Article:
Chambers, D., Cantrell, A., Johnson, M. et al. (2019) Digital and online symptom checkers and assessment services for urgent care to inform a new digital platform : a systematic review. Health Services and Delivery Research, 7 (29). pp. 1-88. ISSN: 2050-4349

https://doi.org/10.3310/hsdr07290

NIHR | National Institute for Health Research

Journals Library

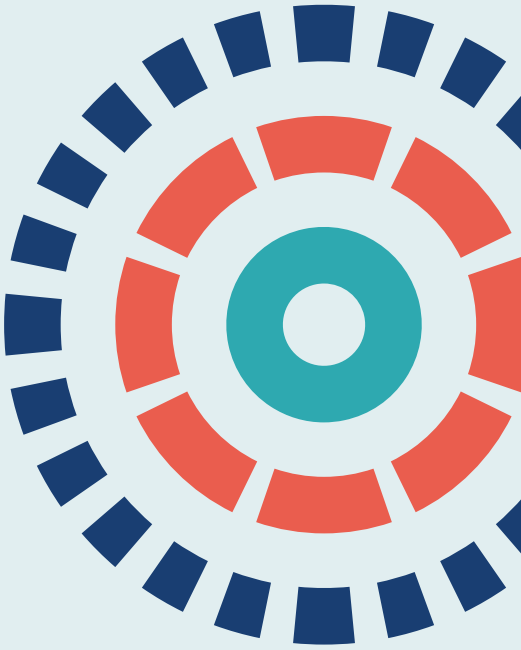
Health Services and Delivery Research

Volume 7 • Issue 29 • August 2019

ISSN 2050-4349

Digital and online symptom checkers and assessment services for urgent care to inform a new digital platform: a systematic review

Duncan Chambers, Anna Cantrell, Maxine Johnson, Louise Preston, Susan K Baxter, Andrew Booth and Janette Turner



DOI 10.3310/hsdr07290

Digital and online symptom checkers and assessment services for urgent care to inform a new digital platform: a systematic review

Duncan Chambers,* Anna Cantrell, Maxine Johnson, Louise Preston, Susan K Baxter, Andrew Booth and Janette Turner

School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, UK

*Corresponding author

Declared competing interests of authors: Andrew Booth is a member of the National Institute for Health Research Complex Review Support Unit Funding Board.

Published August 2019
DOI: 10.3310/hsdr07290

This report should be referenced as follows:

Chambers D, Cantrell A, Johnson M, Preston L, Baxter SK, Booth A, Turner J. Digital and online symptom checkers and assessment services for urgent care to inform a new digital platform: a systematic review. *Health Serv Deliv Res* 2019;**7**(29).

Health Services and Delivery Research

ISSN 2050-4349 (Print)

ISSN 2050-4357 (Online)

This journal is a member of and subscribes to the principles of the Committee on Publication Ethics (COPE) (www.publicationethics.org/).

Editorial contact: journals.library@nihr.ac.uk

The full HS&DR archive is freely available to view online at www.journalslibrary.nihr.ac.uk/hsdr. Print-on-demand copies can be purchased from the report pages of the NIHR Journals Library website: www.journalslibrary.nihr.ac.uk

Criteria for inclusion in the *Health Services and Delivery Research* journal
Reports are published in *Health Services and Delivery Research* (HS&DR) if (1) they have resulted from work for the HS&DR programme or programmes which preceded the HS&DR programme, and (2) they are of a sufficiently high scientific quality as assessed by the reviewers and editors.

HS&DR programme

The Health Services and Delivery Research (HS&DR) programme, part of the National Institute for Health Research (NIHR), was established to fund a broad range of research. It combines the strengths and contributions of two previous NIHR research programmes: the Health Services Research (HSR) programme and the Service Delivery and Organisation (SDO) programme, which were merged in January 2012.

The HS&DR programme aims to produce rigorous and relevant evidence on the quality, access and organisation of health services including costs and outcomes, as well as research on implementation. The programme will enhance the strategic focus on research that matters to the NHS and is keen to support ambitious evaluative research to improve health services.

For more information about the HS&DR programme please visit the website: http://www.nets.nihr.ac.uk/programmes/hsdr

This report

The research reported here is the product of an HS&DR Evidence Synthesis Centre, contracted to provide rapid evidence syntheses on issues of relevance to the health service, and to inform future HS&DR calls for new research around identified gaps in evidence. Other reviews by the Evidence Synthesis Centres are also available in the HS&DR journal. The research reported in this issue of the journal was funded by the HS&DR programme or one of its preceding programmes as project number 16/47/17. The contractual start date was in April 2018. The final report began editorial review in June 2018 and was accepted for publication in October 2018. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HS&DR editors and production house have tried to ensure the accuracy of the authors' report and would like to thank the reviewers for their constructive comments on the final report document. However, they do not accept liability for damages or losses arising from material published in this report.

This report presents independent research funded by the National Institute for Health Research (NIHR). The views and opinions expressed by authors in this publication are those of the authors and do not necessarily reflect those of the NHS, the NIHR, NETSCC, the HS&DR programme or the Department of Health and Social Care. If there are verbatim quotations included in this publication the views and opinions expressed by the interviewees are those of the interviewees and do not necessarily reflect those of the authors, those of the NHS, the NIHR, NETSCC, the HS&DR programme or the Department of Health and Social Care.

Published by the NIHR Journals Library (www.journalslibrary.nihr.ac.uk), produced by Prepress Projects Ltd, Perth, Scotland (www.prepress-projects.co.uk).

Abstract

Digital and online symptom checkers and assessment services for urgent care to inform a new digital platform: a systematic review

Duncan Chambers,* Anna Cantrell, Maxine Johnson, Louise Preston, Susan K Baxter, Andrew Booth and Janette Turner

School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, UK

*Corresponding author d.chambers@sheffield.ac.uk

Background: Digital and online symptom checkers and assessment services are used by patients seeking guidance about health problems. NHS England is planning to introduce a digital platform (NHS111 Online) to operate alongside the NHS111 urgent-care telephone service. This review focuses on digital and online symptom checkers for urgent health problems.

Objectives: This systematic review was commissioned to provide NHS England with an independent review of previous research in this area to inform strategic decision-making and service design.

Data sources: Focused searches of seven bibliographic databases were performed and supplemented by phrase searching for names of symptom checker systems and citation searches of key included studies. The bibliographic databases searched were MEDLINE, EMBASE, The Cochrane Library, CINAHL (Cumulative Index to Nursing and Allied Health Literature), HMIC (Health Management Information Consortium), Web of Science and the Association of Computing Machinery (ACM) Digital Library, from inception up to April 2018.

Review methods: Brief inclusion criteria were (1) population – general population seeking information online or digitally to address an urgent health problem; (2) intervention – any online or digital service designed to assess symptoms, provide health advice and direct patients to appropriate services; and (3) comparator – telephone or face-to-face assessment, comparative performance in tests or simulations (studies with no comparator were included if they reported relevant outcomes). Outcomes of interest included safety, clinical effectiveness, costs or cost-effectiveness, diagnostic and triage accuracy, use of and contacts with health services, compliance with advice received, patient/carer satisfaction, and equity and inclusion. Inclusion was not restricted by study design. Screening studies for inclusion, data extraction and quality assessment were carried out by one reviewer with a sample checked for accuracy and consistency. Final decisions on study inclusion were taken by consensus of the review team. A narrative synthesis of the included studies was performed and structured around the predefined research questions and key outcomes. The overall strength of evidence for each outcome was classified as 'stronger', 'weaker', 'conflicting' or 'insufficient', based on study numbers and design.

Results: In total, 29 publications describing 27 studies were included. Studies were diverse in their design and methodology. The overall strength of the evidence was weak because it was largely based on observational studies and with a substantial component of non-peer-reviewed grey literature. There was little evidence to suggest that symptom checkers are unsafe, but studies evaluating their safety were generally short term and small scale. Diagnostic accuracy was highly variable between different systems but was generally low. Algorithm-based triage tended to be more risk averse than that of health professionals. Inconsistent evidence was found on effects on service use. There was very limited evidence

on patients' reactions to online triage advice. The studies showed that younger and more highly educated people are more likely to use these services. Study participants generally expressed high levels of satisfaction with digital and online triage services, albeit in uncontrolled studies.

Limitations: Findings from symptom checker systems for specific conditions may not be applicable to more general systems and vice versa. Studies of symptom checkers as part of electronic consultation systems in general practice were also included, which is a slightly different setting from a general 'digital 111' service. Most studies were screened by one reviewer.

Conclusions: Major uncertainties surround the probable impact of digital 111 services on most outcomes. It will be important to monitor and evaluate the services using all available data sources and by commissioning high-quality research.

Future work: Priorities for research include comparisons of different systems, rigorous economic evaluations and investigations of patient pathways.

Study registration: The study is registered as PROSPERO CRD42018093564.

Funding: The National Institute for Health Research Health Services and Delivery Research programme.

Contents

List of tables

List of figures

List of boxes

List of abbreviations

A&E	accident and emergency
app	application
CCG	Clinical Commissioning Group
CI	confidence interval
CINAHL	Cumulative Index to Nursing and Allied Health Literature
ED	emergency department
GP	general practitioner
HMIC	Health Management Information Consortium
HSDR	Health Services and Delivery Research
NIHR	National Institute for Health Research
PBRN	practice-based research network
PPI	patient and public involvement
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-analyses
PROSPERO	Prospective Register of Systematic Review Protocols
QUADAS	Quality Assessment of Diagnostic Accuracy Studies
RCT	randomised controlled trial
SORT	Strategy for Off-site Rapid Triage
TIDieR	Template for Intervention Description and Replication
TIDieST	Template for Intervention Description for Systems for Triage

Plain English summary

NHS England is planning to introduce a 'digital 111' service alongside the existing NHS111 urgent-care telephone service. This will allow people to use a website or smartphone app to find out information about a health problem. The new digital 111 service will be like a 'symptom checker' to provide possible diagnoses and suggest a course of action (e.g. self-care, arrange a general practitioner appointment or go to an emergency department). During the research study, we looked for other research on these types of online or digital services that are designed to assess symptoms, provide health advice and direct patients to appropriate services for urgent health problems.

We found 29 relevant publications. There was little evidence to suggest whether symptom checkers are safe or unsafe, and studies evaluating their safety were generally short term and small scale. Symptom checkers were found to be generally less accurate than health professionals, although their performance in studies was variable. Symptom checkers tended to be more cautious in their advice than health professionals. There was some indication that symptom checkers can influence the demand for urgent-care services, but results were inconsistent. There was very limited evidence on patients' reactions to advice from symptom checkers, including information regarding whether or not patients follow the advice.

The studies showed that younger and more highly educated people were more likely to use these services, and women used them more than men. Patients were generally satisfied with digital services, although results came mainly from studies that were not of the highest quality.

The research we found indicates that there is limited knowledge regarding the probable impact of digital 111 services. Findings about patient satisfaction suggest that, once introduced, the use of digital services may increase rapidly. It will be important, therefore, to monitor and evaluate these services using all available data sources and through high-quality research.

Scientific summary

Background

This systematic review deals with digital and online symptom checkers and assessment services used by patients seeking guidance about an urgent health problem. These services generally provide people with possible diagnoses and/or suggest a course of action based on their reported symptoms. NHS England intends to introduce a digital platform (NHS111 Online) to operate alongside the NHS111 telephone service.

The 'digital 111' service is seen as key to reducing demand for the telephone 111 service, enabling resources to be redirected to support 'integrated urgent and emergency care systems', as outlined in the NHS *Five Year Forward View* (NHS England. *Five Year Forward View*. Leeds: NHS England; 2014) and its 2017 update *Next Steps on the NHS Five Year Forward View* (NHS England. *Next Steps on the NHS Five Year Forward View*. Leeds: NHS England; 2017).

Thus, it is anticipated that a digital 111 platform will help to manage demand and increase efficiency in the urgent and emergency care system, complementing the agenda of locally based Sustainability and Transformation Partnerships. However, there is a risk of increasing demand, duplicating health-care contacts and providing advice that is not safe or clinically appropriate.

In 2017, NHS England carried out pilot evaluations of different digital or online triage systems in four regions of England (NHS England. *NHS111 Online Evaluation*. Leeds: NHS England; 2017). The evaluations aimed to assess whether or not digital/online triage was acceptable to users and connected them to appropriate clinical care. The full report of these evaluations is not yet published at the time of writing this report.

Objectives

The objective of this systematic review was to inform further developments of the proposed digital platform by summarising and critiquing the previous research in this area, from both the UK and overseas.

Methods

Focused searches of seven databases were performed, supplemented by phrase searching for names of symptom checker systems (e.g. webGP, iTriage) and citation searches of key included studies. The bibliographic databases searched were MEDLINE, EMBASE, The Cochrane Library, CINAHL (Cumulative Index to Nursing and Allied Health Literature), HMIC (Health Management Information Consortium), Web of Science and the Association of Computing Machinery (ACM) Digital Library, from inception up to April 2018.

Inclusion and exclusion criteria are listed in the sections below.

Population

Members of the general population, including adults and children, who are looking for information online or digitally to address an urgent health problem, which includes issues arising from both acute illness and long-term chronic illness. Non-urgent problems were excluded.

Intervention

The intervention is made up of online or digital services that are designed to assess symptoms, provide health advice and direct patients to appropriate services. This reflects the role of the NHS111 telephone service. Services that provide only health advice were excluded, as were those that offer treatment (e.g. online cognitive–behavioural therapy services).

Comparator

The 'gold standard' comparator is the current practice of telephone assessment (e.g. NHS111) or face-to-face assessment (e.g. in a general practice, urgent-care centre or emergency department). However, studies with other relevant comparators (e.g. comparative performance in tests or simulations) or with no comparator were included if they addressed the research questions.

Outcomes

The main outcomes of interest were:

- safety (e.g. any evidence of adverse events arising from following or ignoring advice from online/digital services)
- clinical effectiveness (any evidence of clinical outcomes associated with the use of online/digital services)
- cost-effectiveness (including costs and resource use)
- accuracy – this refers to the ability to provide a correct assessment and the ability to distinguish between high- and low-acuity (urgency and level of care needed) problems and, hence, direct patients to appropriate services, avoiding over- or undertriage. Direction to appropriate services is also referred to as patient disposition
- impact on service use/diversion (including possible multiple contacts with health services)
- compliance with advice received
- patient/carer satisfaction
- equity and inclusion (e.g. barriers to access, characteristics of patients using the service compared with the general population).

This list is not exhaustive and other relevant outcomes from included studies were extracted.

Study design

We did not restrict inclusion by study design (and included relevant audits or service evaluations in addition to formal research studies) but included studies had to evaluate (quantitatively or qualitatively) some aspect of an online/digital service. Studies were not excluded on grounds of quality. Studies of health-care system from any high-income country were eligible for inclusion.

Excluded

The following types of studies were excluded from the review:

- studies that merely describe services without providing any quantitative or qualitative outcome data
- conceptual papers and projections of possible future developments
- studies conducted in low- or middle-income countries' health-care systems.

Screening studies for inclusion, data extraction and quality assessment (using appropriate tools for different study designs) were carried out by one reviewer with a 10% sample checked for accuracy and consistency. Final decisions on study inclusion were taken by consensus of the review team. To characterise the included digital and online systems as interventions, we identified studies reporting on a particular system and extracted data from all relevant studies using a modification of the Template for Intervention Description and Replication (TIDieR) checklist that we designated Template for Intervention Description for Systems for Triage (TIDieST).

A narrative synthesis of the included studies was performed and structured around the predefined research questions and key outcomes. The overall strength of evidence for each outcome was classified as 'stronger', 'weaker', 'conflicting' or 'insufficient', based on study numbers and design.

Results

We included 29 publications describing 27 studies (nine from the UK). Studies were published between 2006 and 2018 and were diverse in terms of their design and methodology. The overall strength of the evidence base varied between outcomes, but in absolute terms the evidence was weak, being based largely on observational studies and with a substantial component of grey literature.

We were able to extract data on eight systems using the TIDieST checklist. When appropriate, data from multiple studies were combined in one checklist. Some of these systems appear to be no longer in use and we were unable to obtain sufficient data for some systems currently being used and/or evaluated. Four of the included systems were designed to cover a full range of symptoms and four others covered a more limited range (i.e. three for influenza-like illness and one for minor respiratory symptoms). Most systems were accessed through web pages, often linked to health-care providers or government organisations. The 'babylon check' system was the main exception because it was designed for access using a smartphone application (app). Published research studies provided relatively little detail about the systems, possibly reflecting a need for commercial confidentiality (Middleton K, Butt M, Hammerla N, Hamblin S, Mehta K, Parsa A. *Sorting Out Symptoms: Design and Evaluation of the 'Babylon Check' Automated Triage System*. London: Babylon Health; 2016).

We found little evidence to support the hypothesis that digital and online symptom checkers are detrimental to patient safety. However, the studies that reported patient safety outcomes were mostly short term and involved relatively small samples. Some were limited to people with specific types of symptoms (e.g. influenza-like illness or respiratory symptoms) and others recruited from specific population groups (e.g. students), meaning that participants are not representative of all users of urgent-care services.

Study participants generally expressed high levels of satisfaction, albeit in uncontrolled studies. For example, in the NHS England pilot evaluation 70–80% of users were satisfied with their experience at each of the pilot sites. This was based on a sample of > 1500 users, the majority of whom were involved in the London pilot using the 'babylon check' app (NHS England. *NHS111 Online Evaluation*. Leeds: NHS England; 2017).

The evidence from the studies included in the review suggests that digital and online systems have yet to achieve a high level of accuracy in the assessment of specific conditions. This finding applies to both 'general purpose' symptom checkers and those symptom checkers that are limited to particular conditions. However, some of the included studies did not recruit representative populations and others were based on standardised vignettes rather than real-world data. Studies that compared symptom checkers with health professionals tended to use the doctors' clinical diagnosis as the reference standard.

Results for the accuracy of triage were inconsistent between studies. The studies used similar methods to those evaluating diagnostic accuracy. A review of 23 symptom checkers found that triage level was appropriate overall for 57% of patient vignettes, with considerable variation between systems and conditions (Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ* 2015;**351**:h3480). The NHS England evaluation also found differing levels of agreement with clinical expert opinion across the four systems evaluated. In general, algorithm-based triage tended to be more risk averse than the triage of health professionals.

We also found inconsistent evidence on effects on service use, but there was some indication that symptom checkers can influence the pattern of service use. The strongest evidence came from a randomised trial of an intervention that was specifically designed to promote self-care and covered

respiratory symptoms only (Little P, Stuart B, Andreou P, McDermott L, Joseph J, Mullee M, *et al*. Primary care randomised controlled trial of a tailored interactive website for the self-management of respiratory infections (Internet Doctor). *BMJ Open* 2016;**6**:e009769). In this study, the intervention group had fewer contacts with doctors (but more contact with NHS Direct) than the control group, despite having a longer duration and greater severity of illness. The NHS England evaluation found a small shift towards self-care with digital triage compared with telephone triage, and another study reported that the webGP symptom checkers diverted 18% of patients from requesting a general practitioner appointment (Madan A. *WebGP: The Virtual General Practice*. London: Hurley Group; 2014). These are findings from the grey literature and should be interpreted with caution.

There was very limited evidence on patients' reactions to online triage advice and whether the patients follow the advice or seek further help or information. Preliminary findings from the NHS England evaluation suggest that patients may be more likely to seek further advice for more urgent conditions, but further confirmation is required.

Over half of the included studies considered equity and inclusion either directly or by comparing users with non-users of digital triage systems. Not surprisingly, there was a clear consensus that younger and more highly educated people are more likely to use these digital/online services, whereas older and less educated patients are more likely to prefer telephone or face-to-face contact with health-care practitioners. This could have implications for health equity if urgent-care pathways prioritise (or appear to prioritise) requests originating from digital sources.

Limitations

Findings from symptom checker systems for specific conditions may not be applicable to more general systems and vice versa. We have also included studies of symptom checkers as part of electronic consultation systems in general practice, which is, again, a slightly different setting from a general digital 111 service. Most studies were screened by one reviewer.

Conclusions

The current evidence base covers diverse interventions, study designs and outcomes. Major uncertainties surround the probable impact of digital 111 services on most of the important outcomes, but precedent suggests that once introduced their use will increase rapidly. It will be important to monitor and evaluate these services using all available data sources and by commissioning high-quality research.

The studies included in the review suggest that there is a high level of uncertainty about the impact of digital 111 on the urgent-care system and the wider health-care system. The health service may need to respond to short-term increases (or decreases) in demand and/or shifts from one part of the system to another. This may increase pressure on the system, at least in the short term. In the longer term, if the use of the 111 telephone service decreases as planned, there may be opportunities to redeploy staff to fill other roles in the urgent and emergency care system.

Priorities for research include comparisons of different systems, rigorous economic evaluations based on real-world data, investigations of the pathways followed by patients using the new service and investigations of the role of behaviour change theory in the development and implementation of symptom checkers.

Study registration

The study is registered as PROSPERO CRD42018093564.

Funding

Funding for this study was provided by the Health Services and Delivery Research programme of the National Institute for Health Research.

Chapter 1 Background

Digital and online symptom checkers and assessment services are used by patients seeking guidance about an urgent health problem. These services generally provide people with possible alternative diagnoses based on their reported symptoms and/or suggest a course of action [e.g. self-care, attend a general practitioner (GP) appointment or go to an emergency department (ED)].

In England, the NHS111 service provides assessment and triage by telephone for problems that are urgent but not classified as emergencies. The latest data from NHS England[2] show that in April 2018 there were 1,338,253 calls to NHS111, which is an average of approximately 46,100 calls per day. Outcomes of these calls were that 12.1% of calls resulted in ambulances being despatched, 8.7% of patients who called were recommended to attend an ED, 60.7% of patients who called were recommended to attend primary care, 4.6% of patients who called were recommended to attend another service and 14% of patients who called were not recommended to attend another service (e.g. their condition was considered suitable for self-care).

NHS England is planning to introduce a digital platform for NHS111. This would include a possibility for patients to be referred to the NHS111 telephone service for further assessment.

A beta version of the service (referred to as 'NHS111 Online') is available at the following website: https://111.nhs.uk/ (accessed 1 June 2018). The 'digital 111' service is seen as key to reducing demand for the telephone 111 service, enabling resources to be redirected to support 'integrated urgent and emergency care systems' (contains public sector information licensed under the Open Government Licence v3.0 www.nationalarchives.gov.uk/doc/open-government-licence/version/3/), as outlined in the NHS *Five Year Forward View*[3] and its 2017 update *Next Steps on the NHS Five Year Forward View*.[4]

It is thus hoped that a digital 111 platform will help to manage demand and increase efficiency in the urgent and emergency care system, complementing the agenda of locally based Sustainability and Transformation Partnerships. However, there is a risk of increasing demand, duplicating health-care contacts and providing advice that is not safe or clinically appropriate. For example, an evaluation of the NHS111 telephone service at four pilot sites and three control sites found that in its first year the service was not successful in reducing 999 emergency calls or in shifting patients from emergency to urgent care.[5] A recent study of 23 symptom checker algorithms providing diagnostic and triage advice that would form the basis of a digital 111 platform found deficiencies in both their diagnostic capabilities and their triage capabilities.[6]

In 2017, NHS England carried out pilot evaluations of different systems in four regions of England.[7] The evaluations aimed to assess whether or not digital/online triage was acceptable to users and connected them with appropriate clinical care.[7] The full report from NHS England is not yet published at the time of writing this review. The objective of this systematic review was to inform further development of the proposed digital platform by summarising and critiquing the previous research in this area, from both the UK and overseas.

Chapter 2 Review methods

This systematic review was commissioned to provide NHS England with an independent review of previous research in this area to inform strategic decision-making and service design. The research questions that the review addressed are as follows (in relation to digital and online symptom checkers and health advice/triage services):

- Are these services safe and clinically effective? Do they accurately identify both patients with low-acuity problems and patients with high-acuity problems?
- What is the overall impact of these services on health-care demand? In particular, is there evidence that they drive patients towards higher-acuity services (e.g. ambulance and ED use)?
- Do individuals comply with these services and the advice received? Are these services used instead of, or as well as, other elements of urgent and emergency care?
- Are these services cost-effective?
- Are patients/carers satisfied with the services and the advice received?
- What are the implications of these services for equality of access to health services and inclusion/exclusion of disadvantaged groups?

Literature search and screening

Initial scoping searches revealed that a highly sensitive search strategy (designed to retrieve all relevant items), typically conducted for systematic reviews, retrieved a disproportionately high number of references on GPs' decision-making and triage. These references are outside the scope of the NHS Digital 111 review and would have unnecessarily diverted resources from the review focus. Therefore, to optimise retrieval and sifting we devised a three-stage retrieval strategy as an acceptable alternative to comprehensive topic-based searching. This involves:

1. Targeted searches of precise, high-specificity terms (designed to limit the search to items likely to be relevant). *Box 1* shows an example for MEDLINE.
   We searched the following databases –

   - MEDLINE via Ovid
   - EMBASE via Ovid
   - The Cochrane Library via Wiley Online Library
   - CINAHL (Cumulative Index to Nursing and Allied Health Literature) via EBSCO*host*
   - HMIC (Health Management Information Consortium) via OpenAthens
   - Web of Science (Science Citation Index and Social Science Citation Index) via the Web of Knowledge at the Institute for Scientific Information, now maintained by Clarivate Analytics
   - the Association of Computing Machinery (ACM) Digital Library.

   The searches were not restricted by language or date.

2. Phrase searching for named and generic systems. The list below was used for phrase searching and was compiled from a study of symptom checker performance[6] and an examination of internet search results for symptom checkers –

   - Askmd
   - askmygp
   - BetterMedicine
   - DocResponse
   - Doctor Diagnose
   - Drugs.com
   - EarlyDoc
   - Econsult
   - engage consult
   - Esagil
   - Family Doctor
   - FreeMD
   - gp at hand
   - Harvard Medical School Family Health Guide
   - healthdirect
   - Healthline
   - Healthwise
   - Healthy Children
   - Isabel
   - iTriage
   - Mayo Clinic
   - MEDoctor
   - NHS Symptom Checkers
   - online triage
   - push doctor
   - Steps2Care
   - Symcat
   - Symptify
   - Symptomate
   - webgp
   - WebMD.

   The phrase searches were conducted on the databases listed in point 1, with no language or date restrictions.

3. Citation searches of key included studies and reviews. Reference checking of included studies and key reviews. This was complemented by contact with service providers, directly and via websites.

Search results were stored in a reference management system (EndNote; Clarivate Analytics, Philadelphia, PA, USA) and imported into EPPI-Reviewer software for screening (version 4; Evidence for Policy and Practice Information and Co-ordinating Centre, University of London, London, UK). The search results were screened against the inclusion criteria by one reviewer, with a 10% sample screened by a second reviewer. Uncertainties were resolved by discussion among the review team.

BOX 1 Highly focused, specific search strategy on MEDLINE (adapted for different databases)

Database

Ovid MEDLINE(R) Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Ovid MEDLINE(R) Daily and Ovid MEDLINE(R) <1946 to Present>.

Search strategy

- (symptom checker or symptoms checker or symptom checkers or symptoms checkers).tw.
- ('self diagnosis' or 'self referral' or 'self triage' or 'self assessment').tw. (10403)
- TRIAGE/
- 2 or 3
- (online or on-line or web or electronic or automated or internet or digital or app or mobile or smartphone).tw.
- 4 and 5
- ('online diagnosis' or 'web based triage' or 'electronic triage' or etriage).tw.
- 1 or 6 or 7

Inclusion and exclusion criteria

Population

Members of the general population, which includes adults and children, who are looking for information online or digitally to address an urgent health problem, which includes issues arising from both acute illness and long-term chronic illness. Non-urgent problems, such as possible Asperger syndrome or memory loss/early dementia, were excluded.

Intervention

Any online or digital service designed to assess symptoms, provide health advice and direct patients to appropriate services. This reflects the role of the NHS111 telephone service. Services that provide only health advice were excluded, as were those that offer treatment (e.g. online CBT services).

Comparator

The 'gold standard' comparator is the current practice of telephone assessment (e.g. NHS111) or face-to-face assessment (e.g. general practice, urgent-care centre or ED). However, studies with other relevant comparators (e.g. comparative performance in tests or simulations) or with no comparator were included if they addressed the research questions.

Outcomes

The main outcomes of interest were:

- safety (e.g. any evidence of adverse events arising from following or ignoring advice from online/digital services)
- clinical effectiveness (any evidence of clinical outcomes associated with use of online/digital services)
- cost-effectiveness (including costs and resource use)
- accuracy – this refers to the ability to provide a correct diagnosis and distinguish between high and low acuity/urgency problems (and hence direct patients to appropriate services, avoiding over- or undertriage)
- impact on service use/diversion (including possible multiple contacts with health services)
- compliance with advice received
- patient/carer satisfaction
- equity and inclusion (e.g. barriers to access, characteristics of patients using the service compared with the general population).

This list is not exhaustive and other relevant outcomes from included studies were extracted.

Study design

We did not restrict inclusion by study design (and included relevant audits or service evaluations in addition to formal research studies), but included studies had to evaluate (quantitatively or qualitatively) some aspect of an online/digital service. Studies from any high-income country's health-care system were eligible for inclusion.

Excluded

The following types of studies were excluded from the review:

- studies that merely describe services without providing any quantitative or qualitative outcome data
- conceptual papers and projections of possible future developments
- studies conducted in low- or middle-income countries' health-care systems.

Data extraction and quality/strength of evidence assessment

We extracted and tabulated key data from the included studies, including study design, population/setting, results and key limitations. The full data extraction template is provided in *Appendix 1*. Data extraction was performed by one reviewer, with a 10% sample checked for accuracy and consistency by another reviewer.

To characterise the included digital and online systems as interventions, we identified studies reporting on a particular system and extracted data from all relevant studies using a modification of the Template for Intervention Description and Replication (TIDieR) checklist[8] that we designated Template for Intervention Description for Systems for Triage (TIDieST). The checklist is presented in *Appendix 2* and the completed checklists in *Appendix 3*. Characteristics of included systems are summarised in *Chapter 3, Characteristics of included systems*.

Quality (risk-of-bias) assessment was undertaken for peer-reviewed full publications only (i.e. not grey literature publications or conference abstracts). The rationale for this approach was that non-peer-reviewed publications tend to lack the detail required for assessment of risk of bias and/or tend not to follow standard study designs. Randomised controlled trials (RCTs) were assessed using the Cochrane Collaboration risk-of-bias tool.[9] For diagnostic accuracy type studies, we used the Cochrane Collaboration version of Quality Assessment of Diagnostic Accuracy Studies (QUADAS)[10] and for other study designs we used the National Heart, Lung and Blood Institute's tool for observational cohort and cross-sectional studies.[11] Details of quality assessment tools can be found in *Appendix 4*. Quality assessment was performed by one reviewer, with a 10% sample checked for accuracy and consistency by another reviewer. Assessment of the overall strength (quality and relevance) of evidence for each research question is part of the narrative synthesis. Overall strength of the evidence base for key outcomes was assessed using an adaptation of the method described by Baxter *et al.*[12] This involves classifying evidence as 'stronger', 'weaker', 'inconsistent' or 'very limited', based on study numbers and design. Specifically, 'stronger evidence' represented generally consistent findings in multiple studies with a comparator group design or comparative diagnostic accuracy studies, 'weaker evidence' represented generally consistent findings in one study with a comparator group design and several non-comparator studies or multiple non-comparator studies, 'very limited evidence' represented an outcome reported by a single study and, finally, 'inconsistent evidence' represented an outcome for which < 75% of the studies agreed on the direction of effect. All studies included in the review were included in the analysis of the overall strength of evidence.

Evidence synthesis

We performed a narrative synthesis structured around the prespecified research questions and outcomes. This included an 'evidence map' summarising the quantity and strength of evidence for each outcome and identifying gaps that may need to be filled by further research. We did not perform any meta-analyses because the included studies varied widely in terms of design, methodology and outcomes.

Public and patient involvement in the study

The study aimed to be informed by patient and public involvement (PPI) at all stages of the research process. The Sheffield Health Services and Delivery Research (HSDR) evidence synthesis centre patient and public advisory group provided input during the design, analysis and reporting phases, including exploration of the study parameters, discussion regarding the meaning and interpretation of the study findings, drafting of the *Plain English summary* and help with disseminating the findings and maximising the impact of the research.

The advisory group comprised nine members drawn from the Yorkshire and Humber region and two members from other regions of England. Because this study was of a relatively short duration, the group provided input at two advisory group meetings. At the meetings there was discussion regarding the focus of the work, including a presentation on previous research on NHS111 services to provide a context for understanding the current work. The meetings also included the presentation and discussion of the findings of the review in order to explore key messages for patients that could inform the dissemination of the findings.

The discussion during one meeting was structured using a SWOT (strengths, weaknesses, opportunities and threats) analysis approach that revealed a number of potential concerns among patients as well as potential perceived benefits. The group members expressed some concern over the reliability and consistency of symptom checker algorithms, particularly if different systems are being used in different parts of the country. The economic benefits of the systems were also questioned in view of the high costs of programming and system development. Possible equity issues were identified because digital 111 might be less accessible to some groups (people with cognitive impairment were mentioned). The group expressed uncertainty about the impact of this type of service on the wider urgent and emergency care system (e.g. would the creation of a new access point complicate patient pathways and increase demand overall?). The possible vulnerability to external threats leading to the breakdown of the system or the loss of data was identified as a weakness. There was discussion of whether or not patients would feel able to trust the advice they were given. Some group members saw the desire of patients to be able to talk to someone about their problems for reassurance and empathy as a threat to the success of digital 111. Potential benefits included increased access to urgent-care advice at any time and appeal to younger people and to those who might feel anxious or embarrassed about discussing their problem with a health professional.

Involvement of the advisory group was also beneficial in highlighting some issues that had emerged from the systematic review and enabled the reviewers to structure the review findings while taking this into account. The group's uncertainty about the probable impact of digital 111 was reflected in the report's findings and recommendations for ongoing evaluation and further research. The review report also reflects the group's relatively cautious attitude (while recognising the need to update the way services are accessed), which contrasts with the strong belief in some quarters that digital 111 will help to ensure that patients receive appropriate care more quickly while reducing 'inappropriate' visits to EDs and GP appointments. Advisory group members reported that they had limited experience of using symptom checkers for real-life health problems in the urgent-care setting, which echoes the limited available information and uncertainty regarding the probable effects of digital 111 services.

Study registration and outputs

The protocol was registered prospectively with PROSPERO (registration number CRD42018093564) and is also available via the National Institute for Health Research (NIHR) HSDR programme website (www.journalslibrary.nihr.ac.uk/programmes/hsdr/164717/) and the Sheffield HSDR Evidence Synthesis Centre website (https://scharr.dept.shef.ac.uk/hsdr/).

Chapter 3 Review results

Results of literature search

This chapter presents the studies that were included in the review. A PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses) flow diagram (*Figure 1*) details the search process.

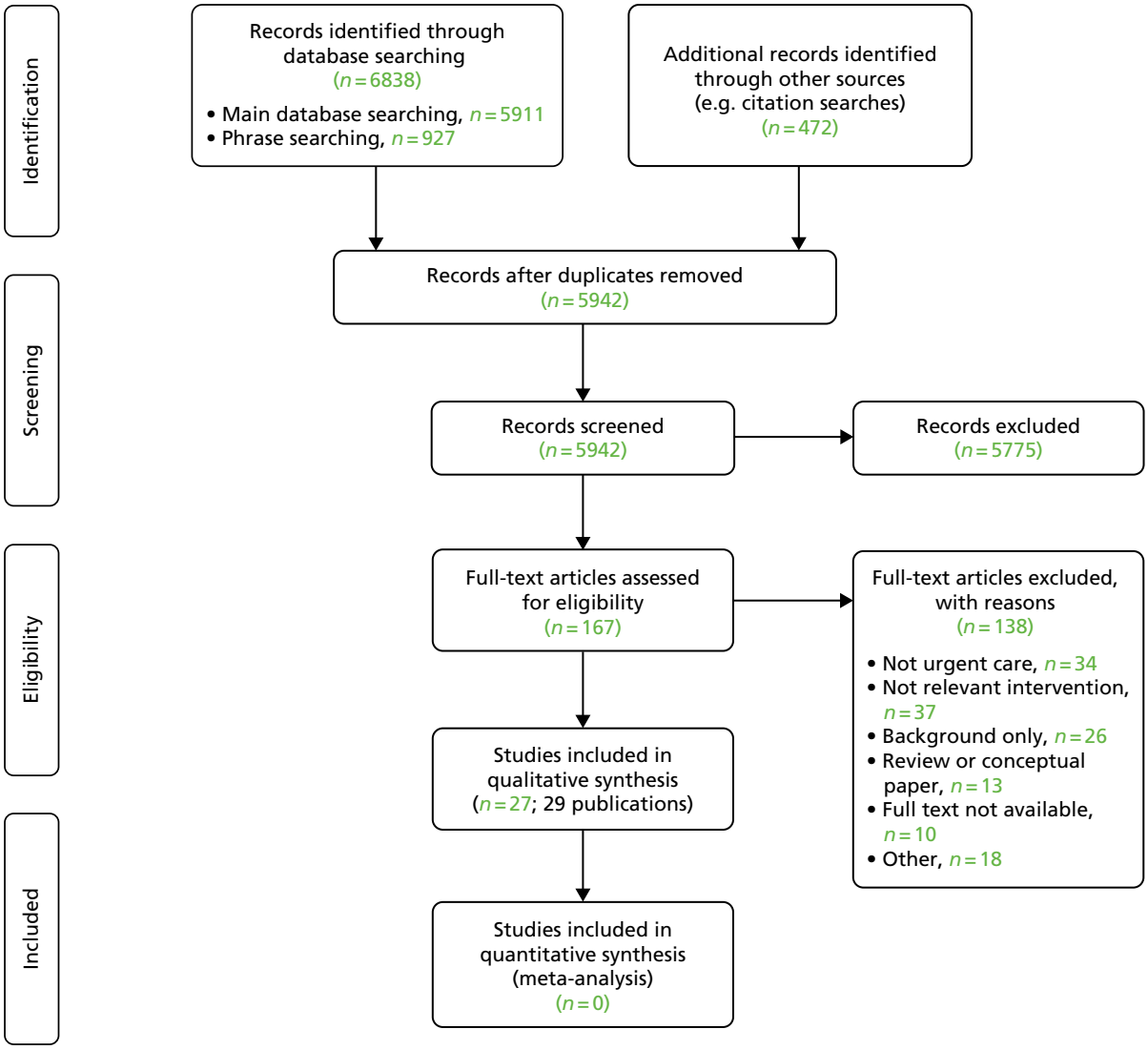


FIGURE 1 The PRISMA flow diagram. Some parts of this chapter have been reproduced from Chambers *et al.*[1] © Author(s) (or their employer(s)) 2019. Re-use permitted under CC BY. Published by BMJ. This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: https://creativecommons.org/licenses/by/4.0/. The figure includes minor additions and formatting changes to the original text.

All titles and abstracts were screened by one researcher from the review team, with a subset (about 10%) of the titles and abstracts screened by one other researcher from the review team. A calculation of the inter-rater agreement was made. A kappa coefficient was calculated, demonstrating moderate agreement between reviewers [$\kappa = 0.582$, 95% confidence interval (CI) 0.274 to 0.889]. Any queries were resolved by discussion. A similar process was followed for final decisions on inclusion/exclusion, based on full-text documents.

Characteristics of included studies

We included 29 publications that represented 27 studies (one study[13] had two associated comments[14,15]). Nine studies were performed in the UK. Seventeen studies (*Table 1*) evaluated symptom checkers as a self-contained intervention, of which eight covered a limited range of symptoms (e.g. respiratory[16,17,25]

TABLE 1 Studies of symptom checkers as a single intervention

Reference	Study design	System type	Comparator	Population/sample
Peer-reviewed papers				
Kellermann *et al.*[16]	Simulation: the developed algorithm was tested against past patient records	Online: SORT was available on two interactive websites	Health professional performance on real-world data. The algorithm was tested against clinicians' decision based on past patient records	Population/condition: specific condition(s); influenza symptoms Sample size (participants/ data set): the algorithm was assessed against patients with influenza-like illnesses visiting EDs in spring before 2009 H1N1 influenza outbreak
Little *et al.*[17]	Experimental RCT	Online: Internet Doctor website	Other: usual GP care without access to the Internet Doctor website	Population/condition: specific condition(s); respiratory infections and associated symptoms Sample size (participants/ data set): 3044 adults registered with a general practice; of these, 852 in the intervention group and 920 in the control group reported one or more RTIs over 20 weeks
Luger *et al.*[18]	Simulation: described as 'human–computer interaction study' using think-aloud protocols	Online: Google (Google Inc., Mountain View, CA, USA) and WebMD (Internet Brands, New York, NY, USA)	Other: comparing two internet-based health tools	Population/condition: general population; older adults (≥ 50 years) Sample size (participants/ data set): 79 participants
Marco-Ruiz *et al.*[19]	Qualitative: qualitative element Other: 1. Online evaluation by users (problem detection) 2. Think-aloud technique by smaller sample of participants (usability)	Online: Erdusyk	None	Population/condition: general population; internet tool users Sample size (participants/ data set): 53 participants completed the evaluation, 15 participants completed the think-aloud phase

TABLE 1 Studies of symptom checkers as a single intervention (*continued*)

Reference	Study design	System type	Comparator	Population/sample
Nagykaldi *et al.*[20]	Uncontrolled: observational	Online: a customised practice website, including a bilingual influenza self-triage module, a downloadable influenza toolkit and an electronic messaging capability. A bilingual seasonal influenza telephone hotline was available as an alternative	None	Population/condition: specific condition(s); influenza Sample size (participants/data set): the website or hotline was available to patients at nine primary care practices through the peak of 2007–8 influenza season. After technology testing, a random convenience sample of patients was selected from participants who had used the website or telephone hotline. Qualitative feedback was obtained from 37 patients and six clinicians
Nijland *et al.*[21]	Uncontrolled: observational and a retrospective analysis of 15 months' data	Online: a web-based triage system (www.dokterdokter.nl)	None	Population/condition: general population Sample size (participants/data set): 13,133 participants started triage, 6538 participants entered a complaint and 3812 participants received medical advice. A total of 192 patients completed a follow-up survey on compliance, of whom 35 reported on actual compliance
Poote *et al.*[22]	Uncontrolled: observational	Online: a prototype self-assessment triage system	Health professional performance on real-world data. GPs' triage rating was compared with the rating from the self-assessment system	Population/condition: general population; students attending a university student health centre with new acute symptoms Sample size (participants/data set): 207 consultations; full data available for 154 students attending a university student health centre with new acute symptoms; seven GPs participated in the study
Anhang Price *et al.*[23]	Uncontrolled: observational	Online: a web-based decision support tool, SORT for Kids, designed to help parents and adult caregivers decide whether or not a child with possible influenza symptoms needs to visit the ED for immediate care	Health professional performance on real-world data. The sensitivity of the algorithm was compared with a gold standard (i.e. evidence from the child's medical records that they received one or more of five ED-specific interventions)	Population/condition: specific condition(s) – influenza in children Sample size (participants/data set): 294 parents or adult caregivers presenting to one of two EDs in the National Capital Region, USA

continued

TABLE 1 Studies of symptom checkers as a single intervention (*continued*)

Reference	Study design	System type	Comparator	Population/sample
Semigran *et al.*[6]	Experimental: described as an audit study	Multiple: 23 symptom checkers were evaluated. Symptom checkers available as apps [via the App Store (Apple Inc., Cupertino, CA, USA) and Google Play (Google Inc., Mountain View, CA, USA)] were identified through searching for 'symptom checker' and 'medical diagnosis', and the first 240 results were screened. Symptom checkers available online were identified through searching Google and Google Scholar (Google Inc., Mountain View, CA, USA) for 'symptom checker' and 'medical diagnosis' and the first 300 results were screened	Other: vignettes had a diagnosis and triage attached to them and these were compared with the symptom checker advice	Population/condition: general population, a single class of illness was examined by the symptom checker; the symptom checker was excluded from the study Sample size (participants/ data set): 23 symptom checkers and 45 standardised patient clinical vignettes. These varied by urgency; of the 45 clinical vignettes, 15 required emergency care, 15 required non-emergency care and 15 required self-care. They also varied by how common/uncommon the condition was
Semigran *et al.*[13]	Experimental: a comparison of physician and symptom checker diagnoses based on clinical vignettes	Multiple: 'Human Dx' is a web- and app-based platform	Health professional performance on test/ simulation. Clinical vignettes: a comparison of 23 symptom checkers with a physician's diagnosis for 45 vignettes	Population/condition: general population; of the 45 condition vignettes there were 15 low-, 15 medium- and 15 high-acuity vignettes. There were 26 common and 19 uncommon condition vignettes Sample size (participants/ data set): 45 vignettes were distributed to 234 physicians (211 physicians were trained in internal medicine and 121 physicians were fellows/ residents)
Sole *et al.*[24]	Uncontrolled: observational, a descriptive comparative study	Online: a web-based triage system (24/7 WebMed)	Health professional performance on real-world data. Data were evaluated from students who had used the web-based triage and then requested an appointment via e-mail (so triage data were available for comparison)	Population/condition: general population Sample size (participants/ data set): students who used the web-based triage system between February and May 2004 (4 months)
Yardley *et al.*[25]	Experimental: an exploratory randomised trial	Online: Internet Doctor website	Other: self-care information provided as a static web page with no symptom checker or triage advice	Population/condition: specific condition(s), minor respiratory symptoms (e.g. cough, sore throat, fever, runny nose) Sample size (participants/ data set): 714 participants (368 participants allocated to Internet Doctor and 346 participants to the control group)

TABLE 1 Studies of symptom checkers as a single intervention (*continued*)

Reference	Study design	System type	Comparator	Population/sample
Reports				
Babylon Health[26]	Uncontrolled: observational No control group but some comparison with NHS111 telephone data	Digital: smartphone app	Health professional performance on real-world data Other: NHS111 data for 12 months from February 2017	Population/condition: general population; participants in the London pilot evaluation of digital 111 services Sample size (participants/data set): 12,299 interactions with the system, of which 5250 were classified as genuine (user agreed to share their data with NHS); 74 cases triaged to urgent and emergency care settings were reviewed in depth
Middleton *et al.*[27]	Simulation	Digital: 'babylon check' automatic triage system	Health professional performance on test/simulation. Twelve clinicians and 17 nurses	Population/condition: general population Sample size (participants/data set): 102 vignettes with professional actors playing patients
Conference abstracts				
Berry *et al.*[28]	Simulation: an evaluation of symptom checker performance on clinical vignettes	Online: 17 symptom checkers	None	Population/condition: specific condition(s); gastrointestinal symptoms Sample size (participants/data set): 10 clinical vignettes multiplied by 17 symptom checkers = 170 diagnoses
Berry *et al.*[29]	Controlled: observational	Online: three online symptom checkers (WebMD, iTriage and FreeMD)	Health professional performance on real-world data	Population/condition: specific condition(s), patients with a cough presenting to an internal medicine clinic Sample size (participants/data set): 116 adult patients
Berry *et al.*[30]	Controlled: observational	Online: three online symptom checkers (WebMD, iTriage and FreeMD)	Health professional performance on real-world data	Population/condition: specific condition(s), abdominal pain Sample size (participants/data set): 49 adult patients presenting with abdominal pain

RTI, respiratory tract infection; SORT, Strategy for Off-Site Rapid Triage.

or gastrointestinal[28,30] symptoms that we considered to be 'urgent'). The remaining studies in this group evaluated symptom checkers that covered a wider range of common urgent-care symptoms. Studies evaluated either a single system[21,22,24,27] or multiple systems.[6,13] We found only one study of a symptom checker specifically intended for the assessment of children's symptoms, a development of the Strategy for Off-Site Rapid Triage (SORT) system for influenza-like illnesses.[23] Two reports with some overlap of content evaluated the 'babylon check' application (app).[26,27]

Five studies[7,31–34] evaluated symptom checkers as part of a broader self-assessment and consultation system (often referred to as electronic consultation or e-consultation). Study characteristics are summarised in *Table 2*. In this type of system, the role of symptom checkers is to help patients decide whether their

TABLE 2 Studies of symptom checkers as part of a larger service/health system

Reference	Study design	System type	Comparator	Population/sample
Peer-reviewed papers				
Carter *et al.*[31]	Uncontrolled: observational, a mixed-methods evaluation	Online: webGP (subsequently known as eConsult)	Other: investigate patient experience by surveying patients who had used webGP and by comparing their experience with controls (patients who had received a face-to-face consultation during the same time period) matched for age and gender	Population/condition: general population, general practices in NHS Northern, Eastern and Western Devon CCG's area Sample size (participants/data set): six practices provided consultations' data, 20 GPs completed case reports (regarding 61 e-consults), 81 patients completed questionnaires, five GPs and five administrators were interviewed
Cowie *et al.*[32]	Uncontrolled: observational, a 6-month evaluation at 11 GPs in Scotland	Online: eConsult, accessed via general practice surgery websites. Service provides self-care assessment and advice, including symptom checkers, triage, signposting to alternative services, access to NHS24 (telephone service) and e-consults allowing the submission of details by e-mail	None	Population/condition: general population; patients registered with participating GPs Sample size (participants/data set): sample size for quantitative analysis not reported; 48 practice staff took part in focus groups or interviews
Nijland *et al.*[33]	Other: online survey	Online: responses of interest related to 'indirect e-consultation' (consulting a GP via secure e-mail with the intervention of a web-based triage system)	None	Population/condition: general population; patients with internet access but no experience of e-consultation Sample size (participants/data set): 1066 patients
Reports				
Madan[34]	Uncontrolled: observational, a report of a 6-month pilot study	Online: webGP (subsequently known as eConsult)	None	Population/condition: general population Sample size (participants/data set): service available to 133,000 patients of 20 London GPs via practice websites

TABLE 2 Studies of symptom checkers as part of a larger service/health system (*continued*)

Reference	Study design	System type	Comparator	Population/sample
NHS England[7]	Uncontrolled: observational, an analysis of data from four pilot studies together with data from other sources	Multiple: pilots featured NHS Pathways (web-based, West Yorkshire), Sense.ly ('voice-activated avatar', West Midlands), Espert 24 (web-based, Suffolk) and babylon (app, North Central London)	None. Authors stated that it was not appropriate to compare pilot sites because of differences in starting date, 'footprints' covered, method of uptake and underlying population differences	Population/condition: general population Sample size (participants/ data set): 10,902 downloads or registrations across all pilot sites between January and June 2017

CCG, Clinical Commissioning Group.

symptoms require a consultation with a doctor or other health professional or can be dealt with by self-care. If a consultation is required, details of the symptoms and a request for an appointment or a call-back can be submitted electronically. This type of study is important because it considers the service in the broader context of the urgent and emergency care system. A limitation is that some studies focused mainly on the 'downstream' elements of the pathway (e.g. consultation with GPs) and provided limited data on the symptom checker element of the system.

A final group of five studies (*Table 3*) examined patient and/or public attitudes to online self-diagnosis in the context of urgent care.[35–39]

The included studies used a wide range of designs, some of which were challenging to classify. Studies that assessed the systems' performance in terms of clinical and service use outcomes, including patient satisfaction, were generally observational or qualitative, although one web-based system was evaluated in two RCTs.[17,25] Diagnostic accuracy was assessed by measuring performance in simulations (e.g. using vignettes to describe symptoms for a known condition) or by comparing the system's performance with that of doctors in diagnosing the cause of symptoms and choosing an appropriate level of triage. This was carried out using either simulated data or real-world data (e.g. by asking patients to complete a symptom checker assessment before seeing a doctor and comparing the two assessments). Risk of bias in studies and the overall strength of evidence for different outcomes are discussed further below (see *Risk-of-bias assessment* and *Overall strength of evidence assessment*, respectively).

The publication status of the included studies also varied. In addition to peer-reviewed journal articles and reports, we included four studies published as conference abstracts[18,28–30] and three reports characterised as grey literature. Two of these reports[27,34] were written and published by the developers of specific systems and may therefore be subject to potential conflicts of interest. The third grey literature report was a draft report of NHS England's ongoing pilot studies of NHS111 online in four regions of England.[7] This report, dated December 2017, was not officially published by NHS England but was easily available online.

Characteristics of included systems

Using studies included in the review, we were able to extract data on eight systems using the TIDieST checklist. When appropriate, data from multiple studies were combined in one checklist. We noted that some of these systems were no longer in use. We were also unable to obtain sufficient data for some systems that are currently being used and/or evaluated. A summary of extracted data is presented below (*Tables 4* and *5*) and full data extractions can be found in *Appendix 3*.

TABLE 3 Studies of patient and public attitudes to online self-diagnosis and symptom checkers

Reference	Study design	System type	Comparator	Population/sample
Backman *et al.*[35]	Qualitative interview, questionnaire and registry of health-care contacts following the index visit	Multiple: participants were asked whether they had used telephone consultation, the internet or any other source to obtain health advice	None	Population/condition: general population; non-urgent ED patients Sample size (participants/data set): 543 patients (396 in primary care and 147 in ED) eligible, of whom 428 (79%) were interviewed and received a questionnaire and a 30-day follow-up health-care contact
Joury *et al.*[36]	Other: cross-sectional search for first 30 websites using 'chest pain' keyword. Quality assessment and content analysis	Online: internet websites that provide patient information on chest pain	None	Population/condition: specific condition(s); chest pain Sample size (participants/data set): 27 websites included
Lanseng and Andreassen[37]	Other: cross-sectional (survey) scenario and questionnaire, adapted version of the Technology Readiness Measurement instrument	Multiple: technology (i.e. the study looked at attitudes to digital technology in general rather than any specific system)	None	Population/condition: general population aged 18–65 years Sample size (participants/data set): TRI – 160 participants randomly selected from one county, of which 132 participants responded manually, 28 participants responded online and six participants responded verbally. Women constituted 46.3% of the sample and nearly half the sample was aged 31–50 years. TAM – 470 inhabitants of an affluent Oslo suburb were sampled. Women constituted 59.1% of the sample
Luger and Suls[38]	Simulation: vignettes simulating either appendicitis symptoms or sinusitis symptoms	Multiple: WebMD symptom checker, Google search, no electronic aid	Other: compared the two online health information sources to no electronic aid	Population/condition: specific condition(s); appendicitis or sinusitis Sample size (participants/data set): undergraduate population of students ($n = 174$, mean age 19.22 years)
North *et al.*[39]	Controlled observational comparison of telephone triage and internet triage (calls and clicks, matching of items)	Multiple: online vs. telephone triage	Other: Ask Mayo Clinic – all assessments in 2009 (70,370 symptom assessments) Categorisation: paediatric (0–17 years) or adult	Population/condition: general population; assessment of 28 online symptoms vs. 20 matched telephone symptoms Sample size (participants/data set): data set from Ask Mayo Clinic. All assessments in 2009 (2,059,299 symptom checker clicks). Categorisation – paediatric (0–17 years) or adult

TAM, Technology Acceptance Model; TRI, Technology Readiness Index.

TABLE 4 Characteristics of systems covering a full range of symptoms

Brief name	Objective of the intervention	Interface	Procedures involved in the intervention	How the intervention is accessed	Tailoring	Modifications/ versions	Simulation/ laboratory testing	Real-world testing	Reference(s)
Babylon check	To provide an automated service allowing patients to check symptoms and receive fast and clear advice on what action to take	'App with a chat bot-style interface'[7]	The user selects a body part and answers a series of multiple-choice questions. This process leads to a list of possible outcomes, of which the highest priority one is presented to the user[27]	Smartphone app	Not reported	Not reported	Tested by the manufacturer in two stages[27]	Included in the NHS England pilot evaluation[26,27]	Middleton *et al.*,[27] NHS England[7] and Babylon Health[26]
Un-named prototype	To enable patients to undertake a self-assessment triage and receive advice on an appropriate course of action based on their symptoms	A simple user interface and a menu from which patients could select their main presenting symptom from a list of several hundred presenting complaints	System-generated age- and gender-specific questions with associated potential answers. Each answer carried a weighting that contributed to the final triage outcome. Triage advice provided by the system consisted of one of six courses of action	In the study evaluating the system, access was via a desktop computer[22]	Not reported	Not reported	Not reported	The system was tested in a university student health centre by Poote *et al.*[22] Students used the system before a face-to-face consultation with a GP. The system's rating of the urgency of the student's condition was compared with that of the GP (who had access to the output from the automated system)	Poote *et al.*[22]
WebGP (subsequently renamed eConsult)	To provide an electronic GP consultation and self-help service for primary care patients	Screenshot of the original webGP available in the 2014 pilot report[34]	WebGP consists of five services: symptom checker, self-help guidance, signposting to other services, information about the 111 telephone service and an e-consultation that allows the patient to complete an online form that is e-mailed to the practice. Details of how the system is integrated into practice procedures are reported to vary between practices	Via practice websites	Not reported	Not reported	Not reported	A 6-month pilot report was produced by the Hurley Group, which was involved in developing the system.[34] Subsequent evaluations have been reported in the UK, including in six practices in Devon[31] and 11 practices across Scotland[32]	Madan *et al.*,[34] Carter *et al.*[31] and Cowie *et al.*[32]

continued

TABLE 4 Characteristics of systems covering a full range of symptoms (*continued*)

Brief name	Objective of the intervention	Interface	Procedures involved in the intervention	How the intervention is accessed	Tailoring	Modifications/ versions	Simulation/ laboratory testing	Real-world testing	Reference(s)
24/7 WebMed	To enhance services provided by a SHS by helping students decide whether or not to seek care	The system is no longer available	The system collected basic demographic data and then users answered a series of questions based on algorithms. The system could analyse > 600 chief complaints and classified assessments into six different levels of urgency. After completing triage, students could request a SHS appointment by e-mail	Via a link from the SHS website	Not reported	Not reported	Not reported	Testing of the system by the SHS at the University of Central Florida was reported by Sole *et al.*[24]	Sole *et al.*[24]

SHS, student health service.

TABLE 5 Characteristics of systems covering specific conditions or symptoms

Brief name	Objective of the intervention	Interface	Procedures involved in the intervention	How the intervention is accessed	Tailoring	Modifications/ versions	Simulation/ laboratory testing	Real-world testing	Reference(s)
Influenza Self-Triage Module	To enhance patient self-management of seasonal influenza and to facilitate patient–provider communication	No longer available	The self-triage module was developed by a practice-based research network multidisciplinary stakeholder group and was provided to primary care practices as part of an influenza management website that was tailored to the needs of each participating practice	Via the websites of participating practices	English and Spanish versions available	Additional questions added to improve patient safety	Not reported	The system was tested in 12 primary care practices during the peak of the 2007–8 influenza season[20]	Nagykaldi *et al.*[20]
SORT	To create a simple but accurate tool that could help minimally trained health-care workers screen large numbers of patients with influenza-like illness	See online figures E1–E3. SORT versions 1.0–3.0. No screenshots. Two interactive websites (www.flu.gov and www.H1N1 ResponseCenter.com)	Three-step process resulting in the assignment of a level of risk and specific recommended action	Websites including www.flu.gov	Potential for adaptation to other acute illnesses (e.g. severe acute respiratory syndrome)	Initially intended for use by minimally trained health-care workers, SORT was modified to allow use by call centres and for self-assessment	Retrospective assessment by the Kaiser Permanente Colorado Institute for Health Research, using their health system's computerised records	Analysis of data from websites that made SORT available	Kellermann *et al.*[16]
SORT for Kids	To help parents and adult caregivers determine if a child with influenza-like illness requires immediate care in an ED	For a screenshot see figure 2 in Anhang Price *et al.*[23]	Four age-specific pathways resulting in assignment to one of three risk levels and associated recommended actions	Accessed via website (URL not provided)	Tailored for different age groups and also takes account of child's general state of health	Not reported	Not reported	SORT for Kids was tested by Anhang Price *et al.*[23] at two paediatric EDs in the National Capital Region of the USA	Anhang Price *et al.*[23]

continued

TABLE 5 Characteristics of systems covering specific conditions or symptoms (*continued*)

Brief name	Objective of the intervention	Interface	Procedures involved in the intervention	How the intervention is accessed	Tailoring	Modifications/ versions	Simulation/ laboratory testing	Real-world testing	Reference(s)
Internet Doctor	To provide tailored advice on self-management of minor respiratory symptoms	A screenshot of the interface is provided in Yardley *et al.*[25] Further details are provided in a multimedia appendix to the paper	The diagnostic pages asked a series of questions about the participant's symptoms. These were completed for one symptom at a time and the algorithm provided advice on whether or not the participant should contact health services (NHS Direct) for that symptom	Via web pages (www.internetdr.org.uk)	Not reported	Not reported	Not reported	A preliminary RCT primarily involving university students,[25] followed by a larger RCT in a UK primary care population[17]	Yardley *et al.*[25] and Little *et al.*[17]

Four of the included systems were designed to cover a full range of symptoms (see *Table 4*) and four covered a more limited range (three for influenza-like illness and one for minor respiratory symptoms; see *Table 5*). Most systems were accessed through web pages, often linked to health-care providers or government organisations. The 'babylon check' system was the main exception, being designed for access using a smartphone app. Published research studies provided relatively little detail about the systems that, in some cases, possibly reflected a need for commercial confidentiality. Details of published studies are included in *Tables 4* and *5* for reference; the findings of these studies are summarised in the following section for each key outcome.

Results by outcome

Safety

Patient safety is an important outcome in evaluating any urgent-care intervention. Misdiagnosis or undertriage of serious conditions could lead to death or serious adverse events. However, none of the six included studies that reported on safety outcomes identified any problems or differences in outcomes between symptom checkers and health professionals (*Table 6*). Most of the studies compared system performance with that of health professionals using real or simulated data. The only study with no comparison group was the 6-month pilot study of webGP,[34] which reported 'no major incidents'.

Limitations of the studies included not being based on real patient data,[27] covering only a limited range of conditions,[16,23] and sampling a young and healthy population (students) that was not representative of the general population of users of the urgent-care system.[22] Studies of e-consultation systems did not generally collect data on those respondents who decided not to seek an appointment, which limited their ability to assess any impact on safety for this group. Overall, the evidence should be interpreted cautiously as indicating no evidence of a detrimental impact on safety rather than evidence of no detrimental effect.

Clinical effectiveness

Only two studies reported on clinical effectiveness outcomes (*Table 7*), making it difficult to draw any firm conclusions. In the study by Little *et al.*,[17] those who used the Internet Doctor website experienced longer illness duration and more days of illness that was rated as moderately bad or worse than the usual-care group, although the difference was not statistically significant. The online intervention in this study was designed to offer self-management advice for respiratory infections only. The pilot study of the webGP system[34] reported that several patients received advice to seek treatment for serious symptoms that might otherwise have been ignored. However, no details or quantitative data were provided.

Costs and cost-effectiveness

Three included studies provided limited data on possible cost savings (*Table 8*). Based on 6 months of pilot data, Madan[34] estimated savings of £11,000 annually for an average general practice (6500 patients) compared with current practice. The report also suggested that there might be a saving to commissioners equivalent to £414,000 annually for a Clinical Commissioning Group (CCG) covering 250,000 patients. These savings were specifically related to the self-reported diversion of patients from GP appointments to self-care and from urgent care to e-consultation and, as such, were associated with the symptom checker part of the system. Using similar methodology, the manufacturers of the 'babylon check' app claimed average savings of > £10 per triage compared with NHS111 by telephone, based on a higher proportion of patients being recommended to self-care.[26] Neither study represents a formal cost-effectiveness analysis. The fact that these studies were produced by system manufacturers should be taken into account when interpreting their findings.

The other study reporting on costs[32] concentrated on potential savings to practices from using e-consultation and found that savings depended on the percentage of face-to-face appointments avoided by use of e-consultation. Although important for the evaluation of e-consultation systems, this study has less relevance to the role of symptom checkers because patients completing an e-consultation have already decided that they need further contact with the health-care system.

TABLE 6 Summary of studies reporting on safety outcomes

Reference	Type of system	Population/condition	Comparator	Main results
Kellermann *et al.*[16]	Online: SORT was available on two interactive websites	Specific condition(s): influenza symptoms	Health professional performance on real-world data. The algorithm was tested against clinicians' decision on past patient records	The algorithm was modified to be more conservative following testing against patient records. The effect of SORT on care seeking and patient safety could not be determined because only web hits were counted. There were no reports of adverse events following the use of SORT, but it is possible that patients could have delayed care seeking, which could have resulted in harm
Little *et al.*[17]	Online: Internet Doctor website	Specific condition(s): respiratory infections and associated symptoms	Other: usual GP care without access to the Internet Doctor website	There was no evidence of increased hospitalisations in the intervention group (risk ratio 0.13, 95% CI 0.02 to 1.01; $p = 0.051$)
Madan[34]	Online: webGP (subsequently known as eConsult)	General population	None	The report stated that 'we have had no significant events' and the only patient complaint related to a delay in processing an e-consultation
Middleton *et al.*[27]	Digital: 'babylon check' automatic triage system	General population	Health professional performance on test/simulation: 12 clinicians and 17 nurses	All outcomes obtained through 'babylon check' were classified as clinically safe, compared with 98% of outcomes obtained through doctors and 97% of outcomes obtained through nurses
Poote *et al.*[22]	Online: prototype self-assessment triage system	General population: students attending a university student health centre with new acute symptoms	Health professional performance on real-world data. GPs' triage rating was compared with the rating from the self-assessment system	The self-assessment system was generally more risk averse than the GPs' triage rating, meaning that the safety of patients did not appear to be compromised
Anhang Price *et al.*[23]	Online: a web-based decision support tool, SORT for Kids, designed to help parents and adult caregivers decide whether or not a child with possible influenza symptoms needs to visit the ED for immediate care	Specific condition(s): influenza in children	Health professional performance on real-world data. The sensitivity of the algorithm was compared with a gold standard (evidence from the child's medical records that they received one or more of the five ED-specific interventions)	10.2% of patients were classified as low risk, 2.4% of patients were classified as intermediate risk and 87.4% of patients were classified as high risk by the SORT for Kids algorithm, based on parents and caregiver responses. SORT classified 14 patients as high risk of the 15 that met explicit criteria for clinical necessity at their ED visit

TABLE 7 Summary of studies reporting on clinical effectiveness

Reference	Type of system	Population/condition	Comparator	Main results
Little *et al.*[17]	Online: Internet Doctor website	Specific condition(s): respiratory infections and associated symptoms	Other: usual GP care without access to the Internet Doctor website	Clinical effectiveness: illness duration [11.3 days of illness in the intervention group vs. 10.9 days of illness in the control group; multivariate estimate of 0.48 days longer (95% CI –0.16 to 1.12 days; $p = 0.141$)] and days of illness rated moderately bad or worse (0.53 days, 95% CI 0.12 to 0.94 days; $p = 0.012$) were slightly longer in the intervention group. The estimate of slower symptom resolution in the intervention group was attenuated when controlling for whether or not individuals had used web pages that advocated the use of ibuprofen
Madan[34]	Online: webGP (subsequently known as eConsult)	General population	None	Clinical effectiveness: report stated that 'a number of patients' received automated advice to seek urgent treatment for serious symptoms that might otherwise have been ignored

Further economic evaluations should fully compare clearly defined alternatives and consider costs and benefits across the whole health-care system, particularly whether or not savings represent genuine cost reductions or merely transfers of costs from one part of the system to another.

Diagnostic accuracy

Eight studies reported at least some data on the diagnostic accuracy of symptom checkers (*Table 9*). One study[18] was excluded from the table because although it contained some data on diagnostic accuracy of Google and WebMD symptom checkers, it was mainly concerned with patients' ability to self-diagnose using the different systems. Most of the other studies attempted to compare the systems' performance with that of health professionals using real patient data or simulations. Four of these studies[23,28–30] evaluated systems designed to diagnose and triage specific symptoms or conditions and only three[6,13,24] evaluated one or several 'general purpose' symptom checkers.

In spite of the diverse methods and comparisons in the included studies, almost all studies agreed that the diagnostic accuracy of symptom checkers was poor in absolute terms (e.g. in evaluating 'vignettes' designed to test knowledge of specific conditions when the correct diagnosis was already known by definition) or relative to that of health professionals. In the most comprehensive evaluation, Semigran *et al.*[6] evaluated 23 symptom checkers across 770 standardised patient evaluations. Overall, the correct diagnosis was made in 34% of cases (95% CI 31% to 37%), although performance varied widely between symptom checkers, high- and low-acuity conditions, and common and rare conditions. When the same authors compared the 23 symptom checkers with physicians using 43 vignettes, physicians were more likely to list the correct diagnosis first (out of three differential diagnoses) (72.1% vs. 34%; $p < 0.001$) as well as among the top-three diagnoses (84.3% vs. 51.2%; $p < 0.001$).[13]

TABLE 8 Summary of studies reporting on costs or cost-effectiveness

Reference	Type of system	Population/condition	Comparator	Main results
Babylon Health[26]	Digital: smartphone app	General population: participants in the London pilot evaluation of digital 111 services	Health professional performance on real-world data. Other: NHS111 data for 12 months from February 2017	Costs/cost-effectiveness: cost savings were calculated by multiplying 'where patients said they would have gone vs. where babylon triaged them to' by savings for each combination (e.g. 3% triaged to a general practice who would have gone to a hospital, making a saving of £107 with a contribution to total savings per triage of £3.22). Based on data from 1373 patients, this report claims an average saving of £10.79 per triage, of which £5.13 comes from referring patients to a GP or self-management instead of sending them to a hospital
Cowie *et al.*[32]	Online: eConsult, accessed via general practice surgery websites. Service provides self-care assessment and advice, including symptom checkers, triage and signposting to alternative services, access to NHS24 (telephone service) and e-consultations allowing submission of details by e-mail	General population: patients registered with participating general practices	None	Costs/cost-effectiveness: potential cost savings/ e-consults depended on the percentage of general practice appointments saved/ e-consultations submitted, from £1.30 per e-consultation at 74% to –£1 per e-consultation at 60%
Madan[34]	Online: webGP (subsequently known as eConsult)	General population	None	Costs/cost-effectiveness: report stated that cost savings during the pilot were equivalent to £420,000 annually, of which approximately half would go to practices (subject to changes in workforce) and the other half to commissioners because of fewer patients attending urgent care. This was equivalent to an annual saving of £414,000 for an average CCG

The only exception to the rule was an evaluation carried out at a student health centre.[24] Using data from 59 participants who used the 24/7 WebMed system and who were subsequently treated at the health centre, the study found good agreement between chief complaint, 24/7 WebMed classification and provider diagnosis (κ-values of 0.89 to 0.94). This study differed from the others in using data from students rather than a general population sample. Data were obtained only from people who were actually treated, raising a potential issue of partial verification bias. In addition, the students' complaints were generally common and uncomplicated, a scenario in which symptom checkers performed relatively well in the study by Semigran *et al.*[13]

TABLE 9 Summary of studies reporting on diagnostic accuracy

Reference	Type of system	Population/condition	Comparator	Main results
Berry *et al.*[28]	Online: 17 symptom checkers	Specific condition(s): gastrointestinal symptoms	None	Accuracy of diagnosis: 57 out of 170 (33%) correct diagnoses were listed on the symptom checkers; 12 correct diagnoses (7%), 24 correct diagnoses (14%) and 41 correct diagnoses (24%) were listed in the top one, top three and top 10, respectively. Clinical vignettes with additional information (e.g. laboratory test results, vitals and/or images) were more likely to be listed
Berry *et al.*[29]	Online: three online symptom checkers (WebMD, iTriage and FreeMD)	Specific condition(s): patients with a cough presenting to an internal medicine clinic	Health professional performance on real-world data	Accuracy of diagnosis: 26 out of 116 patients reported GERD symptoms. A doctor diagnosed five of these patients with GERD. The diagnostic accuracy of the symptom checkers was poor. Providing doctors with symptom checker data alone did not enhance their diagnostic performance. Doctors given the symptom checker data plus notes of the visit did show improved diagnostic performance but were less accurate than the initial in-person diagnosis
Berry *et al.*[30]	Online: three online symptom checkers (WebMD, iTriage and FreeMD)	Specific condition(s): abdominal pain	Health professional performance on real-world data	Accuracy of diagnosis: physician-determined diagnosis was not listed for any of the symptom checkers in the top one for 40 out of 49 (82%) patients or in the top three for 27 out of 49 (55%) patients. Seven (14%) patients and 18 (37%) patients had their diagnosis listed in the top one and top three, respectively, by one symptom checker
Anhang Price *et al.*[23]	Online: a web-based decision support tool, SORT for Kids, designed to help parents and adult caregivers decide whether or not a child with possible influenza symptoms needs to visit the ED for immediate care	Specific condition(s) influenza in children	Health professional performance on real-world data. The sensitivity of the algorithm was compared with a gold standard (evidence from the child's medical records that they received one or more of the five ED-specific interventions)	Accuracy of diagnosis: 10.2% of patients were classified as low risk, 2.4% of patients were classified as intermediate risk and 87.4% of patients were classified as high risk by the SORT for Kids algorithm, based on parents' and caregivers' responses. SORT classified 14 patients as high risk of the 15 that met the explicit criteria for clinical necessity at their ED visit. The algorithm had an overall sensitivity of 93.3% and a low overall specificity of 12.9%

continued

TABLE 9 Summary of studies reporting on diagnostic accuracy (*continued*)

Reference	Type of system	Population/condition	Comparator	Main results
Semigran *et al.*[6]	Multiple: symptom checkers available as apps (via the App Store and Google Play) were identified through searching for 'symptom checker' and 'medical diagnosis', and the first 240 results were screened. Symptom checkers available online were identified through searching Google and Google Scholar for 'symptom checker' and 'medical diagnosis', and the first 300 results were screened. They also asked the developers of two symptom checkers for recommendations of other systems. They identified 143 checkers, 102 of which were excluded because they used the same content as the other checkers, 25 were excluded because they looked at one condition only, 14 were excluded because they did not give a diagnosis or triage advice (only medical advice) and two were not working. Therefore, 23 symptom checkers were evaluated	General population: when a single class of illness was examined by the symptom checker, the symptom checker was excluded from the study	Other: vignettes had a diagnosis and triage attached to them and these were compared with the symptom checker advice	Accuracy of diagnosis: diagnostic performance was assessed across 770 standardised patient evaluations. Overall, the correct diagnosis was made in 34% of cases (95% CI 31% to 37%). Correct diagnosis varied significantly ($p < 0.001$) by urgency of condition. Urgent conditions had 24% of correct diagnosis (95% CI 19% to 30%), non-urgent conditions had 38% of correct diagnosis (95% CI 32% to 34%) and self-care conditions had 40% of correct diagnosis (95% CI 34% to 47%). Correct diagnosis varied non-significantly for common (38%, 95% CI 34% to 43%; $p = 0.004$) as opposed to non-common (28%, 23% to 33%) conditions. Performance varied across symptom checkers. Correct diagnosis first varied from 5% to 50%. Correct diagnosis was one of the first three listed in 51% of cases (95% CI 47% to 54%) and one of the first 20 in 58% of cases (95% CI 55% to 62%). Symptom checkers that included demographic information did not perform any better than those that did not include this information
Semigran *et al.*[13]	Multiple: 'Human Dx' is a web- and app-based platform	General population: of the 45 condition vignettes there were 15 of low acuity, 15 of medium acuity and 15 of high acuity. There were 26 common and 19 uncommon condition vignettes	Health professional performance on test/simulation. Clinical vignettes, a comparison of 23 symptom checkers with physician diagnosis for 45 vignettes	Accuracy of diagnosis: physician diagnosis accuracy was compared with symptom checker diagnostic accuracy for the 45 vignettes using two sample tests of proportion. Physicians listed the correct diagnosis first in 72.1% of cases versus 34.0% of cases for symptom checkers (84.3% vs. 51.2% for correct diagnosis in top-three diagnoses). Physicians were more likely to list the correct diagnosis for high-acuity and uncommon vignettes (as opposed to low-acuity and common vignettes). The symptom checkers were more likely to list the correct diagnosis for low-acuity and common vignettes

TABLE 9 Summary of studies reporting on diagnostic accuracy (*continued*)

Reference	Type of system	Population/condition	Comparator	Main results
Sole *et al.*[24]	Online: a web-based triage system (24/7 WebMed) with the following features: 1. a decision-making tool that allows students to input specific conditions and symptoms to get an assessment of urgency for health-care services 2. information on specific diseases and conditions 3. the ability to request appointments via e-mail The triage system can analyse over 600 distinct complaints. Triage assessments are sorted into six levels of urgency: 1. emergent (call 111) 2. seek care immediately 3. seek care within 12–24 hours 4. seek care within 2–3 days 5. seek care within 1–2 weeks 6. self-care recommended	General population	Health professional performance on real-world data. Data were evaluated from students who had used the web-based triage and then requested an appointment via e-mail (so triage data were available for comparison). Data from the triage system were compared with that from the Student Health Service Medical Record. Data were extracted on: 1. chief complaint 2. 24/7 WebMed classification 3. provider diagnosis	Accuracy of diagnosis: over the 4 months of the study there were 1290 uses of the web-based triage system, 143 out of the 1290 patients requested an appointment via e-mail and 59 out of the 143 patients were actually treated at the health centre. Self-care was recommended across 22.7% of the 1290 uses of the system. Generally, the medical complaints of the students were common and uncomplicated. Of the 59 who were treated at the health centre, the study calculated agreement between chief complaint, 24/7 WebMed classification and provider diagnosis. Chief complaint and WebMed: $\kappa = 0.94$; $p = 0.000$. Chief complaint and provider diagnosis: $\kappa = 0.91$; $p = 0.000$. WebMed and provider diagnosis: $\kappa = 0.89$; $p = 0.000$

GERD, gastro-oesophageal reflux disease.

Accuracy of disposition (triage and signposting to appropriate services)

The ability to distinguish conditions of different degrees of urgency and advise patients on an appropriate course of action is a vital component of symptom checkers and e-consultation services and is closely linked to both safety and service use. Six included studies[6,7,21,22,26,27] reported on this outcome, and one[30] evaluated a 'general purpose' symptom checker (*Table 10*). As with diagnostic accuracy, diverse methodologies and outcome measures were used.

The results overall presented a mixed picture, but most studies indicated that symptom checkers were inferior and/or more cautious in their triage advice than doctors or other health professionals. In their review of 23 symptom checkers, Semigran *et al.*[6] found that the systems provided appropriate triage advice in 57% (95% CI 52% to 61%) of cases. The rate of appropriate triage advice was higher for emergency cases (80%) than non-emergency (55%) or self-care (33%) cases. Performance also varied across the systems evaluated, with correct triage ranging from 33% to 78%. Similarly, the NHS England pilot evaluation[7] of four systems found that agreement with clinical experts varied from 30% to 95%, although the number of responses also varied, reducing the comparability of the results.

TABLE 10 Summary of studies reporting accuracy of disposition

Reference	Type of system	Population/condition	Comparator	Main results
Babylon Health[26]	Digital: smartphone app	General population: participants in the London pilot evaluation of digital 111 services	Health professional performance on real-world data. Other: NHS111 data for 12 months from February 2017	Accuracy of disposition: 74 triages that resulted in referral to urgent/emergency care were reviewed by three babylon doctors, who concluded that patients were signposted appropriately in all cases
Berry *et al.*[30]	Online: three online symptom checkers (WebMD, iTriage and FreeMD)	Specific condition(s): abdominal pain	Health professional performance on real-world data	Accuracy of disposition: diagnoses were classified as emergency, non-emergency and self-care. Overall, 33% of symptom checker diagnoses were at the same level of seriousness as the physicians', 39% of diagnoses were more serious and 30% of diagnoses were less serious
Middleton *et al.*[27]	Digital: 'babylon check' automatic triage system	General population	Health professional performance on test/ simulation. 12 clinicians and 17 nurses	Accuracy of disposition: available outcomes were A&E, general practice urgent, general practice routine, pharmacy and manage at home. An accurate outcome was produced in 88.2% of cases for 'babylon check', 75.5% of cases for doctors and in 73.5% of cases for nurses. When vignettes were delivered by a medical expert, rather than actors, 'babylon check' accuracy improved to 90.2%
NHS England[7]	Multiple: pilots featured NHS Pathways (web-based, West Yorkshire), Sense.ly ('voice-activated avatar', West Midlands), Expert 24 (web-based, Suffolk) and babylon (app, North Central London)	General population	None: authors stated that it was not appropriate to compare pilot sites because of differences in starting date, 'footprints' covered, method of uptake and underlying population	Accuracy of disposition: senior clinicians tested each system under non-laboratory conditions and answered questions about the clinical accuracy of the product and its usability. Percentage clinical agreement was 95% of diagnoses for Expert 24 (43 responses), 84% of diagnoses for babylon (32 responses), 59% of diagnoses for NHS Pathways (41 responses) and 30% of diagnoses for Sense.ly (10 responses)
Nijland *et al.*[21]	Online: web-based triage system (www.dokterdokter.nl)	General population	None	Accuracy of disposition: advice was to visit a doctor in 85% of cases. Authors stated that the frequency of self-care advice was limited, even when the complaint was treatable by self-care. There were also cases (e.g. headache and urinary complaints) when self-care advice was given that seemed to be inappropriate

TABLE 10 Summary of studies reporting accuracy of disposition (*continued*)

Reference	Type of system	Population/condition	Comparator	Main results
Poote *et al.*[22]	Online: prototype self-assessment triage system	General population: students attending a university student health centre with new acute symptoms	Health professional performance on real-world data. GPs' triage rating was compared with the rating from the self-assessment system	Accuracy of disposition: the GP assessment and the self-assessment system classification of level of urgency was the same in 39% of consultations. There was a low association but it was significant ($p = 0.016$). The self-assessment system advised the urgent levels of care-seeking behaviour more than GPs. In over half of the consultations, the system's advice was to seek care more urgently than was advised by the GPs during the face-to-face consultation
Semigran *et al.*[6]	Multiple: symptom checkers available as apps (via the App Store and Google Play) were identified through searching for 'symptom checker' and 'medical diagnosis', and the first 240 results were screened. Symptom checkers available online were identified through searching Google and Google Scholar for 'symptom checker' and 'medical diagnosis', and the first 300 results were screened. They also asked developers of two symptom checkers for recommendations of other systems. They identified 143 checkers, of which 102 were excluded because they used the same content as other checkers, 25 were excluded because they looked at one condition only, 14 were excluded because they did not give a diagnosis or triage advice (only medical advice) and two were not working. Therefore, 23 symptom checkers were evaluated	General population: when a single class of illness was examined by the symptom checker, the symptom checker was excluded from the study	Other: vignettes had a diagnosis and triage attached to them and these were compared with the symptom checker advice	Accuracy of disposition: triage performance was assessed across 532 standardised patient evaluations. Appropriate advice was given in 57% of cases (95% CI 52% to 61%). Appropriate advice was given more often for urgent care than for non-urgent care or self-care (80%, 95% CI 75% to 86%, vs. 55%, 95% CI 47% to 63%, vs. 33%, 95% CI 26% to 40%; $p < 0.001$). Appropriate triage advice was higher for uncommon diagnoses than for common diagnoses (63%, 95% CI 57% to 70%, vs. 52%, 95% CI 46% to 57%; $p = 0.01$). Four symptom checkers never advised self-care, so these were excluded and appropriate advice was given in 61% of cases (95% CI 56% to 66%). Triage advice also varied significantly ($p < 0.001$) based on who was providing the symptom checker [provider groups and physician associations 68% of cases in which the symptom checker gave appropriate triage advice (95% CI 58% to 77%), private companies 59% (95% CI 53% to 65%), health plans or governments 43% (95% CI 34% to 51%)]

A&E, accident and emergency.

For abdominal pain, Berry *et al.*[30] evaluated three symptom checkers and found that 33% of diagnoses were at the same level of urgency as physician diagnoses (emergency, non-emergency or self-care), 39% of cases were diagnosed as more serious and 30% of cases were diagnosed as less serious than the physician's judgement. A similar level of agreement between algorithm and clinician (39%) was reported by Poote *et al.*,[22] whereas the system evaluated by Nijland *et al.*[21] advised patients to visit a doctor in 85% of cases, even when the symptoms were appropriate for self-care.

The only studies to report clearly equal or superior accuracy of disposition using an automated system were the evaluations of 'babylon check' by the company that developed the system. Middleton *et al.*[27] reported that, when using patient vignettes, the app gave an accurate triage outcome in 88.2% of cases, compared with an accurate triage outcome in 75.5% of cases for doctors and in 73.5% of cases for nurses. When vignettes were delivered by a medical professional rather than actors, the accuracy of 'babylon check' increased to > 90%. A later report looked at triage results obtained as part of the NHS England pilot evaluation, concluding that all of the 74 referrals to urgent or emergency care were appropriate.[26] However, this evidence should be treated with some caution as it was published without peer review and all of the authors were affiliated with the developer of the product.

Impact on service use/diversion

The widespread availability of online and digital systems for self-diagnosis and triage is intended to support self-management when appropriate and optimise the use of the urgent and emergency care system. However, unintended impacts are possible and for this reason it is important to monitor effects on service use, particularly the diversion of patients from urgent-care settings to primary care or self-care. Eight studies reported on this outcome (*Table 11*), although one[16] of them merely stated that it was not possible to assess the effect of the intervention (a web-based influenza triage system) on patients' use of health services.

Three of the studies reporting on service use evaluated e-consultation systems rather than symptom checkers as a single intervention. However, the evaluations of webGP/eConsult by Carter *et al.*[31] and Cowie *et al.*[32] reported on the use of the e-consultation part of the system rather than the symptom checkers. By contrast, the pilot evaluation of the system by its developers reported that 18% of users planned to book an appointment but chose not to do so.[34] In addition, 14% of users reported that they would have attended a walk-in centre or other urgent-care service if they had not had access to the webGP system. These are promising findings but few details of the methodology are available from the report, for example the sample size of the survey and whether respondents were representative of webGP users or the general population of participating practices.

Further evidence on service use comes from the informally published draft report[7] of the NHS England pilot evaluation of four online/digital systems in different regions of England. This report compared the recommendations of the digital systems with those of the NHS111 telephone service over a similar time period (the first months of 2017). Compared with the telephone service, the online and digital systems directed a slightly higher proportion of patients to self-care (18% vs. 14%) and a lower proportion of patients to other primary care services, such as general practices, dental services and pharmacists (40% vs. 60%). Again, these findings should be interpreted cautiously as they are based on short-term data and do not take account of differences between the different systems and pilot regions or between online/digital and telephone service users. The manufacturer's data on the 'babylon check' app collected as part of the NHS England evaluation[26] indicated that patients were more likely to be triaged to self-care by the app than by the NHS111 telephone service (40% vs. 14%). This figure includes people who received information leaflets on self-care as well as those who were actively triaged. If the former group is excluded, the figures for the two services are similar (14% of patients triaged by NHS111 vs. 15.6% of patients triaged by 'babylon check').

In their study of self-assessment for students attending a university health centre, Poote *et al.*[22] found that the prototype system that they studied was able to identify a proportion of cases that doctors considered appropriate for self-care, suggesting a potential to reduce service use. Similarly, Little *et al.*'s[17] RCT of a

TABLE 11 Summary of studies reporting on service use/diversion

Reference	Type of system	Population/condition	Comparator	Main results
Babylon Health[26]	Digital: smartphone app	General population: participants in the London pilot evaluation of digital 111 services	Health professional performance on real-world data. Other: NHS111 data for 12 months from February 2017	Service use/diversion: 'babylon check' app directed more patients to self-management compared with NHS 111 [40% vs. 14%; figure includes people who received general self-management advice as well as those who were actively triaged (15.6%)] and fewer to primary care (29% vs. 60%). Figures for urgent/emergency care were similar (23% vs. 21%)
Carter *et al.*[31]	Online: webGP	General population: general practices in NHS Northern, Eastern and Western Devon CCG's area	Other: investigate patient experience by surveying patients who had used webGP and by comparing their experience with controls (patients who had received a face-to-face consultation during the same time period) matched for age and gender	Service use/diversion: use of WebGP by patients during the evaluation was small and the uptake did not have a noticeable impact on practice workload. The data provided by the different practices on consultation workload varied in its completeness. A total of 41 out of 61 (72%) webGP requests were assessed by GPs to need a face-to-face or telephone consultation. The introduction of webGP appeared to be linked to changes in responsibility and workload between practice staff and between practices and patients. A postal survey was completed by 81 out of 231 patients (35.1% response rate). Patients who were e-consulters were slightly younger and more likely to be employed than patients who were face-to-face respondents. Patients appear to have found WebGP broadly acceptable in terms of timeliness and quality/experience of care provided. All respondents presented similar problems
Cowie *et al.*[32]	Online: eConsult, accessed via general practice surgery websites. Service provides self-care assessment and advice, including symptom checkers, triage and signposting to alternative services, access to NHS24 (telephone service) and e-consultations that allow the submission of details by e-mail	General population: patients registered with participating general practices	None	Service use/diversion: use of eConsult (website visits) increased from about 800 visits in April 2017 to almost 1400 visits by August 2017. Website visits and e-consultations submitted varied widely between practices. Data on the use of the symptom checkers were not reported

continued

TABLE 11 Summary of studies reporting on service use/diversion (*continued*)

Reference	Type of system	Population/condition	Comparator	Main results
Kellermann *et al.*[16]	Online: SORT was available on two interactive websites	Specific condition(s): influenza symptoms	Health professional performance on real-world data. The algorithm was tested against clinicians' decision on past patient records	Service use/diversion: it is impossible to assess the effect of SORT on patient care-seeking. SORT could have delayed patients' care-seeking or it could have led to unnecessary trips to the ED
Little *et al.*[17]	Online: Internet Doctor website	Specific condition(s): Respiratory infections and associated symptoms	Other: usual GP care without access to the Internet Doctor website	Service use/diversion: the intervention group reported more contacts with NHS Direct [intervention 44/1734 contacts (2.5%) vs. control 24/1842 contacts (1.3%); multivariate RR 2.53, 95% CI 1.10 to 5.82; $p = 0.029$] but fewer contacts with doctors [intervention 283/1734 contacts (16.3%) vs. control 368/1845 contacts (20.0%); multivariate RR 0.71, 95% CI 0.53 to 0.95; $p = 0.019$]
Luger and Suls[38]	Multiple: WebMD symptom checker, Google search, no electronic aid	Specific condition(s): appendicitis and sinusitis	Other: compared the two online health information sources with no electronic aid	Service use/diversion: outcome measure was the intention to seek health care Results: no significant differences between the two online sources and intention to seek health care. No significant differences between the first- and third-person point-of-view vignette and intention to seek health care. Participants significantly more likely to speak to a pharmacist [$F(1) = 5.85$; $p > 0.05$], friend [$F(1) = 9.30$; $p > 0.01$] or family member [$F(1) = 14.35$; $p > 0.001$] for the appendicitis vignette than for the sinusitis vignette. Participants expressed more intention to see a doctor [$F(1) = 6.81$; $p > 0.01$] or ED [$F(1) = 9.68$; $p > 0.01$] if self-diagnosed via WebMD or Google search than with no electronic aid
Madan[34]	Online: webGP (subsequently known as eConsult)	General population	None	Service use/diversion: report states that about 18% of registered patients used the service in 6 months. There were thousands of 'click-throughs' to symptom checkers and there were 9000 recorded users of self-help and signpost information; 18% of users planned to book an appointment but changed their mind

TABLE 11 Summary of studies reporting on service use/diversion (*continued*)

Reference	Type of system	Population/condition	Comparator	Main results
NHS England[7]	Multiple: pilots featured NHS Pathways (web-based, West Yorkshire), Sense.ly ('voice-activated avatar', West Midlands), Expert 24 (web-based, Suffolk) and 'babylon check' (app, North Central London)	General population	None: authors stated that it was not appropriate to compare pilot sites because of differences in starting date, 'footprints' covered, method of uptake and underlying population	Service use/diversion: reported outcomes across all pilot sites; approximately 40% of patients were directed to a primary care service, 20% of patients were directed to a 111 clinician call-back, 18% of patients were directed to self-care, 20% of patients were directed to 999 (call an ambulance) or an ED and 2% of patients were directed to other services. Figures for the NHS111 telephone service were: 60% of patients were directed to primary care, 21% of patients were directed to 999 (call an ambulance) or an ED, 14% of patients were directed to self-care and 5% of patients were directed to other services
Poote *et al.*[22]	Online: prototype self-assessment triage system	General population: students attending a university student health centre with new acute symptoms	Health professional performance on real-world data: GPs' triage rating was compared with the rating from the self-assessment system	Service use/diversion: the self-assessment system gave advice for higher levels of urgency more often than GPs, which could result in unnecessary use of health services and could potentially increase health professionals' workloads. The system did correctly identify a proportion of patients whom GPs thought required self-care, which could reduce service use

RR, risk ratio.

web-based symptom checker designed to support self-care for respiratory symptoms reported that patients in the intervention group had fewer contacts with doctors than patients in the usual-care control group, despite having a longer duration of illness and more days with relatively severe symptoms. This was balanced by an increase in contacts with the NHS Direct telephone service, and it should be noted that the system under evaluation recommended that people needing treatment should contact NHS Direct rather than go directly to a doctor. Finally, a study of young adults (students)[38] found that the intention to seek treatment for a hypothetical illness was stronger when the diagnosis was made with the aid of WebMD or Google search than with no electronic aid.

In summary, there is some evidence that access to digital/online symptom checkers may influence patients' use of health services and possibly increase self-care. However, most included studies reported surrogate indicators of service use that should be regarded as indirect evidence. The strongest evidence relates to a theory-based intervention designed to support self-care for a limited range of symptoms.[17,25] Evidence for systems covering a broader range of symptoms is much weaker and is derived mainly from informally published non-peer-reviewed sources.

Patient compliance with triage advice

The effectiveness of digital and online symptom checkers will be influenced by both the quality of the triage advice provided and whether or not patients follow that advice. In particular, if patients choose to access further advice or seek appointments not recommended by the system, then the effect could be to increase overall demand on the urgent-care system or the wider health-care system.

Only two of the included studies reported specifically on patients' compliance (or intention to comply) with the advice received (*Table 12*). The NHS England pilot evaluation[7] in four regions asked participants in two of those regions (Suffolk and London) what they intended to do based on the advice received. It appears that the question was asked when patients were aware of the advice from the system, but it was unclear whether the evaluation covered real or hypothetical cases. No quantitative data were provided but the report stated that in the Suffolk pilot 'overall users would have followed the advice given'[7] (contains public sector information licensed under the Open Government Licence v3.0 www.nationalarchives.gov.uk/doc/open-government-licence/version/3/). However, those who were recommended to call 999 or attend an ED were more likely to seek advice from primary care or self-management. Similarly, in the London region there was generally good agreement between advice and intended action but patients who were recommended to call 999 or to go to an ED indicated that they would seek advice from a GP. These findings require further investigation, as delay in seeking treatment for serious conditions could be dangerous for the patient and ultimately costly for the health service. The study is also limited in that it asked about the patients' intention to comply rather than their actual compliance.

In a study of a web-based triage system in the Netherlands, 192 patients were asked about their intention to comply immediately after receiving advice from the system.[21] Thirty-five patients responded to a follow-up survey on actual compliance, of whom 20 (57%) reported that they had followed the advice. Compliance was correlated with intention to comply, which in turn was correlated with the patient's attitude towards the advice received.

The available evidence on patients' willingness to follow digital or online triage advice is thus limited in both quantity and quality. Given the importance of the outcome, this should be a priority for further research and evaluation.

TABLE 12 Summary of studies reporting on compliance with triage advice

Reference	Type of system	Population/condition	Comparator	Main results
NHS England[7]	Multiple: pilots featured NHS Pathways (web-based, West Yorkshire), Sense.ly ('voice-activated avatar', West Midlands), Expert 24 (web-based, Suffolk) and 'babylon check' (app, North Central London)	General population	None: authors stated that it was not appropriate to compare pilot sites because of differences in starting date, 'footprints' covered, method of uptake and underlying population differences	Compliance with advice received: patients in Suffolk and London were asked what they intended to do based on the advice received. Authors stated that there was a high degree of intent to follow the suggested action for primary care, self-care and calling 111. For more urgent or emergency outcomes, there seemed to be a preference to seek further advice
Nijland *et al.*[21]	Online: web-based triage system (www.dokterdokter.nl)	General population	None	Compliance with advice received: 20 out of 35 patients (57%) reported following advice from the system

Patient/carer satisfaction

Ten out of the 27 included studies had an outcome of patient/carer satisfaction. Five of these studies were investigating symptom checkers as a self-contained intervention.[18,20,21,23,25] Four of these studies were evaluating symptom checkers as part of a broader e-consultation system.[7,31,32,34] One study was examining the attitudes of the public to online self-diagnosis within the context of urgent care.[37]

One study reported on the general acceptability of the WebGP website in terms of timeliness, quality and experience for patients. However, general practice staff reported that online consultations interacted negatively with their own systems.[31] Satisfaction with e-consultations was found to be high among respondents (91.4%), although only 6.5% of users responded to the survey.[32] Similar high satisfaction rates were reported by Madan,[34] although response rates were not reported. NHS England piloted a range of e-consultation systems with high rates of satisfaction for each.[7] Sample sizes varied from small (17 respondents in Leeds) to large (1463 respondents in London). A number of different measures were used in the pilot studies. In Leeds and Suffolk, 75–80% of respondents were likely or extremely likely to recommend the service to friends and family, but sample sizes were small (17 in Leeds and 25 in Suffolk). Users in London rated the service on a scale of 1* to 5*, with 68% of users rating the service 4* or 5*.

Two of the studies investigating patient satisfaction considered self-triage tools for influenza.[20,23] One of the studies was investigating the use of SORT for Kids[23] and the other considered satisfaction with a practice website or telephone hotline designed to support self-management.[20] In total, 90% of parents using SORT for Kids found the website 'very easy' to understand and 91.4% of parents found it 'very easy' to use.[23] Qualitative feedback on patients' satisfaction with a practice website or telephone hotline to support self-management of influenza demonstrated a high level of overall satisfaction.[20] Additionally, findings indicated the ease of use, problem-free access and navigation, and the useful and adequate information that was used in a variety of ways by patients.

A study investigating satisfaction with the app Internet Doctor for minor respiratory problems found that overall satisfaction was higher in the Internet Doctor group than in the control (self-care) group ($p = 0.002$).[25] However, there was a high attrition rate for the follow-up after 48 hours (from 714 users to 332 users) and the sample was generated from a population of undergraduate university students, so findings may not be representative of a general population.

A study investigating the experience of older adults searching for online health information[18] found that they regularly commented on the use and navigation of the tools, the layout and features of the tools, and their difficulties with the computer programs. Findings indicate the importance of ease of navigating and troubleshooting symptom checkers to this population, as well as a general lack of confidence in self-diagnosis.

One of the studies[21] investigating a symptom checker as a stand-alone intervention considered patient use and compliance with the medical advice provided and found that patient attitudes towards the advice received were neutral to positive.

Importance of ease of use of the application, navigation and troubleshooting

One included study[37] surveyed two groups of users about their perceptions of symptom checkers. They measured expected usefulness, expected ease of use, trust in service provider, attitude towards using self-diagnosis and intention to use. The study found that there was a positive relationship between expected usefulness and attitude towards using self-diagnosis, as well as between expected ease of use and attitude towards using self-diagnosis, expected usefulness and ease of use, and expected usefulness/ease of use and trust in service provider. A relationship between behavioural intention and expected usefulness was not supported. These findings suggest that trust in the provider of the app is fundamental to the app's perceived usefulness and ease of use and, indirectly, the users' later behavioural intention.

Equity and inclusion

The results for the comparison of users versus non-users will be considered with the outcome equity and inclusion because there is significant overlap in the results of these outcomes. Fourteen out of the 26 included studies investigated the outcome of equity and inclusion or compared users and non-users. Seven of these studies were evaluating symptom checkers as a self-contained intervention.[13,17,19–23] Five of these studies were evaluating an e-consultation system that included a symptom checker[7,31–34] and two of these studies were investigating public attitudes towards symptom checkers.[35,37]

One study[17] reported that patients who were classified as less socioeconomically deprived were more likely to agree to use Internet Doctor than decline participation, although no relationship was found between socioeconomic deprivation and results in this study or between e-consultation use and socioeconomic deprivation in another study.[32]

The association between e-consultation use and education levels was explored in one study.[33] Patients with low to medium levels of education tended to be motivated towards indirect e-consultation (which involves contact with a health professional via e-mail), mainly to reduce uncertainty.

Evidence from included studies suggests that the users of e-consultation were more likely to be young (age range varied between studies),[7,31,32,34] employed[21,31,34] and female[7,21,32,34] than non-users. One study also found a significantly larger use of e-consultation by white patients (78%) than patients of other ethnicities.[32]

A study investigating the use of health information or advice before ED or primary care centres in Sweden compared the characteristics of users with non-users.[35] Only 10% of participants reported ever having consulted internet sources for information or advice on an urgent issue. More ED (49%) than primary care (12%) attendees were reported to have sought health information prior to their visit, although this information was not necessarily from an online source

Patient and public attitudes to online self-diagnosis

Seven cross-sectional papers were included in the review.[18,19,35–39] Study characteristics are summarised in *Table 13*. The studies were carried out in the USA,[18,36,38,39] Norway[19,37] and Sweden[35] and all but one, a conference abstract,[38] were published articles.

Included studies used a range of cross-sectional research methods, including a review of online resources, surveys, interviews and 'think aloud' sessions to assess the quality, usability, acceptability of and the intention to use internet-based health-care information. Given the generally small samples of specific populations (e.g. students, older people, affluent) the generalisation of findings is limited, but the issues identified could be useful when considering the development of a new service.

Six out of the seven papers were assessed for quality, given that the one conference abstract fell outside our criteria for appraisal.[38] The remaining studies were varied in terms of methods and how the methods fit with the appraisal criteria and, therefore, they were not graded; instead, strengths and weaknesses were discussed. All six papers included a clear statement of the research question and four of them presented clear outcome measures, but confounding variables were less clearly dealt with. Only two studies included a participant response rate,[19,35] and both of these indicated that the rate was > 50%, but only five studies recruited participants. Blinding of assessors was either not appropriate or not stated for all of the studies. Other weaknesses included inconsistency of reporting the findings and a general lack of enough detail to make a decision regarding the risk of bias.

TABLE 13 Characteristics of cross-sectional studies

Author, year, country	Population/sample	Focus of study	Methods	System(s)
Backman, 2012, Sweden[35]	396 primary care patients and 147 ED patients ($n = 543$)	The use of health-care information prior to ED or primary care attendance	Interviews Questionnaire Registry	Multiple (not named)
Joury, 2016, USA[36]	27 websites	The assessment of websites providing information about chest pain	Searches of five search engines A suite of scores to assess quality, content and accessibility	Multiple (not named)
Lanseng, 2007, Norway[37]	General population aged 18–65 years ($n = 160$)	Introduction of SST in health diagnosis	Scenario Questionnaire Structural equation modelling: TAM	Multiple (not named)
Luger, 2011, USA[38]	University students (undergraduates) ($n = 174$)	The effect of obtaining online health-care information on the intention to seek health care	Vignettes depicting hypothetical symptoms for appendicitis and sinusitis	WebMD symptom checker, Google search and no aid
Luger, 2014, USA[18]	Adults aged > 50 years ($n = 79$)	Processes used by older adults to self-diagnose using online health information	'Think aloud' while using online symptom checkers for a maximum of 30 minutes	WebMD symptom checker and Google search
Marco-Ruiz, 2017, Norway[19]	Phase 1: 53 participants Phase 2: 15 participants	Assessment of the respiratory disease module of a symptom checker Analysis of components and usability	Phase 1: TAM questionnaire Principal component analysis Phase 2: 'think aloud'	Erdusyk ('Are you ill?')
North, 2011, USA[39]	General population	Comparing telephone and online triage system use	Data assessment and survey	www.mayoclinic.com (internet, website) Ask Mayo Clinic (call centre)

TAM, Technology Acceptance Model; SST, self-service technology.

Aims of the included studies can be distinguished between those that examine characteristics of the supply side (online sources) and those of the demand side (users and potential users), or both. This indicates an interest in:

- the development of online resources for health-care information or advice
- behaviour of users in relation to online health-care information seeking.

These two issues are dealt with separately in the sections below.

Development of resources

On the provider side, websites developed to provide information resources to the public have been reported to vary in quality and content.[36] Developers can certify and comply with the Health on the Net (HON) code[40] to increase the quality and trustworthiness of the website. Trustworthiness may also be increased by not allowing open editing of the information provided. The goal of the provider can have an impact on the website's usefulness; therefore, stating the aim of the website would allow users to make a decision on how useful the website is for their specific purpose.

The process of navigating sites was reported as an issue that could increase frustration, for example if the pathway takes a long time to go through, the signposting is not clear and/or the entered information is not recognised as saved into the system.[19] Users were reported to be more likely to rush the task if the process was taking a long time, which increased the likelihood of self-diagnostic inaccuracy.[19] The ease of the task and the confidence in performing the task were associated with self-diagnostic accuracy; therefore, optimising usability appears to be an important aspect when designing a health information system. Younger participants were reported to better understand navigation and system issues and, therefore, attention to the needs of older users may be required.

User behaviour

It is important that users understand the meaning of the information that they are being given and the questions that they are being asked. Included studies[19,36] identified two barriers to using the services: low readability scores and user-reported language that was too clinical for lay readers. Requests for symptom information could be ambiguous, particularly in relation to measures (e.g. 'how much', 'how many' and 'what colour'). The use of visual aids or more flexible options could help users to better interpret these types of questions.

Previous health knowledge and experience was reported to influence the accuracy of a self-diagnosis online, with prior health experience having a negative effect and prior health knowledge having a positive effect. Luger and Suls[38] suggest that website design should incorporate a process that avoids patient decision-making based on 'pattern rule', whereby cognitive expediency can overrule more thorough, hypothetico-deductive reasoning.

In terms of intention to use, seeking health information was reported to be more likely prior to a visit to accident and emergency (A&E) than to primary care, but the sources of information and advice were not limited to online websites.[35] Men were more likely to attend A&E without prior information, and younger people were just as likely to express the intention to seek help from a health professional for a minor ailment as a serious one. This suggests that, in attempting to avoid unnecessary attendances, it may be prudent to consider how to raise awareness of online health information in different populations. Online use of health information was reported to be more frequent for adult symptom checks than for children, and for chronic rather than acute symptoms, highlighting how users perceive telephone and internet triage systems differently.[39] For example, telephone triage may provide the needed reassurance from a health professional, particularly for parents and in cases of acute symptoms, compared with visiting an online symptom checker.

Strengths and limitations

As well as small, non-diverse samples, this group of studies used a range of cross-sectional methods, some with comparators and some assessing particular systems. Expert assessment rather than lay assessment, as well as the use of scenario techniques for assessing the use of information systems, may lack external validity compared with alternatives. However, there are stated reasons for using these methods. None of the studies met all of the risk-of-bias criteria, mainly because of the methods used and a lack of detailed reporting. The review findings reflect the heterogeneity of the studies and they cannot be combined in anything but a narrative form. However, the study findings provide insight into potential issues that could be considered when designing an internet tool for health information and advice.

Risk-of-bias assessment

We assessed risk of bias in the two included RCTs[17,25] using the Cochrane risk-of-bias tool.[9] Thirteen studies[16,18–21,24,31–33,35–37,39] were assessed with the tool for cross-sectional and cohort studies and four (six publications[6,13–15,22,23]) were assessed with the modified QUADAS tool.[10] Seven grey literature reports and conference abstracts were not formally assessed for risk of bias.[7,27–30,34,38] Identified limitations were extracted for all included studies, the risk-of-bias results are presented in *Appendix 5*.

No serious problems were identified with the two randomised trials, although blinding was somewhat unclearly reported. The studies evaluated using the cross-sectional/cohort study tool were variable both in design and risk of bias. Most studies addressed a clear research question in a defined population. Participation rates, when applicable, varied widely between studies. None of the studies provided a sample size justification, making their statistical power uncertain, and some findings were based on very small samples. Most of the studies had clearly defined outcome measures but none reported on blinding (e.g. of outcome assessment). Adjustment for confounders is important for observational studies but only three of the included studies were judged to have adjusted adequately for confounders (see *Appendix 5*).

The four 'diagnostic' studies assessed using the modified QUADAS tool had some important limitations. Two did not recruit representative samples of patients[22,23] and the other two used patient vignettes, so were judged 'unclear' for this aspect.[6,13] There were also issues with blinding and potential partial verification bias in individual studies.

With the possible exception of the two RCTs,[17,25] the included studies generally had at least a moderate risk of bias. However, the diverse designs and objectives of the studies made risk of bias difficult to assess in some cases with the available tools. Grey literature reports containing relevant data were included in the review but not formally assessed for risk of bias. Reports prepared by individuals with a commercial interest in a specific system and published without independent peer review[27,34] should be treated with particular caution because of possible conflicts of interest.

Overall strength of evidence assessment/evidence map

The overall strength of evidence for key outcomes was assessed using a variant of the scale used by Baxter *et al.*,[12] which relates the evidence for particular outcomes to the size and quality of the whole body of evidence (*Table 14*). The rating scale was as follows: 'stronger evidence' represented generally consistent findings in multiple studies with a comparator group design or comparative diagnostic accuracy studies, 'weaker evidence' represented generally consistent findings in one study with a comparator group design and several non-comparator studies or multiple non-comparator studies, 'very limited evidence' represented an outcome reported by a single study and 'inconsistent evidence' represented an outcome where < 75% of the studies agreed on the direction of effect.

Using this scale, there is relatively strong evidence that digital and online symptom checkers have not been shown to present any major patient safety issues, that their diagnostic accuracy tends to be lower than that of health professionals and that patients who have used these systems generally show high levels of satisfaction (mainly in non-comparative studies). Areas where evidence is lacking or inconsistent include clinical effectiveness and cost-effectiveness, accuracy of disposition to appropriate services and patient compliance with advice received. The implications of these findings will be considered in *Chapter 4*.

TABLE 14 Overall strength of the evidence assessment for key outcomes

Outcome	Relevant studies	Evidence statement	Strength of evidence	Comments
Safety	= **Kellermann *et al.***[16] = **Little *et al.***[17] = **Middleton *et al.***[27] = **Poote *et al.***[22] = **Anhang Price *et al.***[23] Madan[34]	No evidence of a difference in risk between health professionals and symptom checkers	Weaker	Rating changed from 'stronger', based on study numbers and design, to 'weaker' because of low numbers of adverse events reported
Clinical effectiveness	– **Little *et al.***[17] ? Madan[34]	Insufficient evidence to draw any firm conclusions	Very limited	
Costs/cost-effectiveness	+ Babylon Health[26] ± Cowie *et al.*[32] + Madan[34]	Insufficient evidence to draw any firm conclusions	Inconsistent	
Diagnostic accuracy	? Berry *et al.*[28] – **Berry *et al.***[29] – **Berry *et al.***[30] – **Anhang Price *et al.***[23] ? Semigran *et al.*[6] – **Semigran *et al.***[13] = **Sole *et al.***[24]	Symptom checkers appear inferior to health professionals in terms of diagnostic accuracy	Stronger	Mainly for specific conditions or pre-prepared vignettes
Disposition accuracy	= Babylon Health[26] – **Berry *et al.***[30] = **Middleton *et al.***[27] ? Nijland *et al.*[21] – **Poote *et al.***[22] ± Semigran *et al.*[6] ± **NHS England**[7]	Inconsistent findings on accuracy of disposition	Inconsistent	Performance variable between different systems
Service use/diversion	**? Kellermann 2010**[16] ± **Little *et al.***[17] ± **Poote *et al.***[22] **? Carter *et al.***[31] ? Cowie *et al.*[32] + Madan[34] ± **NHS England**[7] + Babylon Health[26] – Luger *et al.*[18]	Inconsistent findings on effects on service use	Inconsistent	

TABLE 14 Overall strength of the evidence assessment for key outcomes (*continued*)

Outcome	Relevant studies	Evidence statement	Strength of evidence	Comments
Compliance	? Nijland *et al.*[21] ? NHS England[7]	No comparative data on compliance	Very limited	
Patient/carer satisfaction	? Nagykaldi *et al.*[20] ? Nijland *et al.*[21] ? Anhang Price *et al.*[23] + Yardley *et al.*[25] ? Carter *et al.*[31] ? Cowie *et al.*[32] ? Madan[34] ? NHS England[7] ? Lanseng and Andreassen[37]	Most studies report high rates of patient satisfaction with symptom checkers and e-consultation systems generally	Weaker	Few studies with comparator data

Notes
Controlled studies in bold.
=, no significant difference in outcomes; +, better outcome with symptom checker; +/–, varying results within study; –, worse outcome with symptom checker; ?, results difficult to interpret in comparative terms.

Chapter 4 Discussion and conclusions

Main findings

The literature search identified 29 publications describing 27 studies that met the inclusion criteria. Studies were published between 2006 and 2018 and were diverse in terms of their design and methodology. Four items of grey literature were included in view of their importance to the topic, although three were produced by the developers of specific systems and should be interpreted with that fact in mind.[26,27,34] At the time of writing this report, NHS England's evaluation of four different systems in different regions was still in progress but an informally published draft report was located online[7] and we have used that in our review while acknowledging that it represents an incomplete preliminary data set.

The overall strength of the evidence base varied between outcomes (see *Table 14*), but in absolute terms the evidence is weak, being based largely on observational studies. A substantial component of grey literature of uncertain quality complicates the interpretation of the evidence. Interpretation of the evidence should also take into account the risks of bias in individual studies, as discussed in *Chapter 3*, *Risk-of-bias assessment*.

We found little evidence to support the hypothesis that digital and online symptom checkers are detrimental to patient safety. Although the evidence was comparatively strong relative to other outcomes assessed in the review, it falls far short of being an adequate assessment of the safety of these systems. The studies that reported on the outcome were mostly short term and involved relatively small samples. Some were limited to people with specific types of symptoms (e.g. influenza-like illness[23] or respiratory infections[17]) and others recruited from specific population groups (e.g. students[22]) who were not representative of typical users of urgent-care services. This body of evidence should, therefore, be interpreted cautiously and not be extrapolated to the possible impact of a nationally available digital urgent-care service being used by millions of people annually. Anecdotal reports of informal tests circulated via social media show symptom checkers making serious errors, but we were unable to locate any research to substantiate these claims.

The evidence on patient satisfaction with digital and online systems also had some limitations, but these findings appear more likely to be generalisable. Study participants generally expressed high levels of satisfaction, albeit in uncontrolled studies. For example, in the NHS England pilot evaluation, 70–80% of users were satisfied with their experience at each of the pilot sites.[7] This was based on a sample of > 1500 users, the great majority of whom were involved in the London pilot using the 'babylon check' app. Despite the unbalanced sample, the findings on satisfaction suggest that any national, digital urgent-care service may be well-used, particularly given the increasing reliance on digital technology in all areas of life. Satisfaction was measured in a number of ways, but generally the studies appeared to be rating usability rather than satisfaction with the advice received or the degree of reassurance provided. Failure to provide the patient with reassurance could result in them seeking further help elsewhere in the urgent and emergency care system.

The evidence from the studies included in the review suggests that digital and online systems have yet to achieve a high level of accuracy in the diagnosis of specific conditions. This finding applies both to 'general purpose' symptom checkers and to those limited to particular conditions. Although the evidence was classified as relatively strong, several caveats should be applied. For example, some of the included studies did not recruit representative populations and others were based on standardised vignettes rather than real-world data. Finally, those studies that compared symptom checkers with health professionals tended to use the doctors' clinical diagnosis as the reference standard, which would bias the comparison in favour of the health professionals. There was an indication from one study that the symptom checker performed better when additional information was available.[29]

The accuracy of signposting patients to the most appropriate level of service is closely related to diagnostic accuracy but results for this outcome were inconsistent between studies. Studies researching the accuracy of signposting used similar methods to those evaluating diagnostic accuracy. A review of 23 symptom checkers found that triage level was appropriate overall for 57% of patient vignettes, with considerable variation between systems and conditions.[6] The NHS England evaluation also found differing levels of agreement with clinical expert opinion across the four systems evaluated.[7] In general, algorithm-based triage tended to be more risk averse than that of health professionals, with 85% of respondents being advised to visit their doctor in one study.[21] Although there is considerable uncertainty about the magnitude of the effect, a national, digital urgent-care service could result in considerable numbers of patients receiving inappropriate advice to visit the ED or request an urgent general practice appointment. Middleton *et al.*[27] claimed that the 'babylon check' app had a high degree of triage accuracy for vignettes when compared with health professionals, but this report was not peer reviewed and so requires further validation.

We also found inconsistent evidence on effects on service use, but there was some indication that symptom checkers can influence the pattern of service use. The strongest evidence comes from a randomised trial of an intervention that was specifically designed to promote self-care and covered respiratory symptoms only.[17] In this study, the intervention group had less contact with doctors (but more contact with NHS Direct) than the control participants despite having a longer duration of illness and illness of a greater severity. It is uncertain whether or not these results are generalisable to systems covering the full range of urgent care. However, the NHS England evaluation found a small shift towards self-care with digital triage compared with telephone triage,[7] and Madan[34] reported that the webGP symptom checkers diverted 18% of patients from requesting a general practice appointment in that consultation. These are findings from the grey literature that have not been peer reviewed and should, therefore, be interpreted with caution.

There is very limited evidence on patients' reactions to online triage advice and whether they follow the advice or seek further help or information. This would have implications for service use. Preliminary findings from the NHS England evaluation[7] suggest that patients may be more likely to seek further advice for more urgent conditions, but further confirmation is required.

Over half of the included studies considered equity and inclusion issues either directly or by comparing users with non-users of digital triage systems. Not surprisingly, studies revealed a clear consensus that younger and more highly educated people are more likely to use these services, whereas older and less educated people are more likely to prefer telephone or face-to-face contact. This could have implications for health equity if urgent-care pathways prioritise (or appear to prioritise) requests originating from digital sources. In primary care, issues have arisen because patients using e-consultation systems to request an appointment following online triage may be seen more quickly than those contacting the practice by telephone.

In summary, the key findings of this systematic review are that considerable uncertainties surround the diagnostic and triage accuracies of digital and online symptom checkers, their overall effects on service use and whether patients are likely to follow their advice or seek further assistance from other sources. Surveys of patients who have used them report high levels of satisfaction with their usability. Increasing

reliance on digital and online services generally suggests that a national digital 111 service may be heavily used and its impact on the health-care system, in whatever direction, could be substantial. This review found limited evidence to suggest whether the new service will stimulate or manage demand for urgent and emergency care.

Strengths and limitations

This systematic review has been undertaken to a timescale determined by the review commissioners using a relatively large team of experienced researchers, including both methodological experts and topic experts. In other words, we have sought to accelerate the systematic review process rather than abbreviate it by omitting some processes. The main exception to this was the use of one reviewer for some processes (with checking of a 10% sample carried out by a second reviewer) rather than full, independent duplication. The evaluation of the screening process suggested that there was only a moderate degree of agreement between reviewers. This reflects a degree of learning by the review team during the screening process. Our initial sift of the search results consciously favoured inclusivity, and items that were found not to meet the inclusion criteria on detailed examination were subsequently discarded. We performed a rigorous search of the literature, including reference checking and citation searching. Rather than a conventional, highly sensitive search (which would have resulted in inefficiencies in the screening process), we combined an initial focused search with subsequent rounds of follow-up searching, including searches for named symptom checker systems. Overall, we consider it unlikely that any significant research in this area was overlooked.

We assessed risk of bias in individual studies using a variety of appropriate checklists, as well as summarising the overall strength of evidence for key outcomes (see *Table 14*). The protocol was developed in association with NHS England and the NIHR HSDR programme team and was registered prospectively on PROSPERO.

The heterogeneous and descriptive nature of the included studies meant that meta-analysis was not feasible for any of the outcomes of interest. Our narrative synthesis approach used a mixture of description and tabulation to summarise the evidence for each of the prespecified outcomes of interest. In addition, we considered the different symptom checker systems as interventions and extracted data on them using the system rather than the study as the unit of analysis (see p. 23 and *Appendix 3*); we also used an adaptation of an established intervention-reporting checklist to assess them.[8] This process highlighted the lack of information available for some systems and it could be developed to form a template for ongoing evaluations of current and new systems.

This was a review of published, including non-peer-reviewed, literature and the coverage of systems is not exhaustive (e.g. we did not extract data from websites). We also did not carry out any original analyses of raw data even when such data were available. The timing of the review meant that the final results of NHS England's pilot evaluation were not available to us. We were able to make use of a draft report that was published online,[7] but we acknowledge that the findings of the final evaluation report, when available, will supersede those of the 2017 draft.

The review inclusion criteria were relatively broad and findings from symptom checker systems for specific conditions may not be applicable to more general systems and vice versa. We have also included studies of symptom checkers as part of e-consultation systems in general practice, which represent a slightly different setting from a general digital 111 service, and this should be kept in mind when interpreting the results.

Implications for service delivery

The implications of this systematic review for service delivery should be considered in the context that a decision has already been taken to introduce a digital 111 service and that the implementation of the service is in progress. The findings can be used to support the ongoing development and evaluation of the service.

Achieving a high level of diagnostic accuracy will be key to the success of a digital 111 service. Failure to provide an accurate diagnosis may result in outcomes such as patient dissatisfaction and unwillingness to use the service again, increased use of other urgent and emergency care services and possible risks to patient safety (although the cautious approach characteristic of most existing systems that we identified may help to mitigate this).

The studies included in the review suggest a high level of uncertainty about the impact of digital 111 on the urgent-care system and the wider health-care system. Some of these uncertainties can be addressed by research and data collection, as discussed below, but the health service may need to respond to short-term increases (or decreases) in demand and/or shifts from one part of the system to another. This may increase pressure on the system, at least in the short term. In the longer term, if the use of the NHS111 telephone service decreases as planned, there may be opportunities to re-deploy staff to fill other roles within the urgent and emergency care system.

Research on the introduction of the NHS111 telephone service from 2010 was outside the scope of this review. However, the telephone service followed a similar pattern of piloting followed by nationwide implementation to that proposed for digital 111. In brief, the evaluation of the pilot sites (compared with control sites where NHS111 was not available) found that NHS111 did not reduce calls to ambulance services,[5] did not improve public perceptions of urgent care provision[41] and identified key groups that were less likely than others to use the service (including older people and men).[42] However, the majority of respondents in the study of acceptability were very satisfied (73%) or quite satisfied (19%) with the service overall.[42] A later study of five NHS111 service providers covering 2014–16 also found high levels of patient satisfaction, but views of other stakeholders were more mixed.[43] Decision-makers may be able to use the lessons learnt from the introduction of the NHS111 telephone service alongside the findings of this review to inform their choices around the implementation of digital 111. For example, some respondents had concerns about the relevance of the questions they were asked,[42] emphasising the importance of flexibility in the underlying algorithm to prevent users giving up in frustration and seeking help elsewhere in the urgent and emergency care system.

Implications for research

The limitations of the evidence base suggest that further rigorous research is urgently needed. We have identified the following priorities for research (in addition to the ongoing collection of data to monitor the use and safety of the digital 111 service):

- Our research suggests that there is a wide variation in performance between symptom checker systems,[6] but we found no studies designed to compare systems directly. Such research will be important for commissioners to make an informed choice on how to implement services at a local level.
- Evidence on costs or cost-effectiveness was classified as very limited; hence, rigorous economic evaluations based on real-world data are needed. A key question for research is whether or not new digital services will reduce overall costs for the health service and whether or not savings are made in one part of the system at the expense of another.
- The introduction of a digital 111 service provides a new point of access for patients needing urgent medical help. Research is needed to investigate the pathways followed by patients using the service and, particularly, if the overall number and level of contacts with the health system can be reduced without affecting the quality and safety of patient care.
- This review located only one study of a service designed to assist parents and other caregivers to assess the acuity of a child's illness.[23] Research is required to ensure that the systems underlying national digital 111 services are suitable for children's illnesses as well as those of adults, or to evaluate a separate 'paediatric digital 111' service.
- Given the example of the Internet Doctor intervention, with a design based on a specific psychological theory,[17,25] the role of behaviour change theory in the development and implementation of symptom checkers should be investigated.

Conclusions

The current evidence base covers a diverse range of interventions, study designs and outcomes. The evidence is generally weak as it is based largely on observational studies and includes some studies that have not been independently peer reviewed. There are major uncertainties around the probable impact of digital 111 services on most of the important outcomes, but evidence on patient satisfaction, together with developments outside the health field, suggests that once introduced the use of these services may increase rapidly. It will be important to monitor and evaluate the services using all available data sources and by commissioning high-quality research.

Acknowledgements

We are grateful to the members of the Sheffield HSDR Evidence Synthesis Centre PPI Group for their contributions. We thank Emma Knowles for attending the PPI meeting and for commenting on the initial draft of this report.

Contributions of authors

Duncan Chambers (Research Fellow in Public Health) was involved in project co-ordination, study selection, data extraction, quality assessment and report writing.

Anna Cantrell (Research Associate in Health Economics and Decision Science) was involved in information retrieval, study selection, data extraction, quality assessment and report writing.

Maxine Johnson (Research Fellow in Public Health) was involved in study selection, data extraction, quality assessment and report writing.

Louise Preston (Research Fellow in Health Economics and Decision Science) was involved in study selection, data extraction, quality assessment and report writing.

Susan K Baxter (Senior Research Fellow in Public Health) was involved in protocol development, study selection, report writing and co-ordinated the PPI meeting.

Andrew Booth (Reader in Evidence-based Information Practice) was involved in information retrieval, study selection and report writing.

Janette Turner (Reader in Emergency and Urgent Care Research) provided topic expert advice and was involved in report writing.

All authors commented on drafts of the protocol and the report.

Publication

Chambers D, Cantrell AJ, Johnson M, Preston L, Baxter SK, Booth A, Turner J. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. *BMJ Open* 2019;**9**:e027743.

Data-sharing statement

No new data have been created in the preparation of this report and, therefore, there is nothing available for access and further sharing. All queries should be submitted to the corresponding author.

References

1. Chambers D, Cantrell AJ, Johnson M, Preston L, Baxter SK, Booth A, Turner J. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. *BMJ Open* 2019;**9**:e027743. https://doi.org/10.1136/bmjopen-2018-027743
2. NHS England. *Statistical Work Areas NHS111: Minimum Data Set*. Leeds: NHS England; 2018. URL: www.england.nhs.uk/statistics/statistical-work-areas/nhs-111-minimum-data-set/statistical-work-areas-nhs-111-minimum-data-set-nhs-111-minimum-data-set-2018–19/ (accessed 31 May 2018).
3. NHS England. *Five Year Forward View*. Leeds: NHS England; 2014.
4. NHS England. *Next Steps on the NHS Five Year Forward View*. Leeds: NHS England; 2017.
5. Turner J, O'Cathain A, Knowles E, Nicholl J. Impact of the urgent care telephone service NHS 111 pilot sites: a controlled before and after study. *BMJ Open* 2013;**3**:e003451. https://doi.org/10.1136/bmjopen-2013-003451
6. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ* 2015;**351**:h3480. https://doi.org/10.1136/bmj.h3480
7. NHS England. *NHS111 Online Evaluation*. Leeds: NHS England; 2017.
8. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, *et al.* Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ* 2014;**348**:g1687. https://doi.org/10.1136/bmj.g1687
9. Higgins JPT, Altman DG, Sterne JAC. Chapter 8: Assessing Risk of Bias in Included Studies. In Higgins JPT, Green S, editors. *Cochrane Handbook for Systematic Reviews of Interventions*. Version 5.1.0 (updated March 2011). London: The Cochrane Collaboration; 2011. URL: www.handbook.cochrane.org (accessed 2 May 2019).
10. Reitsma JB, Rutjes AWS, Whiting P, Vlassov VV, Leeflang MMG, Deeks JJ. Chapter 9: Assessing Methodological Quality. In Deeks JJ, Bossuyt PM, Gatsonis C, editors. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0*. London: The Cochrane Collaboration; 2009. URL: https://methods.cochrane.org/sdt/handbook-dta-reviews (accessed 2 May 2019).
11. National Heart, Lung, and Blood Institute. *Study Quality Assessment Tools*. Bethesda, MD: National Heart, Lung, and Blood Institute; 2019 URL: www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools (accessed 1 May 2019).
12. Baxter S, Johnson M, Chambers D, Sutton A, Goyder E, Booth A. The effects of integrated care: a systematic review of UK and international evidence. *BMC Health Serv Res* 2018;**18**:350. https://doi.org/10.1186/s12913-018-3161-3
13. Semigran HL, Levine DM, Nundy S, Mehrotra A. Comparison of physician and computer diagnostic accuracy. *JAMA Intern Med* 2016;**176**:1860–1. https://doi.org/10.1001/jamainternmed.2016.6001
14. Fraser HSF, Clamp S, Wilson CJ. Limitations of study on symptom checkers. *JAMA Intern Med* 2017;**177**:740–1. https://doi.org/10.1001/jamainternmed.2017.0333
15. Mehrotra A, Semigran HL, Levine DM, Nundy S. Limitations of study on symptom checkers – reply. *JAMA Intern Med* 2017;**177**:741. https://doi.org/10.1001/jamainternmed.2017.0351
16. Kellermann AL, Isakov AP, Parker R, Handrigan MT, Foldy S. Web-based self-triage of influenza-like illness during the 2009 H1N1 influenza pandemic. *Ann Emerg Med* 2010;**56**:288–94.e6. https://doi.org/10.1016/j.annemergmed.2010.04.005

17. Little P, Stuart B, Andreou P, McDermott L, Joseph J, Mullee M, *et al.* Primary care randomised controlled trial of a tailored interactive website for the self-management of respiratory infections (Internet Doctor). *BMJ Open* 2016;**6**:e009769. https://doi.org/10.1136/bmjopen-2015-009769

18. Luger TM, Houston TK, Suls J. Older adult experience of online diagnosis: results from a scenario-based think-aloud protocol. *J Med Internet Res* 2014;**16**:e16. https://doi.org/10.2196/jmir.2924

19. Marco-Ruiz L, Bønes E, de la Asunción E, Gabarron E, Aviles-Solis JC, Lee E, *et al.* Combining multivariate statistics and the think-aloud protocol to assess human–computer interaction barriers in symptom checkers. *J Biomed Inform* 2017;**74**:104–22. https://doi.org/10.1016/j.jbi.2017.09.002

20. Nagykaldi Z, Calmbach W, Dealleaume L, Temte J, Mold J, Ryan JG. Facilitating patient self-management through telephony and web technologies in seasonal influenza. *Inform Prim Care* 2010;**18**:9–16. https://doi.org/10.14236/jhi.v18i1.749

21. Nijland N, Cranen K, Boer H, van Gemert-Pijnen JE, Seydel ER. Patient use and compliance with medical advice delivered by a web-based triage system in primary care. *J Telemed Telecare* 2010;**16**:8–11. https://doi.org/10.1258/jtt.2009.001004

22. Poote AE, French DP, Dale J, Powell J. A study of automated self-assessment in a primary care student health centre setting. *J Telemed Telecare* 2014;**20**:123–7. https://doi.org/10.1177/1357633X14529246

23. Anhang Price R, Fagbuyi D, Harris R, Hanfling D, Place F, Taylor TB, Kellermann AL. Feasibility of web-based self-triage by parents of children with influenza-like illness: a cautionary tale. *JAMA Pediatr* 2013;**167**:112–18. https://doi.org/10.1001/jamapediatrics.2013.1573

24. Sole ML, Stuart PL, Deichen M. Web-based triage in a college health setting. *J Am Coll Health* 2006;**54**:289–94. https://doi.org/10.3200/JACH.54.5.289-294

25. Yardley L, Joseph J, Michie S, Weal M, Wills G, Little P. Evaluation of a web-based intervention providing tailored advice for self-management of minor respiratory symptoms: exploratory randomized controlled trial. *J Med Internet Res* 2010;**12**:e66. https://doi.org/10.2196/jmir.1599

26. Babylon Health. *NHS111 Powered By Babylon: Outcomes Evaluation*. London: Babylon Health; 2017.

27. Middleton K, Butt M, Hammerla N, Hamblin S, Mehta K, Parsa A. *Sorting Out Symptoms: Design and Evaluation of the 'Babylon Check' Automated Triage System*. London: Babylon Health; 2016.

28. Berry AC, Berry BB, Nakshabendi R, Kanar O, Avalos DJ, Myers T, *et al.* Evaluation of accuracy between online symptom checkers for diagnosis of gastrointestinal symptoms from MKSAP clinical vignette board review questions. *Gastroenterology* 2016;**150**:S849–50. https://doi.org/10.1016/s0016-5085(16)32869-4

29. Berry AC, Berry NA, Wang B, Mulekar M, Melvin A, Berry BB. Symptom checkers versus doctors: a prospective, head-to-head comparison for GERD vs. non-GERD cough. *Am J Gastroenterol* 2017;**112**(Suppl. 1):S190. https://doi.org/10.14309/00000434-201710001-00353

30. Berry AC, Cash BD, Mulekar MS, Wang B, Melvin A, Berry BB. Symptom checkers vs. doctors, the ultimate test: a prospective study of patients presenting with abdominal pain. *Gastroenterology* 2017;**152**:S852–3. https://doi.org/10.1016/s0016-5085(17)32937-2

31. Carter M, Fletcher E, Sansom A, Warren FC, Campbell JL. Feasibility, acceptability and effectiveness of an online alternative to face-to-face consultation in general practice: a mixed-methods study of webGP in six Devon practices. *BMJ Open* 2018;**8**:e018688. https://doi.org/10.1136/bmjopen-2017-018688

32. Cowie J, Calveley E, Bowers G, Bowers J. Evaluation of a digital consultation and self-care advice tool in primary care: a multi-methods study. *Int J Environ Res Public Health* 2018;**15**:E896. https://doi.org/10.3390/ijerph15050896

33. Nijland N, van Gemert-Pijnen JE, Boer H, Steehouder MF, Seydel ER. Increasing the use of e-consultation in primary care: results of an online survey among non-users of e-consultation. *Int J Med Inform* 2009;**78**:688–703. https://doi.org/10.1016/j.ijmedinf.2009.06.002

34. Madan A. *WebGP: The Virtual General Practice*. London: Hurley Group; 2014.

35. Backman AS, Lagerlund M, Svensson T, Blomqvist P, Adami J. Use of healthcare information and advice among non-urgent patients visiting emergency department or primary care. *Emerg Med J* 2012;**29**:1004–6. https://doi.org/10.1136/emermed-2011-200464

36. Joury AU, Alshathri M, Alkhunaizi M, Jaleesah N, Pines JM. Internet websites for chest pain symptoms demonstrate highly variable content and quality. *Acad Emerg Med* 2016;**23**:1146–52. https://doi.org/10.1111/acem.13039

37. Lanseng EJ, Andreassen TW. Electronic healthcare: a study of people's readiness and attitude toward performing self-diagnosis. *Int J Service Industry Manage* 2007;**18**:394–417. https://doi.org/10.1108/09564230710778155

38. Luger TM, Suls J. Online health information and intentions to seek healthcare. *Psychosom Med* 2011;**73**:A59.

39. North F, Varkey P, Laing B, Cha SS, Tulledge-Scheitel S. Are e-health web users looking for different symptom information than callers to triage centers? *Telemed J E Health* 2011;**17**:19–24. https://doi.org/10.1089/tmj.2010.0120

40. Health on the Net Foundation. *The HON Code of Conduct for Medical and Health Web Sites (HONcode)*. Geneva: Health on the Net Foundation; 2019. URL: www.healthonnet.org/HONcode/Conduct.html (accessed 1 May 2019).

41. Knowles E, O'Cathain A, Turner J, Nicholl J. Effect of a national urgent care telephone triage service on population perceptions of urgent care provision: controlled before and after study. *BMJ Open* 2016;**6**:e011846. https://doi.org/10.1136/bmjopen-2016-011846

42. O'Cathain A, Knowles E, Turner J, Nicholl J. Acceptability of NHS 111 the telephone service for urgent health care: cross sectional postal survey of users' views. *Fam Pract* 2014;**31**:193–200. https://doi.org/10.1093/fampra/cmt078

43. Pope C, Turnbull J, Jones J, Prichard J, Rowsell A, Halford S. Has the NHS 111 urgent care telephone service been a success? Case study and secondary data analysis in England. *BMJ Open* 2017;**7**:e014815. https://doi.org/10.1136/bmjopen-2016-014815

44. Committee on Health Literacy, Institute of Medicine. *Health Literacy: A Prescription to End Confusion*. Washington, DC: National Academies Press; 2004.

45. Institute of Medicine. *Assessing the Severity of Influenza-Like Illnesses: Clinical Algorithms to Inform and Empower Health Care Professionals and the Public.* Boston, MA: Institute for Healthcare Improvement. URL: www.iom.edu/Activities/PublicHealth/MedPrep/2009-SEP-03.aspx (accessed 25 October 2009).

46. Banks J, Farr M, Salisbury C, Bernard E, Northstone K, Edwards H, Horwood J. Use of an electronic consultation system in primary care: a qualitative interview study. *Br J Gen Pract* 2018;**68**:e1–e8. https://doi.org/10.3399/bjgp17X693509

Appendix 1 Data extraction for digital 111 review

Data to code	Options	Comments
Study ID		Author, year, EndNote number
Study design	Experimental	Broad categories for simplicity
	Controlled observational	
	Uncontrolled observational	
	Simulation	
	Qualitative	
	Literature review	
	Other	
Type of system	Online	'Multiple' for papers evaluating a range of different systems
	Digital	
	Multiple	
Data for TIDieST checklist	Yes	Extraction for TIDieST checklist to be completed separately (system is unit of analysis)
	Limited	
	No	
Population/condition	General population	Add details of condition(s)
	Specific condition(s)	
Sample size	Number of participants	Number of HPs and size of data set
Comparator	HP performance on test/simulation	
	HP performance on real-world data	
	Other	
	None	
Outcomes and results	Safety	Brief results for each outcome, including impact on multiple contacts with health services
	Clinical effectiveness	
	Costs/cost-effectiveness	
	Accuracy of diagnosis	
	Accuracy of disposition	
	Service use/diversion	

APPENDIX 1

Data to code	Options	Comments
	Compliance with advice received	
	Patient/carer satisfaction Equity and inclusion	
	User/non-user characteristics	
	Other	
Limitations		
Authors' conclusions	Conclusions	Free-text summary

HP, health professional; ID, identification.

Appendix 2 Draft checklist for reporting key features of digital and online symptom checkers and health assessment/triage services

The TIDieST (Template for Intervention Description for Systems for Triage) Checklist:*

Information to include when describing an intervention for online self-triage systems

Item number	Item	Relationship to original TIDIER Template
	BRIEF NAME	
1.	Provide the proprietary name or generic name that describes the intervention.	AMENDED
	WHY	
2.	Describe the objective of the intervention (*not the study*).	AMENDED
	WHAT	
3.	Interface: Describe the physical characteristics of the interface, including layout, design and any adaptations for accessibility etcetera. Provide information on where the interface can be viewed (e.g. online, screenshots, demo, URL, research article figures).	AMENDED
4.	Procedures: Describe each of the procedures, activities, and/or processes used in the intervention, including any enabling or support activities.	UNCHANGED
	WHO PROVIDED	
5.	For each category of intervention provider (e.g. psychologist, nursing assistant), describe their expertise, background and any specific training given.	DELETED
	HOW	
6.	Describe how the system is accessed e.g. via Web pages, a remote computer, an app etcetera.	AMENDED

	WHERE	
7.	Describe the type(s) of location(s) where the intervention occurred, including any necessary infrastructure or relevant features.	DELETED
	WHEN and HOW MUCH	
8.	Describe the number of times the intervention was delivered and over what period of time including the number of sessions, their schedule, and their duration, intensity or dose.	DELETED
	TAILORING	
9.	Describe provision for particular disease groups or populations and how these differ from general provision.	AMENDED
	MODIFICATIONS/VERSIONS	
10.‡	If the intervention was modified during the course of the study, describe the changes (what, why, when, and how).	AMENDED
	HOW WELL	
11.	Simulation/Laboratory Testing: How the intervention was tested and by whom.	AMENDED
12.‡	Real world testing: How the intervention was tested and by whom.	AMENDED

** **Authors** - use N/A if an item is not applicable for the intervention being described. **Reviewers** – use '?' if information about the element is not reported/not sufficiently reported.

† If the information is not provided in the primary paper, give details of where this information is available. This may include locations such as a published protocol or other published papers (provide citation details) or a website (provide the URL).

‡ If completing the TIDieR checklist for a protocol, these items are not relevant to the protocol and cannot be described until the study is complete.

* We strongly recommend using this checklist in conjunction with the TIDieR guide[7] which contains an explanation and elaboration for each item.

DOI: 10.3310/hsdr07290

* The focus of TIDieST is on reporting details of the intervention elements (and where relevant, comparison elements) of an online triage system. When a **randomised trial** is being reported, the TIDieST checklist should be used in conjunction with the CONSORT statement (see www.consort-statement.org) as an extension of **Item 5 of the CONSORT 2010 Statement.** When a **clinical trial protocol** is being reported, the TIDieST checklist should be used in conjunction with the SPIRIT statement as an extension of **Item 11 of the SPIRIT 2013 Statement** (see www.spirit-statement.org). For alternate study designs, TIDieST can be used in conjunction with the appropriate checklist for that study design (see www.equator-network.org).

Appendix 3 Data extractions using the TIDiesT checklist

The TIDieST (Template for Intervention Description for Systems for Triage) Checklist:*

Information to include when describing an intervention for online self-triage systems

Item number	Item
	BRIEF NAME
1.	**Provide the proprietary name or generic name that describes the intervention.**
	babylon check
	WHY
2.	**Describe the objective of the intervention (*not the study*).**
	To provide an automated service allowing patients to check symptoms and receive fast and clear advice on what action to take.
	WHAT
3.	**Interface: Describe the physical characteristics of the interface, including layout, design and any adaptations for accessibility etcetera. Provide information on where the interface can be viewed (e.g. online, screenshots, demo, URL, research article figures).**
	The babylon check system is described as an 'app with a chat bot-style interface'.[6] At the time of writing (May 2018), little information on the symptom checker was available on the website of the supplier, Babylon Health (www.babylonhealth.com, accessed 11 May 2018). An app that could be downloaded appeared to relate to the company's 'GP at hand' service for GP appointments.
4.	**Procedures: Describe each of the procedures, activities, and/or processes used in the intervention, including any enabling or support activities.**
	The symptom checker involves the user selecting a body part and answering a series of multiple choice questions. The system collects possible outcomes based on the answers given and whether the triggers doe these outcomes are satisfied. Possible outcomes (recommendations) are discarded if particular features (exemptions) are present. This process leads to a list of possible outcomes, of which the highest priority one is presented to the user.[26]
	HOW

6.	**Describe how the system is accessed e.g. via Web pages, a remote computer, an app etcetera.** The system is accessed via a smartphone app.
	TAILORING
9.	**Describe provision for particular disease groups or populations and how these differ from general provision.** Not reported.
	MODIFICATIONS/VERSIONS
10.	**If the intervention was modified during the course of the study, describe the changes (what, why, when, and how).** Not reported.
	HOW WELL
11.	**Simulation/Laboratory Testing: How the intervention was tested and by whom.** The system was tested by the manufacturer in two stages.[26] The initial validation used 33 clinical scenarios validated by external experts to test babylon check. The system performed significantly better than the average performance of automated triage systems reported in the literature, particularly for non-emergency care and self-care. A further test compared babylon check's performance with that of doctors and nurses using 102 patient vignettes.[26]
12.‡	**Real world testing: How the intervention was tested and by whom.** Babylon check was one of four systems evaluated in ongoing NHS England pilot studies. Preliminary results have been reported.[6]

[redacted]

The TIDieST (Template for Intervention Description for Systems for Triage) Checklist:*

Information to include when describing an intervention for online self-triage systems

Item number	Item
	BRIEF NAME
1.	**Provide the proprietary name or generic name that describes the intervention.**
	Internet Doctor
	WHY
2.	**Describe the objective of the intervention (*not the study*).**
	To provide tailored advice on self-management of minor respiratory symptoms.
	WHAT
3.	**Interface: Describe the physical characteristics of the interface, including layout, design and any adaptations for accessibility etcetera. Provide information on where the interface can be viewed (e.g. online, screenshots, demo, URL, research article figures).**
	A screenshot of the interface is provided in the paper by Yardley et al.[24] The home page explains what the site offers and provides links to details of the medical expert on the team and the medical evidence supporting the advice offered. From the home page, participants could choose to access diagnostic pages, treatment pages or common questions. Further details are provided in a multimedia appendix to the paper.
4.	**Procedures: Describe each of the procedures, activities, and/or processes used in the intervention, including any enabling or support activities.**
	The intervention was created by the research team using LifeGuide software. Advice was based on evidence-based resources and the clinical expertise of members of the research team. The content of the information provided was informed by psychological theory, including Bandura's Social Cognitive Theory and Leventhal's model of self-regulation of illness. The diagnostic pages asked a series of questions about the participant's symptoms. These were completed for one symptom at a time and the algorithm provided advice on whether they should contact health services for that symptom. The treatment pages provided information about natural remedies or over-the-counter medication and advice on how to boost the immune system.
	HOW

6.	**Describe how the system is accessed e.g. via Web pages, a remote computer, an app etcetera.** Via web pages (www.internetdr.org.uk).
	TAILORING
9.	**Describe provision for particular disease groups or populations and how these differ from general provision.** Not reported.
	MODIFICATIONS/VERSIONS
10.	**If the intervention was modified during the course of the study, describe the changes (what, why, when, and how).** Not reported.
	HOW WELL
11.	**Simulation/Laboratory Testing: How the intervention was tested and by whom.** Reports of simulation testing were not available.
12.	**Real world testing: How the intervention was tested and by whom.** The intervention was tested in a preliminary RCT primarily involving university students to assess usage and effects on patient enablement and use of health services.[24] A larger RCT in a UK primary care population evaluated effects on health service contacts for those reporting respiratory infections during the study period, as well as hospitalisations and symptom duration and severity.[16]

The TIDieST (Template for Intervention Description for Systems for Triage) Checklist:*

Information to include when describing an intervention for online self-triage systems

Item number	Item
	BRIEF NAME
1.	**Provide the proprietary name or generic name that describes the intervention.**
	Influenza self-triage module (ISTM)
	WHY
2.	**Describe the objective of the intervention (*not the study*).**
	To enhance patient self-management of seasonal influenza and facilitate patient–provider communication.
	WHAT
3.	**Interface: Describe the physical characteristics of the interface, including layout, design and any adaptations for accessibility etcetera. Provide information on where the interface can be viewed (e.g. online, screenshots, demo, URL, research article figures).**
	The system appears to be no longer available. A search of the producer's website (www.okprn.org) revealed no further information.
4.	**Procedures: Describe each of the procedures, activities, and/or processes used in the intervention, including any enabling or support activities.**
	The self-triage module was developed by a practice-based research network (PBRN) multidisciplinary stakeholder group with input from national experts and clinicians in several PBRNs. Several draft versions were developed and piloted. The module was provided to primary care practices as part of an influenza management website which was tailored to the needs of each participating practice.
	HOW
6.	**Describe how the system is accessed e.g. via Web pages, a remote computer, an app etcetera.**
	The system was accessed via the websites of participating practices.
	TAILORING
9.	**Describe provision for particular disease groups or populations and how these differ from general provision.**

English and Spanish language versions were available.

MODIFICATIONS/VERSIONS

10. If the intervention was modified during the course of the study, describe the changes (what, why, when, and how).

Additional questions were added to improve patient safety (e.g. a question about rash to detect possible meningococcal disease). The official influenza-like illness criteria of the Centers for Disease Control and Prevention were expanded to include additional symptoms such as sore throat, nasal congestion, aching muscles and a runny nose. Additional refinements (details not reported) reduced the time for completion of the protocol (via telephone) to about three minutes.

HOW WELL

11. Simulation/Laboratory Testing: How the intervention was tested and by whom.

Not reported.

12.‡ Real world testing: How the intervention was tested and by whom.

The system was tested in 12 primary care practices during the peak of the 2007–2008 influenza season.[19]

The TIDieST (Template for Intervention Description for Systems for Triage) Checklist:*

Information to include when describing an intervention for online self-triage systems

Item number	Item
	BRIEF NAME
1.	**Provide the proprietary name or generic name that describes the intervention.**
	Un-named prototype adapted from a widely used telephone triage system that supports nurses' decision-making in primary care.[21]
	WHY
2.	**Describe the objective of the intervention (*not the study*).**
	To enable patients to undertake a self-assessment triage and receive advice on an appropriate course of action based on their symptoms.
	WHAT
3.	**Interface: Describe the physical characteristics of the interface, including layout, design and any adaptations for accessibility etcetera. Provide information on where the interface can be viewed (e.g. online, screenshots, demo, URL, research article figures).**
	The system had a simple user interface and menu from which patients could select their main presenting symptom from a list of several hundred presenting complaints.
4.	**Procedures: Describe each of the procedures, activities, and/or processes used in the intervention, including any enabling or support activities.**
	Based on the main complaint, the system generated age- and gender-specific questions with associated potential answers. Each answer carried a weighting which contributed to the final triage outcome. Some answers were linked to further question sets, allowing multiple symptoms to be evaluated. The system had question sets covering the full range of primary care presentations. The triage advice provided by the system consisted of one of six courses of action: call 999; seek GP care immediately; seek care within six hours; seek care within 24 hours; seek a routine appointment; and self-care. The system also created a self-assessment record which summarised the history of the presenting condition.
	HOW

6.	**Describe how the system is accessed e.g. via Web pages, a remote computer, an app etcetera.** In the study evaluating the system, access was via a desktop computer.[21]
	TAILORING
9.	**Describe provision for particular disease groups or populations and how these differ from general provision.** Not reported.
	MODIFICATIONS/VERSIONS
10.	**If the intervention was modified during the course of the study, describe the changes (what, why, when, and how).** Not reported.
	HOW WELL
11.	**Simulation/Laboratory Testing: How the intervention was tested and by whom.** Not reported.
12.‡	**Real world testing: How the intervention was tested and by whom.** The system was tested in a university student health centre by Poote et al.[21] Students used the system before a face-to-face consultation with a GP. The system rating of urgency of the student's condition was compared with that of the GP (who had access to the output from the automated system).

The TIDieST (Template for Intervention Description for Systems for Triage) Checklist:*

Information to include when describing an intervention for online self-triage systems

Item number	Item
	BRIEF NAME
1.	**Provide the proprietary name or generic name that describes the intervention.**
	Strategy for Off-Site Rapid Triage (SORT)[15]
	WHY
2.	**Describe the objective of the intervention (*not the study*).**
	To create a simple but accurate tool that could help minimally trained health care workers screen large numbers of patients with influenza-like illness.
	WHAT
3.	**Interface: Describe the physical characteristics of the interface, including layout, design and any adaptations for accessibility etcetera. Provide information on where the interface can be viewed (e.g. online, screenshots, demo, URL, research article figures).**
	See online Figures E1-E3. SORT versions 1.0-3.0. No screenshots. 2 interactive Web sites, http://www.Flu.gov and http://www.H1N1ResponseCenter.com
4.	**Procedures: Describe each of the procedures, activities, and/or processes used in the intervention, including any enabling or support activities.**
	The group then developed an efficient, 3-step process to assess patients with influenza-like illness. In the first step, patients are screened to determine whether they meet CDC criteria for influenza-like illness. Those who do proceed to the second step, an assessment of illness severity (using questions adopted from the CRB-65 score). Patients with influenza-like illness who have a CRB-65 score of 0 (suggesting relatively mild illness) move on to the third step, a short series of questions designed to determine whether they have a health condition that increases their risk of developing severe complications of influenza.

According to the patient's answers, SORT assigns a level of risk and recommends a specific action. Patients with "**high-risk" influenza-like illness**—in the group's first iteration of the algorithm**, those with a CRB-65 score of 3 or more—would be sent directly to an ED.** "Intermediate-risk" patients—**CRB score of 1 or 2 or comorbid conditions that increase their risk of complications—would be advised to contact their physician or seek care in a walk-in clinic** because early administration of antiviral medication might reduce the chance of complications. **"Low-risk" patients—those with mild disease (CRB-65=0) and no comorbid conditions—would be advised to convalesce at home.**

Support activities involved drafting health literacy friendly instructions and involvement of professional associations.

HOW

6. **Describe how the system is accessed e.g. via Web pages, a remote computer, an app etcetera.**

On October 2, 2009, the CDC adopted a slightly modified version of SORT 3.0 and posted it on the agency's Web site at http://cdc.gov/h1n1flu/clinicians/pdf/adultalgorithm.pdf. In an accompanying disclaimer, the CDC stated that the algorithm was intended for use "by physicians and those working under their supervision." It was also limited to patients older than 18 years.

Five days later, US Department of Health and Human Services (HHS) secretary Katherine Sibelius announced the **posting of an H1N1 self-evaluation application** at http://www.Flu.gov. It closely adheres to the CDC's adult algorithm and used many of the terms and phrases we devised for our demonstration Web site. It is intended for use by adults older than 18 years.

The same day (October 7, 2009), **Microsoft Corporation unveiled its own flu self-assessment application** at http://www.H1N1ResponseCenter.com. Like Flu.gov's application, Microsoft's site closely adheres to the CDC's adult algorithm and uses health-literate language licensed, at no charge, from Emory University. Both HHS and Microsoft encouraged health departments, nongovernmental organizations, private health plans, employers, and other organizations to link to their Web sites free. Many chose to do so.

TAILORING

9. Describe provision for particular disease groups or populations and how these differ from general provision.

Although SORT is designed to assess patients with influenza-like illness, the 3-step approach it uses (screening, severity assessment, associated risk factors) may be used to evaluate many illnesses. SORT-like algorithms for selected public health threats such as severe acute respiratory syndrome could be even be prepared and evaluated in advance and deployed if needed. This method could help reassure a nervous public, particularly in the early phases of an outbreak when many people otherwise rush to the nearest ED.

With additional refinement, Web-based decision-support tools such as SORT may be used to collect important epidemiologic information about disease incidence and severity in non-hospitalized individuals. Information of this type is vital to quickly characterize a new disease's attack rate and virulence.

SORT was subsequently modified for use by caregivers of children with ILI as described by Anhang Price et al.[22]

MODIFICATIONS/VERSIONS

10.‡ If the intervention was modified during the course of the study, describe the changes (what, why, when, and how).

SORT was initially envisioned for use by minimally trained health care workers at off-site flu assessment stations and walk in clinics. But the development group quickly realized that a slightly modified version— one that substitutes symptoms for measured respiratory rate and blood pressure— could be used by call centers or even self-administered through an interactive Web site. Ultimately, both versions were included in the group's work product.

HOW WELL

11. Simulation/Laboratory Testing: How the intervention was tested and by whom.

The Kaiser Permanente Colorado Institute for Health Research performed a retrospective assessment, using their health system's computerized records, to determine how well SORT 3.0 would have performed had it been used to screen patients with influenza-like illness. Between April 1 and June 30, 2009, 2,758 outpatients with influenza like illness visited the Kaiser Permanente Colorado health system. SORT 3.0 categorized 1,540 of these encounters (56%) as low risk. During the next 2 weeks, 7 low-risk patients were hospitalized, but only 2 had problems that were related to the index visit (negative predictive value

99.9%). Intermediate-risk patients were much more likely to be admitted within 2 weeks than low-risk patients (odds ratio 11.9; 95% confidence interval 5.29 to 26.9) (D. Magid, personal communication, August 23, 2009).

Buoyed by these findings, they developed a demonstration Web site with branching logic to depict how patients could use SORT to self-assess their need for care. To ensure that the site was comprehensible to laypeople, they asked experts in health literacy at our institution to translate SORT's clinical terms into plain language.[43] More than 100 lay volunteers of widely varying age, race, and socioeconomic status reviewed draft text and offered suggestions on how to make the content understandable and actionable. Some had an influenza-like illness when they participated; others had recently recovered from the flu.

On September 3 to 4, 2009, they presented draft adult and pediatric SORT algorithms and demonstration Web site at a hastily convened Institute of Medicine workshop titled "Assessing the Severity of Influenza-Like Illnesses: Clinical Algorithms to Inform and Empower Health Care Professionals and the Public."[44] The event, which was sponsored by UnitedHealth Group, attracted national leaders from academia, major clinical societies, public health, law, government, and private industry. Feedback was highly favorable.

As soon as the pediatric algorithm was posted, they began drafting health-literate content to offer the guidance directly to the public through the Web. Unfortunately, **the American Academy of Pediatrics opposed this effort because the algorithm was not prospectively validated. Concerns were also expressed that an interactive pediatric Web site might discourage some parents from contacting their child's medical provider.**

Notwithstanding this disappointment, the overall effort to create, test, and deploy SORT was highly collaborative from beginning to end. Numerous organizations and individuals gave freely of their time and expertise. Recognizing the urgency of the effort, Emory's Office of Technology Transfer readily licensed the technology, at no charge, to any vendor who agreed to provide it free.

12.‡ Real world testing: How the intervention was tested and by whom.

Between October 7, 2009, and February 24, 2010, Flu.gov recorded 721,906 total page views, 320,333 visits to Flu.gov/evaluation (the opening page of the self-evaluation site), and 230,761 completed evaluations to flu.gov/evaluation/index2.html (A. Roszak, personal communication). **To reassure the public that the federal government would respect each user's privacy, HHS did not retain data on site visitors.** As a consequence, we have no additional information. Between October 5 and December 13, 2009, Microsoft's Web site, http://www.H1N1ResponseCenter.com , was visited 1.6 million times. Of the 442,000 visitors (28%) who completed a self-assessment, slightly less than half (N=202,000) chose to share anonymous data with the site. Preliminary analysis indicates that 37% of these visitors provided answers that categorized them as high risk and 13% were **too young to receive guidance**. The other half either did not meet influenza like illness criteria or were assessed as not requiring ED treatment. Microsoft did not identify visitors who used the site multiple times, so **it is possible that some individuals repeatedly entered positive replies.** The Web sites were used approximately 650,000 times. We have no way to determine how many times the CDC's adult and pediatric algorithms were **used by clinicians and call centers**. No adverse events were reported. Microsoft's data suggests that their Web site may have prevented as many as 100,000 ED visits, although the true total is probably less. Because HHS did not record data on visitors to Flu.gov, **we cannot estimate the effect of their self-assessment tool.**

Further Literature

Price RA, Fagbuyi D, Harris R, Hanfling D, Place F, Taylor TB, Kellermann AL. Feasibility of web-based self-triage by parents of children with influenza-like illness: a cautionary tale. JAMA pediatrics. 2013 Feb 1;167(2):112-8. [SORT for Kids]

Abdullah N, Annamalai M, RANI A, AMRY MK. Paging Vs. Scrolling: Navigation Styles For Self-Triage Of Epidemic Diseases. *Journal of Theoretical & Applied Information Technology*. 2016 Jun 20;88(2).

D'Angelo MC, Humphreys KR, Li T, Young ME. The Impact of Medical Terminology in Self-Triage Decision-Making. *Frontiers in Communication.* 2017 Jul 26;2:6.

The TIDieST (Template for Intervention Description for Systems for Triage) Checklist:*

Information to include when describing an intervention for online self-triage systems

Item number	Item
	BRIEF NAME
1.	**Provide the proprietary name or generic name that describes the intervention.**
	WebGP, subsequently renamed eConsult
	WHY
2.	**Describe the objective of the intervention (*not the study*).**
	To provide an electronic GP consultation and self-help service for primary care patients.
	WHAT
3.	**Interface: Describe the physical characteristics of the interface, including layout, design and any adaptations for accessibility etcetera. Provide information on where the interface can be viewed (e.g. online, screenshots, demo, URL, research article figures).**
	The home page (as illustrated in the service developer's pilot report)[33] includes links to self-help guides and videos and photographs of practice staff (intended to boost patient confidence). The symptom checker provides lists of common symptoms in alphabetical order and has a facility to choose from 100 common conditions. Full details of the service are available at https://econsult.net/ (accessed 21 May 2018). Adaptations for accessibility not reported.
4.	**Procedures: Describe each of the procedures, activities, and/or processes used in the intervention, including any enabling or support activities.**
	WebGP consists of five services: symptom checker; self-help guidance; signposting to other services; information about the 111 telephone service; and e-consult, allowing the patient to complete an online form which is e-mailed to the practice. GPs use the information provided to arrange a prescription, arrange a face-to-face appointment (via practice admin team) or undertake a phone consultation.
	Details of how the system is integrated into practice procedures are reported to vary between practices.
	HOW

6.	**Describe how the system is accessed e.g. via Web pages, a remote computer, an app etcetera.**
	The system is accessed through practice websites.
	TAILORING
9.	**Describe provision for particular disease groups or populations and how these differ from general provision.**
	Not reported.
	MODIFICATIONS/VERSIONS
10.	**If the intervention was modified during the course of the study, describe the changes (what, why, when, and how).**
	Not reported.
	HOW WELL
11.	**Simulation/Laboratory Testing: How the intervention was tested and by whom.**
	Not reported.
12.‡	**Real world testing: How the intervention was tested and by whom.**
	A 6-month pilot report was produced by the Hurley Group, which was involved in developing the system.[33] Subsequent evaluations have been reported in the UK, including in six practices in Devon;[30] and 11 practices across Scotland.[31] A further evaluation was excluded from the main review because of lack of information about the symptom checker aspect of the intervention.[45]

The TIDieST (Template for Intervention Description for Systems for Triage) Checklist:*

Information to include when describing an intervention for online self-triage systems

Item number	Item
	BRIEF NAME
1.	**Provide the proprietary name or generic name that describes the intervention.**
	24/7 WebMed
	WHY
2.	**Describe the objective of the intervention (*not the study*).**
	To enhance services offered by a Student Health Service (SHS) by providing a decision tool to help students decide whether to seek care.
	WHAT
3.	**Interface: Describe the physical characteristics of the interface, including layout, design and any adaptations for accessibility etcetera. Provide information on where the interface can be viewed (e.g. online, screenshots, demo, URL, research article figures).**
	The system appears to be no longer available. A Google search for 24/7 WebMed produced no results and the system does not appear on the supplier's current website (www.dshisystems.com).
4.	**Procedures: Describe each of the procedures, activities, and/or processes used in the intervention, including any enabling or support activities.**
	The system collected basic demographic data from users, including zip code, age and gender. Users the answered a series of questions based on algorithms. The system could analyse over 600 chief complaints, stratified by age and gender. The system classified assessments into six different levels of urgency: emergency, call 911; seek care immediately; seek care within 12–24 hours; seek care within 2–3 days; seek care within 1–2 weeks; and self-care recommended. After completing triage, students could request an appointment with SHS by e-mail.
	HOW
6.	**Describe how the system is accessed e.g. via Web pages, a remote computer, an app etcetera.**
	The system was accessed via a link from the SHS website.

TAILORING

9. Describe provision for particular disease groups or populations and how these differ from general provision.

Not reported.

MODIFICATIONS/VERSIONS

10. If the intervention was modified during the course of the study, describe the changes (what, why, when, and how).

Not reported.

HOW WELL

11. Simulation/Laboratory Testing: How the intervention was tested and by whom.

Reports of simulation testing were not available.

12.‡ Real world testing: How the intervention was tested and by whom.

Testing of the system in the setting of SHS at the University of Central Florida was described by Sole et al.[23]

Appendix 4 Quality assessment tools

Quality assessment tool for observational/cross-sectional studies

1. Was the research question clearly stated?
2. Was the study population clearly specified and defined?
3. Was the participation rate of eligible persons at least 50%?
4. Were all the subjects selected or recruited from the same or similar populations (including the same time period)? Were inclusion and exclusion criteria prespecified and applied uniformly to all participants?
5. Was a sample size justification, power description, or variance and effect estimate provided?
6. For exposures that can vary in amount or level, did the study examine different levels of the exposure as related to the outcome?
7. Were the exposure measures (independent variables) clearly defined, valid, reliable and implemented consistently across all study participants?
8. Were the outcome measures (dependent variables) clearly defined, valid, reliable and implemented consistently across all study participants?
9. Were the outcome assessors blinded to the exposure status of participants?
10. Were key potential confounding variables measured and adjusted statistically for their impact on the relationship between exposure(s) and outcome(s)?

Quality assessment tool for diagnostic studies

1. Was the spectrum of patients representative of the patients who will receive the test in practice? (Representative spectrum.)
2. Is the reference standard likely to classify the target condition correctly? (Acceptable reference standard.)
3. Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests? (Acceptable delay between tests.)
4. Did the whole sample or a random selection of the sample receive verification using the intended reference standard? (Partial verification avoided.)
5. Did patients receive the same reference standard irrespective of the index test result? (Differential verification avoided.)
6. Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)? (Incorporation avoided.)
7. Were the reference standard results interpreted without knowledge of the results of the index test? (Index test results blinded.)
8. Were the index test results interpreted without knowledge of the results of the reference standard? (Reference standard results blinded.)
9. Were the same clinical data available when test results were interpreted as would be available when the test is used in practice? (Relevant clinical information.)
10. Were uninterpretable/intermediate test results reported? (Uninterpretable results reported.)
11. Were withdrawals from the study explained? (Withdrawals explained.)
12. Comments.

Appendix 5 Risk-of-bias tables

Risk-of-bias results for randomised trials

Author, date and reference	Selection and performance bias	Detection and attrition bias	Reporting and other bias
Little, 2016[17]	• Random sequence generation: low risk • Allocation concealment: low risk • Blinding of participants and personnel: unclear	• Blinding of outcome assessment: low risk blinded assessment of primary care records • Incomplete outcome data: low risk	• Selective reporting: unclear • Anything else, ideally prespecified: low risk
Yardley, 2010[25]	• Random sequence generation: low risk • Allocation concealment: low risk • Blinding of participants and personnel: low risk	• Blinding of outcome assessment: unclear • Incomplete outcome data: low risk	• Selective reporting: unclear • Anything else, ideally prespecified: low risk

Risk-of-bias results for cohort/cross-sectional studies

Reference	Questions 1–4	Questions 5–7	Questions 8–10
Backman *et al.*[35]	1. **Was the research question clearly stated?** Yes. The aims refer to 'non-urgent' care but the information is sought prior to visiting an ED 2. **Was the study population clearly specified and defined?** Yes 3. **Was the participation rate at least 50%?** Yes, 79% 4. **Were all the subjects selected or recruited from the same or similar populations?** Yes, primary care and ED attendees	5. **Was a sample size justification provided?** No 6. **Did the study examine exposure levels?** Yes. Health advice seeking 7. **Were exposure measures clearly defined?** Unclear. Measures are vague (e.g. 'previous use' of information) and discriminating between types of information	8. **Were outcome measures clearly defined?** Unclear. 'Health care information use in the past'[35] 9. **Were outcome assessors blinded?** Not applicable 10. **Were confounders adjusted for?** Yes. To some extent participant and physician attributes were assessed for influence on the results
Carter *et al.*[31]	1. **Was the research question clearly stated?** Yes 2. **Was the study population clearly specified and defined?** Yes. GPs, practice staff and their patients at six practices in Devon 3. **Was the participation rate at least 50%?** No. The postal survey had a response rate of only 35.1%, but GP judgement of webGP requests was collected and five GPs and five administrators were interviewed	5. **Was a sample size justification provided?** No 6. **Did the study examine exposure levels?** Not applicable 7. **Were exposure measures clearly defined?** Not applicable	8. **Were outcome measures clearly defined?** Yes. Attitudes and experiences of practice staff and patients using webGP 9. **Were outcome assessors blinded?** Not applicable 10. **Were confounders adjusted for?** Not applicable

Reference	Questions 1–4	Questions 5–7	Questions 8–10
	4. **Were all the subjects selected or recruited from the same or similar populations?** Yes. GPs, practice staff and their patients at six practices in Devon		
Cowie *et al.*[32]	1. **Was the research question clearly stated?** Yes 2. **Was the study population clearly specified and defined?** Yes 3. **Was the participation rate at least 50%?** No. Not for patient surveys 4. **Were all the subjects selected or recruited from the same or similar populations?** Yes	5. **Was a sample size justification provided?** No 6. **Did the study examine exposure levels?** No 7. **Were exposure measures clearly defined?** Not applicable	8. **Were outcome measures clearly defined?** Yes 9. **Were outcome assessors blinded?** No 10. **Were confounders adjusted for?** Yes
Joury *et al.*[36]	1. **Was the research question clearly stated?** Yes 2. **Was the study population clearly specified and defined?** Not applicable 3. **Was the participation rate at least 50%?** Not applicable 4. **Were all the subjects selected or recruited from the same or similar populations?** Not applicable	5. **Was a sample size justification provided?** No 6. **Did the study examine exposure levels?** Not applicable 7. **Were exposure measures clearly defined?** Not applicable	8. **Were outcome measures clearly defined?** Yes. Scores used for readability, popularity, content and quality 9. **Were outcome assessors blinded?** Not applicable 10. **Were confounders adjusted for?** Unclear
Kellermann *et al.*[16]	1. **Was the research question clearly stated?** Unclear 2. **Was the study population clearly specified and defined?** Unclear. Patients with influenza-like illness in the USA who accessed one of these websites, either www.flu.gov or www.H1N2ResponseCenter.com 3. **Was the participation rate at least 50%?** Not applicable 4. **Were all the subjects selected or recruited from the same or similar populations?** Unclear. Counted only web hits; no demographic data available on patients. No data on the use of the algorithm by clinicians or call centres	5. **Was a sample size justification provided?** Not applicable 6. **Did the study examine exposure levels?** Not applicable 7. **Were exposure measures clearly defined?** Not applicable	8. **Were outcome measures clearly defined?** Not applicable 9. **Were outcome assessors blinded?** Not applicable 10. **Were confounders adjusted for?** Not applicable

Reference	Questions 1–4	Questions 5–7	Questions 8–10
Lanseng and Andreassen[37]	1. **Was the research question clearly stated?** Yes 2. **Was the study population clearly specified and defined?** Yes 3. **Was the participation rate at least 50%?** Unclear 4. **Were all the subjects selected or recruited from the same or similar populations?** Yes	5. **Was a sample size justification provided?** No 6. **Did the study examine exposure levels?** No 7. **Were exposure measures clearly defined?** Not applicable	8. **Were outcome measures clearly defined?** Yes, use of TRI 9. **Were outcome assessors blinded?** No 10. **Were confounders adjusted for?** Unclear
Luger *et al.*[18]	1. **Was the research question clearly stated?** Yes 2. **Was the study population clearly specified and defined?** Yes 3. **Was the participation rate at least 50%?** Unclear 4. **Were all the subjects selected or recruited from the same or similar populations?** Yes	5. **Was a sample size justification provided?** No 6. **Did the study examine exposure levels?** No 7. **Were exposure measures clearly defined?** Not applicable	8. **Were outcome measures clearly defined?** Yes 9. **Were outcome assessors blinded?** Not applicable 10. **Were confounders adjusted for?** Unclear
Marco-Ruiz *et al.*[19]	1. **Was the research question clearly stated?** Yes 2. **Was the study population clearly specified and defined?** No 3. **Was the participation rate at least 50%?** Yes, 53% 4. **Were all the subjects selected or recruited from the same or similar populations?** Unclear	5. **Was a sample size justification provided?** No 6. **Did the study examine exposure levels?** No 7. **Were exposure measures clearly defined?** Not applicable	8. **Were outcome measures clearly defined?** Not applicable 9. **Were outcome assessors blinded?** Not applicable 10. **Were confounders adjusted for?** Unclear
Nagykaldi *et al.*[20]	1. **Was the research question clearly stated?** Yes 2. **Was the study population clearly specified and defined?** Yes. Study population was patients from 12 primary care practices in the USA 3. **Was the participation rate at least 50%?** Not applicable 4. **Were all the subjects selected or recruited from the same or similar populations?** Yes. All participants were patients from 12 primary care practices who accessed a customised practice website or telephone helpline	5. **Was a sample size justification provided?** Not applicable 6. **Did the study examine exposure levels?** Not applicable 7. **Were exposure measures clearly defined?** Not applicable	8. **Were outcome measures clearly defined?** Yes. Web hits on the customised practice website's influenza self-management web pages. Downloads of self-management influenza toolkit. Completion of influenza self-triage module sessions. Volume of calls to telephone hotlines. Qualitative feedback from patients on their satisfaction with and the utility of self-management websites and the telephone hotline. Qualitative feedback from clinicians around their involvement and their perceptions of patient self-management techniques

Reference	Questions 1–4	Questions 5–7	Questions 8–10
			9. **Were outcome assessors blinded?** Not applicable 10. **Were confounders adjusted for?** Not applicable
Nijland *et al.*[33]	1. **Was the research question clearly stated?** Yes 2. **Was the study population clearly specified and defined?** Yes 3. **Was the participation rate at least 50%?** Unclear 4. **Were all the subjects selected or recruited from the same or similar populations?** Yes	5. **Was a sample size justification provided?** No 6. **Did the study examine exposure levels?** Not applicable 7. **Were exposure measures clearly defined?** Not applicable	8. **Were outcome measures clearly defined?** Yes 9. **Were outcome assessors blinded?** No 10. **Were confounders adjusted for?** Yes. Methods not very clearly reported but appears to be multiple regression
Nijland *et al.*[21]	1. **Was the research question clearly stated?** Yes 2. **Was the study population clearly specified and defined?** Yes 3. **Was the participation rate at least 50%?** No. Low participation rate in the survey relative to the users of the triage system (although unclear how many were invited to participate) 4. **Were all the subjects selected or recruited from the same or similar populations?** Yes	5. **Was a sample size justification provided?** No 6. **Did the study examine exposure levels?** Not applicable 7. **Were exposure measures clearly defined?** Not applicable	8. **Were outcome measures clearly defined?** Yes 9. **Were outcome assessors blinded?** No 10. **Were confounders adjusted for?** Unclear
North *et al.*[39]	1. **Was the research question clearly stated?** Yes 2. **Was the study population clearly specified and defined?** Yes 3. **Was the participation rate at least 50%?** Not applicable 4. **Were all the subjects selected or recruited from the same or similar populations?** Not applicable	5. **Was a sample size justification provided?** Not applicable 6. **Did the study examine exposure levels?** Yes. Self-exposure 7. **Were exposure measures clearly defined?** Not applicable	8. **Were outcome measures clearly defined?** Yes 9. **Were outcome assessors blinded?** Not applicable 10. **Were confounders adjusted for?** Unclear. Some discussion of potential confounders
Sole *et al.*[24]	1. **Was the research question clearly stated?** Yes. The primary purpose of this study was to identify and describe the demographic profile of students who used the newly implemented web-based triage system. The secondary purpose was to compare	5. **Was a sample size justification provided?** No 6. **Did the study examine exposure levels?** Yes 7. **Were exposure measures clearly defined?** Yes	8. **Were outcome measures clearly defined?** Not applicable 9. **Were outcome assessors blinded?** Not applicable 10. **Were confounders adjusted for?** Not applicable

Reference	Questions 1–4	Questions 5–7	Questions 8–10
	web-based triage diagnoses to the diagnoses made in clinic for a subset of students who requested appointments[24] 2. **Was the study population clearly specified and defined?** Yes. Students who used the web-based triage over a 4-month implementation period (1290 students). Then of those students, those who requested an appointment via e-mail (143 students), and then of those 59 students who attended the health centre after requesting an e-mail appointment 3. **Was the participation rate at least 50%?** Not applicable 4. **Were all the subjects selected or recruited from the same or similar populations?** Yes		

TRI, Technology Readiness Index.

Risk-of-bias results for diagnostic studies

Reference	Questions 1–4	Questions 5–8	Questions 9–11
Poote *et al.*[22]	1. **Representative spectrum?** No. Study participants were all patients registered at a student health centre in England and attending with new acute symptoms. If the self-assessment triage system was only for students, to be representative the study population would have needed to include a range of student health centres in different areas. If the self-assessment triage system was for general UK practices, the study population would have needed to include patients of all ages, ethnicities, genders, etc., and from a range of general practices in different areas 2. **Acceptable reference standard?** Yes 3. **Acceptable delay between tests?** Yes	5. **Differential verification avoided?** Not applicable 6. **Was the reference standard independent of the index test?** Unclear. Patients took the assessment from self-triage through to their GP consultation 7. **Index test results blinded?** No. Patients took the assessment from self-triage through to their GP consultation 8. **Reference standard results blinded?** Yes	9. **Relevant clinical information?** Yes 10. **Were uninterpretable results reported?** Not applicable 11. **Were withdrawals from the study explained?** Yes

Reference	Questions 1–4	Questions 5–8	Questions 9–11
	4. **Partial verification avoided?** Yes. All patients who completed self-triage also had a GP consultation during which the GP rated the urgency of their consultation		
Anhang Price *et al.*[23]	1. **Representative spectrum?** No. SORT was only trialled in two EDs in the USA, a larger range of sites would be needed for a representative spectrum. Most patients were from the ED, not home based, so potentially the patients in the sample were sicker than the general population 2. **Acceptable reference standard?** Yes. The sensitivity of the SORT for Kids algorithm in identifying the need for ED care was based on an explicit gold standard: documented evidence that the child received one or more of five ED-specific interventions 3. **Acceptable delay between tests?** Yes 4. **Partial verification avoided?** Yes	5. **Differential verification avoided?** Not applicable? 6. **Was the reference standard independent of the index test?** Yes 7. **Index test results blinded?** Yes 8. **Reference standard results blinded?** Yes	9. **Relevant clinical information?** Yes 10. **Were uninterpretable results reported?** Not applicable 11. **Were withdrawals from the study explained?** No
Semigran *et al.*[6]	1. **Representative spectrum?** Unclear. There were 45 standardised patient vignettes that were divided into three levels of triage urgency and included more and less common conditions. It is not clear how closely this replicates the spectrum of conditions that people use symptom checkers for 2. **Acceptable reference standard?** Yes. The source for each vignette also provided the associated correct diagnosis 3. **Acceptable delay between tests?** Not applicable 4. **Partial verification avoided?** Not applicable	5. **Differential verification avoided?** Not applicable? 6. **Was the reference standard independent of the index test?** Yes 7. **Index test results blinded?** Yes 8. **Reference standard results blinded?** Yes	9. **Relevant clinical information?** Yes. This is the clinical information that would be supplied by the patient, which may or may not differ from the information given by the vignette. Discussion of the true clinical accuracy of symptom checkers.[38] Some standardised patient vignettes contained specific clinical language (e.g. mouth ulcers, tonsils with exudate), and actual patients with the same condition might struggle with the words to use to describe their symptoms or use different terms. Therefore, the authors' analysis represents an indirect assessment of how well symptom checkers would perform with actual patients 10. **Were uninterpretable results reported?** Yes. Inter-rater reliability in capturing symptom checker recommendations

Reference	Questions 1–4	Questions 5–8	Questions 9–11
			for diagnosis and triage was high (Cohen's κ 0.90). In some cases, the authors could not evaluate a vignette because some symptom checkers focused only on children or on adults, or the symptom checker did not list or ask for the key symptom in the vignette. To avoid penalising these symptom checkers, the authors referred to standardised patient vignettes that successfully yielded an output as 'standardised patient evaluations' **11. Were withdrawals from the study explained?** Not applicable
Semigran *et al.*[13]	**1. Representative spectrum?** Unclear. There were 45 standardised patient vignettes that were divided into three levels of triage urgency and included more and less common conditions. It is not clear how closely this replicates the spectrum of conditions that people use symptom checkers for **2. Acceptable reference standard?** Yes **3. Acceptable delay between tests?** Not applicable **4. Partial verification avoided?** No. There was a total of 234 physicians involved in the study. Of the 45 vignettes, each was solved by at least 20 physicians, but it is not clear why they chose the specific vignettes to solve	**5. Differential verification avoided?** Not applicable **6. Was the reference standard independent of the index test?** Not applicable **7. Index test results blinded?** Yes **8. Reference standard results blinded?** Yes	**9. Relevant clinical information?** Yes. The physicians and the symptom checkers used the same vignettes **10. Were uninterpretable results reported?** Not applicable **11. Were withdrawals from the study explained?** No. It is unclear why the physicians chose to solve the specific vignettes

EME
HS&DR
HTA
PGfAR
PHR

Part of the NIHR Journals Library
www.journalslibrary.nihr.ac.uk

Published by the NIHR Journals Library