BMJ Open

How accurate are digital symptom assessment apps for suggesting conditions and urgency advice? A clinical vignettes comparison to GPs

Stephen Gilbert ,[1] Alicia Mehl,[1] Adel Baluch,[1] Caoimhe Cawley,[1] Jean Challiner,[1] Hamish Fraser,[2] Elizabeth Millen,[1] Maryam Montazeri ,[1] Jan Multmeier,[1] Fiona Pick,[1] Claudia Richter,[1] Ewelina Türk,[1] Shubhanan Upadhyay,[1] Vishaal Virani,[1] Nicola Vona,[1] Paul Wicks,[1] Claire Novorol[1]

To cite: Gilbert S, Mehl A, Baluch A, *et al*. How accurate are digital symptom assessment apps for suggesting conditions and urgency advice? A clinical vignettes comparison to GPs. *BMJ Open* 2020;**10**:e040269. doi:10.1136/bmjopen-2020-040269

► Prepublication history and additional materials for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2020-040269).

SG and AM contributed equally.

Received 11 May 2020
Revised 27 October 2020
Accepted 16 November 2020

[1]Ada Health GmbH, Berlin, Germany
[2]Brown Center for Biomedical Informatics, Brown University, Rhode Island, USA

Correspondence to
Dr Stephen Gilbert;
science@ada.com

ABSTRACT

Objectives To compare breadth of condition coverage, accuracy of suggested conditions and appropriateness of urgency advice of eight popular symptom assessment apps.

Design Vignettes study.

Setting 200 primary care vignettes.

Intervention/comparator For eight apps and seven general practitioners (GPs): breadth of coverage and condition-suggestion and urgency advice accuracy measured against the vignettes' gold-standard.

Primary outcome measures (1) Proportion of conditions 'covered' by an app, that is, not excluded because the user was too young/old or pregnant, or not modelled; (2) proportion of vignettes with the correct primary diagnosis among the top 3 conditions suggested; (3) proportion of 'safe' urgency advice (ie, at gold standard level, more conservative, or no more than one level less conservative).

Results Condition-suggestion coverage was highly variable, with some apps not offering a suggestion for many users: in alphabetical order, Ada: 99.0%; Babylon: 51.5%; Buoy: 88.5%; K Health: 74.5%; Mediktor: 80.5%; Symptomate: 61.5%; Your.MD: 64.5%; WebMD: 93.0%. Top-3 suggestion accuracy was GPs (average): 82.1%±5.2%; Ada: 70.5%; Babylon: 32.0%; Buoy: 43.0%; K Health: 36.0%; Mediktor: 36.0%; Symptomate: 27.5%; WebMD: 35.5%; Your.MD: 23.5%. Some apps excluded certain user demographics or conditions and their performance was generally greater with the exclusion of corresponding vignettes. For safe urgency advice, tested GPs had an average of 97.0%±2.5%. For the vignettes with advice provided, only three apps had safety performance within 1 SD of the GPs—Ada: 97.0%; Babylon: 95.1%; Symptomate: 97.8%. One app had a safety performance within 2 SDs of GPs—Your.MD: 92.6%. Three apps had a safety performance outside 2 SDs of GPs—Buoy: 80.0% ($p<0.001$); K Health: 81.3% ($p<0.001$); Mediktor: 87.3% ($p=1.3\times10^{-3}$).

Conclusions The utility of digital symptom assessment apps relies on coverage, accuracy and safety. While no digital tool outperformed GPs, some came close, and the nature of iterative improvements to software offers scalable improvements to care.

Strengths and limitations of this study

- The study included a large number of vignettes which were peer reviewed by independent and experienced primary care physicians to minimise bias.
- General practitioners and apps were tested with vignettes in a manner that simulates real clinical consultations.
- Detailed source data verification was carried out.
- Vignette entry was conducted by professionals as a recent study found that laypeople are less good at entering vignettes for symptoms that they have never experienced.
- Limitations include the lack of a rigorous and comprehensive selection process to choose the eight apps and the lack of real patient experience assessment.

INTRODUCTION

Against the background of an ageing population and rising pressure on medical services, the last decade has seen the internet replace general practitioners (GPs) as the first port of call for health information. A 2010 survey of over 12 000 people from 12 countries reported that 75% of respondents search for health information online,[1] with some two-thirds of patients in 2017 reporting that they 'google' their symptoms before going to the doctor's office.[2] However, online search tools like Google or Bing were not intended to provide medical advice and risk offering irrelevant or misleading information.[3] One potential solution is dedicated symptom assessment applications (ie, apps),[3–6] which use a structured interview or multiple-choice format to ask patients questions about their demographic, relevant medical history, symptoms, and presentation. In the first few screening questions, some symptom assessment apps

exclude patients from using the tool if they are too young, too old, are pregnant, or have certain comorbidities, limiting the 'coverage' of the tool. Exclusion limits the range of users for whom the app can be turned to for advice, but, depending on the market segment the app manufacturer wants to address, having a narrow coverage may be appropriate, and it may in certain circumstances have advantages, for example, if it was a requirement of a regulatory authority within a certain jurisdiction, or, if it was possible to design the app with greater usability by narrowing its focus. Assuming the patient is not excluded, these software tools use a range of computational approaches to suggest one or more conditions that might explain the symptoms (eg, common cold vs pneumonia). Many symptom assessment apps then suggest next steps that patients should take (levels of urgency advice, for example, self-care at home vs seek urgent consultation), often along with evidence-based condition information for the user.

A recent systematic review of the literature identified that rigorous studies are required to show that these apps provide safe and reliable information[4] in the context for which they were designed and for which they have regulatory approval. Most previous studies considered only a single symptom assessment app, focused on specific (often specialty) conditions, had a small number of vignettes (<50), were relatively uncontrolled in the nature of the cases presented, and suffered a high degree of bias.[4] For example, a previous study examined the performance of the Mediktor app in the emergency department (ED) waiting room.[7] While this is a valid setting, most apps were designed and approved for use primarily at home and for newly presenting problems; accordingly, some 38.7% of patients had to be excluded. Few studies have systematically compared symptom assessment apps to one another in this context, which is particularly important as apps may increasingly be used to supplement or replace telephone triage.[4] This is particularly relevant in 2020 due to the COVID-19 pandemic—early in the spread of COVID-19, healthcare facilities risked being overwhelmed and furthering contagion, so communication strategies were needed to provide patients with advice without face-to-face contact.[8 9]

In contrast to deploying apps in a heterogeneous real-world setting, where participants would not have the time to re-enter their symptoms multiple times, and may not receive a verifiable diagnosis, clinical vignettes studies allow direct comparison of interapp and app-to-GP performance.[10–12] Clinical vignettes are created to represent patients, these are reviewed and then assigned gold-standard answers for main and differential diagnoses and for triage. The clinical vignettes are then used to test both apps and GPs. GPs are assessed through mock telephone consultations and apps through their normal question flow.[3 6] Clinical vignettes studies have the advantage of enabling direct GP-to-app comparison, allowing a wide range of case types to be explored, and are generalisable to 'real-life' situations, but are complementary to, not a replacement for, real-patient studies.[4 10 12] Seminal work at Harvard Medical School has established the value of such approaches but has not been updated recently.[3 6]

The objective of the current study was to compare the coverage, suggested condition accuracy, and urgency advice accuracy of GPs and eight popular symptom assessment apps which provide, for a general population, condition suggestions and urgency advice: Ada, Babylon, Buoy, K Health, Mediktor, Symptomate, WebMD, and Your.MD. We had three primary hypotheses:

1. That GPs would have better performance than the apps in the three metrics of (a) condition suggestion accuracy, (b) appropriateness of urgency advice and (c) safety of urgency advice.
2. That performance of each app would be consistent across the three metrics (condition-suggestion accuracy, appropriateness and safety of urgency advice).
3. That apps would differ from one another in their performance across the three metrics.

Exploration of these hypotheses is important for users of the applications and for physicians.

METHODS

The process for clinical vignette creation, review and testing of the GPs and the apps using the vignettes is shown in figure 1.

Clinical vignette creation

An independent primary care clinical expert consultant (JC) was commissioned to lead the creation of 200 clinical vignettes: JC has over 25 years' experience in general practice and emergency practice and has also had many years of experience in creating and customising algorithms for use in telephone triage and for internet-based self-assessment, including for *National Health Service, UK (NHS Direct)*. The vignette creation team also included two GPs (SU and AB—employees of Ada Health), each with over 5 years primary care and ED experience. SU and AB had worked for the Ada Health telehealth service *Dr Chat* but were not involved in the development of Ada's medical intelligence. The vignettes were designed to include both common and less-common conditions relevant to primary care practice, and to include clinical presentations and conditions affecting all body systems. They were created to be fair cases representing real-world situations in which a member of the public might seek medical information or advice from a symptom assessment app, or present to primary care. Most of the clinical vignettes were newly presenting problems experienced by an individual or by a child in their care, and they included some patients with chronic conditions, for example, diabetes, hypertension, and so on (see online supplemental tables 1–3).

The origin of 32.0% of the vignettes (numbers 1–64) was anonymised insights from transcripts of real calls made to NHS Direct (a UK national nurse-led telephone next-steps advice/triage service operational until 2014) which had previously been used as part of an NHS Direct

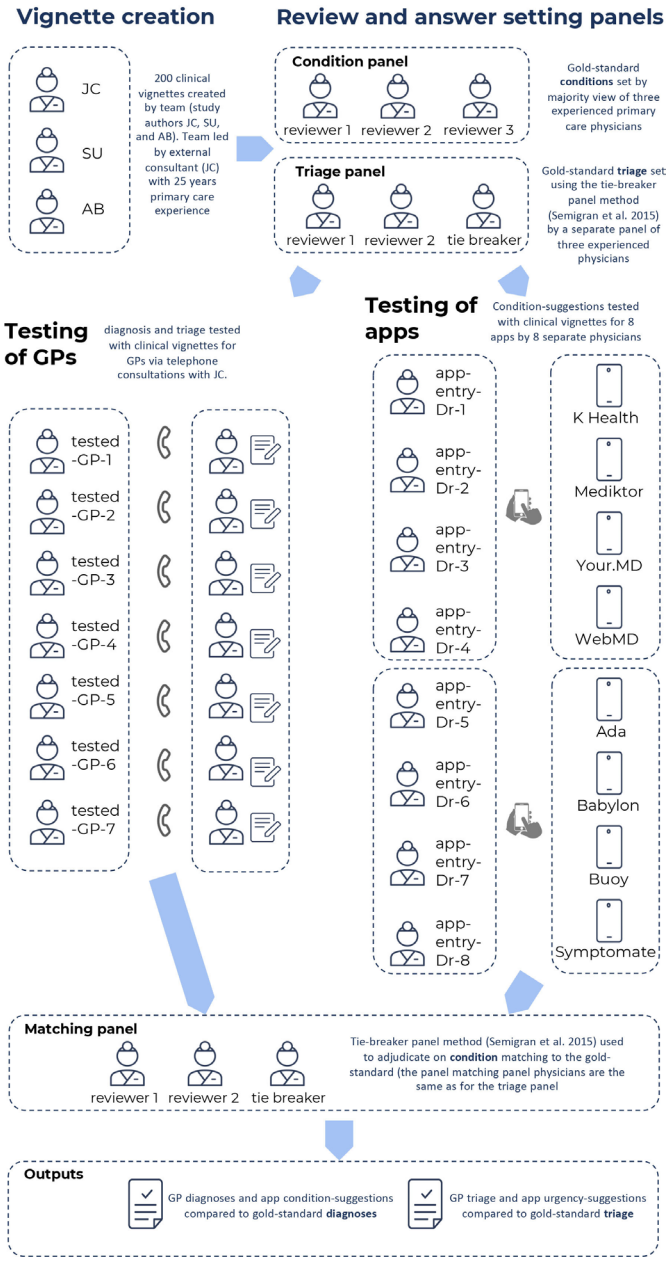


Figure 1 Overview of the study methodology including: (1) vignette creation; (2) vignettes review and answer setting; (3) testing of general practitioners (GPs); and (4) adjudication of matching of condition suggestions to the gold standard.

benchmarking exercise for recommended outcomes (these were used with full consent of NHS Direct). The remaining 68.0% of the vignettes were created by the vignette creation team (JC, SU and AB), including joint assignment of the most appropriate main diagnosis and differential diagnoses, as a starting point for the vignette gold-standard answers. The vignettes included the sex and age of the patient, previous medical history (including factors such as pregnancy, smoking, high blood pressure, diabetes, other illnesses), the named primary complaint, additional information on the primary complaint and current symptoms, and information to be provided only 'if asked' by the tested-GP or symptom assessment app

Table 1 Triage levels assigned to each clinical vignette

Level of urgency advice	Description
1	Call ambulance
2	Go to emergency department
3	See primary care within 4 hours
4	See primary care same day
5	See primary care non-urgent
6	Home care

(see online supplemental table 3). Each vignette was created with a list of gold standard correct conditions, arrived at through the majority decision of the vignette creation panel. This list included a main diagnosis and a list of other differential diagnoses (generally between one and four, but length-varying per vignette, as appropriate to the clinical history).

Vignette review

The vignettes were reviewed externally by a panel of three experienced primary care practitioners, each with more than 20 years primary care experience (see acknowledgements), recruited from the professional network of JC. The role of the review panel was to make changes to improve quality and clarity, and to set the gold-standard main diagnosis and differential diagnoses; this was determined by the majority view.

The gold-standard triage level was set independently of vignette creation, vignette review and vignette diagnosis gold-standard setting – this was done by a separate panel of three experienced primary care practitioners using a tie-breaker panel method based on the matching process set out by.[6] The gold-standard optimal triage was assigned by the panel to a six-point scale (see table 1), independent of the native levels of urgency advice of any of the eight apps. The tested-GPs' triage and the levels of urgency advice of each app were mapped to this scale using the linear mapping set out in online supplemental figure 1.

Assessment of apps and GPs using vignettes

Seven external GPs were tested with the vignettes (the 'tested-GPs'), providing condition suggestions (preliminary diagnoses) for the clinical vignettes after telephone consultations with JC, who had the role of 'patient–actor'–physician. All tested GPs were listed on the GP Register and licensed to practice by the UK General Medical Council and had an average of 11.2 years clinical experience post qualification as a doctor and 5.3 years post qualification as a GP. The seven GPs were recruited from the professional networks of AB, SU and JC. Of these, four had previously worked for the Ada Health telehealth service *Dr Chat* but were no longer employees at Ada. This prior employment did not include any involvement in the development of the Ada symptom assessment app. The other three GPs had no employment connection to any of the app manufacturers. Five of the tested GPs completed telephone consultations for all 200 clinical-vignettes. One

GP completed 130 telephone consultations but had to withdraw due to personal reasons. Another GP completed 100 telephone consultations but had to withdraw due to work commitments. Based on the information provided in the telephone consultation, the GPs were asked to provide a main diagnosis, up to five other differential diagnoses, and a single triage level (appropriate to a telephone triage setting).

Assessment of vignettes by the symptom assessment apps and 'coverage'

The clinical vignettes were entered into eight symptom assessment apps by eight primary care physicians playing the role of 'patient'—(app-entry-Dr-1 to –8 in figure 1). The versions of the symptom assessment mobile apps assessed were the most up to date version available for iOS download between the dates of 19 November 2019 and 9 December 2019. The version of the Buoy online symptom assessment tool used was the version available online between the dates of 19 November 2019 and 16 December 2019. The symptom assessment apps investigated were Ada, Babylon, Buoy, K Health, Mediktor, Symptomate, WebMD, Your.MD (see online supplemental table 4 for a description of these apps). The eight physicians were recruited from the professional network of AB, FP and SU. They were listed on the GP Register and licensed to practice by the UK General Medical Council, with at least 2 years of experience as a GP and had never worked or consulted for Ada Health; these physicians had no other role in this study. Each physician entered 50 randomly assigned vignettes (out of 200) into each of four randomly assigned symptom assessment apps. If the app did not allow entry of the clinical vignette (lack of coverage), the reason for this was recorded, as was the reason for every vignette for which condition suggestions or levels of urgency advice were not provided. If entry was permitted, the physician recorded the symptom assessment app's condition suggestions and levels of urgency advice and saved screenshots of the app's results to allow for source data verification. In this way, each vignette was entered once in each app, with four physicians entering vignettes in each app.

Source data verification

Source data verification was carried out (100% of screenshots compared with spreadsheet data) and any missing or inaccurately transcribed data in the spreadsheets was quantified, recorded in this report and corrected to reflect the screenshot data.

Metrics for assessing condition-suggestion accuracy

We compared the top-1 suggested condition (M1), the top-3 suggested conditions (M3), and the top-5 suggested conditions (M5) provided by the seven tested GPs and the eight apps to the gold-standard main diagnosis. We also calculated the comprehensiveness and relevance of each GP's and each app's suggestions[13]—see table 2 for a description of the metrics used for comparing condition-suggestion accuracy.

Assigning matches between tested-GPs/apps and the gold-standard

Every suggested condition from the tested GPs and the apps was submitted anonymously to an independent panel of experienced primary care physicians who were recruited from the professional network of FP, and who were listed on the GP Register and licensed to practice by the UK General Medical Council, with at least 2 years of experience as a GP and had never worked or consulted for Ada Health. The panel had the role of deciding if the suggested condition matched the gold-standard diagnoses list, unless there was an explicit exact match—that is, identical text of the answer from the tested-GP/app and the gold standard. Matching was decided using a tie-breaker panel method which was based on the method set out by.[6] The panel was presented with the condition suggestions blinded to their source. Panellists were instructed to use their own clinical judgement in interpreting whether condition suggestions were matches to

Table 2 Metrics used in comparison of condition-suggestion accuracy

Abbrev.	Full name	Description
M1 (%)	M1 (Matching-1) accuracy	% of cases where the top-1 condition-suggestion matches the gold-standard main diagnosis.[7]
M3 (%)	M3 (Matching-3) accuracy	% of cases where the top-3 condition-suggestions contain the gold-standard main diagnosis.[7]
M5 (%)	M5 (Matching-5) accuracy	% of cases where the top-5 condition-suggestions contain the gold-standard main diagnosis
COMP (%)	Comprehensiveness	Ratio of the (number of gold standard differentials matched by the suggested differentials) to the (number of gold standard differentials for the vignette), expressed as a mean across all vignettes.[13]
RELE (%)	Relevance	Ratio of the (number of the suggested differentials that match with any of the gold standard differentials for the vignette) to the(number of differentials provided by the tested-GP or the symptom assessment app for the vignette), expressed as a mean across all vignettes.[13]

the gold standard, supported by matching criteria (see online supplemental table 5).

Mapping and comparing levels of urgency advice

Triage suggestions from each GP and levels of urgency advice from each app were mapped to the gold standard triage levels using the simple linear mapping scheme set out in online supplemental figure 1. The degree of deviation of GP triage urgency and of app levels of urgency advice was compared by reporting the percentage of vignettes for which GPs and symptom assessment apps were: (1) overconservative; (2) overconservative but suitable (one level too high); (3) exactly-matched; (4) safe but underconservative (one level too low); or, (5) potentially unsafe.

The WebMD assessment report only provides information on whether each suggested condition is urgent (via an urgency 'flag'). Finer urgency advice on each condition suggestion is available by clicking through to a separate detailed screen on each suggested condition, but unlike the other apps, no overall vignette-level summary urgency advice is provided. Meaningful comparison to the other apps or tested GPs was therefore not possible and WebMD was excluded from the urgency advice analysis in this study. For each app, with the exception of WebMD, the proportion of 'safe' urgency advice, is defined as advice at the gold standard advice level, more conservative, or no more than one level less conservative.

We used confusion matrices in order to fully visualise the severity of misclassification of advice levels.[14] These confusion matrices were weighted in order to represent the relative seriousness of inappropriate urgency advice, either in the direction of being overly conservative (eg, inefficient use of healthcare system resources), or in the direction of being insufficiently conservative (potentially unsafe advice). The weighted confusion matrices were normalised to correct to the number of vignettes for which urgency advice were provided by each app and tested-GP.

Statistical methods

M1, M3 and M5 performance as well as levels of urgency advice were compared using descriptive statistics and tests appropriate for categorical data. χ^2 tests were used to test whether the proportion of correct answers from all apps and from all tested GPs were drawn from the same distribution. In case of a significant difference, two-sided post hoc pairwise Fisher's exact tests[15 16] were used to compare individual app or tested-GP performances. Comprehensiveness and relevance (COMP and RELE) were assessed by Kruskal-Wallis-H-Test (KW-H-Test) applied to all 15 answer datasets (8 apps and 7 tested GPs), followed by post hoc pairwise testing using the two-sided Dunn test,[15] in cases where there was a significant difference on the KW-H-Test. P values were corrected for multiple comparisons using the Benjamini-Hochberg procedure[17] and considered significant if less than 0.05. In figures, error bars for individual app and tested-GP performance represent 95% CI. These were calculated using the Wilson-Score method for categorical data (M1, M3 and M5)[18] and using the percentile bootstrap method for COMP and RELE.[19] The mean app and tested-GP scores were calculated as arithmetic means of the M1, M3, M5, COMP and RELE performance for each app and each tested GP, with error bars that represent the SD.

Patient and public involvement

Patients were not involved in setting the research questions, the design, outcome measures or implementation of the study. They were not asked to advise on interpretation or writing up of results. No patients were advised on dissemination of the study or its main results.

RESULTS

Source data verification

For vignette cases where the app-entry-Drs made data recording errors, these were corrected to match the source verification data saved in the screenshots. Full sets of screenshots were recorded by seven of the eight app-entry-Drs. One app-entry-Dr (#4) did not record all screenshots for K Health, WebMD and for Your.MD and for this reason a subanalysis of the 150 vignettes for which full verification was possible for these apps is provided in online supplemental table 6 and 7. The differences in performance in this subanalysis is relatively minor and might be due to random differences between the 150 and full vignette sets or be due to app-entry-Dr-4 recording error.

App coverage

The apps varied in the proportion of vignettes for which they provided any condition suggestions (see figure 2, online supplemental tables 8–10). The reasons that some apps did not provide condition suggestions included: (1) not included in the apps' regulatory 'Intended Use' or another product design reason (eg, users below a set age limit, or pregnant users); (2) not suggesting conditions for users with severe symptoms (or possible conditions); (3) presenting problem not recognised by the app (even after rewording and use of synonyms); and, (4) some apps did not have coverage for certain medical specialties, for example, mental health. For 12% of the vignettes, the urgency advice from for K Health was not recorded due to app-entry-Dr-4 recording error and was not recorded in source verification data saved in the screenshots. The missing data is labelled in figure 2 and in the later figures describing the appropriateness of urgency advice. A subanalysis of the 150 vignettes for which full data and full verification was possible for K Health is provided in online supplemental table 6.

Suggested conditions: the 'required-answer' approach

The approach adopted in other vignettes studies by authors in refs, semigran HL *et al*,[3 6] Bisson LJ *et al*,[20] Burgess M *et al*,[21] Powley L *et al*,[22] Pulse Today *et al*,[23]

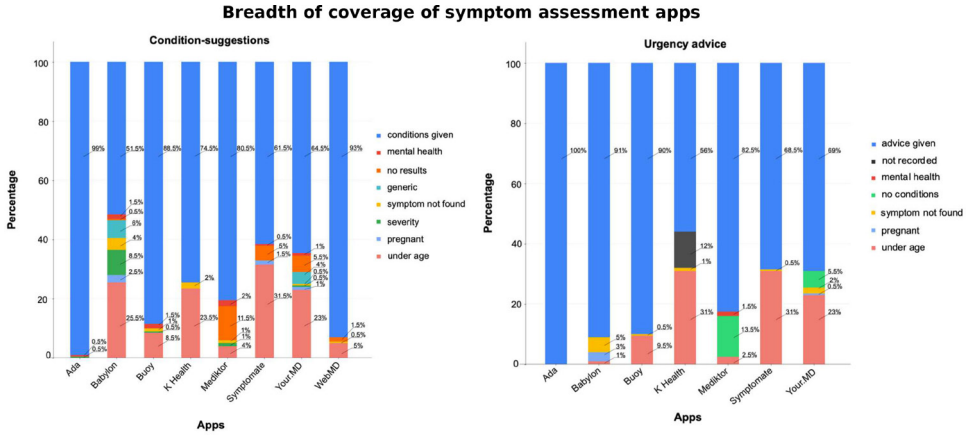


Figure 2 App breadth of coverage—that is, the proportion of vignettes for which condition suggestions and levels of urgency advice were provided. When condition suggestions or urgency were not provided, the principal reason for this is shown, or alternatively 'no results' when no reason was given. conditions given—condition suggestions were provided by the app; mental health—mental health vignettes where no condition suggestions/urgency advice was provided; no results—the app provided a clear statement that no condition suggestion results were found for the vignette (the reason why the app failed to give a condition suggestion for these vignettes is uncertain, but generally these vignettes relate to minor conditions, and in most cases it seems that the app does not have a matching condition modelled); generic—the app gave a generic answer rather than a condition, for example, 'further assessment is needed'; symptom not found—a directly or appropriately matching symptom to the presenting complaint could not be found in the app so the vignette could not be entered; severity—the app did not to give condition-suggestions for very serious symptoms—for example, the app stated only 'Condition causing severe (symptom)'; pregnant,vignettes for which no condition-suggestions/urgency advice was provided by the app as the patient was pregnant; under age—vignettes for which no condition suggestions/urgency advice was provided by the app as the patient was under its specified age limit; advice given—level of urgency advice was provided by the app; not recorded—one app-entry-Dr (#4) did not fully record the levels of urgency advice, and there were no corresponding source data verification screenshots for this subset of data (see online supplemental table 6 for a subanalysis of the 150 vignettes with complete source-data-verified data for K Health on levels of urgency advice); no conditions—no condition suggestions were provided by the app, and, as a result of this, the app did not provide urgency advice. See online supplemental tables 8–10 for details.

Nateqi J *et al*[24] has been to determine the percentage of all vignettes for which the app (or tested GP) provided an appropriate condition-suggestion—here, this analysis method is referred to as the 'required-answer' approach. Results are shown in figure 3. For a full description for each metric, see table 2.

Suggested conditions: the 'provided-answer' approach

For users or physicians choosing or recommending a symptom assessment app, it is relevant to know not only the app accuracy, but also how wide is its coverage and therefore the 'required-answer' analysis in the previous section is the most relevant analysis. An alternative approach is the provided-answer analysis, which is the number of correct suggested conditions provided by an app for each vignette *for which it provides an answer*. In other words, there was no penalty for an app that, for any reason, does not provide condition suggestions for a vignette, for example, children under 2 years old (see online supplemental tables 4 and 10). Both analyses are provided in this study in order to give a fully balanced overview of the performance of all the apps. The results for the provided-answer analysis are shown in figure 4. For a full description for each metric, see table 2.

Levels of urgency advice

The urgency advice performance of each app is summarised in table 3. Tested GPs had safe triage performance of 97.0%±2.5% (where safe is here defined as maximum one level less conservative than gold-standard, expressed per vignette provided with advice)—three apps had safety performance within 1 SD of GPs (mean)—Ada: 97.0%; Babylon: 95.1%; and, Symptomate: 97.8%. One app had a safety performance within 2 SDs of GPs—Your.MD*: 92.6%. Three apps had a safety performance outside 2 SDs of GPs—Buoy: 80.0% ($p<0.001$); K Health*: 81.3% ($p<0.001$); Mediktor: 87.3% ($p=1.3\times10^{-3}$) (*—for two of these apps one app-entry-Dr (#4) did not record all screenshots needed for source data verification—see online supplemental table 6 for a subanalysis of fully verified data, which shows the same trend of results and no significant difference to the data recorded here).

Figure 5 summarises and compares urgency advice performance, including the proportion of vignettes for which some apps did not provide advice.

The visualisation in figure 5 provides a high-level overview of urgency advice performance; however, a limitation of this approach is that the full range of comparisons between gold standard triage and levels of urgency advice is not shown. The full range of overconservative and potentially unsafe urgency advice provided by each app and tested GP is shown in the weighted confusion matrices (figure 6). Low numbers in the matrices (coloured green and yellow) correspond to good urgency advice allocation, high numbers (coloured orange and red) correspond to

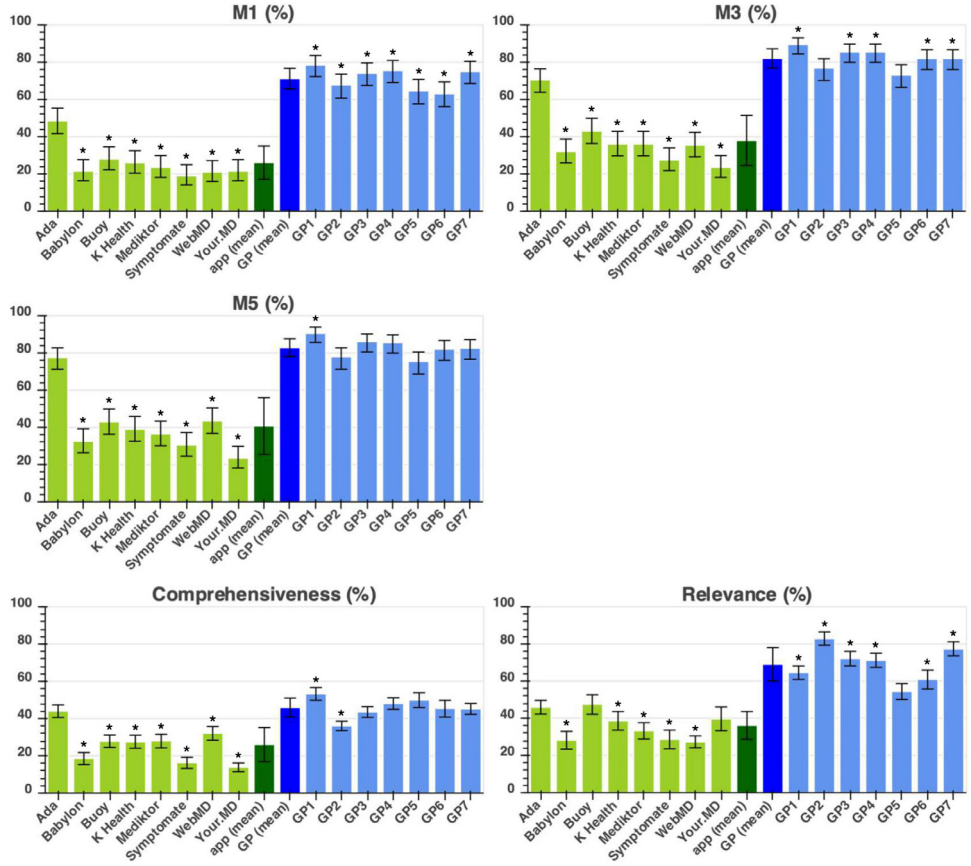


Figure 3 Required answer approach showing the performance metrics (M1, M3, M5, comprehensiveness and relevance—as defined in table 2) of the eight apps and seven tested general practitioners (GPs). App performance is coloured in light green, average (mean) app performance is in dark green, average (mean) tested-GP performance in dark blue, and individual tested-GP performance in light blue. Statistical significance of the difference between the app with highest performance and all other apps/tested GPs is shown with the * symbol indicating: p<0.05. For one of these apps (Your.MD), one app-entry-Dr (#4) did not record all screenshots needed for source data verification—see online supplemental table 7 for a subanalysis of fully verified data, which shows the same trend of results and no significant difference to the data recorded here.

bad urgency advice allocation. In order to visualise the overall urgency advice performance of each app, that is, performance both in urgency advice coverage and in the percentage of safe advice, these measures are plotted against each other in figure 7.

Subanalysis of performance in the NHS-derived and non-NHS-derived vignettes

This study evaluated app and GP performance using 200 vignettes, of which 32.0% were derived from NHS Direct cases and 68.0% were created by the vignette creation team. The performance of each app and average GP performance stratified by vignette source (NHS or non-NHS derived) are shown in online supplemental table 11. The GPs and all apps performed better in providing appropriate urgency advice in the non-NHS vignettes than in the NHS-derived vignettes. In condition-suggestion accuracy, all GPs performed substantially better in M1, M3 and M5 for the non-NHS vignettes (differences in GP mean performance were 15.0%, 11.7% and 10.8%, respectively). Differences in GP performance in COMP and RELE were not large, and performance in COMP was better (difference 3.1%) in the NHS-derived vignettes. Apps differed in their relative condition-suggestion accuracy between the NHS and non-NHS derived vignettes. Ada and Buoy, following the pattern of the GPs, performed substantially better in the non-NHS vignettes, while Symptomate performed similarly in both vignettes sets, and K Health, WebMD and Your.MD performance was relatively better in some and relatively weaker in other metrics in the two sets of vignettes. Mediktor was moderately better in the NHS derived vignettes for all metrics except RELE.

DISCUSSION

Principal findings

In this clinical vignette comparison of symptom assessment apps and GPs, we found that apps varied substantially in coverage, appropriateness of urgency advice and accuracy of suggested conditions.

Synthesising the analyses on the appropriateness of urgency advice (see table 3 and figures 5–7), the apps can be categorised as follows:

1. Levels of safe urgency advice within one SD from the average of GPs and:

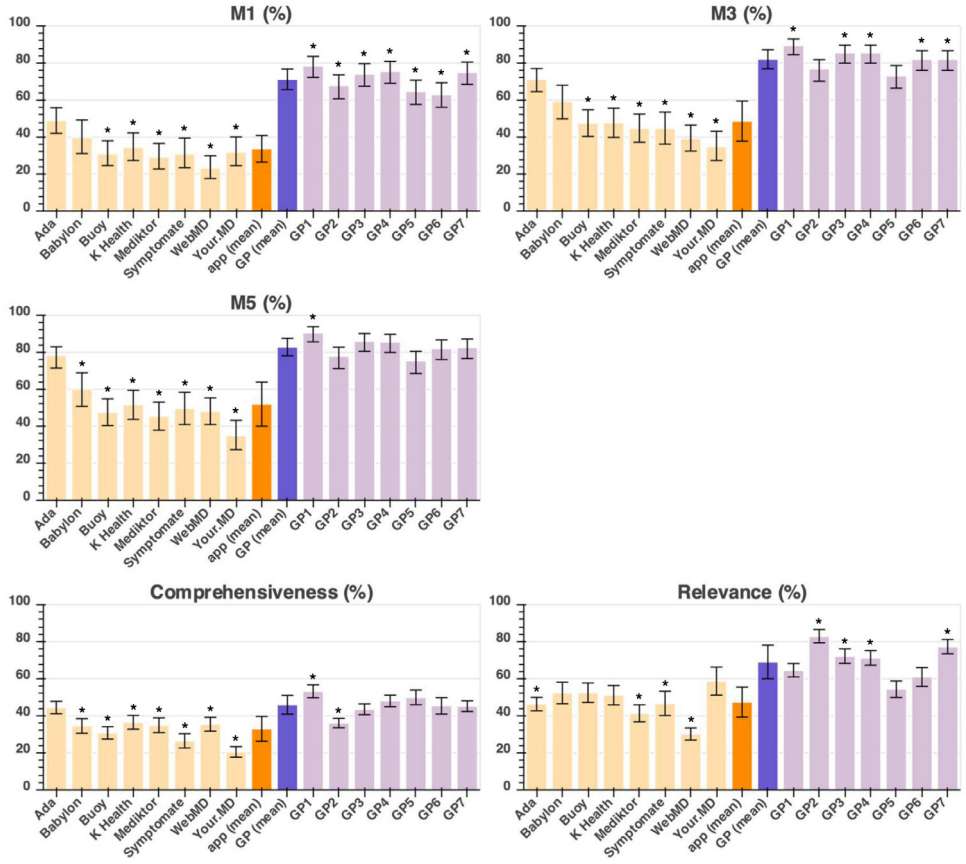


Figure 4 Provided-answer approach showing the performance metrics (M1, M3, M5, comprehensiveness and relevance—as defined in table 2) of the eight apps and seven tested general practitioners (GPs). App performance is coloured in light orange, average (mean) app performance is in dark orange, average (mean) tested-GP performance in dark purple, and individual tested-GP performance in light purple. Statistical significance of the difference between the app with highest performance and all other apps/tested GPs is shown with the * symbol indicating: p<0.05. For one of these apps (Your.MD), one app-entry-Dr (#4) did not record all screenshots needed for source data verification—see online supplemental table 7 for a subanalysis of fully verified data, which shows the same trend of results and no significant difference to the data recorded here.

1a. Full-full or near-full coverage: Ada.
1b. Moderate coverage: Babylon.
1c. Low coverage: Symptomate.
2. Levels of safe urgency advice between one and two SD from the average of GPs and:
2a. Low coverage: Your.MD.
3. Levels of safe urgency advice below three SD from the average of GPs and:
3a. Moderate coverage: Buoy, Mediktor.
3b. Low coverage: K Health.

Condition suggestion coverage varies greatly with a range of 47.5% from highest (Ada; 99.0%) to lowest (Babylon, 51.5%). Although there is no absolute cut-off of what an acceptable condition suggestion coverage is, an app that can provide high coverage along with a high accuracy of condition suggestion and high urgency advice appropriateness, will generally be superior to an app with narrow coverage. There is no identifiable correlation between app M1 or M3 condition-suggestion accuracy or urgency-advice accuracy and the condition-suggestion coverage or urgency-advice coverage.

There was considerable variation in condition-suggestion accuracy between the GPs and between apps. For top-1 condition suggestion (M1), the range of tested GPs was 16.0%, the SD 5.6% and for M3 the range was 15.9% and SD 5.2%. For the apps, the M1 condition-suggestion accuracy range was 29.5%, the SD 8.9% and the M3 range was 47.0% and SD 13.5%. The GPs all outperformed apps for top-1 condition matching. For M3 and M5 (ie, including the gold standard diagnosis in top-3 and top-5 suggestions), the best performing app (Ada) was comparable to tested GPs, with no significant difference between its performance and the performance of several of the tested GPs. The top performing symptom assessment app (Ada) had an M3 27.5% higher than the next best performing app (Buoy, p<0.001) and 47.0% higher than the worst-performing app (Your.MD, p<0.001). There was a significant difference between the top performing app (Ada) and other apps for all condition accuracy measures, with two exceptions for relevance (in the required-answer analysis).

There was also considerable variation in urgency advice performance between the GPs and between apps. The range of tested-GP safe advice was 6.0% and the SD was 2.5%; for the apps, the range of safe advice was 17.8% and the SD 7.4%. Tested GPs had an average safe advice

Table 3 Triage levels assigned to each clinical-vignette, where safe is defined as maximum one level less conservative than gold-standard, expressed per vignette provided with advice.

App/ tested GP	Percentage of safe advice	P value (difference to GP mean)
Ada	97.0	NS
Babylon	95.1	NS
Buoy	80.0	<0.001*
K Health	81.3	<0.001*
Mediktor	87.3	1.3×10^{-3}*
Symptomate	97.8	NS
Your.MD	92.6	NS
App mean±SD.	90.1±7.4	–
GP mean±SD.	97.0±2.5	–
GP1	96.0	NS
GP2	96.9	NS
GP3	94.0	NS
GP4	99.0	NS
GP5	100.0	NS
GP6	93.9	NS
GP7	99.5	NS

*P<0.05. For two of these apps (K Health & Your.MD), one app-entry-Dr (#4) did not record all screenshots needed for source data verification—see online supplemental table 6 for a subanalysis of fully verified data, which shows the same trend of results and no significant difference to the data recorded here). This analysis is for those vignettes for which urgency advice was provided (ie, a 'provided answer) analysis.
GP, general practitioner; NS, no significant difference.

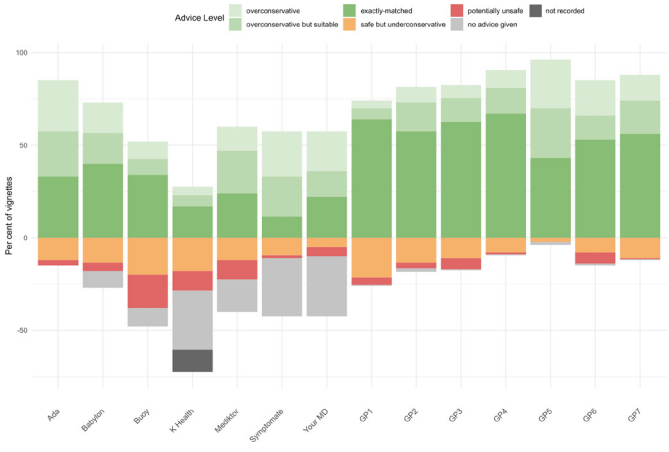


Figure 5 Accuracy of urgency advice displayed as a stacked bar chart centred on the gold standard triage. For two of these apps (K Health & Your.MD), one app-entry-Dr (#4) did not record all screenshots needed for source data verification—see online supplemental table 6 for a subanalysis of fully verified data, which shows the same trend of results and no significant difference to the data recorded here. GP, general practitioner.

performance of 97.1±2.5% and only three apps had safe advice performance within 1 SD of the GPs (mean)—Ada: 97.0%; Babylon: 95.1%; and Symptomate: 97.8%.

The results support acceptance of the hypothesis 1 (a), that GPs have better performance than the apps on condition-suggestion accuracy. Hypothesis 1 (b) and 1 (c) were that GPs would have better performance than the apps in the appropriateness and safety of urgency advice, and these hypotheses are partially rejected, as, while overall GPs performed better in urgency advice than apps, some individual apps performed as well as GPs in urgency advice safety and similarly to GPs in urgency advice accuracy. Hypothesis 2 was that performance of each app would be consistent across the three metrics (condition-suggestion accuracy, appropriateness and safety of urgency advice), and this hypothesis is rejected as the results showed that apps performing well in urgency advice safety or appropriateness did not necessarily have high condition-suggestion accuracy. Hypothesis 3, that apps would differ from one another in their performance across the three metrics. This hypothesis is accepted as there were major differences between apps in all three metrics.

There were relative differences in the performance of the GPs and of the apps in the NHS-derived and non-NHS derived vignettes; however, the overall conclusions of this study are valid for both sets of vignettes, and the performance of each app evaluated is broadly similar irrespective of whether all vignettes are considered or the NHS-derived or non-NHS-derived subsets. The differences in performance likely reflect differences in the case structure complexity in the vignettes, the degree of ambiguity in the vignettes, the individual question flow of the apps, differences in condition coverage of the apps and the different frequencies of disease categories in the vignettes—for example, there were more cardiovascular disease cases in the NHS-derived vignettes, 7/64 (10.9%) compared with 7/136 (5.1%) in the non-NHS-derived vignettes.

Strengths and limitations of this study

The systematic review of Chambers *et al*[4] identified limitations of published studies on the safety and accuracy of symptom assessment apps as: (1) not being based on real patient data; (2) not describing differences in outcomes between symptom assessment apps and health professionals; (3) covering only a limited range of conditions; (4) covering only uncomplicated vignettes; and (5) sampling a young healthy population not representative of the general population of users of the urgent care system. Of these limitations, only one applies to this study—the limitation of being based on clinical vignettes rather than on real-patient data. The effect of this limitation has been minimised through the development of many of the vignettes to be highly realistic through the use of anonymised real patient data collated from NHS Direct transcripts. The use of real patient data with an actual diagnosis is not without

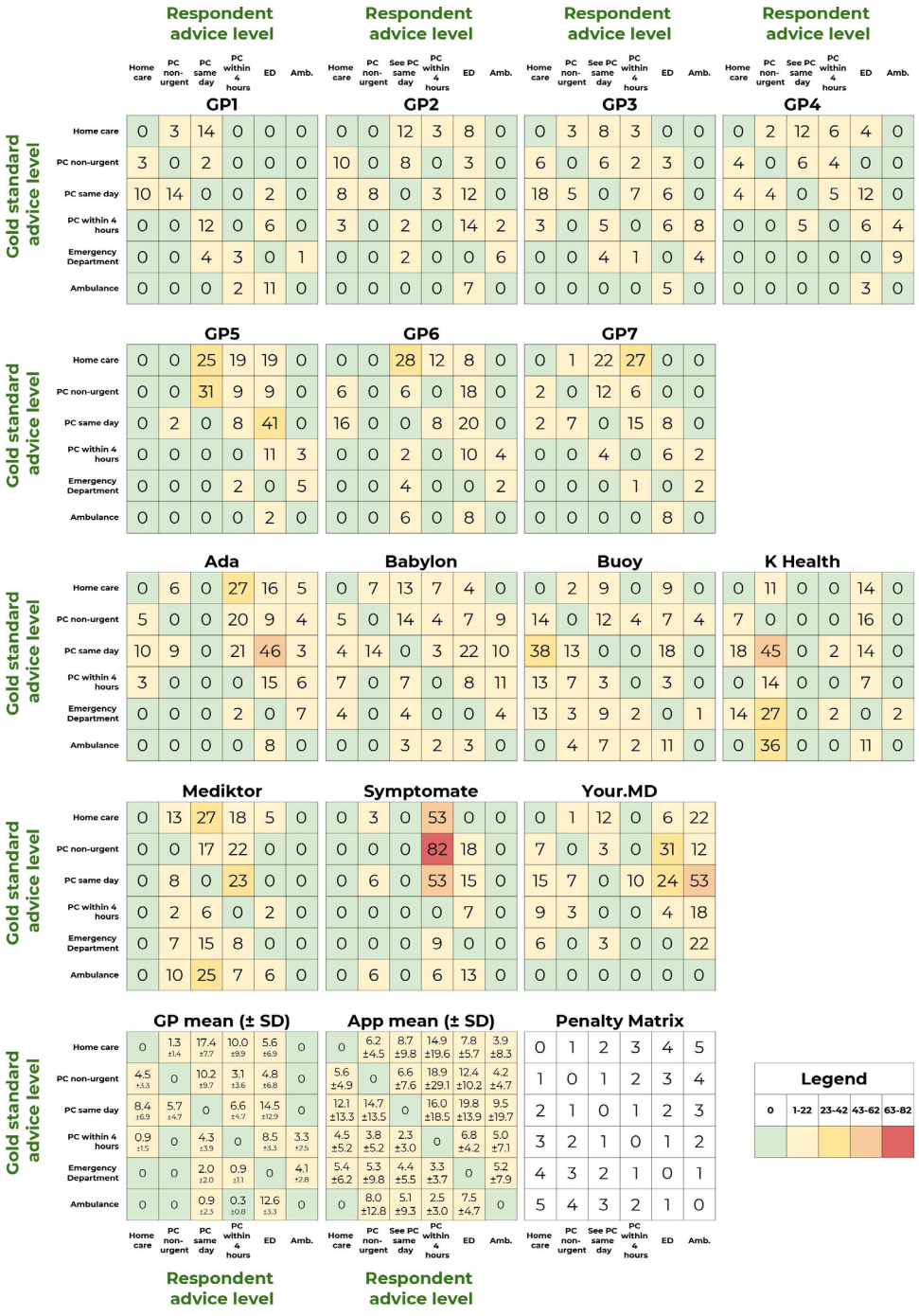


Figure 6 Weighted confusion matrices showing the detailed triage assignments for each app. For two of these apps (K Health & Your.MD), one app-entry-Dr (#4) did not record all screenshots needed for source data verification—see online supplemental table 6 for a subanalysis of fully verified data, which shows the same trend of results and no significant difference to the data recorded here. Amb, ambulance; ED, emergency department; GP, general practitioner; PC, primary care

its limitations in the evaluation of symptom assessment app accuracy as it relies on face-to-face consultation to confirm diagnosis. Very often diagnosis is only provided after physical examination or diagnostic tests, so comparison is confounded as the real patient diagnosis is based on additional information not made available to the app. The vignettes approach has allowed this study to be designed to minimise the limitations (2)–(5) identified by Chambers *et al*.[4] This has been done for limitation (2) through inclusion of a 7-GP comparator group; for limitation (3) by development of vignettes for conditions spanning all body systems and sampling all medical specialisms relevant to primary care presentation; for limitation (4) by designing clinical vignettes including not-only simple and common situations, but also moderately complex and challenging presentations; for limitation (5) through including vignettes spanning from 1 month to 89 years old.

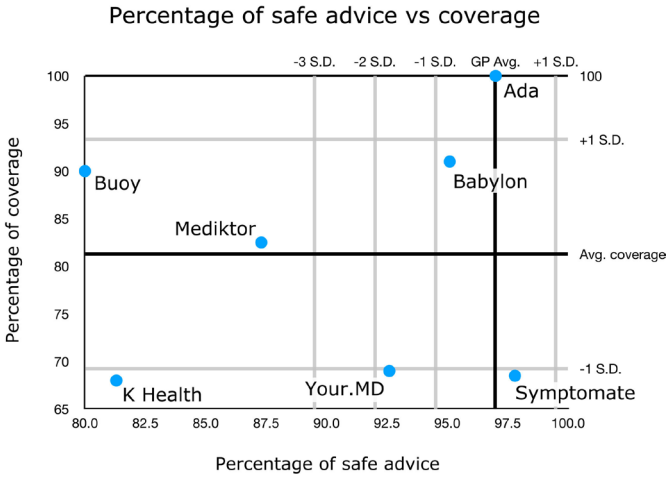


Figure 7 Summary plot of the urgency advice performance of each app. The urgency advice coverage of each app (with respect to app average) is plotted against the percentage of safe advice (with respect to the general practitioner (GP) average). For two of these apps (K Health & Your.MD), one app-entry-Dr (#4) did not record all screenshots needed for source data verification—see online supplemental table 6 for a subanalysis of fully verified data, which shows the same trend of results and no significant difference to the data recorded here.

Relative strengths of this study are the large number of clinical vignettes included (n=200), along with the separation in the design of clinical-vignette writing from the process of deciding on the gold-standard main and secondary differential diagnoses and appropriate levels of urgency advice. Another strength of this study is that GPs were tested with vignettes in a manner that simulates real clinical consulting—in this way the GPs consultation process was assessed, enabling a fair comparison to the apps. Vignettes were entered into the apps by eight additional primary care physicians acting as the user (app-entry-Dr-1–8). A physician also 'acted' as the patient being assessed by the GPs in the phone consultations. It has been argued that lay-person entry is closest to the real intended use of symptom assessment apps[25]; however, it is known that lay-people are less reliable at entering clinical vignettes than healthcare providers.[26] A further strength of this study is that each decision of whether a condition suggestion (from an app or a GP) matched the clinical-vignette's main and other differential diagnoses was made in a rigorous manner following the 3-physician tie-breaker panel approach of Semigran *et al*.[6]

A limitation of this study was that a systematic and comprehensive process was not used to select the symptom assessment apps to be included. Practical considerations in study design necessitated that the study evaluated a total of eight apps, due to the large number of vignettes assessed. The aim of the selection process was to include only apps with similar intended use, and to include those most used, those still in current use which have been evaluated in other studies, those most used within the UK as the study used vignettes based on UK patient data and those most used in the USA, as it is a highly important market for symptom assessment applications. Apps were then selected using a hybrid approach, including, based on the knowledge of available apps of the study team, internet searching and industry sources on usage data of symptom assessment apps. For apps identified using this approach, rigorous exclusion criteria were applied: all apps which did not provide, for a general population, primary care condition suggestions and urgency advice, were excluded. Through the application of this methodology, we have assured that all included apps were appropriate and relevant for inclusion, but it is possible, due to the limiting of the study to eight apps, without a rigorous prioritisation selection procedure, that there was unintentional bias in app inclusion. Nonetheless, the study included all the highest used symptom assessment apps in the UK and the USA, at the time of app selection, based on app usage statistics for the Google Play and Apple iOS app stores. The non-systematic selection criteria used were that, at the time of selection: (1) Babylon and Ada are leading symptom assessment smartphone apps in the UK; (2) K Health, WebMD and Ada are the most used in USA (usage data from Sensely, https://www.sensely.com); (3) Mediktor and Buoy have existing published data[7 27]; and (4) Your.MD has a similar user experience and user interface to Babylon and Ada and has been compared with them in small non-peer reviewed studies.[21 23]

Direct comparison of levels of urgency advice between individual apps and between apps and GPs was challenging because (1) some apps provided no levels of urgency advice for large numbers of vignettes; (2) performing well in one level of urgency advice trades off performance in other levels of urgency advice; (3) the nature of urgency advice reporting was different in WebMD (see the Methods section).

Furthermore, the vignettes may have had a UK bias and some of the symptom assessment apps (eg, Buoy, K Health & WebMD) are primarily used in the USA. The population demographics and the health conditions represented in the vignettes were broadly similar to demographics extracted from UK and NHS England health statistics (see the online supplemental Appendix S1, including online supplemental figures 2 and 3). Ada employees were involved in the vignette creation process, and although it was ensured that the vignette creation was separated from app medical intelligence development, unintentional bias could have resulted in vignette wording that was more accessible to symptom assessment apps than the average real word primary care clinical presentation is. A data acquisition error by one of the app-entry-Drs meant there were unrecorded urgency advice data for 12.0% of vignettes for one app (K Health) and incomplete source data verification screenshots for two other apps (Your.MD and WebMD). The implication for this for the main analysis was investigated in two subanalyses in the data supplement. Future studies could ensure ongoing source data verification rather than waiting until the end of study data collection for review. It

is an unavoidable limitation that software evolves rapidly, and the performance of these apps may have changed significantly (for better or worse) since the time of data collection. Finally, this study was designed, conducted and disseminated by a team that includes employees of Ada Health; future research by independent researchers should seek to replicate these findings and/or develop methods to continually test symptom assessment apps.

Comparisons to the wider literature

The results of this study are qualitatively broadly similar to reported results from other interapp relative performance studies, including one peer-reviewed study[24] and two non-peer-reviewed studies.[21 23] A peer-reviewed study using 45 ear, nose and throat (ENT)—vignettes[24] evaluated M1 and M3 results and found that Ada had substantially better performance than other apps. Overall, Ada was the second-best performing app out of 24 tested apps in the ENT discipline.[24]

A small non-peer-reviewed independent clinical vignettes study tested NHS 111, Babylon, Ada and Your.MD and found similar overall results to this study[23]; they also found that all apps were successful at spotting serious conditions, such as a heart attack, and that they were fast and easy to use. A second small 2017 non-peer-reviewed independent vignettes study,[21] that was carried out by established symptom assessment app academic researchers, tested Babylon, Ada and Your.MD. The trend of the results was similar to those in this study.

In an observational study carried out in a Spanish ED waiting room, the Mediktor symptom assessment app was used for non-urgent emergency cases for patients above 18 years old.[7] The study calculated accuracy with consideration only for those patients whose discharge diagnosis was modelled by the app at the time. For a total of 622 cases, Mediktor's M1 score was reported as 42.9%, M3 score as 75.4% (ie, the symptom assessment app's top-1 (M1), top-3 (M3), or top-10 condition-suggestion(s) matched the discharge diagnosis in this percentage of cases). When Moreno Barriga *et al*[7] reported results are refactored to consider all patient discharge diagnoses (the standard approach) the: M1 is 34.0% and M3 is 63.0%, compared with M1 of 23.5% and M3 of 36.0% for Mediktor in this study (all-vignettes data). The reason for lower Mediktor performance in the current study compared with the study in Moreno Barriga *et al*[7] is not known but it may be related to a different range of conditions or difficulty level than the non-urgent emergency cases presenting to the ED—for example—the vignettes in this study contain many true emergency cases and also many GP or pharmacy/treat-at-home cases which would not be represented by the ED patients included in Moreno Barriga *et al*[7]. In 2017, a 42-vignette evaluation of WebMD[28] determined its accuracy for ophthalmic condition suggestion: M1 was 26.0% and M3 was 38.0%. Urgency advice based on the top diagnosis was appropriate in 39.0% of emergency cases and 88.0% of non-emergency cases.

The manufacturers of the apps Babylon and Your.MD responded to the two non-peer reviewed studies[21 23] observing that their apps have been updated and improved subsequent to the publication of those reports. Nevertheless, the findings with respect to condition-suggestion performance, in the later peer-reviewed study by Nateqi *et al*[24] and in the present study appear to be in line with those from the two non-peer-reviewed studies.

Implications for clinicians and policy-makers

The results of this study are relevant for home users of symptom assessment apps, and to healthcare providers offering advice to their patients on which symptom assessment apps to choose. There are large (and statistically significant) differences between app coverage, suggested condition accuracy and urgency-advice accuracy. One of the biggest challenges in comparing symptom checker apps are the differences in coverage. Some coverage restrictions, such as not allowing symptom assessment for one user subgroup (eg, children), have no negative effect on the app's effective use for other user subgroups (eg, adults). Other situations, such as the inability to search for certain symptoms, providing no condition-suggestions/urgency advice for certain input symptoms, or, excluding comorbidities, mental health or pregnancy are more problematic and can raise concerns about the safety and benefits of the app for users who might be in those groups.

Unanswered questions and future research

Future research should evaluate the performance of the apps compared with real-patient data—multiple separate single-app studies are a very unreliable way to determine the true level of the state of the art of symptom-assessment apps. A positive step in this direction is the ITU/WHO Focus Group AI for Health (FG-AI4H) through which several manufacturers of symptom assessment apps evaluated in this study are working collaboratively to create standardised app benchmarking with independently curated and globally representative datasets.[29] Additional areas that could be explored in such studies are comparative economic impact, understanding user behaviour following an assessment, that is, compliance with urgency advice (extending the approach of Winn *et al*[27]) and impact on health services usage, and, the impact of using the apps to complement a standard GP consult (eg, through diagnostic-decision support). While it has been argued that the accuracy of urgency advice may be the most important output from a health assessment app, the condition suggestions may be valuable to support patient decision-making.[27] To address the effect of patients entering data directly into an app about their own acute conditions, an observational investigation is currently underway in an acute clinical setting in the USA by investigators including coauthors of this study. This includes a survey of users' technological literacy and user experience.

CONCLUSIONS

This study provides useful insights into the relative performance of eight symptom-assessment apps, compared with each other and compared with seven tested GPs, in terms of their coverage, their suggested condition accuracy and the accuracy of their levels of urgency advice. The results show that the best performing of these apps have a high level of urgency advice accuracy which is close to that of GPs. Although not as accurate as GPs in top-1 suggestion of conditions, the best apps are close to GP performance in providing the correct condition in their top-3 and top-5 condition suggestions.

While no digital tool outperformed GPs in this analysis, some came close, and the nature of iterative improvements to software suggests that further improvements will occur with experience and additional evaluation studies.

The findings of this vignettes study on urgency advice are supportive of the use of those symptom-assessment apps, which have urgency advice safety similar to the levels achieved by GPs, in the use case of supplementing telephone triage (a use case described in Chambers *et al*[4]). The findings are also indicative of the future potential of AI-based symptom assessment technology in diagnostic decision support; however, this is an area that requires specific clinical evidence and regulatory approval. Further studies, which include direct use of the symptom assessment apps by patients, are required to confirm clinical performance and safety.

Acknowledgements Paul Taylor (UCL Institute of Health Informatics) independently reviewed and made suggestions on the study protocol, and after study data collection was complete, reviewed and made suggestions in a draft of this manuscript with respect to the analysis approach and the study description. Vignette review was carried out by the following experienced primary care physicians: Alison Grey, Helen Whitworth, Jo Leahy. Study support was provided by Linda Cen (Data Engineer, Ada Health GmbH), Leif Ahlgrimm (Student IT System Administrator, Ada Health GmbH), Neil Rooney (Senior IT System Administrator, Ada Health GmbH). Henry Hoffmann provided helpful suggestions on the final manuscript.

Contributors SG, AM, CC, JC, HF, FP, ET, VV, NV & CN contributed to the planning (study conception, protocol development). SG, AM, AB, CC, JC, FP, CR, SU, VV, NV, & CN contributed to the conduct (coordination of vignette creation, review, coordination of GP or app-testing, condition-matching panel coordination). SG, AM, CC, JC, HF, EM, MM, JM, FP, CR, ET, NV, PW & CN contributed to the data analysis & interpretation. SG, AM, JC, HF, JM, FP, CR, ET, VV, NV, PW & CN contributed to the reporting (report writing). All the authors contributed to commenting on drafts of the report. SG is the guarantor for this work.

Funding This study was funded by Ada Health GmbH. HF has not received any compensation from Ada Health financial or otherwise.

Competing interests All of the authors, with the exception of HF, are or were employees of, contractors for, or hold equity in the manufacturer of one of the tested apps (Ada Health GmbH). See author affiliations. SG, AM, AB, CC, EM, MM, JM, FP, ET, SU, NV and CN are employees or company directors of Ada Health GmbH and some of the listed hold stock options in the company. CR and VV are former employees of Ada Health GmbH. JC and PW have or have had consultancy contracts with Ada Health GmbH. The Ada Health GmbH research team has received research grant funding from Fondation Botnar and the Bill & Melinda Gates Foundation. PW has received speaker fees from Bayer and honoraria from Roche, ARISLA, AMIA, IMI, PSI, and the BMJ.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information with the exception of the case vignettes. The vignettes used in this study can be made available on request to the corresponding author SG, provided that they will be used for genuine scientific purposes, and that these purposes will not compromise their utility in future assessment of symptom assessment applications (for example, by making them publicly available, and therefore accessible to the medical knowledge learning of symptom assessment app manufacturers). They are not publicly available due to planned periodic update of the study analysis, which will be carried out by Ada Health and other independent scientific researchers, in order to monitor comparative change in app performance over time. All vignette access requests will be reviewed and (if successful) granted by the Ada Health Data Governance Board. The vignettes will not be disclosed to the Ada medical intelligence team or to other app developers.

ORCID iDs

Stephen Gilbert http://orcid.org/0000-0002-1997-1689
Maryam Montazeri http://orcid.org/0000-0003-4688-9311

REFERENCES

1 McDaid D, Park A-L. *Online health: untangling the web*, 2011.
2 Van Riel N, Auwerx K, Debbaut P, *et al*. The effect of Dr Google on doctor-patient encounters in primary care: a quantitative, observational, cross-sectional study. *BJGP Open* 2017;1:bjgpopen17X100833.
3 Semigran HL, Levine DM, Nundy S, *et al*. Comparison of physician and computer diagnostic accuracy. *JAMA Intern Med* 2016;176:1860–1.
4 Chambers D, Cantrell AJ, Johnson M, *et al*. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. *BMJ Open* 2019;9:e027743.
5 Millenson ML, Baldwin JL, Zipperer L, *et al*. Beyond Dr. Google: the evidence on consumer-facing digital tools for diagnosis. *Diagnosis* 2018;5:95–105.
6 Semigran HL, Linder JA, Gidengil C, *et al*. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ* 2015;351:h3480.
7 Moreno Barriga E, Pueyo Ferrer I, Sánchez Sánchez M, *et al*. [A new artificial intelligence tool for assessing symptoms in patients seeking emergency department care: the Mediktor application]. *Emergencias* 2017;29:391–6.
8 Greenhalgh T, Wherton J, Shaw S, *et al*. Video consultations for covid-19. *BMJ* 2020;368:m998.
9 Heymann DL, Shindo N. Who scientific and technical Advisory group for infectious hazards COVID-19: what is next for public health? *Lancet* 2020;395:542–5.
10 Converse L, Barrett K, Rich E, *et al*. Methods of observing variations in physicians' decisions: the opportunities of clinical vignettes. *J Gen Intern Med* 2015;30:586–94.
11 Evans SC, Roberts MC, Keeley JW, *et al*. Vignette methodologies for studying clinicians' decision-making: validity, utility, and application in ICD-11 field studies. *Int J Clin Health Psychol* 2015;15:160–70.
12 Veloski J, Tai S, Evans AS, *et al*. Clinical Vignette-Based surveys: a tool for assessing physician practice variation. *Am J Med Qual* 2005;20:151–7.
13 Berner ES, Webster GD, Shugerman AA, *et al*. Performance of four computer-based diagnostic systems. *N Engl J Med* 1994;330:1792–6.

14 Swaminathan S, Qirko K, Smith T, *et al*. A machine learning approach to triaging patients with chronic obstructive pulmonary disease. *PLoS One* 2017;12:e0188532.
15 Nayak BK, Hazra A. How to choose the right statistical test? *Indian J Ophthalmol* 2011;59:85–6.
16 Shan G, Gerstenberger S. Fisher's exact approach for post hoc analysis of a chi-squared test. *PLoS One* 2017;12:e0188709.
17 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B* 1995;57:289–300 https://www.jstor.org/stable/2346101?seq=1
18 Wilson EB. Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc* 1927;22:209–12.
19 Efron B, Tibshirani RJ. *An introduction to the bootstrap (Monographs on statistics and applied probability*. New York: Chapman and Hall/CRC, 1998.
20 Bisson LJ, Komm JT, Bernas GA, *et al*. How accurate are patients at diagnosing the cause of their knee pain with the help of a web-based symptom checker? *Orthop J Sports Med* 2016;4 doi:10.1177/2325967116630286
21 Burgess M. Can you really trust the medical apps on your phone? Wired UK, 2017. Available: https://www.wired.co.uk/article/health-apps-test-ada-yourmd-babylon-accuracy [Accessed 25 Mar 2020].
22 Powley L, McIlroy G, Simons G, *et al*. Are online symptoms checkers useful for patients with inflammatory arthritis? *BMC Musculoskelet Disord* 2016;17:362.
23 Pulse Today. What happened when pulse tested symptom checker apps. Available: http://www.pulsetoday.co.uk/news/analysis/what-happened-when-pulse-tested-symptom-checker-apps/20039333. article [Accessed 25 Mar 2020].
24 Nateqi J, Lin S, Krobath H, *et al*. Vom Symptom zur Diagnose – Tauglichkeit von symptom-checkern. *HNO* 2019;67:334–42.
25 Fraser H, Coiera E, Wong D. Safety of patient-facing digital symptom checkers. *Lancet* 2018;392:2263–4.
26 Jungmann SM, Klan T, Kuhn S, *et al*. Accuracy of a Chatbot (ADA) in the diagnosis of mental disorders: comparative case study with lay and expert users. *JMIR Form Res* 2019;3:e13863.
27 Winn AN, Somai M, Fergestrom N, *et al*. Association of use of online symptom Checkers with patients' plans for seeking care. *JAMA Netw Open* 2019;2:e1918561.
28 Shen C, Nguyen M, Gregor A, *et al*. Accuracy of a popular online symptom checker for ophthalmic diagnoses. *JAMA Ophthalmol* 2019;137:690–2.
29 Wiegand T, Krishnamurthy R, Kuglitsch M, *et al*. WHO and ITU establish benchmarking process for artificial intelligence in health. *Lancet* 2019;394:9–11.

Data Supplement

Supplementary Tables

Age-range	1 month to 89 years
Male / female	43.0% / 57.0%
Body-system/specialism classification	
Haematology	1.5%
Cardiovascular	7.0%
Central Nervous System - non injury	6.0%
Dental	2.0%
Endocrine	3.5%
Ear, Nose, Throat (ENT)	3.5%
Female reproductive system	7.5%
Gastrointestinal	15.0%
Immune system	2.0%
Infection	3.5%
Lower respiratory	8.0%
Male reproductive system	2.5%
Mental health	5.0%
Musculoskeletal - injury	3.5%
Musculoskeletal - non-injury	7.5%
Oncology	0.5%
Ophthalmology	3.5%

Skin	8.0%
Upper respiratory	5.0%
Urinary tract	4.5%

Supplementary Table 1

Summary of the 200 clinical vignettes: age range, sex and condition type. See also **Supplementary Table 2** for a more detailed description of each vignette and **Supplementary Table 3** for the structure of example vignettes.

#	Included in main analysis	Sex M/F	Age	System/specialism	Primary complaint	Advice level	Main-diagnosis
1	YES	F	30 years	Female reproductive system	Abdominal pain	1	ectopic pregnancy
2	NO	F	6 years	Lower Respiratory	Shortness of breath	2	chest infection
3	YES	F	9 months	Lower Respiratory	Shortness of breath	1	bronchiolitis
4	YES	M	66 years	Cardiovascular	Chest pain	1	unstable angina
5	YES	M	35 years	MSK - Injury	Back pain	4	contusion lumbar spine
6	YES	M	23 years	Mental Health	Low mood	3	depression
7	YES	M	26 years	Gastrointestinal	Vomiting	3	viral gastroenteritis
8	YES	F	40 years	CNS - Non injury	Headache	1	brain haemorrhage
9	YES	M	40 years	Skin	Swollen foot	4	cellulitis of foot
10	YES	F	28 years	Dental	Toothache	6	dental abscess
11	YES	M	1 years	Upper respiratory	Lumps in neck	4	tonsillitis
12	NO	F	26 years	Dental	Pain in face	4	dental abscess
13	YES	F	6 years	Gastrointestinal	Itchy bottom	6	pinworm infection
14	YES	F	5 months	Gastrointestinal	Crying baby	5	cow's milk protein intolerance
15	YES	F	33 years	MSK - Injury	Abdominal pain	4	musculoskeletal back pain
16	YES	F	20 years	Immune system	Dizziness	1	allergic reaction
17	YES	M	14 years	Upper respiratory	Sore throat	6	sinusitis
18	YES	M	3 months	Upper respiratory	Crying baby	4	common cold
19	NO	F	27 years	Female reproductive system	Vaginal bleeding	2	early onset of labour
20	YES	F	86 years	Gastrointestinal	Vomiting brown liquid	1	perforated gastric ulcer
21	YES	M	57 years	Gastrointestinal	Diarrhoea	6	gastroenteritis
22	YES	F	29 years	Gastrointestinal	Abdominal pain	2	appendicitis
23	YES	M	2 years	CNS - Non injury	Fever	1	bacterial meningitis

24	NO	M	59 years	ENT	Earache	4	otitis externa
25	NO	F	1 years	Infection - General	Rash	6	varicella zoster
26	NO	M	10 years	Upper respiratory	Cough	6	upper respiratory tract infection
27	YES	M	9 months	Skin	Rash	6	eczema
28	NO	F	64 years	Lower Respiratory	Cough	6	acute bronchitis
29	YES	F	17 years	Lower Respiratory	Muscle aches	4	influenza
30	YES	M	14 years	Gastrointestinal	Diarrhoea	4	appendicitis
31	YES	F	20 years	Female reproductive system	Vaginal discharge	4	threatened miscarriage
32	YES	F	42 years	Mental Health	Low mood	4	postnatal depression
33	YES	F	2 years	ENT	Sticky eye	4	otitis media
34	YES	F	57 years	Mental Health	Aggressive behaviour	3	paranoid psychosis
35	YES	F	72 years	MSK - Injury	Back pain	2	wedge fracture thoracic spine
36	YES	F	89 years	Mental Health	Confusion	3	acute confusional state
37	YES	F	49 years	Mental Health	Suicidal thoughts	3	post traumatic stress disorder
38	NO	F	1 years	Lower Respiratory	Breathing difficulties	3	bronchiolitis
39	YES	M	89 years	Lower Respiratory	Hands shaking	3	influenza
40	YES	F	2 years	Upper respiratory	Earache	6	common cold
41	NO	F	5 years	CNS - Non injury	Fever	1	bacterial meningitis
42	YES	M	68 years	Gastrointestinal	Vomiting blood	2	acute gastritis/gastric erosions
43	YES	F	41 years	Gastrointestinal	Abdominal pain	2	acute cholecystitis
44	NO	F	52 years	Skin	Rash	4	shingles
45	YES	F	7 years	CNS - Non injury	Fever	1	meningococcal septicaemia
46	YES	M	3 years	Infection - general	Blisters on hands	6	hand foot and mouth disease
47	NO	M	30 years	MSK - non injury	Pain in big toe	4	gout
48	NO	F	47 years	MSK - Injury	Back pain	6	muscle strain
49	YES	F	27 years	Upper respiratory	Sore throat	2	quinsy (peritonsillar abscess)

50	NO	M	2 years	ENT	Earache	4	acute otitis media
51	YES	M	62 years	Mental Health	Can't sleep	5	depression
52	YES	M	58 years	Cardiovascular	Chest pain	5	angina
53	NO	F	37 years	Lower Respiratory	Cough	6	acute bronchitis
54	YES	F	67 years	ENT	Dizziness	6	benign positional vertigo
55	YES	F	32 years	Cardiovascular	Shortness of breath	2	pulmonary embolism
56	YES	M	77 years	CNS - Non injury	Drooping of left side of face	1	stroke
57	NO	F	82 years	Cardiovascular	Shortness of breath	5	heart failure (i.e. congestive cardiac failure)
58	YES	F	32 years	Cardiovascular	Irregular heartbeat	5	ventricular extrasystoles
59	YES	M	28 years	CNS - Non injury	Shaking of hands	5	benign essential tremor
60	YES	F	58 years	Lower Respiratory	Cough	5	asthma
61	NO	F	27 years	CNS - Non injury	Headache	5	migraine
62	YES	M	57 years	Cardiovascular	Pain in calves	5	peripheral vascular disease
63	NO	F	41 years	CNS - Non injury	Headache	1	subarachnoid haemorrhage
64	NO	M	32 years	Cardiovascular	Fast pulse	4	paroxysmal tachyarrhythmia
65	YES	F	1 years	Lower Respiratory	Cough	4	viral bronchitis
66	YES	M	13 years	Lower Respiratory	Wheezy	5	asthma
67	YES	F	3 years	Lower Respiratory	Cough	3	pneumonia
68	YES	F	32 years	MSK - non injury	Tingling in hands	5	carpal tunnel syndrome
69	NO	M	42 years	Cardiovascular	Chest pain	1	myocardial infarction
70	NO	M	10 years	CNS - Non injury	Headache	2	intracranial haematoma
71	YES	F	27 years	MSK - non injury	Colour change in fingers	5	Raynaud’s phenomenon of the fingers
72	NO	F	26 years	Cardiovascular	Collapse	5	vasovagal attack
73	YES	M	73 years	Lower Respiratory	Cough	4	lung cancer
74	YES	F	8 years	CNS - Injury	Headache	6	minor head injury

75	YES	M	28 years	MSK - Injury	Chest pain	4	pleurisy
76	NO	F	77 years	Gastrointestinal	Abdominal pain	1	intestinal obstruction
77	YES	F	63 years	Gastrointestinal	Blood on toilet paper	4	colorectal cancer
78	NO	F	37 years	Gastrointestinal	Abdominal pain	5	irritable bowel syndrome
79	YES	M	8 years	Gastrointestinal	Abdominal pain	2	appendicitis
80	YES	F	17 years	Gastrointestinal	Vomiting	6	viral gastroenteritis
81	NO	M	1 months	Gastrointestinal	Vomiting	3	pyloric stenosis
82	NO	M	38 years	Gastrointestinal	Yellow eyes	2	viral hepatitis
83	YES	M	67 years	Gastrointestinal	Difficulty swallowing	4	oesophageal cancer
84	YES	F	68 years	Skin	Redness of leg	4	cellulitis
85	YES	M	37 years	Ophthalmology	Swollen eye	2	periorbital/orbital cellulitis
86	YES	F	63 years	MSK - non injury	Shoulder pain	5	frozen shoulder
87	YES	M	63 years	MSK - non injury	Stiff neck	6	acute torticollis
88	YES	F	82 years	MSK - non injury	Hip pain	5	osteoarthritis of hip
89	YES	M	15 years	MSK - non injury	Knee pain	5	Osgood-Shlatters disease
90	YES	F	35 years	MSK - non injury	Numbness in fingers	5	carpal tunnel syndrome
91	YES	M	50 years	MSK - non injury	Knee pain	4	infected infra-patellar bursa
92	YES	F	50 years	MSK - non injury	Wrist pain	5	tenosynovitis
93	NO	M	67 years	ENT	Discharge from ear	4	otitis externa
94	YES	M	50 years	ENT	Ringing in ear	5	tinnitus
95	YES	F	58 years	Upper respiratory	Hoarse voice	4	laryngitis secondary to acid reflux
96	YES	F	23 years	Mental Health	Hearing voices	3	acute psychosis
97	NO	F	13 years	Mental Health	Heart pounding	5	anxiety disorder
98	NO	F	47 years	Endocrine	Passing lots of urine	4	type 2 diabetes
99	YES	M	9 years	Endocrine	Vomiting	1	diabetic ketoacidosis
100	YES	F	30 years	Endocrine	Sweating	4	Graves' disease

101	NO	F	59 years	Endocrine	Tiredness	5	hypothyroidism
102	YES	M	41 years	Endocrine	Can't get erection	5	pituitary tumour (prolactinoma)
103	YES	F	42 years	Endocrine	Swelling under jaw	5	salivary gland stone
104	YES	F	39 years	Endocrine	Weight gain	4	adrenal adenoma causing Cushing's syndrome
105	YES	M	61 years	Infection - general	Fever	2	malaria
106	YES	F	37 years	Urinary Tract	Painful urination	4	urinary tract infection
107	NO	M	45 years	Ophthalmology	Red eyes	6	viral conjunctivitis
108	YES	F	33 years	Female reproductive system	Abdominal pain	2	ectopic pregnancy
109	YES	F	6 years	Urinary Tract	Abdominal pain	3	urinary tract infection
110	NO	M	26 years	Urinary Tract	Pain in side	2	renal stones
111	YES	M	33 years	Ophthalmology	Red painful eye	3	corneal ulcer
112	YES	M	15 years	Male reproductive system	Testicular pain	2	testicular torsion
113	YES	F	42 years	Female reproductive system	Vaginal discharge	6	vaginal thrush
114	YES	M	57 years	Urinary Tract	Blood in urine	4	urinary tract infection
115	YES	F	14 years	Female reproductive system	Heavy periods	5	menorrhagia
116	YES	F	26 years	Female reproductive system	Vaginal bleeding	6	threatened miscarriage
117	YES	M	79 years	Male reproductive system	Passing urine at night	5	benign prostatic hyperplasia
118	YES	M	27 years	Ophthalmology	Dry eyes	6	dry eye syndrome
119	YES	M	2 years	Ophthalmology	Sticky red eye	6	bacterial conjunctivitis
120	YES	M	19 years	Male reproductive system	Testicular lump	4	hydrocoele
121	YES	F	45 years	Female reproductive system	Vaginal bleeding	4	vaginal atrophy
122	YES	F	62 years	Female reproductive system	Bloating	4	ovarian pathology (cancer)

123	YES	F	56 years	Female reproductive system	Hot flush	5	menopause
124	YES	M	7 years	Gastrointestinal	Constipation	6	constipation
125	NO	F	13 years	Immune system	Sneezing	5	allergic rhinitis
126	YES	M	29 years	Cardiovascular	Palpitations	5	ectopics secondary to caffeine intake
127	YES	F	19 years	Mental Health	Low mood	5	depression
128	YES	M	38 years	Lower Respiratory	Phlegm with blood	4	carcinoma of lung
129	YES	M	32 years	Infection - general	High fever	2	malaria
130	YES	M	20 years	Skin	Rash	4	eczema
131	YES	F	34 years	Skin	Pimples on face	5	acne
132	YES	F	3 years	Infection - general	Rash	6	varicella zoster
133	YES	M	27 years	Immune system	Itchy rash	6	acute urticaria
134	YES	M	33 years	Skin	White toenail	5	fungal nail infection
135	YES	F	37 years	Blood	Tiredness	5	iron deficiency anaemia
136	YES	F	11 years	Blood	Lump in neck	4	lymphoma
137	NO	M	43 years	Blood	Bruising	4	clotting disorder
138	NO	F	33 years	Cardiovascular	Chest pain	3	pulmonary embolism
139	YES	M	40 years	Gastrointestinal	Stomach pain	3	diverticulitis
140	YES	M	50 years	Gastrointestinal	Stomach pain	1	diverticular abscess
141	YES	M	64 years	Male reproductive system	Difficulty passing urine	4	benign prostatic hyperplasia
142	YES	F	64 years	Oncology	Lump in breast	4	breast cancer
143	NO	F	38 years	CNS - Non injury	Headache	5	tension headache
144	NO	F	25 years	Female reproductive system	Tummy bloating before period	5	premenstrual syndrome
145	NO	F	35 years	Gastrointestinal	Abdominal pain	5	irritable bowel syndrome
146	YES	F	36 years	Skin	Hair falling out	5	alopecia
147	YES	F	39 years	Skin	Flaky patches of skin	5	psoriasis

148	YES	F	3 years	Urinary Tract	Crying on weeing	3	urinary tract infection
149	YES	F	3 years	Lower Respiratory	Cough	3	pneumonia
150	NO	M	3 years	Male reproductive system	End of penis red	4	balanitis
151	YES	F	12 years	Skin	Rash	4	scabies
152	YES	F	32 years	MSK - non injury	Heel pain	5	plantar fasciitis
153	YES	M	30 years	MSK - non injury	Shin pain	6	shin splints
154	YES	M	40 years	CNS - Non injury	Back pain	1	cauda equina syndrome
155	YES	F	30 years	Female reproductive system	Irregular periods	5	polycystic ovary syndrome
156	NO	F	12 years	MSK - non injury	Hip pain	4	slipped upper femoral epiphysis
157	YES	M	21 years	Lower Respiratory	Shortness of breath	1	spontaneous simple pneumothorax
158	YES	F	70 years	Lower Respiratory	Shortness of breath	4	asbestosis
159	YES	F	33 years	Cardiovascular	Chest pain	2	pulmonary embolism
160	YES	F	47 years	Cardiovascular	Lump in lower leg	4	thrombophlebitis
161	NO	F	4 years	Upper respiratory	Runny nose	6	upper respiratory tract infection
162	YES	M	10 years	Skin	Rash on chin	4	impetigo
163	YES	M	40 years	Urinary Tract	Pain in left side	2	renal calculus
164	NO	M	62 years	Skin	Cracked skin between toes	6	athlete's foot
165	YES	M	52 years	Gastrointestinal	Blood on toilet paper	5	internal haemorrhoids
166	YES	M	39 years	ENT	Blocked nose	5	nasal polyps
167	NO	F	21 years	MSK - injury	Swollen ankle	4	ankle sprain
168	YES	F	11 years	MSK - non injury	Back ache	5	idiopathic scoliosis
169	YES	M	18 years	MSK - non injury	Knee pain	2	dislocated patella
170	YES	M	24 years	Skin	Rash on hands and arms	4	contact dermatitis
171	NO	F	2 years	Ophthalmology	Sticky eyes	6	viral conjunctivitis

172	YES	M	16 years	Immune system	Runny nose	6	allergic rhinitis
173	YES	M	68 years	Urinary tract	Blood in urine	4	bladder cancer
174	YES	F	36 years	Skin	Rash under breasts	4	candidal intertrigo
175	YES	F	32 years	Female reproductive system	Bleeding after sex	4	cervical cancer
176	YES	F	27 years	CNS - Non injury	Headaches	5	cluster headaches
177	NO	M	23 years	Gastrointestinal	Diarrhoea	4	coeliac disease
178	YES	M	6 months	Skin	Flaky scalp	6	cradle cap
179	YES	M	10 months	Upper respiratory	Difficulty breathing	3	croup
180	NO	F	28 years	Female reproductive system	Abdominal pain	5	endometriosis
181	YES	F	23 years	Female reproductive system	Blisters on vulva	4	genital herpes
182	YES	F	10 months	Gastrointestinal	Crying baby	3	constipation
183	YES	M	2 years	Upper respiratory	Cough	6	common cold
184	YES	F	5 years	Infection - general	Rash	4	rubella
185	NO	M	50 years	Dental	Toothache	4	dental abscess
186	NO	M	36 years	Ophthalmology	Blurred vision	2	corneal ulcer
187	NO	F	6 months	Gastrointestinal	Diarrhoea	3	viral gastroenteritis
188	YES	M	27 years	Gastrointestinal	Abdominal pain	1	acute pancreatitis
189	NO	M	2 months	Urinary Tract	Vomiting	3	urinary tract infection
190	YES	F	7 months	Gastrointestinal	Vomiting	2	viral gastroenteritis
191	YES	M	9 years	Dental	Toothache	5	dental caries
192	YES	F	33 years	Mental Health	Worried about pregnancy	4	anxiety
193	YES	M	24 years	Infection - general	Fever	4	glandular fever
194	YES	F	4 months	Gastrointestinal	Diarrhoea	4	viral gastroenteritis
195	YES	F	15 years	MSK - injury	Back pain	5	mechanical low back pain
196	YES	F	22 years	Urinary Tract	Pain on passing urine	3	urinary tract infection
197	YES	F	10 months	Gastrointestinal	Vomiting	4	viral gastroenteritis

198	NO	F	4 years	Cardiovascular	Fever	3	Kawasaki disease
199	YES	F	60 years	Skin	Crusted patch of skin	4	basal cell carcinoma
200	YES	F	56 years	Gastrointestinal	Pain in groin	4	femoral hernia

Supplementary Table 2

Individual details of the 200 clinical-vignettes: age range, sex, condition type (system/specialism), level of urgency advice, and main diagnosis. See also **Supplementary Table 1** for a summary.

Category	Example vignette #79 (see Supplementary Table 2)	Example vignette #86 (see Supplementary Table 2)
Sex	male	female
Age	8 years old	63 years old
Previous medical history	None of note.	None of note.
Primary complaint	abdominal pain	shoulder pain
Additional information on primary complaint and current symptoms	abdominal pain and fever. Temp 39.5C. Started to get pain in the middle of his abdomen last night before he went to bed (6 hours ago). Had some infant paracetamol suspension and went to sleep. Now he has woken up and is crying with pain. Feels feverish. Says his tummy hurts on the right side now. Severity 8/10. Tummy hurts to touch. Feels hot. Looks flushed. Constant pain - no radiation.	painful shoulder. Unable to move it for 24 hours. Onset of pain in right shoulder about 6 months ago. Gradually got worse. In the past week it has been much worse and now unable to move the shoulder. Right shoulder appears higher than the other. Background ache. Severe pain if she tries to move the shoulder.
If asked	location of pain: Right lower abdomen Character: constant Diarrhoea: NO constipation - NO blood in stools - NO mucus - NO urinary or testicular symptoms - NO vomiting - NO dehydration - NO fever - YES dry Mouth - NO	history of injury to the shoulder: NO redness of the skin over the shoulder: NO painful to touch: NO deformity: NO numbness, tingling or pain in the arm: NO weakness of the arm: NO – just can't lift it up because of pain. does the shoulder feel as if it is coming out of socket?: NO
Progression	getting worse	getting worse
Levels of urgency advice	**see primary care within 4 hours**	**see primary care non-urgent**
Main diagnosis	**appendicitis**	**frozen shoulder**
Differential diagnoses (not ordered)	**urinary tract infection** **mesenteric adenitis** **inguinal hernia** **testicular torsion** **intestinal malrotation** **ureteric colic**	**osteoarthritis of shoulder** **bursitis of shoulder** **shoulder impingement syndrome** **dislocated shoulder**

Body-system/ specialism classification	**gastrointestinal**	**musculoskeletal - non-injury**

Supplementary Table 3

Two example vignettes showing the vignette structure. The items in grey shading are the vignette gold-standard answer and vignette body-system/specialism classification.

Name of symptom assessment app	Provider	Platform addressed in this study	Regional availability	Questions on pre-existing factors/ attributes	Number of questions
Ada	Ada Health GmbH, Berlin, Germany	App	Available everywhere.	-Gender -Date of birth -Pregnant -Smoker -High blood pressure -Diabetes	Variable
Babylon	Babylon Healthcare Services Ltd., London, UK	App	Services provided in compliance with UK law and regulation-- outside of the UK, user must check if lawful. Not all services available outside of UK **UK version tested in this study.**	-Gender -Date of birth	Variable
Buoy	Buoy Health, Inc., Boston, MA, USA	Website (Note - the term 'app' is loosely used in this study as a generic term for symptom assessment tools - strictly Buoy is not an app, but instead a website, though it can be put onto phone homescreens	For US residents **US version tested in this study.**	-Gender -Age (between 2-120 years)	Variable
K Health	K Health Inc., New York, NY, USA	App	Services available outside the US & Israel but are solely directed to those in the US and/or Israel. **US version tested in this study.**	-Gender -Age -Chronic conditions relevant to symptoms	Around 20 questions but variable
Mediktor	Teckel Medical S.L., Barcelona, Spain	App	Available everywhere	-Gender -Age	Variable
Symptomate	Infermedica, Wrocław, Poland	App	Available everywhere	-Gender -Age (Minimum age 18) -Overweight -Smoker	19 after region where live

				-Recently injured -High cholesterol -Hypertension -Diabetes	
Your.MD	YOUR.MD AS Oslo, Norway and subsidiary: Your.MD Ltd. London, UK	App	Available everywhere	-Gender -Year of birth (-Minimum age 16) -Asthma -Diabetes -Smoker	Variable
WebMD	WebMD, LLC Atlanta, GA, USA	App (website also available)	Available everywhere	-Gender -Age -Past medical history -Medications	4 (excluding age and gender)

Supplementary Table 4

Summary of the 8 apps evaluated in this study

Conditions should be considered a reasonable match when:
- the suggested condition is the same as the gold standard (GS).
- alternative names for the same condition are used.
- the suggested condition is a more precise description of the gold standard condition.
- the condition is an umbrella term including the other condition.
- it is reasonable to assume two different doctors might use the two different descriptions to label the same conditions.
- one condition causes the other (directly and explicitly)
- the suggested condition conveys the nature of the GS condition, is reasonably related to the GS condition (but is less precise) - similarity is so clear that in primary care medicine practice, the conditions would be considered a near match.
- the suggested condition is both highly related to the GS condition and shares symptoms to a high degree with the GS condition.

Supplementary Table 5

Matching criteria that were considered by panel physicians in deciding if condition-suggestions (from GPs or from apps) match the gold-standard diagnoses. Note - the matching criteria apply equally to all apps and to the tested-GPs.

Levels of urgency advice	**all vignettes for K Health (200)**	**150 vignettes with source verified data for K Health**	**all vignettes for Your.MD (200)**	**150 vignettes with source verified data for Your.MD**
no advice given (%) [not recorded]	32.0 [12.0]	33.3	32.5	38.7
potentially unsafe (%)	10.5	14.0	5.0	4.7
safe but underconservative (%)	18.0	23.3	5.0	4.0
exactly matched (%)	17.0	20.0	22.0	23.3
overconservative but suitable (%)	6.0	6.7	14.0	15.3
overconservative (%)	4.5	2.7	21.5	14.0
statistical testing (%) Fisher's exact test for significant difference between the urgency advice allocation	***p*-value**	0.59 (no significant difference)	***p*-value**	0.69 (no significant difference)
% of safe advice (where safe is defined as maximum one level less conservative than gold-standard, expressed per vignette provided with advice)**.**	81.3*	79.0	92.6	92.3

Supplementary Table 6

Urgency advice subanalysis for K Health and Your.MD. One app-entry-Dr (#4) did not record complete and source data verifiable data for K Health and Your.MD. For K Health, 12.0% of urgency advice was not recorded by app-entry-Dr (#4) and no screenshot was saved. For Your.MD, all urgency advice was recorded but in 25% of vignettes, screenshots were not saved for source data verification. A subanalysis was conducted to show the K Health and Your.MD performance in the fully verified 150 vignettes and to allow comparison to their performance in the full set of 200 vignettes. There are some differences in Your.MD's performance in the 200-vignette and in the 150-vignette analyses, but these are not statistically significant. It is not possible to determine what proportion of the differences present was due to unrecorded or unverified data, and what proportion was due to random differences between the 150 and 200 vignettes. * - in the case of K Health, % of safe advice (where safe is defined as maximum one level less conservative than gold-standard, expressed per vignette with available advice).

Condition-suggestion performance	**all vignettes for WebMD (200)**	**150 vignettes with source verified data for WebMD**	**all vignettes for Your.MD (200)**	**150 vignettes with source verified data for Your.MD**
M1 (%)	21.0	25.3	21.5	18.7
M3 (%)	35.5	38.0	23.5	20.0
M5 (%)	43.5	45.3	23.5	20.0
COMP. (%)	38.9	42.5	14.3	12.0
RELE. (%)	33.4	34.8	40.2	35.1
statistical testing (%) Fisher's exact test for significant difference between the urgency advice allocation	***p*-value**	0.88 (no significant difference)	***p*-value**	0.88 (no significant difference)

Supplementary Table 7

Condition-suggestion subanalysis for WebMD and Your.MD. One app-entry-Dr (#4) did not record complete source data verification screenshots for WebMD or for Your.MD and therefore verification of entered data was not possible for 50 out of 200 of the randomly assigned vignettes for these apps. A subanalysis was conducted to show the performance of WebMD and Your.MD in the fully verified 150 vignettes and to allow comparison to their performance in the full set of 200 vignettes. There are some differences in performance between the 200-vignette and the 150-vignette analyses, but these are not statistically significant. It is not possible to determine what proportion of small differences present were due data entry error in the 50-vignettes, and what proportion was due to random differences between the 150 and 200 vignettes.

Scenarios for which there was no urgency advice							
	Under the age limit[1] % (specified age limit)	Pregnant %	Symptom Not Found [2] %	No condition-suggestions [3] %	Not Recorded[4] %	Mental health[5] %	Total for which no urgency advice (i.e. 100% - breadth of coverage
Ada	0	0	0	0	0	0	0
Babylon	1.0 (16+)*	3.0	5.0	0	0	0	9.0
Buoy	9.5 (13+)	0	0.5	0	0	0	10.0
K Health	31.0 (18+)	0	1.0	0	12.0	0	44.0
Mediktor	2.5 (1+)	0	0	13.5	0	1.5	17.5
Symptomate	31.0 (18+)	0	0.5	0	0	0	31.5
Your.MD	23.0 (16+)	0.5	2.0	5.5	0	0	31.0

Supplementary Table 8

Coverage of each app and reasons why apps did not provide condition-suggestions for some vignettes.

* - Babylon provides levels of urgency advice to these patients such as "go to the nearest ED", but it does not provide condition-suggestions for patients under 16 years of age.

[1] - The app did not provide a condition-suggestion for those under its specified age limit.

[2] - A matching symptom or appropriately matching symptom to the presenting complaint could not be found in the app, so the vignette could not be entered.

[3] - No condition-suggestions were provided by the app, and, as a result of this, the app did not provide urgency advice.

[4] - One app-entry-Dr (#4) did not fully record the levels of urgency advice, and there were no corresponding source data verification screenshots for this subset of data. See **Supplementary Table 6** for a subanalysis of the 150 vignettes with complete source-data-verified data for K Health on levels of urgency advice.

[5] - These were mental health scenarios where no level of urgency advice was provided. Note: for the other apps, where question flow was stopped due to mental health, the level of urgency advice of 'seek emergency care' was none-the-less provided.

Scenarios for which there were no condition-suggestions								
	Under the age limit[1] % (specified age limit)	Pregnant %	Symptom Not Found[2] %	Condition Severity[3] %	Generic[4] %	Clear Statement No Results[5] %	Mental health [6] %	Total for which no condition suggestions (i.e. 100% - breadth of coverage)
Ada	0	0	0	0.5	0	0	0.5	1.0
Babylon	25.5 (16+)	2.5	4.0	8.5	6.0	0.5	1.5	48.5
Buoy	8.5 (13+)	0	1.0	0.5	0	0	1.5	11.5
K Health	23.5 (18+)	0	2.0	0	0	0	0	25.5
Mediktor	4.0 (1+)	0	1.0	1.0	0	11.5	2.0	19.5
Symptomate	31.5 (18+)	1.5	0	0	0	5.0	0.5	38.5
Your.MD	23 (16+)	1.0	0.5	0.5	4.0	5.5	1.0	35.5
WebMD	5.0	0	0.5	0	0	1.5	0	7.0

Supplementary Table 9

The breadth of condition-suggestion coverage is shown for each app along with the reasons why apps did not provide condition-suggestions for some vignettes.

[1] - The app does not provide a condition-suggestion for users under a specified age limit.

[2] - A directly or appropriately matching symptom to the presenting complaint could not be found in the app, so the vignette could not be entered.

[3] - The app did not to give condition-suggestions for very serious symptoms - e.g. the app stated only 'Condition causing severe [*symptom*]'

[4] - The app gave a generic answer rather than a condition, e.g. "further assessment is needed"

[5] - The app provided a clear statement that no condition-suggestion results were found for the vignette. The reason why the app failed to give a condition-suggestion for these vignettes is uncertain, but generally these vignettes relate to minor conditions, and in most cases, it seems that the app does not have a matching condition modelled.

[6] - These were mental health scenarios where no conditions were provided at all

Classification of reason	App	Number of vignettes for which this was the reason for no condition-suggestion
Mental Health "If she is struggling with suicidal thoughts, please stay calm, don't leave her alone and make sure to look for professional help right now…(**Ada**)" "I'm sorry you're having thoughts of self-harm or suicide. I'm stopping your assessment as this is beyond my capability…(**Babylon**)" "Some of the symptoms you reported might need to be checked out by a GP within the next 6 hours…(**Babylon**)" "When experiencing thoughts of suicide, she should go immediately to the ER…(**Buoy**)" "Seeing or hearing things is a sign of a possible mental health issue. She should…(**Buoy**)" "Mental or medical condition causing hallucinations. (**Buoy**)" "Go to the hospital right away. It is very likely you need medical care. (**Mediktor**)" "Thank you for telling me, Guest. As you know, I am not a doctor, but there are people that can help you with these thoughts…(**Your.MD**)" "[Symptom] are worrying symptoms. You should see a doctor within 48 hours…(**Your.MD**)"	Ada	1
	Babylon	3
	Buoy	3
	Mediktor	4
	Symptomate	1
	Your.MD	2
Generic "Your symptoms require further medical assessment. (**Babylon**)." "This needs looking into…(**Babylon**)" "Range of possible causes…(**Babylon**)" "Symptoms can be managed at home…(**Babylon** - note condition not named)" "Ok guest, here is some information you might find helpful: [symptom/related condition information] (**Your.MD**)"	Babylon	12
	Your.MD	8
Emergency/ very severe "Call the local emergency number for a life-threatening medical emergency (**Ada**)" "Severe chest pain…[Symptom] might be a sign of something serious...(**Buoy**)" "Condition causing…[serious symptom] (**Buoy**)" "[Symptoms] are worrying symptoms. You should see a doctor within 48 hours (**Your.MD**)" "Symptoms should not be ignored (**Babylon**)" "Could be signs of something serious…(**Babylon**)"	Ada	1
	Buoy	18
	Your.MD	1
	Babylon	16

Clear statement of no results "No results (**Symptomate**)" "I wasn't able to find any causes for your symptoms…(**Your.MD**)" "No prediagnosis was reached. Please retry the assessment describing it differently or check with a health professional (**Mediktor**)" "No conditions found, try adding more symptoms (**WebMD**)" "The symptoms you describe are quite complex…(**Babylon**)"	Symptomate	10
	Your.MD	11
	Mediktor	23
	WebMD)	3
	Babylon	1

Supplementary Table 10

Detailed analysis of why apps did not provide condition-suggestions for some vignettes

	GP mean			App mean			Ada			Babylon		
Vignette group	**All (200)**	**NHS (64)**	**Non-NHS (136)**	**All (200)**	**NHS (64)**	**Non-NHS (136)**	**All (200)**	**NHS (64)**	**Non-NHS (136)**	**All (200)**	**NHS (64)**	**Non-NHS (136)**
Condition-suggestion performance												
M1 (%)	71.2±5.6	62.7±7.8	77.7±2.8	26.1±8.9	23.2±6.4	27.5±10.4	48.5 (41.7-55.4)	39.1 (28.1-51.3)	52.9 (44.6-61.1)	21.5 (16.4-27.7)	18.8 (11.1-30.0)	22.8 (16.5-30.5)
M3 (%)	82±5.2	75.2±5.9	86.9±5.3	38.0±13.5	35.9±9.5	39.0±15.6	70.5 (63.8-76.4)	57.8 (45.6-69.1)	76.5 (68.7-82.8)	32.0 (25.9-38.8)	29.7 (19.9-41.8)	33.1 (25.7-41.4)
M5 (%)	82.8±4.7	76.6±5.5	87.4±5.0	40.8±15.2	39.6±12.6	41.3±16.8	77.5 (71.2-82.7)	67.2 (55.0-77.4)	82.4 (5.1-7.8)	32.5 (26.4-39.3)	29.7 (19.9-41.8)	33.8 (26.4-42.1)
COMP. (%)	45.9±5.0	48.0±6	44.9±4.9	26.1±9.1	26.5±9.7	25.9±9.0	44.0 (40.7-47.4)	43.5 (36.6-50.4)	44.2 (40.4-47.9)	18.7 (15.4-21.8)	15.1 (9.2-20.5)	20.3 (16.4-24.1)
RELE. (%)	69±9.0	64.9±6.9	70.3±11.2	36.2±7.5	32.5±6.8	37.9±7.9	45.9 (42.4-49.6)	38.4 (32.3-44.5)	49.5 (45.1-53.8)	28.3 (23.5-32.9)	22.4 (14.0-30.2)	31.0 (25.0-36.8)
Urgency-advice performance												
no advice given (%) [not recorded]	0.9±0.6	2.0±1.5	0.5±1.0	20.6±15.8	19.6±16.1	21.1±15.9	0.0	0.0	0.0	9.0	9.4	8.8
potentially unsafe (%)	2.9±2.5	3.3±3.4	2.4±1.9	7.6±5.8	10.5±6.1	6.2±5.7	3.0	6.3	1.5	4.5	7.8	2.9
safe but underconservative (%)	10.8±5.9	7.1±3.5	13.0±7.2	12.9±5.0	10.3±5.6	14.1±6.6	12.0	7.8	14.0	13.5	6.3	16.9
exactly matched (%)	57.6±8.0	48.2±11.3	63.5±8.3	25.9±10.2	24.6±7.9	26.6±11.5	33.0	32.8	33.1	40.0	34.4	42.6
overconservative but suitable (%)	15.2±6.3	17.6±7.1	14.6±10.1	16.3±7.2	13.8±4.7	17.4±8.7	24.5	20.3	26.5	16.5	14.1	17.6
overconservative (%)	12.6±7.7	21.0±10.4	5.4±7.7	16.7±8.3	21.2±9.5	14.6±8.5	27.5	32.8	25.0	16.5	28.1	11.0
% of safe advice	97.0±2.5	96.6±3.5	97.6±1.9	90.1±7.4	86.5±7.9	91.8±7.2	97.0	93.8	98.5	95.1	91.4	96.8

	Buoy			K Health			Mediktor			Symptomate		
Vignette group	**All (200)**	**NHS (64)**	**Non-NHS (136)**	**All (200)**	**NHS (64)**	**Non-NHS (136)**	**All (200)**	**NHS (64)**	**Non-NHS (136)**	**All (200)**	**NHS (64)**	**Non-NHS (136)**
Condition-suggestion performance												
M1 (%)	28.0 (22.2-34.6)	18.8 (11.1-30.0)	32.4 (25.1-40.6)	26.0 (20.4-32.5)	23.4 (14.7-35.1)	27.2 (20.4-35.2)	23.5 (18.2-29.8)	25.0 (16.0-36.8)	22.8 (16.5-30.5)	19.0 (14.2-25)	18.8 (11.1-30.0)	19.1 (13.4-26.5)
M3 (%)	43.0 (36.3-49.9)	35.9 (25.3-48.2)	46.3 (38.2-54.7)	36.0 (29.7-42.9)	39.1 (28.1-51.3)	34.6 (27.1-42.9)	36.0 (29.7-42.9)	39.1 (28.1-51.3)	34.6 (27.1-42.9)	27.5 (21.8-34.1)	28.1 (18.6-40.1)	27.2 (20.4-35.2)
M5 (%)	43.0 (36.3-49.9)	35.9 (25.3-48.2)	46.3 (38.2-54.7)	39.0 (32.5-45.9)	42.2 (30.9-54.4)	37.5 (29.8-45.9)	36.5 (30.1-43.4)	39.1 (28.1-51.3)	35.3 (27.8-43.6)	30.5 (24.5-37.2)	29.7 (19.9-41.8)	30.9 (23.7-39.1)
COMP. (%)	27.9 (24.5-31.1)	28.8 (22.5-35.1)	27.4 (23.6-31.1)	27.5 (24-31)	30.7 (23.5-37.5)	26.1 (22.1-30)	28.1 (24.3-31.7)	29.9 (22.9-36.7)	27.2 (22.7-31.6)	16.3 (13.3-19.2)	15.9 (10.4-20.9)	16.5 (12.9-20.0)
RELE. (%)	47.5 (42.3-52.7)	42.7 (33.6-51.6)	49.8 (43.4-56.2)	38.6 (33.7-43.6)	35.5 (27.1-43.7)	40.1 (33.7-46.2)	33.3 (28.9-37.5)	29.6 (22.6-36.3)	35 (29.4-40.3)	28.7 (23.6-33.8)	26.9 (17.8-35.5)	29.6 (23-35.7)
Urgency-advice performance												
no advice given (%) [not recorded]	10.0	10.9	9.6	44.0	46.9	42.6	17.5	12.5	19.9	31.5	32.8	30.9
potentially unsafe (%)	18.0	21.9	16.2	10.5	12.5	9.6	10.5	14.1	8.8	1.5	4.7	0.0
safe but underconservative (%)	20.0	12.5	23.5	18.0	12.5	20.6	12.0	18.8	8.8	9.5	12.5	8.1
exactly matched (%)	34.0	28.1	36.8	17.0	15.6	17.6	24.0	21.9	25.0	11.5	14.1	10.3
overconservative but suitable (%)	8.5	10.9	7.4	6.0	6.3	5.9	23.0	18.8	25.0	21.5	12.5	25.7
overconservative (%)	9.5	15.6	6.6	4.5	6.3	3.7	13.0	14.1	12.5	24.5	23.4	25.0
% of safe advice	80.0	75.4	82.1	81.3	76.5	83.3	87.3	83.9	89.0	97.8	93.0	100.0

	WebMD			Your.MD		
Vignette group	**All (200)**	**NHS (64)**	**Non-NHS (136)**	**All (200)**	**NHS (64)**	**Non-NHS (136)**
Condition-suggestion performance						
M1 (%)	21.0 (15.9-27.2)	20.3 (12.3-31.7)	21.3 (15.3-28.9)	21.5 (16.4-27.7)	21.9 (13.5-33.4)	21.3 (15.3-28.9)
M3 (%)	35.5 (29.2-42.3)	32.8 (22.6-45)	36.8 (29.1-45.1)	23.5 (18.2-29.8)	25.0 (16.0-36.8)	22.8 (16.5-30.5)
M5 (%)	43.5 (36.8-50.4)	48.4 (36.6-60.4)	41.2 (33.3-49.6)	23.5 (18.2-29.8)	25.0 (16.0-36.8)	22.8 (16.5-30.5)
COMP. (%)	32.2 (28.4-35.8)	33.2 (26.7-39.7)	31.7 (27.2-36.1)	13.9 (11.5-16.2)	14.9 (10.7-19.1)	13.4 (10.5-16.1)
RELE. (%)	27.3 (24.0-30.5)	25.7 (20.3-30.9)	28.1 (23.9-32.1)	39.7 (33.3-45.9)	38.7 (27.7-49.5)	40.1 (32.5-47.8)
Urgency-advice performance						
no advice given (%) [not recorded]	n.d.	n.d.	n.d.	32.5	25.0	36.0
potentially unsafe (%)	n.d.	n.d.	n.d.	5.0	6.3	4.4
safe but underconservative (%)	n.d.	n.d.	n.d.	5.0	1.6	6.6
exactly matched (%)	n.d.	n.d.	n.d.	22.0	25.0	20.6
overconservative but suitable (%)	n.d.	n.d.	n.d.	14.0	14.1	14.0
overconservative (%)	n.d.	n.d.	n.d.	21.5	28.1	18.4
% of safe advice	n.d.	n.d.	n.d.	92.6	91.7	93.1

Supplementary Table 11

Condition-suggestion and urgency-advice performance for all apps for the complete set of 200 vignettes, and subanalyses for the 32.0% of vignettes sourced from NHS Direct and the 68.0% of vignettes created by the vignette creation team (95% confidence intervals are shown in brackets). The mean app and mean GP performance (± S.D.) are also shown for each set of vignettes. n.d. – not determined (this applies to WebMD which was not included in the urgency-advice comparison).

Appendix S1:

Comparison of the vignettes to UK health statistics

We explored the demographic properties and frequency of health conditions of the vignettes to investigate how our results would generalize beyond the artificially generated vignette patient, population and spectrum of conditions (see main manuscript section on vignette creation for details). The age and sex distribution of the vignettes was compared to the latest available data UK population [1]. The age and sex distribution of the chosen vignettes corresponds broadly to that of the UK population (see **Supplementary Fig. 2).** Vignettes with young adults as subjects are proportionately more frequent than young adults in the UK population, but those groups constitute the current main user group for digital symptom assessment tools and there is therefore good justification for their frequent inclusion in vignettes. Vignettes with young children as subjects are also proportionately more frequent than young children in the UK population, but young children are a high health dependency population, where the use of a symptom assessment application by the caregiver for advice is a highly relevant use case, and a use case in which condition suggestion and urgency advice accuracy is highly important. Over-representation of this demographic in the vignettes is therefore justified. Importantly, the vignettes sample the whole range of ages found in the UK population.

To compare the frequency of health conditions found in the vignettes to that in the English population, we compared the relative frequency of condition groups of the vignettes to the prevalence of condition groups reported by Quality and Outcomes Framework (QOF) for the NHS in England (Quality and Outcomes Framework) [2]. The QOF reports the prevalence of conditions belonging to six groups: (i) cardiovascular conditions (atrial fibrillation, cardiovascular disease primary prevention, coronary heart disease, heart failure, hypertension, peripheral arterial disease, as well as stroke and transient ischaemic attack); (ii) respiratory conditions (asthma, COPD); (iii) lifestyle related conditions (obesity, smoking); (iv) high dependency and long term care conditions (cancer, chronic kidney disease, diabetes mellitus, palliative care); (v) mental health and neurological conditions (dementia, depression, epilepsy, learning disabilities, mental health); and, (vi) musculoskeletal conditions (osteoporosis, rheumatoid arthritis). The condition

classification system adopted by the QOF differs substantially to the classification system used to subdivide the vignettes (see **Supplementary Table 1)**, and some of the QOF groups were not explicitly included in specific vignettes as they overlap other QOF groups e.g. lifestyle related conditions such as obesity and smoking related conditions were not represented by individual vignette main diagnoses, but these were relevant to the clinical history provided in some vignettes. Therefore, in order to compare the condition coverage of the vignettes to the QOF data, we adapted the condition groups of the QOF and matched them to the vignette condition groups (as described in **Supplementary Table 1**). The distribution of the condition categories among the vignettes shows the vignette creation goal, of inclusion of a diverse range of conditions in the vignettes spanning the conditions seen by GPs, was achieved (see **Supplementary Fig. 3)**. With the exception of cardiovascular conditions, the proportion of condition categories found among the vignettes and among the English NHS population are broadly similar. Although there is a mismatch between the relative frequency of cardiovascular conditions in the vignettes and in the English NHS population, it should be noted that there are a broad range of conditions included this QOF category (e.g. hypertension, and primary prevention for cardiovascular disease). It should also be noted that the proportions reported by the QOF are not necessarily comparable to an *ideal* proportion of vignettes - the QOF reports condition prevalence based on all patients who fulfil the requirements for the condition list in the respective year of the report, whereas new occurrences of a disease are more likely to lead to the use of a digital symptom assessment application. More generally, population morbidity data collected at doctors' practices based on confirmed diagnoses will always lack the unknown proportion of un- or misdiagnosed conditions which may be picked up by digital symptom assessment applications. It is therefore not necessarily optimal to their intended evaluation goal, for vignettes used to assess symptom assessment applications to have a perfect match in condition frequency to the QOF prevalence data.

References

1. Overview of the UK population—Office for National Statistics. (2019). Retrieved October 15, 2020, from https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/overviewoftheukpopulation/august2019.
2. Quality and Outcomes Framework. (2019-20). NHS Digital. Retrieved October 15, 2020, from https://digital.nhs.uk/data-and-information/publications/statistical/quality-and-outcomes-framework-achievement-prevalence-and-exceptions-data/2019-20.

Supplementary Figures

Supplementary Figure 1

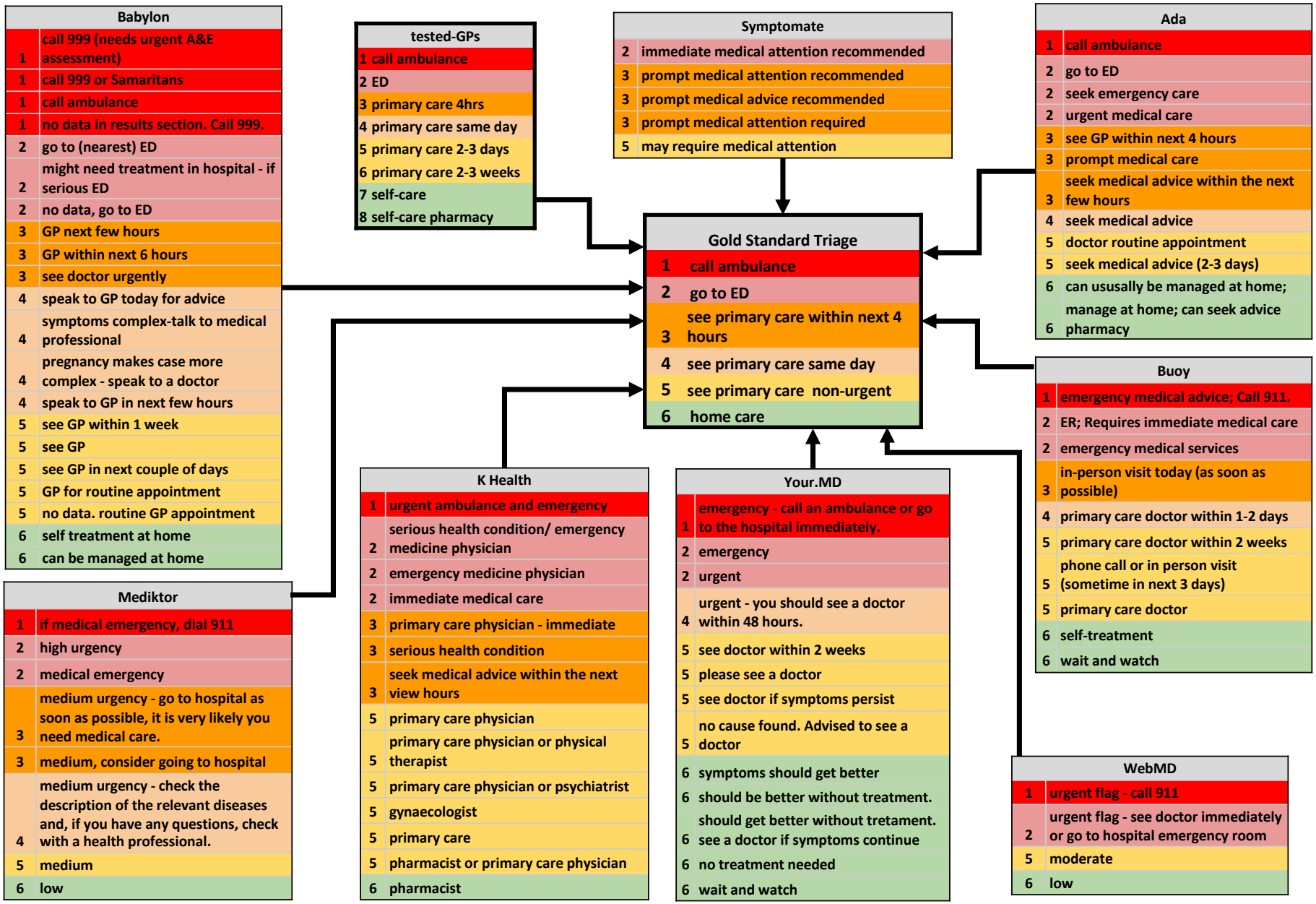
Mapping between the native levels of urgency advice provided by each app and tested-GP and the study defined common levels of urgency advice (as shown in main manuscript **Table 1**).

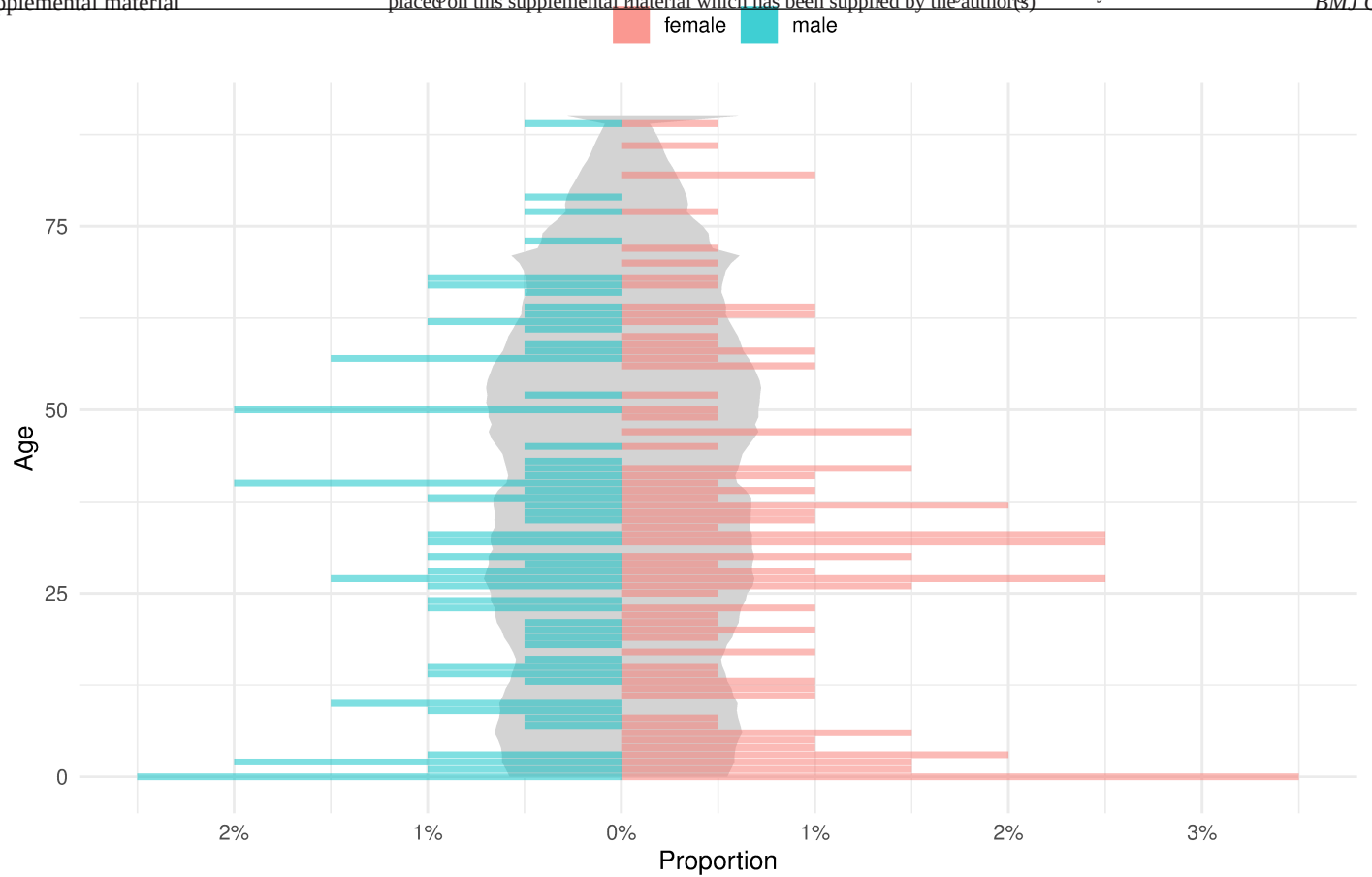
Supplementary Figure 2

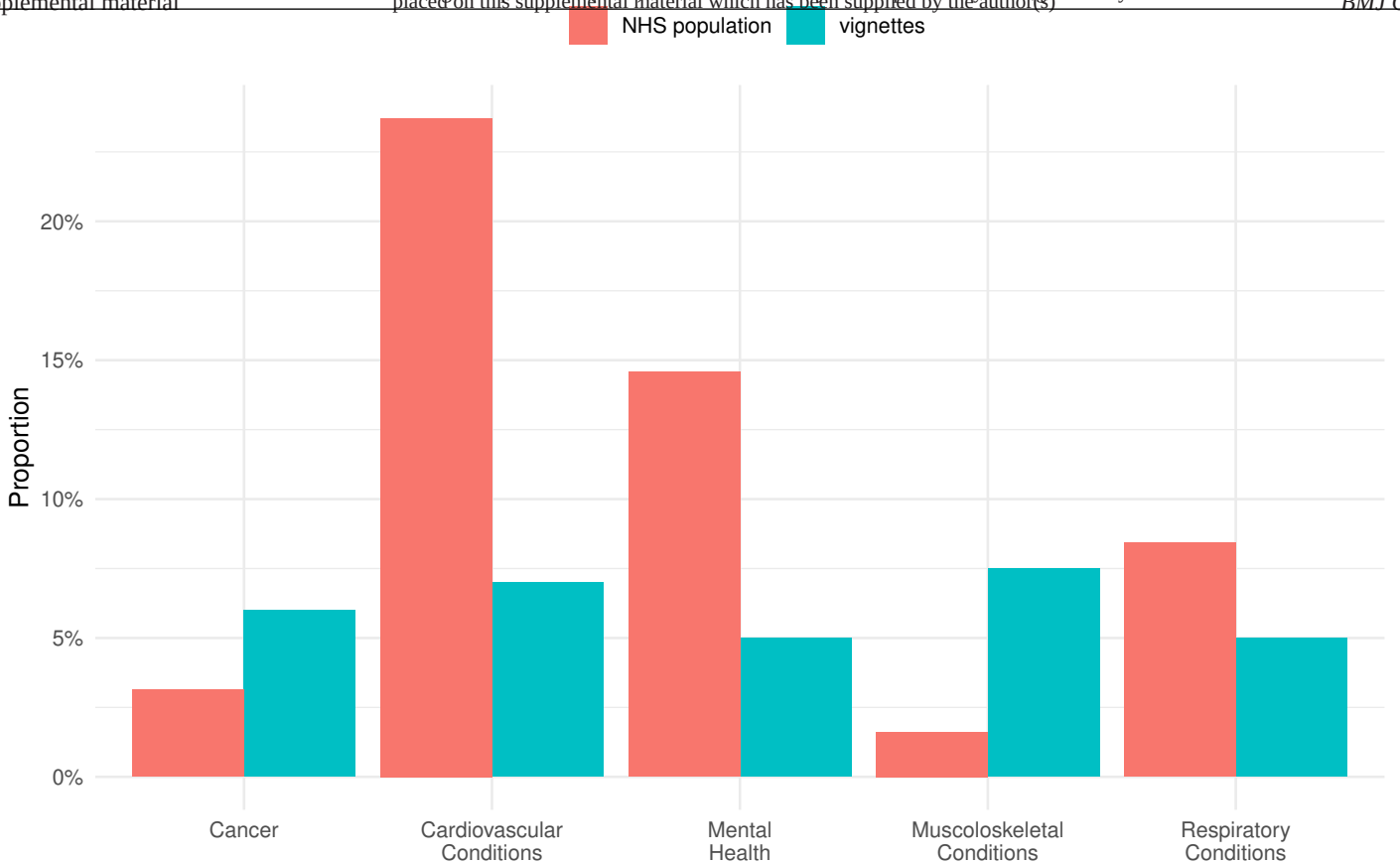
Distribution of age and sex in the study vignettes (bars) compared to the demography of the UK population (shaded area), expressed as proportion of the vignettes or the total population, respectively.

Supplementary Figure 3

Proportion of cases per condition category reported in NHS prevalence data versus vignette conditions.


Babylon
1 call 999 (needs urgent A&E assessment)
1 call 999 or Samaritans
1 call ambulance
1 no data in results section. Call 999.
2 go to (nearest) ED
2 might need treatment in hospital - if serious ED
2 no data, go to ED
3 GP next few hours
3 GP within next 6 hours
3 see doctor urgently
4 speak to GP today for advice
4 symptoms complex-talk to medical professional
4 pregnancy makes case more complex - speak to a doctor
4 speak to GP in next few hours
5 see GP within 1 week
5 see GP
5 see GP in next couple of days
5 GP for routine appointment
5 no data. routine GP appointment
6 self treatment at home
6 can be managed at home
tested-GPs
1 call ambulance
2 ED
3 primary care 4hrs
4 primary care same day
5 primary care 2-3 days
6 primary care 2-3 weeks
7 self-care
8 self-care pharmacy
Symptomate
2 immediate medical attention recommended
3 prompt medical attention recommended
3 prompt medical advice recommended
3 prompt medical attention required
5 may require medical attention
Ada
1 call ambulance
2 go to ED
2 seek emergency care
2 urgent medical care
3 see GP within next 4 hours
3 prompt medical care
3 seek medical advice within the next few hours
4 seek medical advice
5 doctor routine appointment
5 seek medical advice (2-3 days)
6 can ususally be managed at home;
6 manage at home; can seek advice pharmacy
Gold Standard Triage
1 call ambulance
2 go to ED
3 see primary care within next 4 hours
4 see primary care same day
5 see primary care non-urgent
6 home care
Buoy
1 emergency medical advice; Call 911.
2 ER; Requires immediate medical care
2 emergency medical services
3 in-person visit today (as soon as possible)
4 primary care doctor within 1-2 days
5 primary care doctor within 2 weeks
5 phone call or in person visit (sometime in next 3 days)
5 primary care doctor
6 self-treatment
6 wait and watch
Mediktor
1 if medical emergency, dial 911
2 high urgency
2 medical emergency
3 medium urgency - go to hospital as soon as possible, it is very likely you need medical care.
3 medium, consider going to hospital
4 medium urgency - check the description of the relevant diseases and, if you have any questions, check with a health professional.
5 medium
6 low
K Health
1 urgent ambulance and emergency
2 serious health condition/ emergency medicine physician
2 emergency medicine physician
2 immediate medical care
3 primary care physician - immediate
3 serious health condition
3 seek medical advice within the next view hours
5 primary care physician
5 primary care physician or physical therapist
5 primary care physician or psychiatrist
5 gynaecologist
5 primary care
5 pharmacist or primary care physician
6 pharmacist
Your.MD
1 emergency - call an ambulance or go to the hospital immediately.
2 emergency
2 urgent
4 urgent - you should see a doctor within 48 hours.
5 see doctor within 2 weeks
5 please see a doctor
5 see doctor if symptoms persist
5 no cause found. Advised to see a doctor
6 symptoms should get better
6 should be better without treatment.
6 should get better without tretament. see a doctor if symptoms continue
6 no treatment needed
6 wait and watch
WebMD
1 urgent flag - call 911
2 urgent flag - see doctor immediately or go to hospital emergency room
5 moderate
6 low


female
male
75
50
25
0
Age
2%
1%
0%
1%
2%
3%
Proportion


NHS population
vignettes
Proportion
20%
15%
10%
5%
0%
Cancer
Cardiovascular Conditions
Mental Health
Muscoloskeletal Conditions
Respiratory Conditions