food but has lost her appetite lately. She likes to check her weight regularly and has also noticed that it has dropped from 55kg to 52kg in the last month. On further questioning, she reveals that her bowels are opening less regularly in the last few weeks, and she finds it more difficult to pass stool. She has no other symptoms and examination findings are normal.

QCANCER RISK: 8.78%

Appropriate response: Referral – risk >3%

NAME: Norman England (male)

Age: 70

BMI: 26.7

Smoking: Current smoker, 5 cigarettes/day

Alcohol intake: 14 units/week

Norman England has come to see you because he has noticed that his stools have become loose and he is opening his bowels more frequently over the last four weeks. He usually loves his wife's cooking but doesn't feel like eating anymore. He has lost about 4kg of weight in the last 2 months. He has also noticed that he is getting some abdominal pain. He tells you that his father had bowel cancer when he was of a similar age. He has no other symptoms and examination findings are normal.

QCANCER RISK: 22.82%

Appropriate response: Referral – risk >3%

NAME: Jane Tarley (female)

Age: 75

BMI: 24.9

Smoking: Currently smokes 20 cigarettes/day

Alcohol intake: Nil

Jane Tarley is your next patient. She has a background of COPD for which she takes inhalers. She has become aware of some abdominal pain in the last month. She tells you that she is off her food, and thinks that she has lost weight. You weigh her and note that she has lost about 3kg in the last 2 months. Jane also tells you that she has had some blood in her stool most days in the last 2 weeks. She has not had anything like this before. She tells you that her brother was recently diagnosed with bowel cancer. She has no other symptoms and examination findings are normal.

QCANCER RISK: 40.14%

Appropriate response: Referral – risk >3%

NAME: Olivia Fielding (female)

Age: 88

BMI: 23.1

Smoking: ex-smoker

Alcohol intake: nil

Olivia Fielding comes in to see you today. She is usually well and has no significant past medical history. She tells you that her family have commented that she appears to have lost quite a bit of weight over the past 6 months. She has noticed that her clothes feel looser. On further questioning, you discover that she has had abdominal pain for most days in the last 3 months. Her stools seem to be 'more runny' in the last few weeks. You organise some blood tests, which reveal microcytic anaemia (Hb 10.1) and low ferritin. She has no other symptoms and examination findings are normal.

QCANCER RISK: 20.76%

Appropriate response: Referral – risk >3%

Supplement 2: Bayesian analyses

For the non-significant statistical tests, we used the Bayes factor (BF) (Dienes, 2014; Rouder et al. 2009) to distinguish between data insensitivity and evidence for the null or the alternative hypothesis. We applied the Dienes and Mclatchie Bayes factor calculator (2018) adopted to the R environment. We report BFs in the following format: $BF_{H(0, SD)}$, where H indicates that we modelled the predictions of the alternative hypotheses with half-normal distributions. The values within the parentheses indicate the parameters of the half-normal priors: 0 for the mode and SD for the SD of the distribution (for more information on how we defined the SDs for each hypothesis, see below). While the BF is a continuous measure of evidence, we used it for hypothesis testing by applying the conventional threshold of 3 for substantial evidence for the alternative hypothesis, and the threshold of 1/3 for substantial evidence for the null hypothesis. We interpreted BFs between 3 and 1/3 to indicate data insensitivity (Jeffreys, 1961). To ascertain the robustness of our conclusions to the parameters of our models, we report Robustness Regions (RRs) for each BF, where we indicate the range of SDs for which we would have arrived to the same qualitative conclusion (Dienes, 2019).

The predictions of all alternative hypotheses (also known as prior distributions) were modelled with half-normal distributions with a mode of zero; hence, we only needed to identify the SD of these distributions for the various hypotheses. Notably, the SD of such a distribution represents the effect size one would expect if the alternative hypothesis were true (Dienes, 2019). We used two heuristics to identify the expected effect size for each hypothesis. If there was a relevant effect in the literature, we used the effect size found in a previous study. If there was no relevant effect in the literature, we used the "room-to-move heuristic", which recommends using the half of the maximum possible effect size as the expected effect size (Dienes, 2019). We used the largest relevant effect size with a significant test for the maximum possible effect size. Here, we list the applied heuristic for each hypothesis:

H1: all tests were statistically significant, so we did not run a Bayesian analysis
H2: we used the relevant effect size from Kostopoulou et al. (2022), OR = 1.26, as the expected effect size.
H3 and H6: perceived global patient benefits were the strongest predictor of referral likelihood with a coefficient of $b = 0.30$. We used this value as the maximum possible effect size for the non-significant predictors; hence the expected effect size was 0.15 for all non-significant predictors.
H4: the test was statistically significant, so we did not run a Bayesian analysis.
H5 and H7: Since the test of H4 was significant, we used the effect size of $b = 0.004$ as a potential maximum effect for H5 and H7. First, we halved the effect size and then converted it to the measurement scales used for H5 and H7 by multiplying it by 2 and 50, respectively. The expected effect sizes for H5 and H7 were $b = 0.004$ and $b = 0.10$ respectively.

Dienes, Z. (2019). How Do I Know What My Theory Predicts? *Advances in Methods and Practices in Psychological Science*, *2*(4), 364–377. https://doi.org/10.1177/2515245919876960

Dienes Z. Using Bayes to get the most out of non-significant results. *Front Psychol*. 2014;5:781.

Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon Bull Rev*. 2009;16:225-237.

Jeffreys H. *The Theory of Probability*. Oxford University Press; 1961.

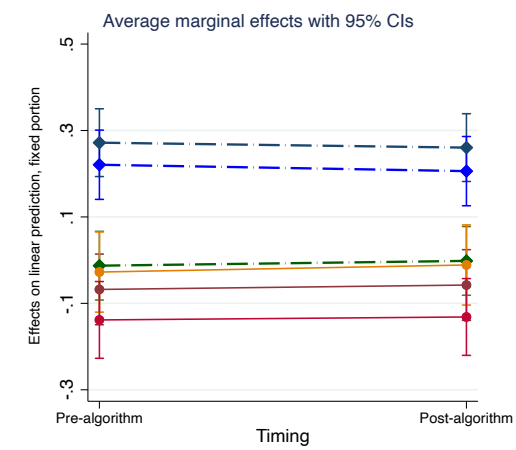
Kostopoulou, O., Kavleen, A., & Palfi, B. (2022). Using cancer risk algorithms to improve risk estimates and referral decisions. *Communications Medicine*, *2*(2). https://doi.org/10.1038/s43856-021-00069-1

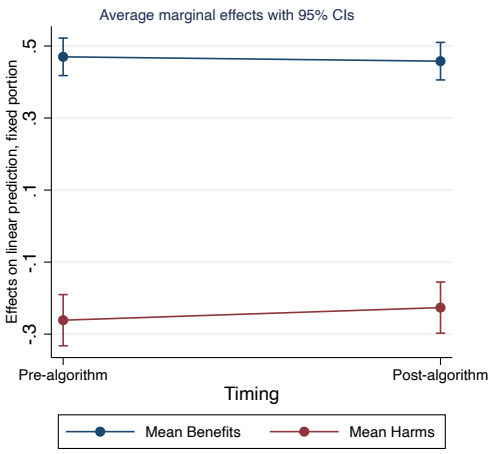
⇒ Supplement 3: Perceived harms and benefits as predictors of referral responses

Regression coefficients, 95% confidence intervals, and p values of multiple regression models predicting referral responses pre-algorithm and post-algorithm.

Stakeholder	Predictor	Pre-algorithm	Post-algorithm	Predictor x Timing (pre-post) interaction
Patient	Benefits	*b*=0.32 [0.23, 0.40] *p*<0.001	*b*=0.23 [0.14, 0.31] p<0.001	*b*=0.01 [-0.08, 0.10] p=0.752 $BF_{H(0, 0.15)}$ = 0.23
	Harms	*b*=-0.08 [-0.17, 0.006] *p*=0.067	*b*=-0.04 [-0.13, 0.05] *p*=0.351	*b*=0.02 [-0.07, 0.12] p=0.600 $BF_{H(0, 0.15)}$ = 0.47
NHS/Society	Benefits	*b*=0.23 [0.15, 0.32] *p*<0.001	*b*=0.17 [0.08, 0.26] *p*<0.001	*b*=-0.06 [-0.15, 0.03] p=0.163 $BF_{H(0, 0.15)}$ = 1.29
	Harms	*b*=-0.15 [-0.25, -0.06] *p*=0.002	*b*=-0.12 [-0.22, -0.03] *p*=0.012	*b*=-0.01 [-0.11, 0.10] p=0.928 $BF_{H(0, 0.15)}$ = 0.32
GP/Practice	Benefits	*b*=-0.0004 [-0.09, 0.09] *p*=0.993	*b*=-0.009 [-0.10, 0.08] *p*=0.845	*b*=0.04 [-0.04, 0.13] p=0.330 $BF_{H(0, 0.15)}$ = 0.15
	Harms	*b*=-0.04 [-0.14, 0.06] *p*=0.427	*b*=0.01 [-0.09, 0.11] *p*=0.809	*b*=-0.001 [-0.10, 0.10] p=0.991 $BF_{H(0, 0.15)}$ = 0.32

NB. The Bayesian analyses of the interactions revealed evidence for the null for the impact of the algorithm on patient benefits, GP benefits, GP harms, and NHS harms, while the rest of the analyses showed data insensitivity.





⇒ Supplement 4: GP comments

GP comments written at the end of the survey, which suggest that GPs interpreted the harm/benefit questions in different ways, and some had difficulty thinking about referrals in those terms.

"Grading at the end is difficult - is it a potential harm or benefit for a patient to be diagnosed with cancer?" (GP 177)

"I liked that you had a section on harms of overdiagnosis." (GP 22)

"Very interesting to think about the potential benefits to patient and society following referrals. The potential benefits are theoretically higher for the patient if their risk of cancer is low, and they are referred for those rarer diagnoses. Conversely the benefit to society for referral reaches a peak with higher risk cases." (pilot participant 2)

"I didn't really understand the second part of the survey where we had to say how negligible, low, medium and high-risk patients being a referred on a 2WW pathway would be harmful/beneficial to the NHS/practice." (GP 145)

"I found the benefit to GP/practice question difficult to answer. I wasn't sure what I was meant to be thinking about when answering this question." (pilot participant 1)

⇒ Supplement 5: Ordinal logistic regression analyses

Given that our main outcome variables, referral responses, risk judgements and perceived harms and benefits were measured on scales with limited width, where distances between the units are not necessarily equal, we repeated the regression analyses assuming that our outcome variables are ordinal, to assess the robustness of our conclusions. We ran cumulative link mixed models using the ordinal R package (Christensen, 2023), and in each regression model, we included the same random effects variables as we did in their linear counterparts. For the ordinal predictors of these models, we report the Odds ratios (ORs) of the linear trends of the polynomial contrasts. Note that these ORs cannot be interpreted in the usual way of linear regression (e.g., one unit change in the predictor leads to X change in the outcome measure). They indicate a trend (increasing or decreasing) across the ordered levels of the predictor. For example an OR>1 indicates the increase in the odds of being in a higher category of the outcome variable for each step along the linear trend of the predictor (e.g., from negligible to high in a steady increase).

Algorithm impact on risk and referral responses

In two separate regression models, we regressed referral responses and risk judgements on Timing (pre- vs. post-algorithm). In line with our main findings, we found that both the likelihood of referral and risk judgements reduced post-algorithm (referral *OR* 0.84 [0.73, 0.97] p=0.015, and risk judgement OR 0.53 [0.45, 0.61] p<0.001).

Perceived harms and benefits as a function of risk level

Regression coefficients of the linear trends reported as Odds Ratios (ORs), 95% Confidence Intervals (CIs), and p values of a multiple ordinal regression model predicting perceived harms and benefits as a function of risk level (risk judgement).

Outcome variable	OR	95% CI	p
Benefits to patients	4761	1620, 13995	<0.001
Harms to patients	0.013	0.008, 0.021	<0.001
Benefits to NHS/Society	978	453, 2107	<0.001
Harms to NHS/Society	0.009	0.005, 0.015	<0.001
Benefits to GP/Practice	966	438, 2134	<0.001
Harms to GP/Practice	0.033	0.020, 0.053	<0.001

Perceived harms and benefits as predictors of referral responses

Regression coefficients of the linear trends reported as Odds Ratios (ORs), 95% Confidence Intervals (CIs), and p values of a multiple ordinal regression model predicting referral responses.

Stakeholder	Predictor	OR	95% CI	p
Patient	Benefits	8.89	5.18, 15.24	<0.001
	Harms	0.47	0.28, 0.81	0.006
NHS/Society	Benefits	3.16	1.91, 5.25	<0.001
	Harms	0.21	0.11, 0.39	<0.001
GP/Practice	Benefits	1.39	0.81, 2.40	0.232
	Harms	1.65	0.68, 4.00	0.272
	Timing (post-algorithm)	1.11	0.96, 1.28	0.177

Perceived harms and benefits as predictors of referral responses pre- and post-algorithm

Regression coefficients of the linear trends reported as Odds Ratios (ORs), 95% Confidence Intervals (CIs), and p values of multiple ordinal regression models predicting referral responses pre-algorithm and post-algorithm.

Stakeholder	Predictor	Pre-algorithm	Post-algorithm
Patient	Benefits	*OR*=9.88 [4.59, 21.25] *p*<0.001	*OR*=6.52 [3.19, 13.35] p<0.001
	Harms	*OR*=0.46 [0.22, 0.99] *p*=0.048	*OR*=0.47 [0.23, 0.97] *p*=0.042
NHS/Society	Benefits	*OR*=4.06 [2.06, 7.98] *p*<0.001	*OR*=1.98 [1.00, 3.93] *p*= 0.049
	Harms	*OR*=0.21 [0.09, 0.54] *p*=0.001	*OR*=0.26 [0.11, 0.59] *p*=0.001
GP/Practice	Benefits	*OR*=1.12 [0.56, 2.25] *p*=0.746	*OR*=1.65 [0.81, 3.37] *p*=0.167
	Harms	*OR*=3.11 [0.78, 12.39] *p*=0.107	*OR*=1.27 [0.41, 3.93] *p*=0.672

Christensen, R. H. B. (2023). ordinal—regression models for ordinal data. *R package version*, 2023.*12-4.1*. https://cran.r-project.org/web/packages/ordinal/index.html