BMJ Health & Care Informatics

Women's attitudes to the use of AI image readers: a case study from a national breast screening programme

Niamh Lennox-Chhugani [1], Yan Chen,[2] Veronica Pearson,[3] Bernadette Trzcinski,[4] Jonathan James[5]

To cite: Lennox-Chhugani N, Chen Y, Pearson V, *et al*. Women's attitudes to the use of AI image readers: a case study from a national breast screening programme. *BMJ Health Care Inform* 2021;**28**:e100293. doi:10.1136/bmjhci-2020-100293

Received 03 December 2020
Revised 01 March 2021
Accepted 08 March 2021

[1]Research & Implementation, TaoHealth Ltd, London, UK
[2]School of Medicine, University of Nottingham, Nottingham, UK
[3]East Midlands Imaging Network, Nottingham University Hospitals NHS Trust, Nottingham, UK
[4]Breast Screening Service, United Lincolnshire Hospitals NHS Trust, Lincoln, UK
[5]Nottingham Breast Institute, Nottingham University Hospitals NHS Trust, Nottingham, UK

Correspondence to
Dr Niamh Lennox-Chhugani; nlennoxchhugani@taohealth.co.uk

ABSTRACT

Background Researchers and developers are evaluating the use of mammogram readers that use artificial intelligence (AI) in clinical settings.

Objectives This study examines the attitudes of women, both current and future users of breast screening, towards the use of AI in mammogram reading.

Methods We used a cross-sectional, mixed methods study design with data from the survey responses and focus groups. We researched in four National Health Service hospitals in England. There we approached female workers over the age of 18 years and their immediate friends and family. We collected 4096 responses.

Results Through descriptive statistical analysis, we learnt that women of screening age (≥50 years) were less likely than women under screening age to use technology apps for healthcare advice (likelihood ratio=0.85, 95% CI 0.82 to 0.89, $p<0.001$). They were also less likely than women under screening age to agree that AI can have a positive effect on society (likelihood ratio=0.89, 95% CI 0.84 to 0.95, $p<0.001$). However, they were more likely to feel positive about AI used to read mammograms (likelihood ratio=1.09, 95% CI 1.02 to 1.17, $p=0.009$).

Discussion and Conclusions Women of screening age are ready to accept the use of AI in breast screening but are less likely to use other AI-based health applications. A large number of women are undecided, or had mixed views, about the use of AI generally and they remain to be convinced that it can be trusted.

INTRODUCTION

Population breast screening in England aims to detect breast cancer earlier, thus improving outcomes for women between the ages of 50 and 70 years. The National Health Service (NHS) Breast Screening Programme (NHSBSP) invites more than 2 million women for a test every year nationally. In the light of the high volume of images to be read, artificial intelligence (AI) is focusing on the development of image reading technology.[1–3] As studies confirm the diagnostic accuracy of AI products in breast cancer diagnosis, there is an emerging concern among clinicians that AI image reading may not be sufficiently focused on patients. 'Clinically meaningful endpoints such as survival, symptoms and need for treatment' could mitigate the risks of overtreatment and false positives.[4]

In a healthcare context, where shared decision-making is increasing,[5] patients are seeking a greater understanding of how a diagnosis is arrived at. Regulators of AI technology are starting to acknowledge the importance of being seen as trustworthy on uptake and adoption.[6]

Public attitudes to the use of AI and machine learning in healthcare are evolving. Social attitudes to the use of AI to support diagnosis are positive but people still want human involvement.[7–11] Specifically in radiology, people want to be fully informed about the use of AI and want to retain human interaction in the diagnostic process.[12 13] However, they hold positive views about the use of such technology to support clinician diagnosis and deliver faster, more precise and unbiased results.

The public are not passive recipients of care. They are essential stakeholders in the healthcare system. Their willingness to adopt new innovations can enable or constrain spread and scale.[14] There is a need to understand how acceptable AI is in breast cancer screening services as well as the many ethical, social and legal implications of its use.[15] A few qualitative studies, although with small sample sizes, have explored public perception of the use of AI in medicine.[16–18] A recent survey conducted in the Netherlands involving 922 participants examined the perception of the use of AI to read mammograms. It found that the women surveyed did not support the use of AI without a human reader.[19] If the benefits of AI are to be delivered in breast screening and the disbenefits minimised, then the public should be actively engaged in the design, development and monitoring of this technology.[20 21]

Our study seeks to address the gap in the research into public attitudes towards AI. We did this as part of a wider real-world testing of AI tools in the NHSBSP in England. The researchers developed a short survey which collected both quantitative and qualitative data. The researchers followed up with focus group discussions to understand the attitudes of a sample of women to the use of AI in breast screening. The NHSBSP currently invites women between the ages of 50 and 70 years for screening every 3 years. Mammograms are double read by two human readers.

This paper focuses on women's attitudes to the possible future use of an AI second reader in the NHSBSP.

MATERIALS AND METHODS

This study was a prospective mixed method design. The study was conducted in four NHS trusts providing acute care in the East Midlands of England. All participants gave electronic informed consent to participate in the survey and focus groups.

Survey tool development and testing

We developed an open e-survey according to good practice guidelines,[22] including the Checklist for Reporting Results of Internet E-Surveys.[23] This is used for the development, administration and reporting of web-based surveys. Our research question set out in the study protocol was how do the attitudes of women to the use of AI in the breast screening process affect the adoption and spread of these innovations? We conducted a review of the literature on the influence of adopter attitudes to AI in general and innovation adoption in healthcare specifically. Based on this review, we developed a set of open and closed questions. These were tested with a small sample group of women (n=10) for question clarity, underlying assumptions (bias), question sensitivity, problems with Likert scale labels, question order and online user experience.

The final version of the survey had six sections:

1. Personal attributes, which included age.
2. Experience of breast cancer (direct or indirect).
3. Knowledge and experience of breast screening.
4. Use of AI-based technology in everyday life.
5. Attitudes towards AI-based technology in general.
6. Attitudes towards the use of AI in breast screening (figure 1).

The survey tool was submitted for ethical approval along with the study protocol.

Data collection

The chosen sampling strategy was non-probability sampling. This was chosen because the topic being explored was under-researched and the study was exploratory rather than testing a hypothesis. The sample size for the survey was calculated based on a 1% response rate from the ≥18 years female population of the East Midlands of England, a confidence level of 95% and a margin of error of 2% (n=2435). This was submitted to the Health Research Authority as part of the ethical approval process. The survey was set up on a dedicated General Data Protection Regulation-compliant online survey platform and information was shared via a range of site communication channels with women over the age of 18 years working or volunteering at four acute hospital sites in the East Midlands and their friends and relatives. As one of the largest and most diverse employers in the region, the NHS workforce provided a good proxy for the wider population. Respondents were recruited between 4 December 2019 and 29 February 2020.

Information was gathered on age, ethnicity and employment status. This enabled us to identify any representation gaps in the sample cohort and guided targeted recruitment for the survey and focus groups. Focus group participants were recruited from the general population with a greater representation of women from black and minority ethnic groups since these were slightly underrepresented in the survey. Due to COVID-19 restrictions, focus groups were conducted using a secure online video conferencing platform.

Data analysis

The survey responses were analysed using descriptive statistics to understand the current status of women's views on AI-based technology generally and in the breast screening programme specifically. Likelihood ratios were used to determine the significance of differences between women under screening age and of screening age.

NVivo (NVivo is a qualitative and mixed methods data analysis software tool used by academics and professional

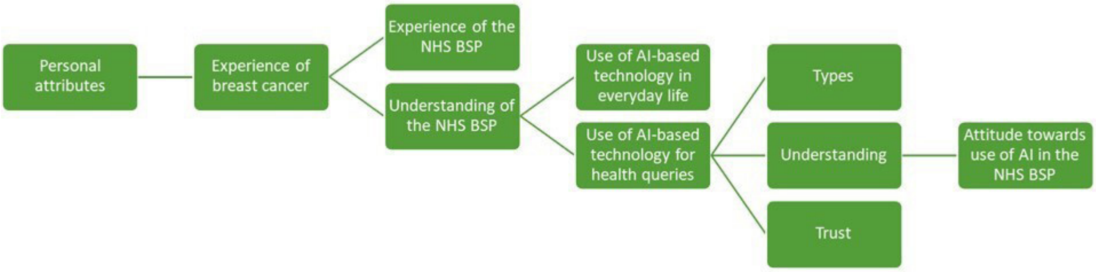
Personal attributes
Experience of breast cancer
Experience of the NHS BSP
Understanding of the NHS BSP
Use of AI-based technology in everyday life
Use of AI-based technology for health queries
Types
Understanding
Trust
Attitude towards use of AI in the NHS BSP

Figure 1 Survey map: the topic covered in the survey in the order the questions were presented. NHSBSP, NHS Breast Screening Programme.

Table 1 Age bands of the survey participants

Respondents age profile		
Age band (years)	**No**	**Proportion**
18–19	21	0.51%
20–29	606	14.79%
30–39	776	18.95%
40–49	946	23.10%
50–59	1293	31.57%
60–69	372	9.08%
70+	82	2.00%
Grand total	4096	100%

researchers globally) software (QSR International, UK) was used to organise and visualise qualitative data from surveys (open-ended questions with free-text responses) and focus group transcripts. A hierarchical thematic framework was used to classify and organise data according to key themes, concepts and emergent categories. This approach allows us to explore data in depth while simultaneously maintaining an effective and transparent audit trail. This enhances the rigour of the analytical processes and the credibility of the findings.

RESULTS

Sample

The survey was distributed to a population of 23 332 men and women working at four NHS trusts in the East Midlands. Of the consenting participants (n=4132), 4096 were identified as women. The respondents (n=4096) covered all the age bands targeted, with the largest group from the 50–59 years age band. Most women who took part were in paid employment (92.8%, 3802/4096) with the remainder retired, self-employed, carer of dependents or volunteers. The ethnicity profile of the respondents was like that of the profile for the East Midlands except for Asian/Asian British which was under-represented (2.88% in the survey responses as opposed to 6.5% in the East Midlands population). This guided the purposive sampling strategy for the focus groups where 20% of women recruited were Asian/Asian British.

The 4096 women were segmented into two groups: 1747 (42.7%) were or had recently been of screening age and 2349 (57.3%) were under screening age (<50 years) and, thus, future users of the programme (table 1).

Differences in self-reported technology use

Women of screening age were less likely to use technology platforms or applications for healthcare advice, 64.9% (1134/1747), than women under screening age, 76.2% (1790/2349)–likelihood ratio=0.85, 95% CI 0.82 to 0.89, p<0.001. Women of screening age were also less likely to trust the recommendations of these platforms, 57% (997/1747), than women under screening age, 61% (1449/2349)–likelihood ratio=0.93, 95% CI 0.88 to 0.97, p=0.003 (figure 2). These differences replicate the results of similar studies of attitudes to technology across whole populations.[24 25]

Differences in attitudes towards the effect of AI on society

Women of screening age were less likely to agree that AI can have a positive effect on society, 47.1% (822/1747), than women under screening age, 52.9% (1242/2349)—likelihood ratio=0.89, 95% CI 0.84 to 0.95, p<0.001. Women of screening age were also more likely to be undecided on the issue, 47.7% (834/1747), than women under screening age, 41.3% (969/2349)—likelihood ratio=1.16, 95% CI 1.06 to 1.27, p=0.001, 95% CI 1.08 to 1.24, p<0.001 (figure 3). The likelihood of disagreeing

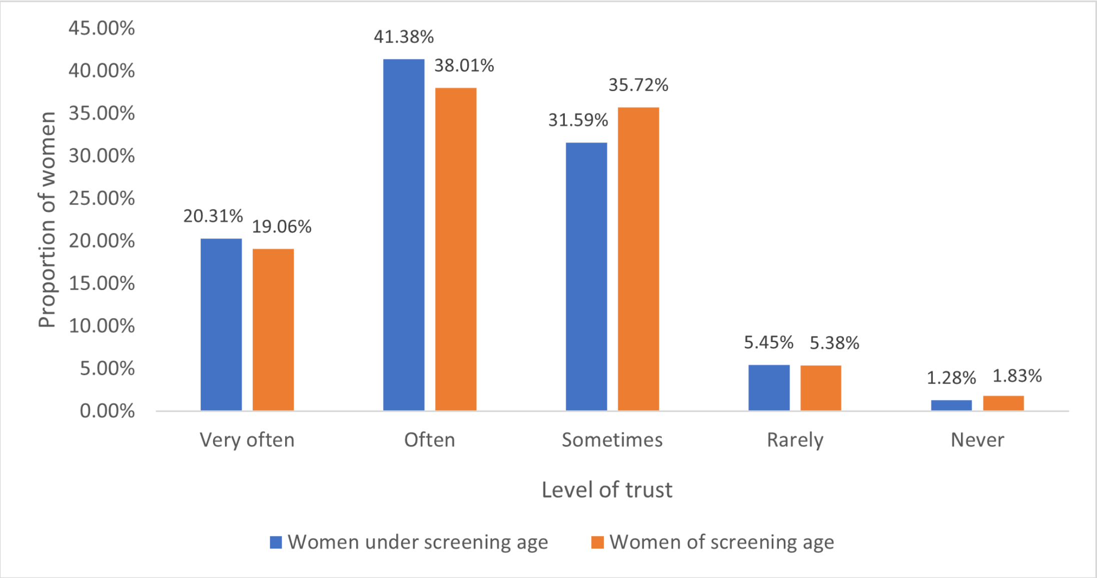

Figure 2 The self-reported level of trust that women under and of screening age had in everyday artificial intelligence-powered applications when seeking health advice.

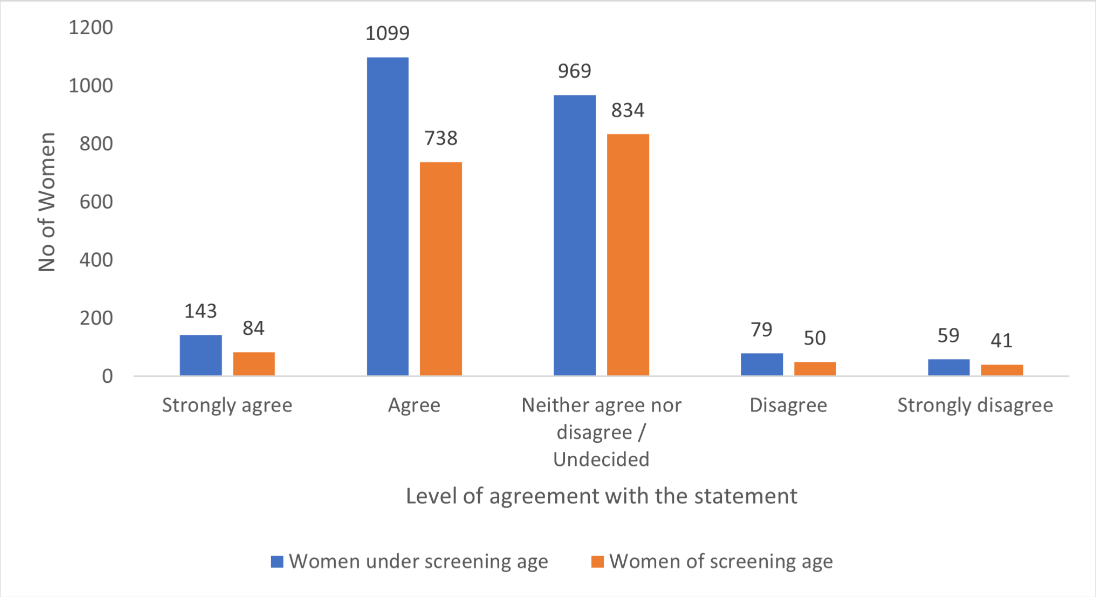

Figure 3 The self-reported level of agreement with the statement ‘artificial intelligence can have a positive effect on society’ for women under and of screening age.

that AI can have a positive effect on society was similar among women of screening age, 5.2% (91/1747), and women under screening age, 5.9% (138/2349)—likelihood ratio=0.89, 95% CI 0.69 to 1.15, p=0.359.

Sentiment analysis of free-text responses on the issue of whether AI can have a positive effect on society found that many women, who had a negative or mixed view of the effect of AI in society, were unsure of why they felt this way (n=96). However, they described AI as an inevitable part of their lives in the future (n=20). Those who did express a view cited:

1. Concern about the reliability and safety of technology (n=123).
2. A lack of trust in the technology itself or the systems that sit around it (n=65).
3. A fear about a combination of over-reliance on AI and job losses that might ensue (n=32).
4. Concern about the absence of the human touch in interactions (n=46).

Differences in attitudes towards the use of AI in breast screening

Women's baseline understanding of the current process of reading mammograms was weak. Only 22% of women under screening age and 27% of women of screening age identified that two human readers blind read all screening mammograms in the NHSBSP. Sentiment analysis of free-text responses (n=3987) showed that the largest proportion of women overall were positive about using AI in breast screening, 47.2% (1880/3987). The next largest group expressed mixed or undecided views, 35.9% (1432/3987) and 17.9% (675/3987) expressed a negative view. A further 109 women did not provide a free-text response, 2.7% of the total 4096 survey respondents (figure 4). Women of screening age were more likely to feel positive about using AI to read mammograms, 49.5% (849/1714), than women under screening age, 45.4% (1031/2273)—likelihood ratio=1.09, 95% CI 1.02 to 1.17, p=0.009. This finding was confirmed by the finding that women of screening age were less likely to have mixed or neutral feelings on the issue, 34.1% (584/1714), than women under screening age, 37.3% (848/2273)—likelihood ratio=0.91, 95% CI 0.84 to 0.99, p=0.036. Women of screening age, 16.0% (281/1714), and women under screening age, 17.3% (394/2273), were similarly likely to have negative views on the use of AI in breast screening—likelihood ratio=0.95, 95% CI 0.82 to 1.09, p=0.434.

Thematic analysis of the free-text data focusing on the perceived benefits of using AI in the breast screening programme showed that women were most likely to say that they were not sure what these would be (n=543). When they did express a view, the most frequently mentioned perceived benefits were:

1. Increased efficiency (n=162).
2. Improved reliability (n=263).
3. Greater safety (n=139).

A significant number of women expressed the view that AI in breast screening would and should happen (n=847) in the future.

Overall, women of screening age are less likely to use AI for health advice in their everyday life or have a positive view of its effect on society but are more likely to have a positive view on the use of AI in breast screening (table 2).

Detailed understanding of attitudes towards to use of AI in breast screening

A total of 25 women took part in six focus groups conducted during July 2020. Overall, 19/25 had either experienced a breast cancer diagnosis themselves or knew someone who had and 18/25 had attended a breast cancer screening appointment. Overall, 15/25 of the women who took part knew that two readers looked at

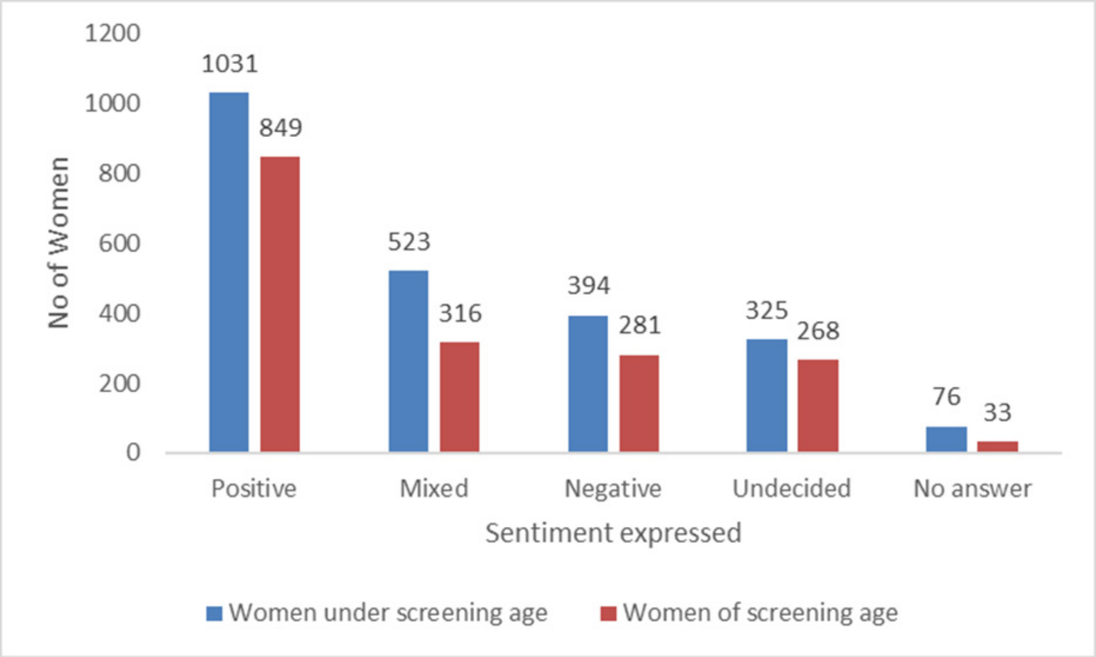

Figure 4 The sentiment expressed in free text by women under and of screening age when asked how they felt about artificial intelligence being used to read mammograms in breast screening.

mammograms. Therefore, they were a more informed group than the general population surveyed.

Many of the women who took part expressed the view that the use of AI in healthcare and specifically in the breast screening programme was inevitable. Some saw a positive contribution being made by AI generally. They identified the following key benefits from using AI in breast screening:

1. Increased efficiency.
2. Improved reliability.
3. Improved outcomes and improved safety/fewer errors.

They also hypothesised that introducing AI into the breast screening programme might:

1. Release staff for higher value patient-centred activities.
2. Save money for the service.
3. Help to address the workforce shortage within the breast screening programme.

The main concerns that were expressed by the women were:

1. The absence of the 'human touch' in the diagnostic process.
2. A lack of clarity on how the AI tools will be governed.
3. Potential discriminatory bias.
4. A lack of clarity on how data privacy will be protected.

When asked what kind of actions they thought would mitigate some of their concerns, the women suggested that breast screening process would always need to involve humans. For some women this meant human oversight of the AI technology which undertakes most of the activity including decision-making. For others, the human role is pre-eminent, with AI used only to augment clinical activity and decision-making. The women assumed that this technology would never be used without clear evidence of its effectiveness. They expected the impact on equity of access to breast screening to be closely monitored through governance processes.

Women were divided on whether or not they would want to be informed if AI tools were being used as part of the breast screening process. However, they agreed overall that women should be given information about the role of AI in breast screening as part of the process of informed consent when taking part in the breast screening programme.

DISCUSSION

As the use of AI in the field of radiology accelerates rapidly,[26–29] attention has focused on the performance and safety of the algorithms being used. Real-world deployment of these tools is imminent and a greater understanding of radiologist and radiographer attitudes to the technology in different countries across the globe is needed.[30–38]

This large-scale study, aimed at understanding the attitudes of healthy users to the use of this technology in diagnosis, has shown that women of screening age are open to the use of AI in breast screening. However, they are less likely than women under screening age to use other AI-based health applications. These differences replicate the results of similar studies of attitudes to technology across whole populations.[24 25] There are large proportions of women in both groups who are undecided or hold mixed views about the use of AI. They cite a lack of understanding and trust in the technology and a desire to know more. This bears out the findings of recent smaller scale studies.[16–18] Women of all ages see human interaction in diagnosis as critical to their experience of high-quality care.

Table 2 Survey results summary

				Likelihood ratio			
Topic	**Metric**	**Women of screening age**	**Women under screening age**	**Women of screening age/ women under screening age**	**Lower bound**	**Upper bound**	**P value**
Do you use healthcare apps if you feel unwell?	Likelihood of using technology platforms or applications for healthcare advice	64.9% (1134/1747)	76.2% (1790/2349)	0.85	0.82	0.89	>0.001*
	Likelihood of trusting the recommendations of these platforms	57.1% (997/1747)	61.7% (1449/2349)	0.93	0.88	0.97	.003*
Artificial intelligence (AI) can have a positive effect on society.	Likelihood of agreeing that AI can have a positive effect on society	47.1% (822/1747)	52.9% (1242/2349)	0.89	0.84	0.95	>0.001*
	Likelihood of being undecided on whether AI can have a positive effect on society	47.7% (834/1747)	41.3% (969/2349)	1.16	1.08	1.24	>0.001*
	Likelihood of disagreeing that AI can have a positive effect on society	5.2% (91/1747)	5.9% (138/2349)	0.89	0.69	1.15	0.359
How would you feel about AI being used to read mammograms?	Likelihood of feeling positive about the use of AI in reading mammograms	49.5% (849/1714)	45.4% (1031/2273)	1.09	1.02	1.17	.009*
	Likelihood of mixed/ neutral feelings about the use of AI in reading mammograms	34.1% (584/1714)	37.3% (848/2273)	0.91	0.84	0.99	.036*
	Likelihood of negative feelings about the use of AI in reading mammograms	16.4% (281/1714)	17.3% (394/2273)	0.95	0.82	1.09	0.434

*statistically significant at $\alpha = 0.05$.

Women of screening age have an immediate interest in screening that is as accurate, quick and reliable as possible. Previous studies[7 11] found that those who are identified as 'patients' are more likely to perceive positive effects of new technology than those who are identified as 'healthy users'.

In this case, women of screening age share 'patient' attributes as they are currently part of the NHSBSP. The openness of women of screening age to the use of AI in breast screening is moderated by:

1. A desire to understand more about the technology.[39]
2. The evidence to support its performance.[40]
3. Its use to augment and not replace clinical interaction and decision-making.[13]

These moderators are evident in the literature on the adoption of digital health technology generally. Clinical adoption of novel digital technology, including AI, relies on robust evidence of accuracy through high-quality clinical trials.[41] There is little evidence yet of a similar direct relationship for public adoption of AI in health. This goes some way to explain the large number of respondents who were equivocal or undecided in their attitudes towards the use of AI in breast screening.

Mass media stories and the views of the clinical professionals they are interacting with are more influential than direct exposure to evidence of accuracy.[42 43] Several women responding to the survey highlighted the positive media representation of the *Nature* article on the performance of AI in breast image reading.[1] This influenced their perception of AI in breast screening positively. Women's views on the importance of retaining human interaction in the diagnostic process confirm the findings of previous studies.[9 12]

The response rate to the survey was substantially greater than targeted in the study protocol (4096/2435). However, women of Asian ethnicity were under-represented (3% in survey, 6% in East Midlands' population). To address this, this group was successfully targeted for inclusion in the focus groups by design. Women in paid employment were also over-represented because NHS employees were used as a proxy for the general population. Some of the potential selection biases introduced by the non-probability sampling method were addressed by the mixed methods design of the wider study and purposive sampling for the focus groups. The authors recommend future survey administration should use probability sampling. The survey itself is not a psychometrically tested tool. This limits the generalisability of the findings, although adherence to accepted standards for research survey development have minimised this limitation.

Women invited to population breast screening are important stakeholders in the service and how it is delivered.[44] This study demonstrates that women of screening age are open to the use of AI in breast cancer screening. However, there are large proportions of women who are undecided or have mixed views about the use of AI and remain to be convinced that it can be trusted. Understanding their attitudes will be an important factor in the acceptance and adoption of the AI-based technology. Regulators of health technology are starting to understand this.[45] Attitudes change over time in response to multiple intrinsic and extrinsic factors. Education and dissemination of information about the use of AI in the clinical pathway will need to be considered.

Twitter Niamh Lennox-Chhugani @taohealth1

Acknowledgements The authors would like to thank the East Midlands Imaging Network Project team for their support in accessing participants and documents to conduct this research. The Patient and Public Involvement Group for the AI in Breast Screening Programme was involved in the design of data collection tools and research communications. Representatives of the public were part of the overall project governance including this research. Particular thanks to Christine Foster, one of the authors' patient representatives, for her support in editing this article and enhancing its readability.

Contributors NL-C led the development of the research study, collected the survey and conducted the focus groups, analysed the data and wrote the first draft of the paper. JJ and BT made a substantial contribution to all stages of the project. YC contributed to statistical analysis of the survey data. All authors commented on the draft paper.

Funding This research was funded by the Office of Life Sciences through Innovate UK project grant (104655) as part of the NHS Test Beds Programme (Wave 2).

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval This study received ethical approval from the UK Health Research Authority on 5 July 2019 (Project ID 262287).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information. Datasets are available upon reasonable request.

ORCID iD

Niamh Lennox-Chhugani http://orcid.org/0000-0002-1297-0237

REFERENCES

1 McKinney SM, Sieniek M, Godbole V, *et al*. International evaluation of an AI system for breast cancer screening. *Nature* 2020;577:89–94.
2 Kim H-E, Kim HH, Han B-K, *et al*. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit Health* 2020;2:e138–48.
3 Salim M, Wåhlin E, Dembrower K, *et al*. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. *JAMA Oncol* 2020;6:1581–8.
4 Oren O, Gersh BJ, Bhatt DL. Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. *Lancet Digit Health* 2020;2:e486–8.
5 Beach MC, Sugarman J. Realizing shared decision-making in practice. *JAMA* 2019;322:811–2.
6 European Commission. *White paper on artificial intelligence – a European approach to excellence and trust*. Brussels: European Commission, 2020.
7 Mori I. *Public views of machine learning*. London: The Royal Society, 2017.
8 Mori I. Future data-driven technologies and the implications for use of patient data. London Academy of Medical Sciences; 2018.
9 Tran V-T, Riveros C, Ravaud P. Patients' views of wearable devices and AI in healthcare: findings from the ComPaRe e-cohort. *NPJ Digit Med* 2019;2:53.
10 Ada Lovelace Institute. No green lights, no red lines: public perspectives on COVID-19 technologies. London Ada Lovelace Institute; 2020.
11 Safi S, Danzer G, Schmailzl KJ. Empirical research on acceptance of digital technologies in medicine among patients and healthy users: questionnaire study. *JMIR Hum Factors* 2019;6:e13472.
12 Ongena YP, Haan M, Yakar D, *et al*. Patients' views on the implementation of artificial intelligence in radiology: development and validation of a standardized questionnaire. *Eur Radiol* 2020;30:1033–40.
13 Jutzi TB, Krieghoff-Henning EI, Holland-Letz T, *et al*. Artificial intelligence in skin cancer diagnostics: the patients' perspective. *Front Med* 2020;7:233.
14 Lennon MR, Bouamrane M-M, Devlin AM, *et al*. Readiness for delivering digital health at scale: lessons from a longitudinal qualitative evaluation of a national digital health innovation program in the United Kingdom. *J Med Internet Res* 2017;19:e42.
15 Fuchsjäger M. Is the future of breast imaging with AI? *Eur Radiol* 2019;29:4822–4.
16 McCradden MD, Sarker T, Paprica PA. Conditionally positive: a qualitative study of public perceptions about using health data for artificial intelligence research. *BMJ Open* 2020;10:e039798.
17 Nelson CA, Pérez-Chada LM, Creadore A, *et al*. Patient perspectives on the use of artificial intelligence for skin cancer screening: a qualitative study. *JAMA Dermatol* 2020;156:501–12.
18 Kovarik CL. Patient perspectives on the use of artificial intelligence. *JAMA Dermatol* 2020;156:493–4.
19 Ongena YP, Yakar D, Haan M, *et al*. Artificial intelligence in screening mammography: a population survey of women's preferences. *J Am Coll Radiol* 2021;18:79–86.
20 Kirsch A. Explain to whom? putting the user in the center of Explainable AI. HAL Id: hal-01845135. *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 co-located with 16th International Conference of the Italian Association for Artificial Intelligence*; Nov 14-17, Bari, Italy, 2017.
21 Katell M, Young M, Dailey D. Toward situated interventions for algorithmic equity: lessons from the field. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (ACM FAT* '20). Association for Computing Machinery*; Jan 27-30, Barcelona, Spain, 2020:45–55.
22 Kelley K, Clark B, Brown V, *et al*. Good practice in the conduct and reporting of survey research. *Int J Qual Health Care* 2003;15:261–6.
23 Eysenbach G. Improving the quality of web surveys: the checklist for reporting results of Internet E-Surveys (cherries). *J Med Internet Res* 2004;6:e34.
24 Gnambs T. Attitudes towards emergent autonomous robots in Austria and Germany. *Elektrotech Inftech* 2019;136:296–300.
25 Lee CC, Czaja SJ, Moxley JH, *et al*. Attitudes toward computers across adulthood from 1994 to 2013. *Gerontologist* 2019;59:22–33.
26 Allen B, Dreyer K, McGinty GB. Integrating artificial intelligence into radiologic practice: a look to the future. *J Am Coll Radiol* 2020;17:280–3.
27 Nagendran M, Chen Y, Lovejoy CA, *et al*. Artificial intelligence versus clinicians: systematic review of design, reporting Standards, and claims of deep learning studies. *BMJ* 2020;368:m689.
28 Liu X, Faes L, Kale AU, *et al*. A comparison of deep learning performance against health-care professionals in detecting diseases

from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 2019;1:e271–97.

29 Shen J, Zhang CJP, Jiang B, *et al*. Artificial intelligence versus clinicians in disease diagnosis: systematic review. *JMIR Med Inform* 2019;7:e10010.

30 Waymel Q, Badr S, Demondion X, *et al*. Impact of the rise of artificial intelligence in radiology: what do radiologists think? *Diagn Interv Imaging* 2019;100:327–36.

31 Laï M-C, Brian M, Mamzer M-F. Perceptions of artificial intelligence in healthcare: findings from a qualitative survey study among actors in France. *J Transl Med* 2020;18:14.

32 Gong B, Nugent JP, Guest W, *et al*. Influence of artificial intelligence on Canadian medical students' preference for radiology specialty: ANational survey study. *Acad Radiol* 2019;26:566–77.

33 Pinto Dos Santos D, Giese D, Brodehl S, *et al*. Medical students' attitude towards artificial intelligence: a multicentre survey. *Eur Radiol* 2019;29:1640–6.

34 Park CJ, Yi PH, Siegel EL. Medical student perspectives on the impact of artificial intelligence on the practice of medicine. *Curr Probl Diagn Radiol* 2020. ;;S0363-0188:30124–9.

35 van Hoek J, Huber A, Leichtle A, *et al*. A survey on the future of radiology among radiologists, medical students and surgeons: students and surgeons tend to be more skeptical about artificial intelligence and radiologists may fear that other disciplines take over. *Eur J Radiol* 2019;121:108742.

36 Abdullah R, Fakieh B. Health care employees' perceptions of the use of artificial intelligence applications: survey study. *J Med Internet Res* 2020;22:e17620.

37 Oh S, Kim JH, Choi S-W, *et al*. Physician confidence in artificial intelligence: an online mobile survey. *J Med Internet Res* 2019;21:e12422.

38 Sit C, Srinivasan R, Amlani A, *et al*. Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: a multicentre survey. *Insights Imaging* 2020;11:14.

39 Safi S, Thiessen T, Schmailzl KJ. Acceptance and resistance of new digital technologies in medicine: qualitative study. *JMIR Res Protoc* 2018;7:e11072.

40 Hennemann S, Beutel ME, Zwerenz R. Drivers and barriers to acceptance of web-based aftercare of patients in inpatient routine care: a cross-sectional survey. *J Med Internet Res* 2016;18:e337.

41 Kelly CJ, Karthikesalingam A, Suleyman M, *et al*. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019;17:195.

42 van Bekkum JE, Hilton S. Primary care nurses' experiences of how the mass media influence frontline healthcare in the UK. *BMC Fam Pract* 2013;14:178.

43 Boutron I, Haneef R, Yavchitz A, *et al*. Three randomized controlled trials evaluating the impact of "spin" in health news stories reporting studies of pharmacologic treatments on patients'/caregivers' interpretation of treatment benefit. *BMC Med* 2019;17:105.

44 Richards M. Report of the independent review of adult screening programmes in England, 2019. Available: www.england.nhs.uk/wp-content/uploads/2019/02/report-of-the-independent-review-of-adult-screening-programme-in-england.pdf

45 Care Quality Commission. Using machine learning in diagnostic services. A report with recommendations from CQC’s regulatory sandbox. London CQC; 2020.

BMJ Health & Care Informatics

Survey protocol for exploring video and phone use in Aotearoa New Zealand general practice: considerations for future telehealth

Karen Day,[1] Inga Hunter,[2] Vasudha Rao,[2] Greig Russell,[2] Rachel Roskvist,[1] Fiona Moir,[1] Emily Gill,[1] Bert van der Werf[1]

To cite: Day K, Hunter I, Rao V, *et al*. Survey protocol for exploring video and phone use in Aotearoa New Zealand general practice: considerations for future telehealth. *BMJ Health Care Inform* 2021;**28**:e100309. doi:10.1136/bmjhci-2020-100309

Received 15 December 2020
Revised 01 February 2021
Accepted 22 February 2021

[1]School of Population Health, The University of Auckland, Auckland, New Zealand
[2]School of Management, Massey University, Palmerston North, New Zealand

Correspondence to
Dr Karen Day;
k.day@auckland.ac.nz

ABSTRACT

Introduction Telehealth became the most practical option for general practice consultations in Aotearoa New Zealand (NZ) as a result of the national lockdowns in response to the COVID-19 pandemic. What is the consumer experience of access to telehealth and how do consumers and providers perceive this mode of care delivery going forward?

Methods and analysis A national survey of general practice consumers and providers who used telehealth services since the national lockdowns in 2020 will be distributed. It is based on the Unified Theory of Acceptance and Use of Technology framework of technology acceptance and the access to care framework. The data will be statistically analysed to create a foundation for in-depth research on the use of telehealth services in NZ general practice services, with a specific focus on consumer experiences and health outcomes.

Ethics and dissemination Ethics approval was granted by the Auckland Health Research Ethics Committee on 13/11/2020, reference AH2539. The survey will be disseminated online.

BACKGROUND

Primary healthcare is provided from the community by a broad range of health professionals and aims to reduce the causes, development and severity of diseases by providing treatment and education including the promotion of self-care.[1 2] Within Aotearoa New Zealand (NZ), most primary healthcare services and funding models are provided by clinicians working within general practice, such as doctors, nurses and nurse practitioners. The NZ health system is, in the majority, tax funded, but most consumers are required to make copayments for services rendered within general practice.

Telehealth (care at a distance using information and communications technologies)[1] became the most practical option for general practice consultations in NZ during COVID-19 alert levels 3 and 4[2] after 23 March 2020. Telehealth was not a frequently or routinely used consultation modality in primary care, and specifically in general practice, up to this point in NZ.

Traditional telehealth research highlights that, in comparison with in-person consultations, telehealth has lower costs for both consumer and provider and that there is no difference in service utilisation or disease progression for people with long term conditions,[3] along with the convenience of phone or video consultations. However, the introduction of video for telehealth (as opposed to the phone) has been accompanied by disruption in processes, and concerns in clinical quality and accountability, and patient privacy.[3 4]

The priority in NZ, at the onset of the pandemic, was to limit exposure to and possible spread of COVID-19 while accessing and/or providing care. Within days, general practices set up telehealth processes (and associated software), and patients were triaged into video and/or phone appointments or in-person appointments, where physical examinations of patients were required and could be done safely.[3]

Continuity of care, as a process measure of access to care, remained a priority during this time. Continuity of care is the longitudinal therapeutic relationship between a clinician and patient,[5] which is essential for patient-centred[6] and person-focused care.[7] Consumer experience research describes how patients prefer continuity of care with the same provider, are unaware when a telehealth option is available and tend to revert to 'how we've always done things' when under pressure.[8] Person-focused care recognises the longitudinal relationship between clinician and patient that incorporates multiple interactions about a combination of long-term and short-term health issues over time.[7] This approach, in turn, assumes the inclusion of different modes of interaction, such as in-person clinic visits, video and phone discussions and consultations, email/secure message correspondence and patient portal interactions.

Penchansky and Thomas[9] describe access to care in terms of dimensions of accessibility, availability, affordability, service design, acceptability, implementation and design. Saurman[10] adds awareness (knowing that a service is available) as the final dimension. Telehealth is one way to enable access to care but could potentially also become a barrier in terms of equity,[11] where one assumes the availability of technologies and skills to be able to participate in, for example, a video consultation. The International Covenant on Economic, Social and Cultural Rights treaty[12] outlines the right to equitable healthcare. NZ has an obligation under Te Tiriti o Waitangi (the Treaty of Waitangi) to ensure that improved health is equitably accessed for both Māori and non-Māori.[11 13] Changes resulting from the introduction of telehealth must consider whether the new processes will result in improved health outcomes and be accessible equitably. With these considerations in mind, upscaling telehealth from sudden unplanned emergency use to business as usual requires understanding of the consumer experience in the postemergency COVID-19 period.

The intention of consumers and providers of general practice services to use telehealth technology after having experienced it is important to understand for future adoption of video and/or phone during consultations. The Unified Theory of Acceptance and Use of Technology (UTAUT)[14] asserts that if there is perceived ease of use, perceived usefulness and positive social norm (peer support for adoption), one can predict user acceptance. To make sense of the intent to use theory (UTAUT), we will contextualise the findings in the theory of access to care by Penchansky and Thomas.[9]

AIM AND OBJECTIVES

In the context of NZ general practice and the COVID-19 pandemic, our research question is, 'What is the consumer experience of access to telehealth and how do consumers perceive this mode of care delivery going forward?'. This also includes the providers' experience because telehealth in this context consists of real-time (synchronous) interactions between consumer and provider, that is, consultations via video and/or phone (mobile and/or landline).

Research aim

To explore the use of video and phone consultations in general practice since 23 March 2020 to describe: (1) factors (negative and positive) about consumers' access to telehealth; and (2) perceptions of consumers and providers regarding future use of telehealth in NZ general practice.

Research objectives

To achieve the research aim we will:

- Conduct a scoping literature review.[15]
- Design two questionnaires (consumer and provider) to gather data on the use of video and phone consultations to describe the consumer's perspective (unknown at this stage) and the providers' perspectives (to provide context to the consumer's experiences).
- Use the UTAUT to measure the acceptability of technology (video and phone) and attitudes to future use of telehealth.
- Contextualise the UTAUT findings using the access to care theory[9] and equity as lenses.
- Publish the findings as an exploratory descriptive study to establish a base for future research.

METHODS

Our study is a prospective observational study. In the absence of a standardised checklist for our study, we have adopted the Strengthening the Reporting of Observational Studies in Epidemiology statement[16] to guide the study design. The statement covers the reporting of cohort, cross-sectional and case–control studies.

Since most NZ primary healthcare services and associated funding models operate within general practice, our sampling strategy consists of a national survey to describe the use of video and/or phone for NZ general practice consultations since the lockdown periods are associated with the COVID-19 pandemic. Consumers and general practice providers (doctors and nurses) will be invited to participate. The NZ lockdown period moved from alert level 3 (23 March 2020) to level 4 (25 March), returning to level 3 (27 April) and progressing to level 2 (13 May) and level 1 (8 June). A second short term and regionally targeted lockdown occurred in August moving the country to level 2 nationally with Auckland at level 3 temporarily. Just before the lockdowns started, general practitioners (GPs) were directed to conduct at least 70% of their consultations via video or phone or a combination of both. We have created two questionnaires (one for consumers and one for general practice providers, ie, GPs, nurse practitioners and registered nurses).

Data will be collected between 1 December 2020 and 30 June 2021. Ideally, we would have conducted the survey as close to the first lockdown as possible but the research team and governance establishment processes, questionnaire design, and ethical approval process caused delays.

This initial telehealth survey is designed to target providers and consumers who use general practice services, as this is how the bulk of NZ primary healthcare is delivered. However, follow-up surveys are planned. The next survey will cover telehealth in allied health services in primary care and a separate one will cover midwifery. This should enable the ability to develop a multidimensional understanding of telehealth in primary care in NZ.

Since a survey cannot be designed to cover all aspects of a research question, there is also a need for follow-up qualitative studies to explore and examine nuances that cannot be detected by a questionnaire, for example, sensitive aspects of consultations, the patient–clinician relationship, and decisions and policies about whether to use an in-person, video or phone modality for a consultation.[17 18]

Participants

Anyone (consumers) can participate if they are 18 years or older; have had at least one general practice consultation with their GP, nurse or nurse practitioner by phone or video after 23 March 2020; are able to understand English well enough to complete the survey; currently reside in New Zealand; and are able to confirm that they have understood what the study is about and agree to participate. Any general practice provider (doctor or nurse) can participate if they have conducted a consultation via phone or video after 23 March 2020.

Since this is a national study, and due to the increase in workload in general practice resulting from the changes brought about by the pandemic, individual consumers will not be identified or recruited by their general practice providers. We will disseminate the survey via the Royal New Zealand College of General Practitioners, the College of Nurses Aotearoa (NZ), the New Zealand Telehealth Leadership Group and other organisations that regularly communicate with clinicians. To recruit consumers, we will use social media such as Facebook and Twitter; the news media; and flyers and posters in healthcare services, for example, general practices. The questionnaires will be delivered online, and participants will self-select to complete them.

Variables

The questionnaire design has been informed by the UTAUT[14] and theory of access to care.[9] The UTAUT variables are perceived usefulness, perceived ease of use and effect of social influence on use. The access to care variables include accessibility, availability, affordability, service design, acceptability, implementation and design. The questionnaire contains questions about demographics, how a person accessed a telehealth consultation (eg, how they made appointments) and how the appointment occurred (eg, by phone or video or combination of both). It also contains questions about the acceptability of healthcare via telehealth and intention to use video and/or phone again for consultations (using the UTAUT). Some health outcome questions are also included, for example, able to make an appointment.

The most concerning potential confounder is the passing of time and possible normalisation of telehealth in general practice, or conversely a reversion to prelockdown preferences for in-person consultations. Some people may have forgotten or have imperfect memories of their experiences of lockdown, which in turn may skew the results. These confounders may be sources of bias, especially because of self-selection and self-reporting required to respond to the survey. However, we are also measuring intent to continue using telehealth in the future (as well as collecting data on past experiences), and future intentions will not be affected by time or memory.

Power calculation

The NZ population is 4 900 600[19] with the Pacific population as the smallest at 8.3% of the total population. The Maori population accounts for 16.7% of New Zealanders.[19]

To calculate the power needed, we will assume that the percentage of Pacific people (because this is the smallest subpopulation) is 8.3%, and the main survey outcome is a binary value (satisfied with telehealth or not but that can be any other binary question as well). We will need 2000 total individuals to get a coefficient of variation (CV) of a maximum of 5% when at least 70% in the Pacific population is satisfied with telehealth. The other subpopulations are larger and therefore their CV is smaller, everything else being equal. The CV decreases as the satisfaction percentage increases.

Data analysis

As a descriptive study, the analysis will be divided into two parts. The analysis will be completed using 'R' and various applicable specialist libraries. The general format for the analysis will follow that of Wickham and Grolemund.[20] The first phase is the descriptive analytics and exploratory data analysis. The results of the online questionnaire will be summarised to provide an overview of telehealth users' experiences using packages such as 'skimr'[21] and special packages for Likert scales such as 'likert'[22] within the context of an opinionated data framework.[23] Descriptive statistics to describe the central tendency and variation across each sample will be described then will be presented by packages such as 'finalfit'.[24] The exploratory data analysis will divide the population based on the outcome parameters such as whether the consumer was able to get their needs met via a telehealth consultation. Routine tests such as χ^2 and analysis of variance along with similar non-parametric tests such as the Mann-Whitney U tests will also be undertaken.[25]

The second phase will be to develop explanatory statistical models, depending on the results of the exploratory data analysis. The approach will be on building exploratory analytic models to understand what factors explain the outcome parameters, such as whether the patient achieved a satisfactory outcome or not. Approaches such as logistic regression[25] or a random forest analysis via 'ranger'[26] may well be applicable. This will enable the relative variable importance in contributing to the outcome variance to be considered via importance plots.[27]

LIMITATIONS

The main value of the results will be to inform future studies on telehealth in NZ general practice specifically but also primary care more generally. The survey is limited to people who self-select to participate, written English capacity and relies on recall of experiences of telehealth since lockdown, which may create bias resulting in the inability to generalise the results. Further bias is introduced by the online nature of the survey as those without access to social media or other dissemination methods will be excluded. Often these, people are most negatively affected by increased telehealth. Bias has been mitigated by the use of theory to frame the survey, that is, the UTAUT model of user acceptance[14] and the access to care framework.[9 10] Additional mitigation measures include hard copy posters in GP waiting rooms and community venues. Hard copy questionnaires have been formatted so that participants who used telehealth but are unable to complete the survey online will be able to participate.

Some consumers who may benefit from a telehealth service may not be able to use it due to lack of access to appropriate technology, privacy or may lack the skills required to use the technology effectively. Others who may have attempted to make a telehealth appointment and abandoned it due to lack of skills or appropriate technology may want to participate in this research. Since the research is limited to those who actually experienced a telehealth appointment, the voice of those who were not successful will not be heard until follow-up studies have been completed.

The results of this study will help inform future research that addresses the above limitations. The design of follow-up qualitative and quantitative studies will aim to capture additional experiences, including but not limited to, other languages, those living with disability and those who could not access telehealth.

CONCLUSION

This protocol describes the first of several studies on telehealth in NZ general practice and other primary care services as a response to the changes brought about by the COVID-19 pandemic. The next step after analysing the data will be to establish how processes and workflows for providers are changed to accommodate telehealth as a 'business as usual' option. An in-depth investigation into the consumer experience will be designed to establish what can be done to enhance consumers' health outcomes via telehealth and enrich their ability for self-care.

Twitter Karen Day @KarenJDay

Acknowledgements The authors acknowledge the endorsement of this research protocol by the New Zealand Telehealth Leadership Group and the support of the project's advisory group in terms of their time, advice and feedback on the protocol development, the questionnaire design and assistance in disseminating the questionnaires.

Contributors IH and KD are coprincipal investigators, and the other authors are coinvestigators. The whole team designed the study. BvdW provided the statistical content. KD converted the study protocol used for ethical approval application into the submitted manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval Ethical approval was granted by the Auckland Health Research Ethics Committee on 13/11/2020, reference 2539.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Wootton R. Twenty years of telemedicine in chronic disease management–an evidence synthesis. *J Telemed Telecare* 2012;18:211–20.
2 Hollander JE, Carr BG. Virtually perfect? telemedicine for Covid-19. *N Engl J Med Overseas Ed* 2020;382:1679–81.
3 Greenhalgh T, Wherton J, Shaw S, *et al*. Video consultations for covid-19. *BMJ* 2020;368:m998.
4 Day K, Kerr P. The potential of telehealth for 'business as usual' in outpatient clinics. *J Telemed Telecare* 2012;18:138–41.
5 Wright M, Mainous A. Can continuity of care in primary care be sustained in the modern health system? *Aust J Gen Pract* 2018;47:667–9.
6 Bodenheimer T, Ghorob A, Willard-Grace R, *et al*. The 10 building blocks of high-performing primary care. *Ann Fam Med* 2014;12:166–71.
7 Starfield B. Is patient-centered care the same as person-focused care? *Perm J* 2011;15:63.
8 Portnoy J, Waller M, Elliott T. Telemedicine in the era of COVID-19. *J Allergy Clin Immunol Pract* 2020;8:1489–91.
9 Penchansky R, Thomas JW. The concept of access: definition and relationship to consumer satisfaction. *Med Care* 1981;19:127–40.
10 Saurman E. Improving access: modifying Penchansky and Thomas's theory of access. *J Health Serv Res Policy* 2016;21:36–9.
11 McLeod M, Gurney J, Harris R, *et al*. COVID-19: we must not forget about Indigenous health and equity 2020.
12 WHO. Gender, equity and human rights 2020, 2020. Available: https://www.who.int/gender-equity-rights/understanding/human-rights-definition/en/ [Accessed 16 Jul 2020].
13 Robson B, Harris R. *Hauora: M ori standards of health IV. A study of the years 2000–2005*. Wellington: Te Ropu Rangahau Hauora a Eru Pomare, 2007.
14 Venkatesh V, Thong JYL, Chan FKY, *et al*. Extending the two-stage information systems continuance model: incorporating UTAUT predictors and the role of context. *Information Systems Journal* 2011;21:527–55.
15 Paré G, Trudel M-C, Jaana M, *et al*. Synthesizing information systems knowledge: a typology of literature reviews. *Inf Manage* 2015;52:183–99.
16 von Elm E, Altman DG, Egger M, *et al*. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Int J Surg* 2014;12:1495–9.
17 Powell RE, Henstenburg JM, Cooper G, *et al*. Patient perceptions of telehealth primary care video visits. *Ann Fam Med* 2017;15:225–9.
18 Donaghy E, Atherton H, Hammersley V, *et al*. Acceptability, benefits, and challenges of video consulting: a qualitative study in primary care. *Br J Gen Pract* 2019;69:e586–94.
19 StatsNZ. Estimated resident population (2018-base): at 30 June 2018: StatsNZ, 2020. Available: https://www.stats.govt.nz/information-releases/estimated-resident-population-2018-base-at-30-june-2018 [Accessed 13 Nov 2020].
20 Wickham H, Grolemund G. R for data science: import, tidy, transform, visualize, and model data. " O'Reilly Media, Inc." 2016.
21 Waring E. Skimr: compact and flexible summaries of data. R package version 2.1.2, 2020. Available: https://CRAN.R-project.org/package=skimr
22 Bryer J. Speerschneider KJRpv. likert: analysis and visualization likert items 2016.
23 Wickham H, Averick M, Bryan J, *et al*. Welcome to the Tidyverse. *JOSS* 2019;4:1686.
24 Harrison E, Drake T, RJRpv O. R-project. org/package= finalfit. finalfit: quickly create elegant regression results tables and plots when modelling, 2019. Available: https://CRAN
25 R Core Team. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2020. https://www.R-project.org/
26 Wright MN, AJapa Z. Ranger: a fast implementation of random forests for high dimensional data in C++ and R 2015.
27 Greenwell BM, Boehmke B. Variable importance Plots-An introduction to the VIP package. *The R Journal* 2020;12:1.

BMJ Health & Care Informatics

Performance of national COVID-19 'symptom checkers': a comparative case simulation study

Fatma Mansab,[1,2] Sohail Bhatti,[1] Daniel Goyal [1,3,4]

To cite: Mansab F, Bhatti S, Goyal D. Performance of national COVID-19 'symptom checkers': a comparative case simulation study. *BMJ Health Care Inform* 2021;**28**:e100187. doi:10.1136/bmjhci-2020-100187

► Additional material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2020-100187).

Received 03 June 2020
Revised 03 February 2021
Accepted 06 February 2021

[1]Postgraduate School of Medicine, Department of Public Health, Gibraltar Health Authority, Gibraltar, Gibraltar
[2]University of Gibraltar, Gibraltar, Gibraltar
[3]Deparment of Medicine, Gibraltar Health Authority, Gibraltar, Gibraltar
[4]Department of Health Systems, University of Gibraltar, Gibraltar, Gibraltar

Correspondence to
Dr Daniel Goyal;
daniel.goyal@gha.gi

ABSTRACT

Objectives Identifying those individuals requiring medical care is a basic tenet of the pandemic response. Here, we examine the COVID-19 community triage pathways employed by four nations, specifically comparing the safety and efficacy of national online 'symptom checkers' used within the triage pathway.

Methods A simulation study was conducted on current, nationwide, patient-led symptom checkers from four countries (Singapore, Japan, USA and UK). 52 cases were simulated to approximate typical COVID-19 presentations (mild, moderate, severe and critical) and COVID-19 mimickers (eg, sepsis and bacterial pneumonia). The same simulations were applied to each of the four country's symptom checkers, and the recommendations to refer on for medical care or to stay home were recorded and compared.

Results The symptom checkers from Singapore and Japan advised onward healthcare contact for the majority of simulations (88% and 77%, respectively). The USA and UK symptom checkers triaged 38% and 44% of cases to healthcare contact, respectively. Both the US and UK symptom checkers consistently failed to identify severe COVID-19, bacterial pneumonia and sepsis, triaging such cases to stay home.

Conclusion Our results suggest that whilst 'symptom checkers' may be of use to the healthcare COVID-19 response, there is the potential for such patient-led assessment tools to worsen outcomes by delaying appropriate clinical assessment. The key features of the well-performing symptom checkers are discussed.

Summary box

What is already known?

- The availability and use of symptom checkers are increasing.
- Symptom checkers are currently in use at a national level to help in the healthcare response to COVID-19.
- There is limited evidence to support the effectiveness or safety of symptom checkers as triage tools during a pandemic response.

What does this paper add?

- This study compares performance of symptom checkers across different countries, revealing marked variation between national symptom checkers.
- The symptom checkers employed by Japan and Singapore are twice as likely to triage cases onward for clinical assessment than those of the USA or UK.
- The US and UK symptom checkers frequently triaged simulated cases of sepsis, bacterial pneumonia and severe COVID-19 to stay home with no further healthcare contact.
- We discuss the key aspects of the well-performing triage systems.

INTRODUCTION

COVID-19 is a new infection in humans. The symptom profile, disease progression and complication rates are still relatively unknown.[1] From the available evidence, four broad categories of illness have been postulated. 'Mild COVID-19' makes up over 80% of cases and is typically a self-limiting infection similar to the common cold, resolving without intervention. 'moderate COVID-19' typically has features of viral pneumonia in the absence of hypoxia, progressing to 'severe COVID-19' typically when patients require oxygen therapy. 'Critical COVID-19', where ventilatory support is typically required, occurs in less than 5% of cases.[2] The rate of disease progression is not fixed: early intervention and various management strategies can reduce the rate of progression to critical illness and death.[2–6]

While the infection fatality rate is yet to be determined, COVID-19 is associated with a substantive mortality. Over a period of 5 months, COVID-19 has led to more than 300 000 deaths, with more than half these deaths occurring within the last month.[7]

The risk of mortality is affected by a number of risk factors. Coexisting health problems such as diabetes, heart disease and cancer have been implicated as conferring a higher risk of mortality in COVID-19.[8] Age appears to be the most striking and consistent risk factor for COVID-19 related mortality.[9] Based on current data, the mortality rate in patients under 50 years of age is thought to be less than 1.1%, rising to around 14% in those over 80 years of age.[10]

Variation in mortality also seems to exist between countries.[11] Initially, this variation was thought to be predominantly related to the method of recording deaths and the total number of tests conducted (ie, the detection of milder cases).[12] As the pandemic spreads across the globe, it is becoming increasingly clear that how a country responds to the pandemic impacts the number of deaths their locality will experience.[6 11]

The national response to the COVID-19 pandemic has many important tenets. On the public health side, infection control initiatives attempt, in part, to mitigate the surge of infections that can accompany new pathogens where there is little circulating immunity. This reduces mortality by preventing the healthcare services from being overwhelmed, thus permitting improved access to medical management for those who need it.[6] The clinical response to COVID-19 also centres on access to treatment. To successfully reduce the mortality rate, those patients who are developing more severe disease must be identified.[3]

Identifying those patients with COVID-19 that require treatment is challenging. First, COVID-19 has a broad range of presentations that can mimic common conditions that rarely require clinical assessment (eg, the common cold).[1] Second, there are no clinical signs or symptoms that reliably predict who will progress to severe disease.[3] As such, the clinical community is left with a large number of potential cases without any clear symptom indicators for: (1) who has the disease and (2) who is developing more severe disease. The problem is compounded further as more serious, life-threatening conditions (eg, bacterial pneumonia and sepsis) can mimic any stage of COVID-19 disease.[13 14]

National 'Symptom Checkers' have been implemented in many countries in the hope of reducing this burden faced by healthcare services. Symptom checkers are self-assessment tools. The individual—typically online or via computer application—enters their symptoms into a predetermined platform and from there a predetermined algorithm produces an outcome (usually advice). This is a form of self-led triage. It is hoped that such self-directed assessments will enable the identification of potential cases[15] and will correctly triage those individuals who would benefit from clinical assessment and/or management into further care.[16] For such a hope to be realised, symptom checkers must be able to determine mild conditions from severe conditions.

While self-triage has been used for some years in non-emergency conditions to varying degrees of success,[17] self-triage has never before been used in a pandemic setting and as yet the efficacy and safety has not be formally studied. Caution must be exercised as, to date, studies examining symptom checkers have had mixed and disappointing results in general—demonstrating poor diagnostic performance (34%–58%) and questionable triage performance (55%–80%).[18] The stakes are high, in that a failure to triage serious medical conditions (such as severe COVID-19, bacterial pneumonia or sepsis) in for further assessment will inevitably lead to delayed treatment and higher mortality.[19–22]

Here, we test the performance of four nationwide symptom checkers from four nations to ascertain how safe and efficient each symptom checker is in differentiating mild from severe COVID-19 cases, and how well they detect time-sensitive COVID-19 mimickers such as bacterial pneumonia and sepsis.

METHODOLOGY

Five countries were initially selected for analysis. Three (Singapore, Japan and Norway) were selected as they maintained low case fatality rates (CFRs) despite a demonstrable surge of cases in the preceding 2 months. Two countries (the UK and the USA) were selected due to concern regarding high CFRs.

Public health guidelines from each country were reviewed. Access was obtained to any available government sponsored online patient-led triage system (Singapore: 'Singapore COVID-19 Symptom Checker',[23] Japan: 'Stop COVID-19 Symptom Checker',[24] USA: 'CDC Coronavirus Symptom Checker'[25] and the UK: '111 COVID-19 Symptom Checker'[26]). Whereas the NHS '111' COVID-19 Symptom Checker was and continues to be heavily used (with over 500 000 assessments completed on average each month[27]), there was no available data as to the usage of the other symptom checkers.

For the purpose of this analysis, data were extracted only from those countries with symptom checkers (Singapore, Japan, UK and USA), in an effort to compare the performance of symptom checkers specifically.

Case scenarios

Fifty-two standardised cases were designed simulating common COVID-19 related presentations with varying severity or risk factors.

Case scenarios included four distinct presentations: (1) cough and fever; (2) comorbidity, cough and fever; (3) immunosuppression, cough and fever and (4) shortness of breath and fever. These distinct presentations were then varied in relation to one or more of the following: (1) duration of symptoms; (2) age of patient and (3) severity of symptoms. The symptoms chosen for analysis are considered common in COVID-19: history of fever (50%–90%), dry cough (60%–86%) and shortness of breath (53%–80%).[3 28]

'Fever' was chosen as a core symptom of COVID-19 due to its high discriminatory value for infection. Even though it may only be present in less than half of COVID-19 cases at presentation,[28] the presence of fever permits greater focus on infective causes in relation to shortness of breath and cough. Fever also presents commonly in sepsis and pneumonia,[29] which are two of the key diagnoses that triage systems need to detect to prevent excess mortality.

Fever has also been shown to relate to disease severity and mortality outcomes in COVID-19.[30]

'Cough' is a non-specific symptom covering a wide range of conditions. Combined with fever, cough raises the possibility of chest infection, including COVID-19 and bacterial pneumonia (one of the critical differential diagnoses in COVID-19). Detecting possible bacterial pneumonia is a prerequisite to a functioning triage system given the time critical need for antibiotic initiation to prevent unnecessary deaths.[30]

'Shortness of breath' is generally accepted as a marker of COVID-19 disease progression,[31] although there are other reasons for shortness of breath, and specifically in COVID-19, patients may not experience shortness of breath despite being hypoxic—so called silent hypoxia.[32]

'Duration' was chosen as a severity marker as the prolongation of fever, cough and/or shortness of breath within the context of COVID-19 or a COVID-19 mimicker (pneumonia, sepsis and so on) carries a worse prognosis. In particular, an unremitting, persistent fever warrants further assessment in regard to COVID-19[30] but also in relation to sepsis.[29]

'Age' is a well-defined risk factor for severe complications of COVID-19.[9 10] As such, it was deemed useful to include age as a variable in the case simulations to test whether the symptom checker accounted for age when determining risk.

'Severity' of symptoms relates to duration of fever, cough and shortness of breath. Shortness of breath had its own severity scale and was crucial for staging level of complicated COVID-19, severity of pneumonia and sepsis.[29 30] Mild shortness of breath was defined as shortness of breath during activities that did not stop one completing the activity. Moderate shortness of breath was defined differently depending on age. That is, respiratory reserve was considered to be less in adults aged >70 years of age in comparison with the younger age groups, and as such, we defined moderate shortness of breath in those >70 years of age as preventing the completion of most tasks, while for younger cases, moderate shortness of breath would still permit most tasks to be completed. Severe shortness of breath was defined as shortness of breath at rest.

The immunosuppression case simulations related to the development of cough and fever 4 days after chemotherapy, simulating potential neutropaenic sepsis. Neutropenic sepsis is a medical emergency requiring immediate medical attention, and early antibiotic therapy - door to needle time for sepsis should be less than 1 hour, and for neutropaenic sepsis less than 30 min.[33 34]

Except for the paediatric case, hypertension was chosen as the comorbidity due to its discriminatory value between mild and severe comorbidities. There is evidence that hypertension may be an independent risk factor for poorer outcomes in COVID-19; however, it remains, as do many of the proposed 'high-risk' comorbidities, unproven.[8] Differentiating symptom checkers that account for milder comorbidities or make allowances for the uncertainty that remains in the evidence base for at-risk groups was deemed useful in regard to symptom checkers' safety performance.

Where equivocal answers existed, such as for breathless: 'yes', 'I'm not sure' or 'no', the equivocal answer ('I'm not sure') was interpreted as mild symptoms. Unless stated in the specific case scenario, any question pertaining to comorbidity was answered as 'no'. All other variations were as described for each case scenario (online supplemental data).

The combination of symptoms, duration and other severity markers were varied to simulate many of the common presentations of COVID-19 and COVID-19 mimickers. Upper respiratory tract infection (URTI) and mild COVID-19 were represented in scenario 1; moderate COVID-19, bacterial pneumonia and sepsis were represented in scenarios 1 and 4; severe COVID-19, septic shock and critical COVID-19 are represented in scenario 4; and neutropaenic sepsis in scenario 3 (see online supplemental tables 1–4)

Statistical analysis

The primary outcome was total number of cases referred onward for further clinical assessment, which was converted into a percentage ratio and then compared between countries.

RESULTS

The key baseline population and testing data are presented in table 1. Notably, the highest rate of testing for COVID-19 was by Singapore with the lowest being

Table 1 Key population and COVID-19 testing data from each of the four countries

Population data	Singapore	Japan	USA	UK
Total tests (per million)	20 815	1166	16 507	9867
Total tests (thousands)	122	147	5500	669
Population (millions)	5.8	126.5	331	67.9
Confirmed COVID-19 positive	12 693	13 182	899 281	148 381
Cases per thousand inhabitants	2.2	0.1	2.7	2.3
Case fatality rate (%)	<0.1	2.7	5.6	13.6
Physicians per 10 000 head of capita	24	24	25	28

Table 2 Total number (percentage) of case simulations referred on by country

	Case fatality rate %	Total cases referred onwards	Cough+fever	Comorbidity +cough+fever	Immunosuppressed +cough+fever	Shortness of breath +fever
		n=52 (%)	n=16 (%)	n=12 (%)	n=12 (%)	n=12 (%)
Singapore	0.1	46 (88)	10 (63)	12 (100)	12 (100)	12 (100)
Japan	2.7	40 (77)	12 (75)	8 (67)	8 (75)	12 (100)
USA	5.6	20 (38)	0	3 (25)	12 (100)	5 (42)
UK	13.6	23 (44)	0	3 (25)	12 (100)	8 (67)

Distinct scenarios are included. Variation within each scenario is not detailed here (see online supplemental data).

Japan. The UK had the highest reported physicians per capita, while Japan and Singapore had the lowest. Cases per thousand inhabitants varied greatly, with Singapore and the UK maintaining similar rates. From the available statistics, Singapore had the lowest CFR (<0.1%) and the UK had the highest CFR (13.6%) currently. All population and testing data were extracted from The WHO as of 26 April 2020.

Fifty-two case scenarios were applied to each country's patient-led triage systems. The results for each scenario are presented in tabulated format (online supplemental data). Singapore had the highest overall referral rate at 88%, and the USA had the lowest at 38% (table 2).

From the cases not referred, the USA and UK triaged a significant number of cases to 'stay home' that would typically have required early clinical assessment. The US triage system (CDC Coronavirus Symptom Checker) frequently triaged home case simulations with possible severe COVID-19, bacterial pneumonia and sepsis and triaged possible neutropaenic sepsis to healthcare contact within 24 hours. The UK's 111 COVID-19 Symptom Checker frequently triaged possible severe COVID-19 and bacterial pneumonia to stay at home with no follow-up and is likely to have delayed treatment for sepsis, severe COVID-19 and neutropaenic sepsis. It is of note that while Japan's symptom checker generally performed well, our simulation revealed a potential delay to treatment for neutropaenic sepsis. Indeed, all four symptom checkers failed to triage the simulation for neutropaenic sepsis into the 'emergency department' (table 3).

High CFR versus low CFR countries

The main differences in triage criteria extrapolated from the national symptom checkers relating to COVID-19 between the low CFR countries and the high CFR countries are presented at table 4.

DISCUSSION

This case simulation study examined the symptom trackers from four countries. Following application of 52 standardised case simulations to each country's symptom checker, the percentage of onward referrals were calculated. The low case fatality nations' (Singapore and Japan) symptom checkers triaged in twice as many cases for direct clinical assessment than the higher case fatality nations (the USA and UK). Of clinical concern was the failure of both the US and UK symptom checkers to triage cases simulating bacterial pneumonia, sepsis and severe COVID-19 on to any healthcare contact.

The upside of symptom checkers, particularly during a pandemic is difficult to ignore. By reducing physical patient contacts, symptom checkers can potentially save valuable resources and avoid further viral transmission. While telephone and telemedicine triaging also protects staff and reduces transmission, such services require more healthcare staff than symptom checkers and hence carries a greater financial and human resource burden.

Evidence to date also suggests the majority of cases of COVID-19 resolve after a short, self-limiting viral illness.[1] There are, though, no discriminatory signs or symptoms.[2]

Table 3 Tabulated view of likely triage outcome of specific diagnosis in each country

	URTI (mild)	COVID-19 (mild)	COVID-19 (moderate)	Bacterial pneumonia	Sepsis	COVID-19 (severe)	Neutropaenic sepsis	Septic shock	COVID-19 (critical)
USA	🟥	🟥	🟥	🟥	🟥	🟧	🟧	🟧	🟩
UK	🟥	🟥	🟥	🟥	🟧	🟧	🟧	🟩	🟩
Japan	🟥	🟧	🟩	🟩	🟩	🟩	🟧	🟩	🟩
Singapore	🟥	🟧	🟩	🟩	🟩	🟩	🟧	🟩	🟩

Columns indicate clinical diagnosis and rows represent the likely consequence of the country's triage response. Red indicates cases that would have likely been dismissed (stay home) by the patient-led triage system. Orange indicates cases that were likely to have been triaged to delayed clinical contact or to stay at home. Green indicates diagnoses likely to have been captured and triaged to clinical care.
URTI, upper respiratory tract infection.

Table 4 Differences in triage criteria between low and high case fatality countries

Triage criteria	Low CFR country	High CFR country
Duration of symptoms	Singapore and Japan recommend clinical assessment after day 4 of symptoms.	For both USA and the UK, duration of symptoms did not affect the triage advice in any case simulation completed.
Age	Singapore triages all patients over the age of 65 years with viral symptoms to clinical assessment. Japan recommend all 'older adults' to seek medical attention if viral symptoms persist more than 2 days.	Age (adults) did not appear to affect the recommendations in either the USA or UK triage systems.
Comorbidity	Singapore triaged all patients with any comorbidity directly to specialist clinic. Japan recommend patients with any comorbidity be assessed if symptoms are not improved after the second day.	The USA is more likely to triage patients with specific comorbidities to further care. The UK only considered patients with severe, high-risk comorbidities in their triage process.
Shortness of breath	Singapore and Japan all advise immediate clinical contact if a patient develops shortness of breath.	Both US and UK systems attempt to qualify the severity of shortness of breath. The USA and UK advise patients with 'mild' shortness of breath to remain at home.
Severity and safety-net advice	Singapore and Japan are explicit and repetitive about the need to make clinical contact if there are any worsening of symptoms.	The UK system's advice to seek medical care if symptoms worsen is distant to the initial recommendation to remain at home. Guidance is provided on how to manage 'breathlessness' at home. Both the USA and UK focused on 'stay home'.

CFR, case fatality rate; GP, general practitioner.

COVID-19 can present like the common cold or influenza or indeed bacterial pneumonia. COVID-19 can also progress quickly[6 35] and can even present with asymptomatic hypoxia.[32] Sifting through the mild colds and self-limiting flus and trying to determine who will have a mild course of COVID-19 and also trying not to miss bacterial pneumonia, sepsis and signs of COVID-19 pneumonia is a challenge for even trained clinicians let alone an automated system.

It is here where Singapore's symptom checker performs well. The checker is presented on a single webpage, more akin to an online risk calculator. There are six inputs required from the patient and one of three outputs generated. The algorithm powering the symptom tracker is not complicated. Age over 65 years, or the presence of any health condition, or duration of symptoms over 4 days triggers the advice to seek medical assessment. Any degree of shortness of breath is triaged directly to the emergency department. The Singapore COVID-19 Symptom Checker, if used by the public, is likely to reduce healthcare contacts by the young, fit patients who are early on in the illness, thus off-loading the healthcare burden to some degree while maintaining a relatively low risk to the public.

The UK '111' symptom checker performs poorly in this regard. The algorithm is complex, attempting to quantify symptoms such as shortness of breath and the overall severity of illness by asking subjective, qualitative questions with multiple choices. The '111' symptom checker seems to take on a much broader clinical role and attempts to triage out cases that would typically be triaged in or out of care based on an actual clinical assessment. For example, a 72 year old person who presents with a 7-day history of fever and cough is triaged by the '111' symptom checker to stay at home with no clinical, nursing or healthcare contact. Faced with such a clinical scenario, clinicians would typically insist on at least basic clinical observations (pulse, temperature, oxygen levels and so on) before considering triaging such a patient to stay at home. The differential in this case includes sepsis, bacterial pneumonia and COVID-19 pneumonia, and while it remains possible that fever can persist for 7 days in mild/moderate COVID-19, complications or alternative diagnoses are much more likely.

The qualifying questions used by the '111' symptom checker to discriminate between severity will have insufficient discriminatory value in such cases. Furthermore, the wording of the question encourages the self-reporting towards lower categories of illness:

Are you so ill that you have stopped doing all of your usual daily activities?

a. 'Yes - Ive stopped doing everything I usually do'.
b. 'I feel ill but can do some of my usual activities'.
c. 'No - I feel well enough to do most of my usual activities'.

(Extracted question from '111' Coronavirus Symptom Checker).

It is the use of absolute and equivocal qualifiers that prevent the severity-qualifying question from achieving any useable clinical triage information: the use of 'all' in the question, 'everything' in the affirmative answer, and even the negative answer stipulates 'most'. Our case simulation demonstrated that answering B, the moderately severe answer, still triages patients to self-isolate with no healthcare contact. As such, patients with cough and fever for 7 days would have to be so severely unwell that they are unable to do anything they usually do to be triaged to any clinical contact.

Our case simulation study indicates that both the '111 COVID-19 Symptom Checker' and the 'CDC Coronavirus Symptom Checker', if used as the sole initial point of healthcare contact, are likely to delay presentations of serious medical conditions to appropriate care, and as such, are likely to confer an increased risk of morbidity and mortality. Both symptom checkers maintain a high threshold for referring onward to clinical contact, triaging the majority of patients to stay home with no clinical contact. Again, beyond the mortality impact, there is no evidence that such an approach actually reduces healthcare burden. Indeed, beyond the established evidence in pneumonia generally,[19–22] there is direct evidence that early correction of hypoxia in COVID-19 prevents progression to mechanical ventilation,[5] consistent with basic medical principles. Programming symptom checkers to aggressively triage patients to stay home may well lead to patients presenting to healthcare later, requiring more intensive healthcare to recover, and as such, symptom checkers 'set' to keep patients at home may actually increase the burden on intensive care facilities and perpetuate a healthcare crisis.

Symptom checkers are currently being used in the pandemic for two purposes: (1) identifying potential cases for testing/surveillance and (2) identifying 'unwell' patients who require medical attention. Both functions are likely to be enhanced by the use of symptom checkers when the intention is to 'catch' more patients or reach more cases. That is, when symptom checkers are used to identify more cases than would otherwise be detected and to direct more patients to medical care than would otherwise make healthcare contact, then symptom checkers are merely providing an additional 'safety-net', and therefore, in such a healthcare support role, the risk of harm from their use is expected to be relatively minimal. Conversely, if symptom checkers are being used to replace the assessment of patients by trained personnel and are programmed to try and prevent further healthcare contact, then, as our case simulation study highlights, there are real concerns about the potential risk of harm from such an unproven approach.

Considering that the efficacy of symptom checkers have not been established,[17] caution would be advisable. Delay in the correction of hypoxia, failure to commence thromboprophylaxis and missing the opportunity for earlier initiation of steroids in the hypoxic patient with COVID-19, are all likely to carry a considerable morbidity and mortality cost.

If we are to accept the lesser option of an automated, self-directed triage system over the standard of care offered by the dynamic, experienced clinical assessment, then we must be mindful of what we are asking of the 'symptom checker'. Based on our independent case simulation study, symptom checkers do not appear advanced enough to fulfil the 'stay home' intent with any sufficient level of safety. They may though be sufficient enough to assist in the improved identification of at risk patients requiring further clinical assessment, and some form of symptom checker may even be able to contribute to the increased ongoing vigilance required for all patients diagnosed with COVID-19. Evidence, though, should be provided before replacing actual clinical contact with an online self-directed triage system.

Strengths and limitations

This case simulation study was conducted using 52 standardised simulated cases. The cases were designed to test specific COVID-19 related scenarios and as such were symptom-based without the need for subjective interpretation. Nonetheless, there remains a risk of bias, particularly when facing subjective questions. The majority of simulations were though more quantitative, for example, duration, age and symptoms, and unlikely to be affected meaningfully by any bias.

The UK data are pooled from all four nations (England, Wales, Scotland and Northern Ireland). England (making up 90% of the total UK population) uses the same '111' COVID-19 patient-led triage system analysed here, whereas Wales, Scotland and Northern Ireland have implemented their own individual patient-led triage systems. It was beyond the scope of this initial investigation to examine each triage system separately. A similar situation applies to the USA, where some individual states have implemented their own triage systems.

CONCLUSION

In this case simulation study, the UK and USA patient-led triage systems (COVID-19 Symptom Checkers) maintained a high disease-severity threshold for onward referral to healthcare assessment. Particular concerns were advising no clinical contact for elderly patients with COVID-19 related symptoms or patients who had developed shortness of breath or any patient with persistent fever. The low CFRcountries (Singapore and Japan) used symptom checkers to reduce clinical demand while maintaining a lower health risk to patients. Our study indicates that while symptom checkers may be of use in the healthcare response to COVID-19, the 'CDC Coronavirus Symptom Checker' and the '111 COVID-19 Symptom Checker', if used as the sole point of initial healthcare contact, are likely to confer a tangible risk of delaying the presentation of time-critical acute illnesses. Our results support the recommendation that symptom checkers should be subjected to the same level of evidenced-based quality assurance as other diagnostic tests prior to implementation.

Contributors All authors contributed to the conception, methodology and analysis. FM led the study and undertook the majority of the write-up and revisions. DG undertook edits and contributed to the write-up and revisions of the manuscript. All authors had final review and edit.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information. The data generated in this study are based on case simulations and are available to use with citation.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iD

Daniel Goyal http://orcid.org/0000-0003-0418-8859

REFERENCES

1 Koh J, Shah SU, Chua PEY, *et al*. Epidemiological and clinical characteristics of cases during the early phase of COVID-19 pandemic: a systematic review and meta-analysis. *Front Med* 2020;7:295.

2 National Institutes of Health. COVID-19 treatment guidelines panel. coronavirus disease 2019 (COVID-19) treatment guidelines. Available: https://www.covid19treatmentguidelines.nih.gov/ [Accessed 05 Oct 2020].

3 Wiersinga WJ, Rhodes A, Cheng AC, *et al*. Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (COVID-19): a review. *JAMA* 2020;324:782–93.

4 RECOVERY Collaborative Group, Horby P, Lim WS, *et al*. Dexamethasone in Hospitalized Patients with Covid-19 - Preliminary Report. *N Engl J Med* 2020. doi:10.1056/NEJMoa2021436. [Epub ahead of print: 17 Jul 2020].

5 Sun Q, Qiu H, Huang M, *et al*. Lower mortality of COVID-19 by early recognition and intervention: experience from Jiangsu Province. *Ann Intensive Care* 2020;10:33.

6 Ji Y, Ma Z, Peppelenbosch MP, *et al*. Potential association between COVID-19 mortality and health-care resource availability. *Lancet Glob Health* 2020;8:e480.

7 WHO. Coronavirus disease 2019 (COVID-19) situation report – 97. Data as received by WHO from national authorities by 10:00 CEST; 2020.

8 Ssentongo P, Ssentongo AE, Heilbrunn ES, *et al*. Association of cardiovascular disease and 10 other pre-existing comorbidities with COVID-19 mortality: a systematic review and meta-analysis. *PLoS One* 2020;15:e0238215.

9 Dhama K, Patel SK, Kumar R, *et al*. Geriatric population during the COVID-19 pandemic: problems, considerations, Exigencies, and beyond. *Front Public Health* 2020;8:574198.

10 Bonanad C, García-Blas S, Tarazona-Santabalbina F, *et al*. The effect of age on mortality in patients with COVID-19: a meta-analysis with 611,583 subjects. *J Am Med Dir Assoc* 2020;21:915–8.

11 Bilinski A, Emanuel EJ. COVID-19 and excess all-cause mortality in the US and 18 comparison countries. *JAMA* 2020;324:2100–2.

12 Rajgor DD, Lee MH, Archuleta S, *et al*. The many estimates of the COVID-19 case fatality rate [published online ahead of print, 2020 Mar 27]. *Lancet Infect Dis* 2020;S1473-3099:30244–9.

13 Nickel CH, Bingisser R. Mimics and chameleons of COVID-19. *Swiss Med Wkly* 2020;150:w20231.

14 Coleman JJ, Manavi K, Marson EJ, *et al*. COVID-19: to be or not to be; that is the diagnostic question. *Postgrad Med J* 2020;96:392–8.

15 Mehl A, Bergey F, Cawley C, *et al*. Syndromic surveillance insights from a symptom assessment APP before and during COVID-19 measures in Germany and the United Kingdom: results from repeated cross-sectional analyses. *JMIR Mhealth Uhealth* 2020;8:e21364.

16 Judson TJ, Odisho AY, Neinstein AB, *et al*. Rapid design and implementation of an integrated patient self-triage and self-scheduling tool for COVID-19. *J Am Med Inform Assoc* 2020;27:860–6.

17 Chambers D, Cantrell AJ, Johnson M, *et al*. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. *BMJ Open* 2019;9:e027743.

18 Semigran HL, Linder JA, Gidengil C, *et al*. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ* 2015;351:h3480.

19 Chalupka AN, Talmor D. The economics of sepsis. *Crit Care Clin* 2012;28:57–76.

20 Burdick H, Pino E, Gabel-Comeau D, *et al*. Effect of a sepsis prediction algorithm on patient mortality, length of stay and readmission: a prospective multicentre clinical outcomes evaluation of real-world patient data from US hospitals. *BMJ Health Care Inform* 2020;27:e100109.

21 Hu W-P, Zhang F-Y, Zhang J, *et al*. Initial diagnosis and management of adult community-acquired pneumonia: a 5-day prospective study in Shanghai. *J Thorac Dis* 2020;12:1417–26.

22 Blot SI, Rodriguez A, Solé-Violán J, *et al*. Effects of delayed oxygenation assessment on time to antibiotic delivery and mortality in patients with severe community-acquired pneumonia. *Crit Care Med* 2007;35:2509–14.

23 Singapore COVID-19 symptom Checker. Available: https://sgcovidcheck.gov.sg

24 Tokyo Government COVID-19 task force website. How to get help if you suspect having COVID-19. Available: https://stopcovid19.metro.tokyo.lg.jp/en/flow/

25 CDC, USA. Coronavirus symptom Checker. Available: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html

26 NHS, UK. COVID-19 symptom Checker. Available: https://111.nhs.uk/covid-19/

27 NHS Digital. Potential coronavirus (COVID-19) symptoms reported through NHS pathways and 111 online. Available: https://digital.nhs.uk/data-and-information/publications/statistical/mi-potential-covid-19-symptoms-reported-through-nhs-pathways-and-111-online/latest [Accessed 10 Oct 2020].

28 Richardson S, Hirsch JS, Narasimhan M, *et al*. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. *JAMA* 2020;323:2052.

29 Tidswell R, Singer M. Sepsis - thoughtful management for the non-expert. *Clin Med* 2018;18:62–8.

30 Zheng Z, Peng F, Xu B, *et al*. Risk factors of critical & mortal COVID-19 cases: A systematic literature review and meta-analysis. *J Infect* 2020;81:e16–25.

31 Keeley P, Buchanan D, Carolan C, *et al*. Symptom burden and clinical profile of COVID-19 deaths: a rapid systematic review and evidence summary. *BMJ Support Palliat Care* 2020;10:381–4.

32 Wilkerson RG, Adler JD, Shah NG, *et al*. Silent hypoxia: a harbinger of clinical deterioration in patients with COVID-19. *Am J Emerg Med* 2020;38:2243.e5–2243.e6.

33 Rhodes A, Evans LE, Alhazzani W, *et al*. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock: 2016. *Intensive Care Med* 2017;43:304–77.

34 Kochanek M, Schalk E, von Bergwelt-Baildon M, *et al*. Management of sepsis in neutropenic cancer patients: 2018 guidelines from the infectious diseases Working Party (AGIHO) and intensive care Working Party (iCHOP) of the German Society of hematology and medical oncology (DGHO). *Ann Hematol* 2019;98:1051–69.

35 Goh KJ, Choong MC, Cheong EH, *et al*. Rapid progression to acute respiratory distress syndrome: review of current understanding of critical illness from COVID-19 infection. *Ann Acad Med Singap* 2020;49:108–18.

BMJ Health & Care Informatics

COVID-19: are the elderly prepared for virtual healthcare?

Ahmed Ezzat,[1] Harpreet Sood,[2] Josephine Holt,[1] Hashim Ahmed,[1] Matthieu Komorowski [3]

To cite: Ezzat A, Sood H, Holt J, *et al*. COVID-19: are the elderly prepared for virtual healthcare? *BMJ Health Care Inform* 2021;**28**:e100334. doi:10.1136/bmjhci-2021-100334

Accepted 02 March 2021

[1]Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London, UK
[2]Health Education England, London, UK
[3]Intensive Care Unit, Charing Cross Hospital,Imperial College Healthcare NHS Trust, London, UK

Correspondence to
Dr Ahmed Ezzat; a.ezzat@nhs.net

We follow with interest the unprecedented shift towards virtual healthcare during the COVID-19 pandemic. We echo concerns reported by colleagues of the fine balance between a need for global initiatives in cutting traditional red tape to enable rapid deployment of virtual health infrastructures versus the potential risks to quality of patient care that might occur when the patient is not physically in the same room as the clinician.[1]

We have also highlighted the potential clinical, socioeconomic and environmental benefits of virtual consultations in secondary care.[2] However, with approximately one million consultations occurring each day in primary care in England alone, and 60% of primary care service users above the age of 60 years, there is a concern that elderly populations throughout the world will be disadvantaged in access to these virtual services due to lower proficiency with the tools of communication technology.[3 4] For instance, the UK Office of National Statistics reports only 50% ownership of smartphones in those aged 55 and over versus up to 95% in the 16–24 age group.[5]

Besides the potential for accrued health risks from reduced physical interactions with healthcare services, we are concerned that older patients are at greater risk of social isolation and potential worsening mental health. If the sudden transition to remote consultations, rightly necessitated by a pandemic, is continued, then these patients may miss out on essential physical encounters including those crucial for chronic disease management. Indeed, for many patients who live alone, visits to health clinics constitute a social occasion to connect face to face with peers and carers.

So, when planning for a digital healthcare system beyond the immediacy of the COVID-19 pandemic, and which incorporates a greater role for remote consultations, careful consideration as to the impact to elderly and vulnerable populations is required. This is especially relevant during a second COVID-19 peak of infections and new local lockdown measures. Health policies can support this necessary technological revolution in this age group with special consideration to their premorbid conditions, such as arthritis, visual impairment, or cognitive impairment. If planned well and delivered robustly, this virtual shift could also represent a fortuitous opportunity to widen access for the elderly in both ownership and use of information and communication technology. Access to virtual consultations, in the ageing population, emphasises the quality of life benefits, fostered through the adoption of technology. It is part of a growing trend of elderly people who not only receive heathcare but also stay in touch with friends and family at distance and reduce isolation.

Twitter Matthieu Komorowski @matkomorowski

Contributors All authors contributed equally.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iD
Matthieu Komorowski http://orcid.org/0000-0003-0559-5747

REFERENCES

1 Webster P. Virtual health care in the era of COVID-19. *Lancet* 2020;395:1180–1.
2 Edison MA, Connor MJ, Miah S. Understanding Virtual Urology Clinics: A Systematic Review [published online ahead of print, 2020 May 28]. *BJU Int* 2020.
3 Available: https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/demographics/age-groups/latest [Accessed on 03/06/2020].
4 Available: https://www.england.nhs.uk/five-year-forward-view/next-steps-on-the-nhs-five-year-forward-view/primary-care/ [Accessed on 03/06/2020].
5 Available: https://www.ofcom.org.uk/__data/assets/pdf_file/0022/117256/CMR-2018-narrative-report.pdf [Accessed on 17/10/2020].

BMJ Ezzat A, *et al*. *BMJ Health Care Inform* 2021;**28**:e100334. doi:10.1136/bmjhci-2021-100334 1

BMJ Health & Care Informatics

Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review

Mustafa Khanbhai ,[1] Patrick Anyadi,[1] Joshua Symons,[2] Kelsey Flott,[1] Ara Darzi,[3] Erik Mayer[1]

To cite: Khanbhai M, Anyadi P, Symons J, *et al*. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. *BMJ Health Care Inform* 2021;**28**:e100262. doi:10.1136/bmjhci-2020-100262

Received 19 October 2020
Revised 03 January 2021
Accepted 12 January 2021

[1]Patient Safety Translational Research Centre, Imperial College of Science Technology and Medicine, London, UK
[2]Big Data and Analytical Unit, Imperial College of Science Technology and Medicine, London, UK
[3]Institute of Global Health Innovation, Imperial College of Science Technology and Medicine, London, UK

Correspondence to
Mustafa Khanbhai;
m.khanbhai@imperial.ac.uk

ABSTRACT

Objectives Unstructured free-text patient feedback contains rich information, and analysing these data manually would require a lot of personnel resources which are not available in most healthcare organisations. To undertake a systematic review of the literature on the use of natural language processing (NLP) and machine learning (ML) to process and analyse free-text patient experience data.

Methods Databases were systematically searched to identify articles published between January 2000 and December 2019 examining NLP to analyse free-text patient feedback. Due to the heterogeneous nature of the studies, a narrative synthesis was deemed most appropriate. Data related to the study purpose, corpus, methodology, performance metrics and indicators of quality were recorded.

Results Nineteen articles were included. The majority (80%) of studies applied language analysis techniques on patient feedback from social media sites (unsolicited) followed by structured surveys (solicited). Supervised learning was frequently used (n=9), followed by unsupervised (n=6) and semisupervised (n=3). Comments extracted from social media were analysed using an unsupervised approach, and free-text comments held within structured surveys were analysed using a supervised approach. Reported performance metrics included the precision, recall and F-measure, with support vector machine and Naïve Bayes being the best performing ML classifiers.

Conclusion NLP and ML have emerged as an important tool for processing unstructured free text. Both supervised and unsupervised approaches have their role depending on the data source. With the advancement of data analysis tools, these techniques may be useful to healthcare organisations to generate insight from the volumes of unstructured free-text data.

Summary

What is already known?

- The ability to analyse and interpret free-text patient experience feedback falls short due to the resource intensity required to manually extract crucial information.
- A semiautomated process to rapidly identify and categorise comments from free-text responses may overcome some of the barriers encountered, and this has proven successful in other industries.

What does this paper add?

- Natural language processing and machine learning (ML) have emerged as an important tool for processing unstructured free text from patient experience feedback.
- Comments extracted from social media were commonly analysed using an unsupervised approach, and free-text comments held within structured surveys were analysed using a supervised approach.
- Healthcare organisations can use the various ML approaches depending on the source of patient experience free-text data, that is, solicited or unsolicited (social media), to gain near real-time insight into patient experience.

BACKGROUND

Over the last decade, there has been a renewed effort focusing on patient experiences, demonstrating the importance of integrating patients' perceptions and needs into care delivery.[1 2] As healthcare providers continue to become patient-centric, it is essential that stakeholders are able to measure, report and improve experience of patients under their care. Policy discourse has progressed from being curious about patients' feedback, to actually collecting and using the output to drive quality improvement (QI).

In the English National Health Service (NHS), USA and many European health systems patient experience data are abundant and publicly available.[3 4] NHS England commissions the Friends and Family Test (FFT), a continuous improvement tool allowing patients and people who use NHS services to feedback on their experience.[5] It asks users to rate services, or experiences, on a numerical scale such as the Likert scale. In addition to quantitative metrics, experience surveys such as the FFT also include qualitative data in the form of patient narratives. Evidence suggests that when staff are presented with both patient narratives and quantitative data, they tend to pay more attention to the narratives.[6] Patient narratives can even complement quantitative data by providing information on experiences not covered by quantitative data,[7 8] and give more detail that may help contextualise responses to structured questions. These free-text comments can be especially valuable if they are reported and analysed with the same scientific rigour already accorded to closed questions.[9 10] However, this process is limited by human resource and the lack of a systematic way to extract the useful insights from patient free-text comments to facilitate QI.[11 12]

Natural language processing (NLP) and machine learning (ML)

A potential solution to mitigate the resource constraints of qualitative analysis is NLP. NLP is currently the most widely used 'big data' analytical technique in healthcare,[13] and is defined as 'any computer-based algorithm that handles, augments and transforms natural language so that it can be represented for computation.'[14] NLP is used to extract information (ie, convert unstructured text into a structured form), perform syntactic processing (eg, tokenisation), capture meaning (ie, ascribe a concept to a word or group of words) and identify relationships (ie, ascribe relationships between concepts) from natural language free text through the use of defined language rules and relevant domain knowledge.[14–16] With regards to text analytics, the term ML refers to the application of a combination of statistical techniques in the form of algorithms that are able to complete diverse computation tasks,[17] including detect patterns including sentiment, entities, parts of speech and other phenomena within a text.[18]

Text analysis

Topic or text analysis is a method used to analyse large quantities of unstructured data, and the output reveals the main topics of each text.[19 20] ML enables topic analysis through automation using various algorithms, which largely falls under two main approaches, supervised and unsupervised.[21] The difference between these two main classes is the existence of labels in the training data subset.[22] Supervised ML involves predetermined output attribute besides the use of input attributes.[23] The algorithms attempt to predict and classify the predetermined attribute, and their accuracies and misclassification alongside other performance measures are dependent on the counts of the predetermined attribute correctly predicted or classified or otherwise.[22] In healthcare, Doing-Harris *et al*[24] identified the most common topics in free-text patient comments collected by healthcare services by designing automatic topic classifiers using a supervised approach. Conversely, unsupervised learning involves pattern recognition without the involvement of a target attribute.[22] Unsupervised algorithms identify inherent groupings within the unlabelled data and subsequently assign label to each data value.[25] Topics within a text can be detected using topic analysis models, simply by counting words and grouping similar words. Besides discovering the most frequently discussed topics in a given narrative, a topic model can be used to generate new insights within the free text.[26] Other studies have scraped patient experience data within comments from social media to detect topics using an unsupervised approach.[27 28]

Sentiment analysis

Sentiment analysis, also known as opinion mining, helps determine the emotive context within free-text data.[29 30] Sentiment analysis looks at users' expressions and in turn associates emotions within the analysed comments.[31] In patient feedback, it uses patterns among words to classify a comment into a complaint, or praise. This automated process benefits healthcare organisations by providing quick results when compared with a manual approach and is mostly free of human bias, however, reliability depends on the method used.[27 32 33] Studies have measured the sentiment of comments on the main NHS (NHS choices) over a 2-year period.[27 34] They found a strong agreement between the quantitative online rating of healthcare providers and analysis of sentiment using their individual automated approach.

NLP and patient experience feedback

Patient experience is mostly in natural language and in narrative free text. Most healthcare organisations hold large datasets pertaining to patient experience. In the Englanish NHS almost 30 million pieces of feedback have been collected, and the total rises by over a million a month, which according to NHS England is the 'biggest source of patient opinion in the world'.[5] Analysing these data manually would require a lot of personnel resources which are not available in most healthcare organisations.[5 35] Patient narratives contain multiple sentiments and may be about more than one care aspect; therefore, it is a challenge to extract information from such comments.[36] The advent of NLP and ML makes it far more feasible to analyse these data and can provide useful insights and complement structured data from surveys and other quality indicators.[37 38]

Outside of a healthcare organisation, there is an abundance of patient feedback on social media platforms such as Facebook, Twitter, and in the UK, NHS Choices and Care Opinion and other patient networks. This type of feedback gives information on non-traditional metrics, highlighting what patients truly value in their experiences by offering nuances that is often lacking in structured surveys.[39] Sentiment analysis has been applied ad hoc to online sources, such as blogs and social media[7 27 33 34] demonstrating in principle the utility of sentiment analysis for patient experience. There appears to be an appetite to explore the possibilities offered by NLP and ML within healthcare organisations to turn patient experience data into insight that can drive care delivery.[40 41] However, healthcare services need to be cognizant of what

NLP methodology to use depending on the source of patient experience feedback.[5] To date, no systematic review related to the automated extraction of information from patient experience feedback using NLP has been published. In this paper, we sought to review the body of literature and report the state of the science on the use of NLP and ML to process and analyse information from patient experience free-text feedback.

The aim of this study is to systematically review the literature on the use of NLP and ML to process and analyse free-text patient experience data. The objectives were to describe: (1) purpose and data source; (2) information (patient experience theme) extraction and sentiment analysis; (3) NLP methodology and performance metrics and (4) assess the studies for indicators of quality.

METHODS

Search strategy

The following databases were searched from January 2000 and December 2019; MEDLINE, EMBASE, PsycINFO, The Cochrane Library (Cochrane Database of Systematic Reviews, Cochrane Central Register of Controlled Trials, Cochrane Methodology Register), Global Health, Health Management Information Consortium, CINAHL and Web of Science. Grey literature and Google Scholar were used to extract articles that were not retrieved in the databases searched. Owing to the diversity of terms used inferring patient experience, combinations of search terms were used. The search terms, derived from the Medical Subject Headings vocabulary (US National Library of Medicine) for the database queries that were used can be found below. A review of the protocol was not published.

"natural language processing" OR "NLP" OR "text mining" OR "sentiment analysis" OR "opinion mining" OR "text classification" OR "document classification" OR "topic modelling" OR "machine learning" "supervised machine learning" OR "unsupervised machine learning" AND "feedback" OR "surveys and questionnaires" OR "data collection" OR "health care surveys" OR "assessment" OR "evaluation" AND "patient centred care" OR "patient satisfaction" OR "patient experience".

Inclusion criteria

To be eligible for inclusion in the review, the primary requirement was that the article needed to focus on the description, evaluation or use of NLP algorithm or pipeline to process or analyse patient experience data. The review included randomised controlled trials, non-randomised controlled trials, case–control studies, prospective and retrospective cohort studies and qualitative studies. Queries were limited to English language but not date constraints. We excluded studies that gathered patient-reported outcome measurements, symptom monitoring, symptom information, quality of life measures and ecological momentary assessment without patient experience data. Conference abstracts were excluded, as there was limited detail in the methodology to score against quality indicators.

Study selection

The research adhered to the guideline presented in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2009 checklist.[42] The initial search returned 1007 papers; after removing duplicates 241 papers were retained. The titles and abstract were screened by two reviewers (MK and PA) independently, and discrepancies were resolved by a third reviewer (EM). Thirty-one articles were identified as potentially eligible for inclusion. Full-text articles were retrieved and assessed for inclusion by the same reviewers, of which 19 were retained for final inclusion. The main reason for exclusion was the articles reported other patient-reported feedback and not patient experience. Figure 1 illustrates the PRISMA flowchart representing the study selection process and reasons for exclusion.

Data collection process

We developed a data collection tool with the following data fields: department of corresponding authors, country of study, study purpose, data source, solicited feedback, time period, information extraction method, data processing, ML classifiers, text analysis approach, software, performance, key findings and limitations. Two reviewers (MK and PA) independently completed the data collection, and met to compare the results, and discrepancies were resolved by a third reviewer (EM).

Data synthesis

Due to the heterogeneous nature of the studies, a narrative synthesis was deemed most appropriate. A formal quality assessment was not conducted, as relevant reporting standards have not been established for NLP articles. Instead, we report indicators of quality guided by elements reported in previous NLP-focused systematic reviews.[43–46] We included information related to the study purpose, corpus (eg, data source and number of comments), NLP (eg, methodology and software used and performance metrics). Two reviewers (MK and PA) independently evaluated indicators of quality in each study, disagreements in evaluation were resolved by discussion with a third reviewer (EM). Inter-rater agreement Cohen's kappa was calculated. In the reviewed studies, we assessed the NLP methodology and the rationale for its use. The key NLP approaches were summarised based on text analysis incorporating either text classification or topic modelling depending on the corpus available and evaluation was done as to whether sentiment analysis was performed using existing or bespoke software.

Performance metrics

To understand how well an automated ML algorithm performs, there are a number of statistical values that help determine its performance with the given data.[18] Algorithm performance is measured as recall (proportion of all true positive observations that are correct,

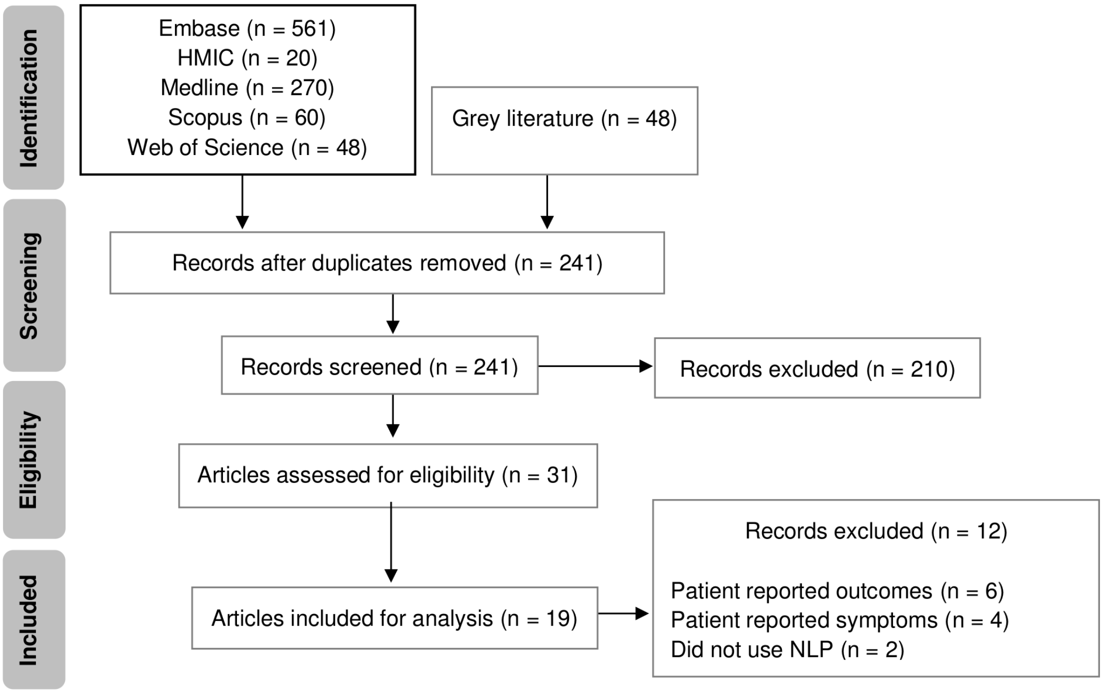
Identification
Screening
Eligibility
Included
Embase (n = 561)
HMIC (n = 20)
Medline (n = 270)
Scopus (n = 60)
Web of Science (n = 48)
Grey literature (n = 48)
Records after duplicates removed (n = 241)
Records screened (n = 241)
Records excluded (n = 210)
Articles assessed for eligibility (n = 31)
Records excluded (n = 12)
Patient reported outcomes (n = 6)
Patient reported symptoms (n = 4)
Did not use NLP (n = 2)
Articles included for analysis (n = 19)

Figure 1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2009 flowchart. NLP, natural language processing.

that is, true positives/(true positives+false negatives)), precision (ratio of correctly predicted positive observations to the total predicted positive observations) and by the F-score which describes overall performance, representing the harmonic mean of precision and recall.[43] K-fold cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it. This ensures that the results are not by chance, and therefore ensures the validity of the algorithms performance. We look all the recorded performance metrics in each of the included studies in order to gain a better understanding of how the data and ML approach can influence the performance.

RESULTS

Study characteristics

Year of publication ranged from 2012 to 2020 with almost 80% (15/19) of articles published in the last 5 years. The study purpose of the 19 articles was similar, in that they applied language analysis techniques on free-text patient experience feedback to extract information, which included themes or topics and sentiment. The feedback was either solicited[24 47–50] or unsolicited.[6 26–28 32 34 51–58] Six studies were from the UK,[26–28 48 49 55] two from Spain,[58] of which one included Dutch reviews[54] and the rest were conducted in the USA,[6 24 32 34 47 50 52 53 56 57] of which one[51] looked at Chinese language reviews translated in English. The authors of all except one study[47] were from a healthcare informatics department.

Data source

The majority (15/19) of the feedback used for language analysis was extracted from social media sites, such as Twitter,[28 52] Facebook[6] and healthcare specific forums, for example, NHS Choices,[26 27 55] Yelp,[56 57] RateMDs,[32 34 53] Haodf,[51] Masquemedicos,[54 58] Zorgkaart Nederland.[54] RateMDs and Yelp are US platforms that provide information, reviews and ratings on everything from cleanliness of hospital and care centre facilities, to clinician knowledge, as well as giving patients the ability to share personal experiences of care. NHS Choices is a UK-based platform that allows patients, carers and friends to comment on their experience of care received in any NHS institution. Haodf, Masquemedicos and Zorgkaart Nederland are platforms that incorporate patient experiences in Chinese, Spanish and Dutch, respectively. Five studies used the accompanying free text from structured patient feedback surveys; Press Ganey,[24 50] vendor supplied (HCAHPS and comments),[47] bespoke cancer experience survey with free-text comments,[48] Cancer Patient Experience Survey.[49] The initial dataset in terms of number of reviews captured to perform language analysis varied significantly from 734 reviews[58] to 773 279 reviews.[51] Where provided, the number of words, characters or sentences within the reviews varied. Table 1 gives an overview of the length of comments provided as either range, mean or median.

Software

The most common coding environment, sometimes used in combination, was Python (n=5)[24 49 50 52 53] followed by R (n=3),[26 48 55] Waikato Environment for Knowledge

Table 1 The length of comments provided in five of the 19 studies, arranged in descending order according to the total number of comments

Author	Data source	No. of comments	Length of comments
Hao *et al*[51]	Haodf	773 279	Mean 95.75 characters
Rastegar-Mojarad *et al*[57]	Yelp	79 173	Median 635 characters
Wallace *et al*[32]	RateMDs	58 100	Median 41 words
Wagland *et al*[48]	Cancer survey	5636	1–225 words
Plaza-del-Arco *et al*[58]	Masquemedicos	734	Mean 44 words

Analysis (n=2),[27, 34] Machine Learning for Language Toolkit (n=2),[53, 56] RapidMiner (n=2),[6, 58] and C++ (n=1).[54]

Language analysis approach

Studies used a variety of approaches to develop their language analysis methodology. The two most common approaches were supervised (n=9)[6, 27, 28, 34, 47, 48, 50, 52, 54] and unsupervised learning (n=6),[24, 26, 51, 53, 55, 56] followed by a combination, that is, (semisupervised) (n=3),[32, 57, 58] rule-based (n=1)[49] and dictionary look-up (n=1)[54] (figure 2). Sentiment analysis with a combination of text analysis was performed in 10 studies,[24, 26, 28, 32, 47–49, 52, 53, 57] sentiment analysis alone was performed in four[6, 28, 50, 54] and text analysis alone in four studies.[51, 55, 56, 58] We describe the details of the two approaches, sentiment analysis and text analysis, which incorporated text classification and topic modelling, categorised as supervised and unsupervised learning, respectively.

Supervised learning

Manual classification into topics or sentiment was performed in those studies that used a supervised approach. The most common approach was manual classification of a subset of comments as the training set. The percentage of total number of comments used for manual classification varied in each study, as did the number of raters. Sentiment was generally expressed as positive, negative and neutral. Five studies did not perform manual classification and employed existing software to perform the sentiment analysis, that is, TheySayLtd,[28] TextBlob,[52] SentiWordNet,[57] DICTION,[53] Keras.[50] We split the supervised approach based on sentiment analysis (table 2A) and text classification (table 2B), where we document the percentage of total comments manually classified into categories for sentiment and topics for text classification, the number of raters including the inter-rater agreement and the classifier(s) used for ML. In addition, where reported, we also highlight the configuration employed during the data processing steps. Support vector machine (SVM) was the most commonly used classifier (n=6) followed by Naïve Bayes (NB) (n=5).

Unsupervised learning

Topic modelling is an approach that automatically detects topics within a given comment. Seven studies[24, 26, 32, 51, 53, 55, 56] used this approach and majority of the studies (n=6)[24, 26, 51, 53, 55, 56] used latent Dirichlet allocation (LDA). One study[32] used a variation, factorial LDA, however this was a semisupervised approach as it involved some manual coding. LDA is a generative model of text that assumes words in a document reflect a mixture of latent topics (each word is associated with a single topic). For the output to be understandable, the number of topics has to be chosen, and table 3 demonstrates the variation in topics determined while employing LDA.

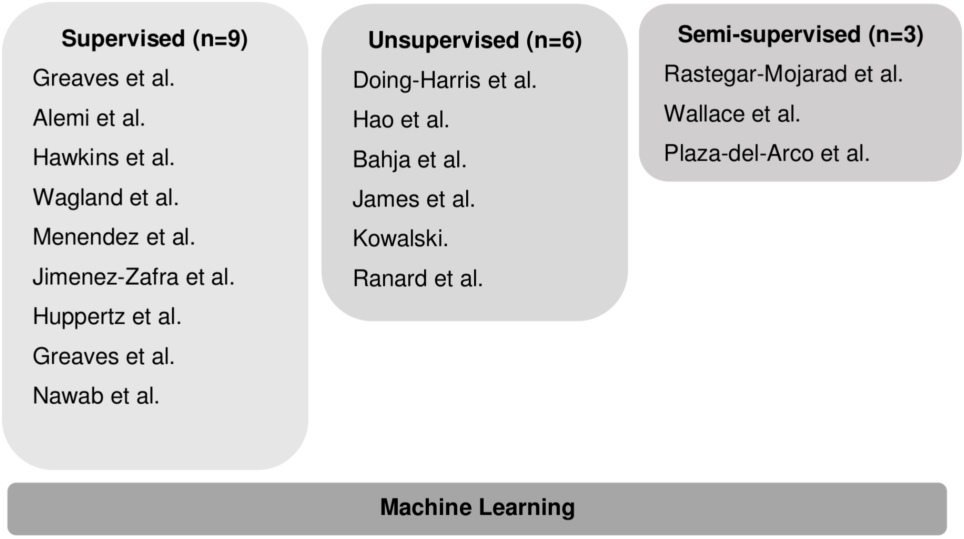

Figure 2 Most common approaches used to analyse free-text patient experience data identified in the systematic review.

Table 2A Studies that performed sentiment analysis using supervised approach, including the number of raters and associated inter-rater agreement expressed as Cohen's kappa (κ), classifiers and configuration applied where reported. Studies are reported in chronological order

					Sentiment categories				Classifier							
Author	Data source	Comments classified	No. of raters	κ	Positive	Negative	Mixed	Neutral	SVM	NB	DT	B	RF	GL	KN	Configuration
Alemi *et al*[34]	RateMDs	100%* (n=955)	NR	NR	✓	✓	✓	✓	✓	✓	✓	✓				Sparsity rule Information gain SVM: RBF kernel
Greaves *et al*[7]	NHS choices	17.56%† (1000/5695)	2	0.76	✓	✓	✓	✓	✓	✓	✓	✓				Prior polarity Information gain SVM: RBF kernel
Wagland *et al*[48]	Cancer experience	14.19% (800/5634)	3	0.64–0.87	✓	✓			✓			✓	✓	✓		NR
Bahja *et al*[26]	NHS choices	75% (56 818/76 151)	N/A	N/A	✓	✓			✓	✓						Sparsity rule Ratings in binary sentiment
Jimenez-Zafra *et al*[54]	COPOS and COPOD‡	100% (n=156 975 COPOD and n=743 COPOS)	N/A	N/A	✓	✓			✓							Ratings in binary sentiment SVM: linear kernel
Huppertz *et al*[6]	Facebook	0.88% (508/57 986)	3	NR	✓	✓			✓	✓					✓	NR
Doing-Harris *et al*[24]	Press Ganey	0.58% (300/51 235)	3	0.73	✓	✓				✓						NR
Menendez *et al*[47]	Vendor supplied	100% (132/132)	NR	NR	✓	✓	✓	✓								NR

*Classified as praise (positive), complaint (negative), praise and complaint (mixed), neither (neutral).
†Only n-grams classified.
‡Also used dictionary lookup and cross domain method.
B, bagging; COPOD, corpus of patient opinions in Dutch; COPOS, corpus of patient opinions in Spanish; DT, decision trees; GL, generalised linear model; KN, k-nearest neighbour; NB, Naïve Bayes; NR, not reported; RBF, radial basis function; RF, random forest; SVM, support vector machine.

Table 2B Studies that performed text classification using supervised approach, including the number of rater and associated inter-rater agreement expressed as Cohen's kappa (κ), classifiers and configuration applied where reported. Studies are reported in chronological order

Author	Data source	Comments classified	No. of raters	κ	No. of themes	Classifier							Configuration
						SVM	NB	DT	B	RF	GL	KN	
Alemi *et al*[10 34]	RateMDs	100% (n=955)	NR	NR	9	✓	✓	✓	✓				Sparsity rule SVM: RBF kernel
Greaves *et al*[7]	NHS choices	*17.56% (1000/5695)	2	0.76	3	✓	✓	✓	✓				Prior polarity Information gain SVM: RBF kernel
Wagland *et al*[48]	Cancer experience	14.19% (800/5634)	3	0.64–0.87	11	✓			✓	✓	✓		NR
Doing-Harris *et al*[24]	Press Ganey	0.58% (300/51 235)	3	0.73	7		✓						NR
Hawkins *et al*[52]	Twitter	7511/11 602†	AMT	0.18–0.52	10	✓	✓						NR

*Only n-grams classified.

†Tweets classified as pertaining to patient experience only.

AMT, Amazon Mechanical Turk; B, bagging; DT, decision trees; GL, generalised linear model; KN, k-nearest neighbour; NB, Naive Bayes; NR, not reported; RBF, radial basis function; RF, random forest; SVM, support vector machine.

Table 3 The number of topics arranged in descending determined in each study using latent Dirichlet allocation as a type of unsupervised learning approach

Author	Data source	No. of topics
Kowalksi	NHS choices	60
Ranard *et al*[56]	Yelp	50
Bahja *et al*[26]	NHS choices	30
Doing-Harris *et al*[24]	Press Ganey	30
Hao *et al*[51]	Haodf	10
James *et al*[53]	RateMDs	6

Performance

Seven studies did not report performance of the NLP algorithm or pipeline.[28 32 47 51 53 56 57] The remaining 12 studies reported one or more evaluation metrics such as accuracy, sensitivity, recall, specificity, precision, F-measure. The higher the F1 score the better, with 0 being the worst possible and one being the best. In the studies that employed a supervised approach, SVM and NB was the preferred classifier as it produced better results compared with other classifier demonstrated by the F1 score with sentiment analysis and text classification. Table 4 demonstrates the performance measure reported as F-measure or accuracy of the best performing classifiers for sentiment and text analysis using only supervised approach, and the k-fold cross-validation where reported in 12 studies, of which only five studies reported multiple fold validation.

Indicators of quality

Inter-rater agreement (Cohen's kappa) was calculated as 0.91 suggesting an almost perfect agreement. The individual evaluation with a description on each domain is detailed in table 5. Specifically, clarity of the study purpose statement, and presence of information related to the dataset, the number of comments analysed. information extraction and data processing, adequate description of NLP methodology and evaluation metrics. All studies had at least four of the seven quality indicators. Twelve studies addressed all seven indicators of quality,[6 24 26 27 34 48–50 52 54 55 58] and three studies addressed only four.[28 47 57]

DISCUSSION

In this systematic review, we identified 19 studies that evaluated various NLP and ML approaches to analyse free-text patient experience data. The majority of the studies dealt with documents written in English, perhaps because platforms for expressing emotions, opinions or comments related to health issues are mainly orientated towards Anglophones.[58] Three studies[51 54 58] were conducted using non-English free-text comments, however Hao *et al*[51] and Jimenez-Zafra *et al*[54] translated comments to English that were initially written in Chinese and Spanish, respectively. Accurate and automated analysis is challenging due to

Table 4 Performance metrics in the studies used supervised learning (sentiment analysis and text classification). SVM and NB were the preferred classifier as it produced better results demonstrated by the F1 score. Only five studies reported multiple fold validation

		Sentiment analysis		**Text classification**	
Author	**k-fold cross-validation**	**Classifier**	**Performance**	**Classifier**	**Performance**
Alemi *et al*[34]*†	Five repetitions of twofold cross-validation	SVM	Positive 0.89 Negative 0.64	SVM	Staff related 0.85 Doctor listens 0.34
		NB	Positive 0.94 Negative 0.68	NB	Staff related 0.80 Doctor listens 0.37
Doing-Harris *et al*[24]*	NR	NB	0.84	NB	Explanation 0.74 Friendliness 0.40
Greaves *et al*[27]	Single-fold cross-validation	NB SVM	0.89 0.84	NB SVM	Dignity and respect 0.85 Cleanliness 0.84 Dignity and respect 0.8 Cleanliness 0.84
Hawkins *et al*[52]	10-fold cross-validation	–	–	SVM	0.89‡
Jimenez-Zafra *et al*[54]	10-fold cross-validation	SVM	COPOD 0.86 COPOS 0.71	–	–
Huppertz *et al*[6]	NR	SVM	0.87‡	–	–
Wagland *et al*[48]	Single-fold cross-validation 10-fold cross-validation	SVM	0.80	–	–
		SVM	0.83	–	–
Bahja *et al*[26]	Single-fold cross-validation 4-fold cross-validation	SVM NB	0.84 0.78	–	–
		SVM NB	0.81 0.78	–	–

*Best and worst performing category, respectively.
†Classified as praise (positive), complaint (negative).
‡Reported as overall accuracy.
COPOD, corpus of patient opinions in Dutch; COPOS, corpus of patient opinions in Spanish; NB, Naïve Bayes; NR, not reported; SVM, support vector machine.

the subjectivity, complexity and creativity of the language used, and translating into other language may lose these subtleties. The type of patient feedback data used and choice of ML algorithm can affect the outcome of language analysis and classification. We show how studies used various ML approaches.

The two most common approaches were supervised and unsupervised learning for text and sentiment analysis. Briefly, text analysis identifies the topic mentioned within a given comment, whereas sentiment analysis identifies the emotion conveyed. Of the two approaches, the most common approach used was supervised learning, involving manual classification of a subset of data by themes[24 27 34 48 52] and sentiment.[6 24 26 27 34 48 52 54] Comprehensive reading of all comments within the dataset remains the 'gold standard' method for analysing free-text comments, and is currently the only way to ensure all relevant comments are coded and analysed.[48] This demonstrates that language analysis via an ML approach is only as good as the learning set that is used to inform it. The studies that used a supervised approach in this review demonstrated that there were at least two independent reviewers involved in manual coding, however, there was no consistency in the percentage of total comments coded, how the data was split into training and test set, and the k-fold cross-validation used. Within supervised learning, the most common classifier was SVM followed by NB. SVM and NB have been widely used for document classification, which consistently yield good classification performance.

NLP has problems processing noisy data, reducing overall accuracy.[18 59] Pre-processing of textual data is the first and an important step in processing of text that has been proven to improve performance of text classification models. The goal of pre-processing is to standardise the text.[59] We noted that pre-processing varied in the studies in this review. In addition to the standard pre-processing steps, that is, conversion to lowercase, stemming, stop word elimination, Alemi *et al*[34] used sparsity rule and information gain, Greaves *et al*[27] used information gain and prior polarity and Bahja *et al*[26] used sparsity rule alone. Plaza-del-Arco *et al*[58] used a combination of stopper and stemmer, and found that the accuracy was best (87.88%) with stemmer alone, however, F-measure was best (71.35%) when no stemmer or stopper was applied. However, despite these pre-processing steps, no

Table 5 Evaluation of studies and performance metrics

Author*	Defined purpose†	Data source described	Number of comments specified	Data processing described	Language analysis approach described	Evaluation metrics reported‡	Inclusion of comparative evaluation§
Alemi *et al*[34]	✓	✓	✓	✓	✓	✓	
Greaves *et al*[27]	✓	✓	✓	✓	✓	✓	✓
Greaves *et al*[28]	✓	✓	✓	✓			✓
Wallace *et al*[32]	✓	✓	✓	✓	✓		✓
Rastegar-Mojarad *et al*[57]	✓	✓	✓	✓			✓
Hawkins *et al*[52]	✓	✓	✓	✓	✓	✓	✓
Wagland *et al*[48]	✓	✓	✓	✓	✓	✓	
Hao *et al*[51]	✓	✓	✓	✓	✓		
James *et al*[53]	✓	✓	✓	✓	✓		
Bahja *et al*[26]	✓	✓	✓	✓	✓	✓	
Plaza-del-Arco *et al*[58]	✓	✓	✓	✓	✓	✓	
Doing-Harris *et al*[24]	✓	✓	✓	✓	✓	✓	
Huppertz *et al*[6]	✓	✓	✓	✓	✓	✓	
Kowalski	✓	✓	✓	✓	✓	✓	
Ranard *et al*[56]	✓	✓	✓	✓	✓		✓
Jimenez-Zafra *et al*[54]	✓	✓	✓	✓	✓	✓	
Menendez *et al*[47]	✓	✓	✓	✓			
Rivas *et al*[49]	✓	✓	✓	✓	✓	✓	
Nawab *et al*[50]	✓	✓	✓	✓	✓	✓	

✓ Indicates the presence of information in the article.
*Studies have been arranged in chronological order.
†Indicates reviewer judgement of clear statement of the study purpose.
‡Evaluation metrics include F-measure or accuracy.
§Comparison includes association with other survey data.

consensus could be found over a preferred supervised ML classification method to use for sentiment or text classification in the patient feedback domain.

The most interesting finding in this review was that the ML approach employed corresponded to the data source. The choice of approach is based on the performance metrics of the algorithm results, which depends on three factors.[21] First, identifying patterns is dependent on the quality of the data available. In text classification or sentiment analysis, the diversity of comments affects the accuracy of the machine prediction. More diversity decreases the ability of the ML algorithm to accurately classify the comment.[6] Second, each ML algorithm is governed by different sequential sets of rules for classifying semantic or syntactic relationships within the given text, and certain algorithms may suit some datasets better than others. Third, the larger the training sets used the higher the accuracy of the algorithms at identifying similar comments within the wider dataset, but trade-offs with time and human coding are necessary to ensure the method is resource-efficient.[21] We found that comments extracted from social media were commonly analysed using an unsupervised approach[26 32 51 53 55 56] and free-text comments held within structured surveys were analysed using a supervised approach.[6 27 28 34 47 48 50 52 54]

There is little evidence in the literature on the statistical properties for the minimum text size needed to perform language analysis, principally because of the difficulty of natural language understanding and the content and context of a text corpus.[6] The studies that reported text size demonstrate that the average character count was around 40 words. The domain of patient feedback from free-text complementing structured surveys appears fixed in its nature, making it attractive data for supervised learning.[31] Just as the domain is fixed, the perspective of a patient feedback document is also fixed[31]: there is limited vocabulary that is useful for commenting about health service, and therefore it is possible to anticipate the meaning of various phrases and automatically classify the comments.[34] Rastegar-Mojarad *et al*[57] also observed that a small (25%) vocabulary set covered a majority (92%) of the content of their patients comments, consistent with a study[60] exploring consumer health vocabulary used by consumers and healthcare professionals. This suggests that patients use certain vocabulary when expressing their experience within free-text comments.

The overall domain of patient feedback is the healthcare system,[31] and this study revealed the content of reviews tend to focus on a small collection of aspects associated with this as demonstrated by the topics used for text classification in the studies.[24 27 34 48 52] In contrast, the studies[26 32 51 53 55 56] that performed topic modelling, did so on the premise that patient feedback comments contain a multitude of different topics. Topic modelling can be useful in evaluating how close results come to what humans with domain knowledge have determined the topics to be, and if this unsupervised approach finds new topics not identified by humans.[49] LDA was used to extract a number of topics from the free-text reviews as they occur in the data without any prior assumption about what patients care about. The topics identified by six studies that used LDA did not generate any new topics, which is in keeping with the earlier finding that consumer healthcare reporting has limited vocabulary. This finding was supported by Doing-Harris *et al*,[24] who showed that their topic modelling results echo topic classification results, demonstrating that no unexpected topics were found in topic modelling.

Other factors should be taken into account when employing LDA. LDA is mainly based on frequency on co-occurrence of words under similar topics.[51] Topics discovered using LDA may not match the true topics in the data, and short documents, such as free-text comments, may result in poor performance of LDA.[49] In addition to the short comments, studies in this review also demonstrate that majority of the comments on social media tend to be positive, in contrast to the negative reviews which are longer but less frequent. Wagland *et al*[48] found that the content of positive comments was usually much less specific that for negative comments. Therefore an unsupervised approach to short positive reviews may not detect new topics, and the low frequency of negative reviews may not highlight new topics either. To mitigate this, there is a role of using a supervised approach to identify subcategories for negative reviews.[48]

Choice of the number of topics for LDA model also affects the quality of the output.[25 56] If topics are too few, their content gives insight into only very general patterns in the text which are not very useful. Too many topics, on the other hand, make it difficult to find common themes with numerous topics. An LDA topic model with an optimal number of topics should demonstrate meaningful patterns without producing many insignificant topics. The number of topics identified in the studies reviewed[26 32 51 53 55 56] was not consistent and ranged from 6 to 60, demonstrating that deciding on the optimal number is challenging. Performance of the LDA models is affected by semantic coherence (the rate at which topic's most common words tend to occur together in the same reviews) and exclusivity (the rate at which most common terms are exclusive to individual topics). Both measures are useful guidance of which model to choose,[55] however, of the six studies that used LDA, only one study[55] reported LDA performance measures.

Sentiment analysis was commonly conducted using a supervised approach (n=8).[6 24 26 27 34 47 48 54] Even though pre-classified, understanding what the comments both negative and positive are specifically talking about still requires reading through the comments. NLP makes this process efficient by identifying trends in the comment by sentiment. This review identified the most common approach to sentiment classification was to categorise the comment into a single category, that is, positive or negative. However, this implies that there must be polarity associated with a document, which is not always the case. This fails to capture the mixed sentiments or neutral sentiments which could provide useful insights into patient experience. Nawab *et al*[50] demonstrated that splitting the mixed sentiments by sentences revealed distinct sentiments. Therefore, although the percentage of mixed or neutral sentiment is low compared with the overall dataset, analysis of comments within these mixed and neutral sentiment can provide useful information and therefore should not be discarded.

Greaves *et al*[27] and Bahja *et al*[26] used the associated star rating within the NHS Choices data to directly train the sentiment tool. This approach is able to make use of the implicit notion that if a patient says they would recommend a hospital based on star rating, they are then implying a positive sentiment, and conversely if not a negative sentiment, therefore automatically extracting a nominal categorisation. This automated classification removes the need for manual classification and eliminates potential biases of reviewer assignment of comments, but it makes an assumption that star ratings correlate with the sentiment. This is supported by Kowalski,[55] who demonstrated intuitive relationships between topics' meanings and star rating across the analysed NHS Choices dataset. In contrast, Alemi *et al*[34] found that sentiment in comments from RateMDs are not reflected in the overall rating, for example 6% of the patients who gave highest overall rating still included a complaint in their comments, and 33% of patients who gave lowest overall rating included praise. This suggests that the sentiment may not always correlate with the star rating, and therefore researchers need to recognise that the approach used for classification may have implications on validity.

With regard to sentiment analysis of Twitter dataset, Greaves *et al*[28] found no associations when comparing Twitter data to conventional metrics such as patient experience, Hawkins *et al*[52] found no correlation between twitter sentiment and HCAHPS score, suggesting twitter sentiment must be treated cautiously in understanding quality. Therefore, although star ratings can be informative and in line quantitative measures of quality, they may not be sufficiently granular to help evaluate service quality based solely on the star rating without considering the textual content.[53]

Studies in this review demonstrate that NLP and ML have emerged as an important tool for processing patient experience unstructured free-text data and generating structured output. However, most of the work has been done on extracting information from social media.[6 26–28 32 34 51–58] Healthcare organisations have raised concerns about the accuracy or comments expressed on social media,[61] making policymakers reluctant to endorse narrative information as a legitimate tool. Even though most administrators remove malicious messages manually, anyone can comment on the website and intentionally distort how potential patients evaluate healthcare services. The validity and reliability of NLP is further limited by the fact that most patients do not post reviews online. Kowalski[55] found that healthcare services in England received fewer that 20 reviews over a period of three and a half years. For a limited amount of data, NLP may not be very expedient, and with a smaller number of comments the results may not be as fruitful and there may not be enough raw data to detect a specific pattern.[50] Furthermore, rating posted in social media reviews is not adjusted for user characteristics or medical risk, whereas structured survey scores are patient mix adjusted.[6]

Limitations

We focused on indicators of quality of the included articles rather than assessing the quality of the studies, as relevant formal standards have yet to be established for NLP articles. Due to the heterogenous nature of the studies, and various approaches taken with regard to pre-processing, manual classification and performance of classifiers, it is challenging to make any comparative statements.

CONCLUSION

Studies in this review demonstrate that NLP and ML have emerged as an important tool for processing unstructured free-text patient experience data. Both supervised and unsupervised approaches have their role in language analysis depending on the data source. Supervised learning is time consuming due to the manual coding required, and is beneficial in analysing free-text comments commonly found in structured surveys. As the volume of comments posted on social media continues to rise, manual classification for supervised learning may not be feasible due to time constraints and topic modelling may be a satisfactory approach. To ensure that every patients' voice is heard, healthcare organisations must react and mould their language analysis strategy in line with the various patient feedback platforms.

Acknowledgements We thank Jacqueline Cousins (Library Manager and Liaison Librarian at Imperial College London) for the support improving the composition of the search terms and procedural aspects of the search strategy.

Contributors MK, JS and EM contributed to conception and design of the work. MK and PA contributed to database searching. MK, PA and EM contributed to full-text screening. MK, PA, KF, JS and EM contributed to data analysis and interpretation.

Funding This work is supported by the National Institute for Health Research (NIHR) Imperial Patient Safety Translation Research Centre. Infrastructure support was provided by the NIHR Imperial Biomedical Research Centre.

Disclaimer The study funder(s) did not play a role in study design; in the collection, analysis and interpretation of data; in writing of the report; and in the decision to submit the article for publication. In addition, researchers were independent from funders, and all authors had full access to all the data included in this study and can take responsibility for the integrity of the data and accuracy of the data analysis.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data sharing not applicable as no datasets generated and/or analysed for this study.

ORCID iD
Mustafa Khanbhai http://orcid.org/0000-0002-4434-1785

REFERENCES

1 Darzi A. High quality care for all: NHS next stage review final report department of health, 2008. Available: www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_085825
2 Coulter AFR, Cornwell J. *Measures of patients' experience in hospital: purpose, methods and uses*. Kings Fund, 2009.
3 Coulter A. What do patients and the public want from primary care? *BMJ* 2005;331:1199–201.
4 Coulter A, Cleary PD. Patients' experiences with hospital care in five countries. *Health Aff* 2001;20:244–52.
5 NHS England. *The friends and family test*. Publication Gateway Ref, 2014.
6 Huppertz JW, Otto P. Predicting HCAHPS scores from hospitals' social media Pages: a sentiment analysis. *Health Care Manage Rev* 2018;43:359–67.
7 Greaves F, Ramirez-Cano D, Millett C, *et al*. Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. *BMJ Qual Saf* 2013;22:251–5.
8 López A, Detz A, Ratanawongsa N, *et al*. What patients say about their doctors online: a qualitative content analysis. *J Gen Intern Med* 2012;27:685–92.
9 Trigg L. Patients' opinions of health care providers for supporting choice and quality improvement. *J Health Serv Res Policy* 2011;16:102–7.
10 Cognetta-Rieke C, Guney S. Analytical insights from patient narratives: the next step for better patient experience. *J Patient Exp* 2014;1:20–2.
11 Robert G, Cornwell J. Rethinking policy approaches to measuring and improving patient experience. *J Health Serv Res Policy* 2013;18:67–9.
12 Ipsos-MORI. *Real time patient feedback: information patients need and value, research report prepared for West Midlands strategic health authority*, 2008.
13 Mehta N, Pandit A. Concurrence of big data analytics and healthcare: a systematic review. *Int J Med Inform* 2018;114:57–65.
14 Yim W-W, Yetisgen M, Harris WP, *et al*. Natural language processing in oncology: a review. *JAMA Oncol* 2016;2:797–804.
15 Fleuren WWM, Alkema W. Application of text mining in the biomedical domain. *Methods* 2015;74:97–106.
16 Wang Y, Wang L, Rastegar-Mojarad M, *et al*. Clinical information extraction applications: a literature review. *J Biomed Inform* 2018;77:34–49.
17 Gibbons C, Richards S, Valderas JM, *et al*. Supervised machine learning algorithms can classify Open-Text feedback of doctor performance with Human-Level accuracy. *J Med Internet Res* 2017;19:e65.
18 Chowdhury GG, Cronin B. Natural language processing. *Ann Rev Info Sci Tech* 2002;37:51–89.

19 Hotho A, Nurnberger A, Paass G. A brief survey of text mining. *Ldv Forum* 2005;20:19–62.

20 Feldman R, Sanger J. *The text mining Handbook: advanced approaches in analyzing unstructures data*. Cambridge University Press, 2007.

21 Collingwood L, Wilkerson J. Tradeoffs in accuracy and efficiency in supervised learning methods. *J Inf Technol* 2012;9:298–318.

22 Alloghani M, Al-Jumeily D, Mustafina J. A systematic review on supervised and unsupervised machine learning algorithms for data science. In: *Supervised and unsupervised learning for data science*. Springer, Cham, 2020.

23 Kotsiantis SB. Supervised machine learning: a review of classification techniques. *Informatica* 2007;31:249–68.

24 Doing-Harris K, Mowery DL, Daniels C, *et al*. Understanding patient satisfaction with received healthcare services: a natural language processing approach. *AMIA Annu Symp Proc* 2016;2016:524–33.

25 Blei DNA, Jordan M. Latent Dirichlet allocation. *J Mach Learn Res* 2003;3:993–1022.

26 Bahja MLM. Identifying patient experience from online resources via sentiment analysis and topic modelling. *Association for Computing Machinery* 2016;6.

27 Greaves F, Ramirez-Cano D, Millett C, *et al*. Use of sentiment analysis for capturing patient experience from free-text comments posted online. *J Med Internet Res* 2013;15:e239.

28 Greaves F, Laverty AA, Cano DR, *et al*. Tweets about Hospital quality: a mixed methods study. *BMJ Qual Saf* 2014;23:838–46.

29 Liu B. *Sentiment analysis and opinion mining*. San Rafael, California: Morgan & Claypool Publishers, 2012: 5. 1–167.

30 Pang B, Lee L. Opinion mining and sentiment analysis. *FNT in Information Retrieval* 2008;2:1–135.

31 Smith P. *Sentiment analysis of patient feedback*. University of Birmingham, 2015.

32 Wallace BC, Paul MJ, Sarkar U, *et al*. A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. *J Am Med Inform Assoc* 2014;21:1098–103.

33 Gohil S, Vuik S, Darzi A. Sentiment analysis of health care Tweets: review of the methods used. *JMIR Public Health Surveill* 2018;4:e43.

34 Alemi F, Torii M, Clementz L, *et al*. Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments. *Qual Manag Health Care* 2012;21:9–19.

35 Sheard L, Marsh C, O'Hara J, *et al*. The Patient Feedback Response Framework - Understanding why UK hospital staff find it difficult to make improvements based on patient feedback: A qualitative study. *Soc Sci Med* 2017;178:19–27.

36 The Power of Information. *Putting all of US in control of the health and care information we need*. London: Department of Health, 2012.

37 Griffiths A, Leaver MP. Wisdom of patients: predicting the quality of care using aggregated patient feedback. *BMJ Qual Saf* 2018;27:110–8.

38 Gibbons C, Greaves F. Lending a hand: could machine learning help hospital staff make better use of patient feedback? *BMJ Qual Saf* 2018;27:93–5.

39 Rozenblum R, Greaves F, Bates DW. The role of social media around patient experience and engagement. *BMJ Qual Saf* 2017;26:845–8.

40 Department of Health. *What matters: a guide to using patient feedback to transform services*, 2009.

41 Francis R. *Report of the mid Staffordshire NHS Foundation trust public inquiry*, 2013.

42 Moher D, Liberati A, Tetzlaff J, *et al*. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009;6:e1000097.

43 Pons E, Braun LMM, Hunink MGM, *et al*. Natural language processing in radiology: a systematic review. *Radiology* 2016;279:329–43.

44 Mishra R, Bian J, Fiszman M, *et al*. Text summarization in the biomedical domain: a systematic review of recent research. *J Biomed Inform* 2014;52:457–67.

45 Dreisbach C, Koleck TA, Bourne PE, *et al*. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *Int J Med Inform* 2019;125:37–46.

46 Koleck TA, Dreisbach C, Bourne PE, *et al*. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *J Am Med Inform Assoc* 2019;26:364–79.

47 Menendez ME, Shaker J, Lawler SM, *et al*. Negative Patient-Experience comments after total shoulder arthroplasty. *J Bone Joint Surg Am* 2019;101:330–7.

48 Wagland R, Recio-Saucedo A, Simon M, *et al*. Development and testing of a text-mining approach to analyse patients' comments on their experiences of colorectal cancer care. *BMJ Qual Saf* 2016;25:604–14.

49 Rivas C, Tkacz D, Antao L, *et al*. *Automated analysis of free-text comments and dashboard representations in patient experience surveys: a multimethod co-design study. health services and delivery research*. Southampton (UK), 2019.

50 Nawab K, Ramsey G, Schreiber R. Natural language processing to extract meaningful information from patient experience feedback. *Appl Clin Inform* 2020;11:242–52.

51 Hao H, Zhang K. The voice of Chinese health consumers: a text mining approach to web-based physician reviews. *J Med Internet Res* 2016;18:e108.

52 Hawkins JB, Brownstein JS, Tuli G, *et al*. Measuring patient-perceived quality of care in US hospitals using Twitter. *BMJ Qual Saf* 2016;25:404–13.

53 James TL, Villacis Calderon ED, Cook DF. Exploring patient perceptions of healthcare service quality through analysis of unstructured feedback. *Expert Syst Appl* 2017;71:479–92.

54 Jimenez-Zafra SM M-VM, Maks I, Izquierdo R. Analysis of patient satisfaction in Dutch and Spanish online reviews 2017;58:101–8.

55 Kowalski R. Patients' written reviews as a resource for public healthcare management in England. *Procedia Comput Sci* 2017;113:545–50.

56 Ranard BL, Werner RM, Antanavicius T, *et al*. Yelp reviews of hospital care can supplement and inform traditional surveys of the patient experience of care. *Health Aff* 2016;35:697–705.

57 Rastegar-Mojarad M, Ye Z, Wall D, *et al*. Collecting and analyzing patient experiences of health care from social media. *JMIR Res Protoc* 2015;4:e78.

58 Plaza-del-Arco F M-VT, Jimenez-Zafra SM, Molina-Gonzalez D. COPOS: corpus of patient opinions in Spanish. Application of sentiment analysis techniques. *Procesamiento del Lenguaje Natural* 2016;57:83–90.

59 Haddi E, Liu X, Shi Y, Xiaohui L, Yong S. The role of text pre-processing in sentiment analysis. *Procedia Comput Sci* 2013;17:26–32.

60 Zeng QT, Tse T. Exploring and developing consumer health vocabularies. *J Am Med Inform Assoc* 2006;13:24–9.

61 McCartney M. Will doctor rating sites improve the quality of care? no. *BMJ* 2009;338:b1033.