Australian Government
Department of Health
Therapeutic Goods Administration

Actual and potential harm caused by medical software

A rapid literature review of safety and performance issues

Version 1.0, July 2020

TGA Health Safety Regulation

Contents

1. Introduction

The advent of widespread software resources, particularly those accessible through smart phones and smartphone apps has led to public expectations that digital health and medical information will be easy to access via apps and that such information will also be reliable. In 2018, over a third of adults were self-diagnosing health conditions by going online, and by 2020 over half a million health apps were available in major app stores. This is despite there being no way for consumers (or their healthcare providers) to verify the safety, accuracy or effectiveness of this software.

At the same time, the increasing adoption of digital health technologies and the enhancement of traditional medical software used in or as medical devices has increased its complexity and usage. This includes the rapidly developing software specialisation of artificial intelligence. The current COVID-19 pandemic is also thought to have accelerated the implementation and adoption of digital health, making remote consultations and diagnosis much more frequent, so consumers, medical professionals and patients will particularly need medical apps to be reliable and safe.

2. Methods

A rapid search of published articles using MEDLINE and Pubmed was conducted by the TGA of studies and reviews that specifically reported on medical software safety and efficacy over the last seven years were identified. The search include both mobile apps and medical software used for purposes which included screening, diagnosis, management or monitoring and operating medical equipment. Other relevant papers were identified using Google Scholar.

Therefore while many of the studies cited relate specifically to challenges with software as a medical device, and software controlling medical devices, some additional reports of safety and performance issues with medical software have also been included for context.

3. Software role in device recalls

An analysis of medical device recalls by the TGA in the five years to April 2020 showed that software defects were one of the most common reasons for hospital or retail level medical device recalls. Over 20 % of all device recalls in that period were due to software faults – for example, this equated to 50 recalls in the six month period from 1 October 2019 to 1 April 2020. While the recall reports are largely from software problems when the software is integrated with a medical device they illustrate the significant and negative health impacts from faults that have required recalls.

Software faults required recalls in a number of devices that perform critical diagnostic and in patient functions. Some examples reported to the TGA include:

- Electrocardiographs: Either loss of data or improper signal analysis, incorrect patient assignment affecting clinical assessment of ECG arrhythmias
- Implanted cardiac device: loss of diagnostic information, unable to communicate remotely leading to battery depletion
- Patient monitors: Loss of data, incorrect display, inaccurate blood pressure readings, SpO2 values freezing, fail alarms not activated, inability to measure high pulse rates, cybersecurity vulnerabilities
- Incorrect drug dosage calculations: inaccurate patient height and weight calculations in bedside monitor

Data from medical device recall databases may significantly under-represent software errors since the information reported as part of the recall may be too brief to show the details behind the recall. Root cause analysis may also not have identified software as the source of error where it causes other components to fail. The effects of software errors can be subtle or difficult to detect even though they cause other components to fail. In many cases, patients may also be unaware how to report problems.

A study of U.S. recall data by Zhang et al. (2019) also reported 140 recalls per year in that country for user interface type errors. As this data reflects only one type of software-related recall this suggests that the total number of recalls for other types of software error is much higher. The user interface errors examined in this study occurred in devices used for medical imaging, picture archiving and communication systems, cardiovascular, radiation therapy, microbiology, clinical chemistry, general hospital applications and other unspecified applications.

An earlier US study (Ronquillo and Zuckermann, 2017) found that a total of 627 software devices (1.4 million units) were subject to recalls, with 12 of these devices (190,596 units) subject to the highestrisk recall level by the US FDA. Eleven of the devices recalled as high risk had entered the market through the FDA review process that at that stage did not require evidence of safety or effectiveness, and one device was completely exempt from regulatory review. The largest high-risk recall categories were anesthesiology and general hospital, with one each in cardiovascular and neurology. Five electronic medical record systems (9,347 units) were recalled for software defects classified as posing a moderate risk to patient safety. The defects included incorrect drug dosage calculations, treatment files, and medication order status, as well as failed allergy interaction warnings and the display of medical information for the wrong patients.

4. Software safety – summary of reviews and meta-analyses

4.1 Safety concerns

Several studies have highlighted apps that could compromise patient safety and are potentially dangerous. Lewis et al. (2014) identified a range of different kinds of risks that medical apps can contribute to and important contextual variables that can modify these risks. For example, certain apps designed for opioid dosage conversion or melanoma detection demonstrate dangerously poor accuracy, while a number of other medical apps do not follow evidence-based guidelines (Ferrero et al. 2013, Huffey et al. 2013, Wolff et al. 2013, Brierbrier et al. 2014).

One issue highlighted by these studies is that many app developers have little or no formal medical training and do not involve clinicians in the development process and may therefore be unaware of patient safety issues raised by inappropriate app content or functioning (Hamilton and Brady 2012, Huckvale 2012, Rodrigues 2013).

Buechi et al. (2017) assessed published studies on the diagnostic value of health apps using inbuilt smartphone sensors, several of which were melanoma screening apps. Quality assessment revealed a high risk of bias in all studies, some concerns with specificity and sensitivity and rather limited diagnostic evidence of available health apps on Apple's and Google's app stores. This contributed to serious safety concerns in particular through missing malignant cancers.

Van Velthoven et al. 2018, in their review concluded that apps have great potential to improve the quality of care and reduce costs, but this has not yet been achieved. They found many low-quality, unsafe health apps, resulting in different types of risks. Of the 74 studies identified by

Akbar et al (2020), nearly half related to disease management. A total of 80 safety concerns were identified, 67 related to the quality of information presented including incorrect or incomplete information, variation in content, and incorrect or inappropriate response to consumer needs. The remaining 13 related to app functionality including gaps in features, lack of validation for user input, delayed processing, failure to respond to health dangers, and faulty alarms.

4.2 Limited clinical validation

Clinical validation to demonstrate efficacy is still uncommon for digital health products (Powell et al. 2014). Mathews et al. (2019) emphasised that regardless of what part of the continuum of care (prevention, detection, or management) the product addresses, a validation study must compare it to relevant clinical gold standards. Particularly for studies aiming to demonstrate the clinical impact of the product, these may take the form of accepted care quality metrics, such as measures of clinical outcomes (e.g. presence of disease or complications, clinical functional scores) and process (e.g. laboratory values, adherence to treatment guidelines).

Grundy et al. (2016) undertook a systematic review on methods for searching, data extraction, and analysis of mobile health apps and found that many studies of app content are largely descriptive, and most use surrogate and one-dimensional outcome measures for app content quality. Few studies have been based on clinical data.

Gordon et al. (2020) discuss the importance of validation, and how to ensure apps that are used in clinical practice have been appropriately validated. While some apps have been rigorously studied, there is a general dearth of evidence for health apps, both because a small percentage have been studied, and evidence tends to be low quality for those that have been studied. (Mathews et al. 2019) (Larsen et al. 2019, Byambasuren et al. 2018). While consumer ratings of health apps might seem to be a scalable methodology of validating apps, prior work suggests that consumer ratings poorly reflect clinical utility and usability (Singh et al. 2016).

4.3 Error characteristics

The technical categorisation of software errors in the literature revealed a diverse range of types of problems. These included functional errors for example, where incorrect values or data were represented, incorrect outputs were inferred or calculated, or the software did not work as described. Wyatt (2018) found that four out of five melanoma screening apps performed so poorly that they could pose a public health hazard by falsely reassuring users about a suspicious mole. The failure of software to validate data entered by users was significant contributors to problems in calculations (Akbar et al. 2020).

Resilience and reliability errors included crashing, freezing or intermittent function, or instances where the software did not respond or perform in the expected time. Alarms failing to function or operating at an incorrect time were another example in this category. Usability or accessibility problems, including navigation, flow, inconsistent formats, have led to incorrect usage especially by population groups who may be disadvantaged, have special needs or not be digitally literate. Bonoto et al. (2017) showed considerable variation in efficacy between differing populations groups segmented by education, age and other factors.

Data errors included lack of or incorrect validation of user entry, data integrity, quality and inconsistency. Privacy and security problems were identified with harvesting of identifiable patient data, lack of encryption, trackers, user surveillance and hacking.

4.4 Quality management challenges

Several studies (Boulos et al. 2014, McMillan et al. 2015) have reported concern among health care professionals about the quality of apps for patient or professional use, how patients use apps and whether they reveal this use in a consultation. Lewis et al. (2014) described how app risk is associated with both app complexity and its functions. They point out that risk is related to the context of app use, including the user's knowledge and the clinical setting. Paradoxically, this risk may be higher in community settings rather than in clinical settings such as intensive care units, where patients are constantly monitored and a "crash team" is on hand.

Buechi et al. (2017) reviewed methodological quality for diagnostic quality of apps, and identified possible bias due to patient selection, conduct or interpretation of the test, patient flow and timing in all of the eleven studies surveyed that investigated diagnostic quality. They also found that bias due to the use of chosen reference standards was possible in six of the eleven studies reviewed. In another study by Zhang et al. (2019), the authors showed that many apps had disappeared or been removed from sale during the study period, making it complex to obtain or examine consistent data on errors or broader safety concerns.

A number of recent studies have specifically focussed on mobile apps. A key difficulty in measuring and tracking errors is the rapidity of evolution of software development; consequently, the life cycle of an individual app can be short. The low barriers to entry for software developers now mean that products be brought to market quickly by vendors without sufficient resources to conduct quality management, or support an error remediation service. Huckvale et al. (2015) found that over a quarter of apps in their review, including those from "reputable" providers were withdrawn during the period of their study. Lewis and Wyatt (2014) also found that there is little reporting of faulty apps or clinical incidents associated with app use.

5. Issues with specific types of products

5.1 Symptom checkers and diagnostic apps

MIllenson et al. (2017) conducted a scoping review of the peer-reviewed and grey literature from 2014 to 2017. The greatest number of articles reviewed focused on dermatology-related diagnostic apps. Other diagnostic areas included infectious diseases, sexually transmitted infections, mental health issues, neurology, oncology, orthopaedics, knee pain, hand surgery, eye and vision issues, otolaryngology, rheumatology and urology. The most common clinical areas evaluated were dermatology (particularly malignancy) and general diagnostic and triage advice for a range of conditions. Only about half of the so called evaluations examined actual performance.

Apps were found to vary widely in functionality, accuracy, safety and effectiveness, and they concluded that *"The current evidence base on direct to consumer, interactive diagnostic apps is sparse in scope, uneven in the information provided and inconclusive with respect to safety and effectiveness, with no studies of clinical risks and benefits involving real-world consumer use. Given that diagnostic apps are rapidly evolving, rigorous and standardized evaluations are essential to inform decisions"*.

Peer-reviewed results of general symptom checkers for particular diseases showed few favourable results. In one study, the diagnosis suggested by a symptom checker matched a final diagnosis related to hand surgery just 33% of the time (Hageman et al. 2015), while in another, a symptom checker provided "frequently inaccurate" advice related to inflammatory joint disease (Poweley et al. 2016). An app for knee pain diagnoses had an accuracy rate of 58% (Bisson et al. 2016). Apps to screen for Alzheimer's disease were all rated "poor to very poor", and the authors noted that one tested app even concluded the user had the condition no matter what data were entered (Robillard et al. 2015).

However, when specialty symptom checkers used data directly entered from sensors, they sometimes showed more promise, albeit with significant variability in the findings. For example, while one study warned of substantial potential for patient harm from a dermatology app's misleading results (Ferrero et al. 2013), another study of that same app using a different methodology two years later found an accuracy rate of 81% in detecting melanoma, a sensitivity of 73% and a specificity of 39.3% (Maier et al. 2015). Some study designs raised questions of evaluator bias against the interactive apps. Among the criticisms were whether a particular evaluation overweighed relatively rare diagnoses (Fraser et al. 2017).

Actual diagnostic performance varied widely. A study of 23 general symptom checkers by Semigran et al. (2015) found an 80% rate of appropriate triage advice in emergent cases, but just 33% for appropriate self-care suggestions. An audit of 23 symptom checker apps in a US clinic by Rowland et al. (2020) found that for emergency cases, diagnostic accuracy was only 34% and triage advice was considered appropriate in only 55% of non-emergent cases. Diagnostic symptom checker apps targeted at specific symptoms such as hand or knee pain have been shown to frequently deliver inaccurate advice. Published evidence suggests that symptom checker apps are often risk averse and may lead to unnecessary consultations in the non-emergency setting.

5.2 Diabetes management software

Six out of 13 apps for diabetes management reviewed by Bonoto et al. (2017) provided a statistical significant reduction (P<.05) in randomised controlled trials of glycated haemoglobin (HbA1c) at the end of studies in the intervention group. It was concluded that use of these six apps by diabetic patients could help improve the control of HbA1c and strengthen the perception of self-care. However a recent review by Fleming et al (2020) confirmed that very few diabetes apps have been assessed for clinical or diagnostic performance. Although there are almost half a million mobile health apps available for download, there are far fewer randomized controlled trials, case-control studies, and cohort studies that have evaluated whether app-based interventions improve health.

Drincic et al. (2016) reviewed several mobile diabetes apps available in the U.S. or EU. They found only 14 apps with clinical outcomes data published in peer-reviewed literature or that have been approved by the FDA in the U.S. or received a CE mark in Europe. Those particular apps were found to positively affect outcomes, such as HbA1c, hypoglycemia incidence, and diabetes self-care measures, in the short term. Fu et al (2017) reviewed seven usability studies and 10 clinical effectiveness studies of diabetes mobile apps. Usability problem ratings ranged from moderate to catastrophic. Top usability problems are multi-steps task, limited functionality and interaction, and difficult system navigation. More recently, a 2018 comprehensive study for the U.S. Agency for Healthcare Research and Quality found only 11 RCTs (clinical vs. control) reporting health outcomes among the hundreds of commercial apps for diabetes self-management. Of these 11 RCTs, only five were associated with clinically significant but small improvements in HbA1Ac. However none of the studies demonstrated improvements in quality of life, blood pressure, weight, or BMI. None of the studies were considered to be high quality.

A joint study by the European Association for the Study of Diabetes and the American Diabetes Association in 2020 (Fleming et al 2020) found inadequate evidence for most on accuracy and clinical validity. The authors observed that there are few randomized control trials, case-control studies and cohort studies. Reasons for this include the constant feature evolution and improvement, the inability to devise placebo effect, and the financial cost and resources required to conduct studies relative to commercial value of products during its short life cycle.

5.3 Melanoma/ skin analysis software

Kassianos et al (2015) reviewed currently available apps for the detection of melanoma aimed at general community, patient and generalist clinician users. Thirty-nine apps were identified with the majority available only for Apple users. Four apps provided a risk assessment to patients about the probability that a lesion was malignant or benign, and one app calculated users' future risk of melanoma. None of the apps appeared to have been validated for diagnostic accuracy or utility using established research methods, so there was no assurance that the prediction of melanoma in those tested was at all reliable.

A 2017 systematic review (Buechi et al. 2017) identified 30 manuscripts investigating 35 apps most commonly intended to screen photographs for melanoma or diagnose tremor through analyses of movements. A pooled estimate of diagnostic sensitivity for such apps was reported as 82% and pooled specificity as 89%. They highlighted that this specificity and sensitivity is not at a level to suggest that the apps overall could replace a clinical consultation. A in depth study by Nabil et al. (2017) also demonstrated poor agreement between a mobile phone application for the analysis of skin lesions and the clinical diagnosis of the dermatologist.

Wang et al (2016) reviewed available dermatology-related apps, noting that one in five patients under the age of 50 have used a smartphone to help diagnose a skin problem. Their review showed that the diagnostic performance of apps are inferior to in person consultations with one study showing that three out of four applications incorrectly classified 30% or more melanomas as low-risk lesions. Thissen et al. (2017) evaluated a SkinVision app. On the 108 cases used for evaluation the algorithm scored 80% sensitivity and 78% specificity in detecting (pre)malignant conditions. They concluded that although it was less accurate than the dermatologist's clinical eye, the app may offer support to other professionals who are less familiar with differentiating between benign and malignant lesions

Ngoo et al. (2018) assessed three melanoma smartphone apps in making clinical decisions about risk, compared with lesion assessment by specialist trained dermatologists. The apps' sensitivity and specificity ranged from 21 to 72% and 27 to 100%, respectively, when compared with the specialists' decisions. Two apps were unable to analyse 14 and 18% of lesions submitted. The low sensitivity in detecting lesions that are suspicious to a trained specialist may mean false reassurance is being given to patients.

5.4 Asthma

Huckvale et al. (2015) systematically assessed the content of 191 apps available from 2011 to 2015 examining the comprehensiveness of asthma information, consistency with the evidence base for asthma self-management and adherence to best practice principles for trustworthy content. They found that newer apps were no more likely than those available in 2011 to include comprehensive information, such as the use of action plans, or offer guidance consistent with evidence; 39% of those intended to manage acute asthma, recommended self-care procedures unsupported by evidence. The findings underline the need for coordinated quality assurance processes that can adapt to changing clinical and information governance-related risks, ensure compliance with the evidence base and reflect local variations in clinical practice.

5.5 Cardiovascular measurements

Coppetti et al. (2017) found substantial performance differences between four studied heart rate measuring apps. The two contact photoplethysmography-based apps had higher feasibility and better accuracy for heart rate measurement than the two non-contact photoplethysmographybased apps.

Plante et al (2018) explored a top-selling blood pressure-measuring app that underreported elevated blood pressure. However, its iTunes app store user ratings and reviews were generally positive. The authors concluded that *"these data suggest reassuring app results from an inaccurate blood pressure-measuring app may have improved user experience, which may have led to more positive user reviews and greater sales. Systematic underreporting of elevated blood pressure may have been a contributor to the app's success. Further studies are needed to confirm whether falsely reassuring output from other mobile health apps improve user experience and drives uptake"*.

5.6 Medicine dose

Bierbrier et al. (2014) assessed 13 medical calculation apps and found 11 provided high accuracy, two of the products provided a number of errors. Approximately half of the errors (8 of 17) were clinically significant resulting in a significant change in prognosis. Many apps claim to advise patients about drug doses or risks. However, even apps intended to help clinicians calculate drug doses have been found to give misleading results (e.g. opiate calculators, Haffey et al. 2014). Wyatt (2018) recommended that clinicians should avoid recommending apps for dosage adjustment or risk assessment unless they have personally checked the app's accuracy, or read a published independent evaluation of accuracy.

6. Conclusion

The tendency for journals to publish positive findings and for negative reports to struggle to be published means that there is only limited research into the safety of software when used in a medical setting and that studies indicating performance failures for apps are often not published (Buechi et al. 2017).

While a number of authors noted benefits and future promise of apps for improving health outcomes, there were clear examples of potential and actual harm across the range of different conditions for which software is intended for use. Safety concerns were identified in direct-to-consumers apps and situations where medical professionals directly use or guide use of the software.

In many cases, particularly where regulation does not currently apply, the responsibility of adverse consequences from apps falls on individual clinicians.

7. References

Akbar S, Coiera E and Magrabi F (2020) Safety concerns with consumer-facing mobile health applications and their consequences: a scoping review. *J Amer Med Info Ass*, 27: 330–340.

Bierbrier R, Lo V and Wu RC (2014). Evaluation of the accuracy of smartphone medical calculation apps. *J Med Internet Res*. 16:e32. doi: 10.2196/jmir.3062.

Bisson LJ, Komm JT, Bernas GA, Fineberg MS, Marzo JM, Rauh MA, et al (2016). How accurate are patients at diagnosing the cause of their knee pain with the help of a web-based symptom checker? *Orthop J Sports Med* 4:2325967116630286.

Bonoto BC, de Araujo VE, Godoi IP, Pires de Lemos LL, Godman B, Bennie M, Diniz LM and Guerra AA (2017) Efficacy of mobile apps to support the care of patients with diabetes mellitus: a systematic review and meta-analysis of randomized controlled trials. *JMIR Mhealth Uhealth* 5, e4 doi: 10.2196/mhealth.6309.

Boulos MN, Brewer AC, Karimkhani C, Buller DB andDellavalle RP (2014). Mobile medical and health apps: state of the art, concerns, regulatory control and certification. *Online J Public Health Inform*. 5:229.

Buechi R, Faes L, Bachmann LM, Thiel MA, Bodmer, NS, Schmid MK, Job O and Lienhard KR (2017) Evidence assessing the diagnostic performance of medical smartphone apps: a systematic review and exploratory meta-analysis. *BMJ Open* 7: 1–8.

Byambasuren O, Sanders S, Beller E. and Glasziou P (2018). Prescribable mHealth apps identified from an overview of systematic reviews. *NPJ Digit. Med*. 1: 12.

Chavez S, Fedele D, Guo Y, Bernier A, Smith M, Warnick J and Modave F (2017). Mobile apps for the management of diabetes. *Diabetes Care* 40:e145–e146.

Coppetti T, Brauchlin A, Müggler S, Attinger-Toller A, Templin C, Schönrath F, et al. (2017) Accuracy of smartphone apps for heart rate measurement. *Eur J Prevent Cardiol* 24:1287–1293.

Ferrero NA, Morrell DS, Burkhart CN (2013). Skin scan: a demonstration of the need for FDA regulation of medical apps on iPhone. *J Am Acad Dermatol* 68:515–516.

Fleming GA, Petrie JR, Bergenstal RM, Holl RW, Peters AL and Heinemann L (2020) Diabetes Digital App Technology: Benefits, Challenges, and Recommendations. A Consensus Report by the European Association for the Study of Diabetes (EASD) and the American Diabetes Association (ADA) Diabetes Technology Working Group. *Diabetes Care* 43: 250-260.

Fraser HS, Clamp S, and Wilson CJ (2017). Limitations of study on symptom checkers. *JAMA Intern Med* 177:740–741.

Fu H, McMahon SK, Gross CR, Adam TJ and Wyman JF (2017) Usability and clinical efficacy of diabetes mobile applications for adults with type 2 diabetes: a systematic review. *Diabetes Res Clin Pract* 131:70– 81.

Gordon WJ, Landman A, Zhang H and Bates DW (2020) Beyond validation: getting health apps into clinical practice *npj Digit. Med*. 3:14. https://www.nature.com/articles/s41746-019-0212-z

Grundy QH, Wang Z and Bero LA (2016) Challenges in assessing mobile health app quality: a systematic review of prevalent and innovative methods. *Am J Prevent Med* 51:1051–1059.

Haffey F, Brady RR and Maxwell S (2013). A comparison of the reliability of smartphone apps for opioid conversion. *Drug Saf*. 36:111–1117.

Hageman MG, Anderson J, Blok R, Bossen JK and Ring D (2015). Internet self-diagnosis in hand surgery. *Hand (NY)* 10:565–9.

Huckvale K, Car M, Morrison C and Car J (2012). Apps for asthma self-management: a systematic assessment of content and tools. *BMC Med*. 10:144.

Huckvale K, Morrison C, Ouyang J, Ghaghda A and Car J (2015). The evolution of mobile apps for asthma: an updated systematic assessment of content and tools. *BMC Med*. 13:58.

Jutel A and Lupton D (2015). Digitizing diagnosis: a review of mobile applications in the diagnostic process. *Diagnosis* 2: 89–96.

Kassianos AP, Emery JD, Murchie P and Walter FM (2015). Smartphone applications for melanoma detection by community, patient and generalist clinician users: a review. *Brit J Dermatol* 172:1507–18.

Larson RS (2018) A path to better-quality mHealth apps. *JMIR Mhealth Uhealth*. 6:e10414.

Lewis TL and Wyatt JC (2014) mHealth and mobile medical apps: a framework to assess risk and promote safer use. *J Med Internet Res*. 16:e210.

Mathews SC, McShea MJ, Hanley CL, Ravitz A, Labrique AB and Cohen AB (2019). Digital health: a path to validation. *npj Digit. Med*. 2, 38 (2019). https://www.nature.com/articles/s41746-019-0111-3

McMillan B, Hickey E, Mitchell C and Patel M (2015) The need for quality assurance of health apps. *BMJ*. 351:h5915.

Milea D et al. (2020) Artificial Intelligence to Detect Papilledema from Ocular Fundus Photographs *New Eng J Med* 382:1687-1695.

Millenson ML, Baldwin JL, Zipperer L and Singh H (2017) Beyond Dr. google: the evidence about consumer-facing, digital tools for diagnosis. *Diagnosis* https://doi.org/10.1515/dx-2018-0009

Nabil R, Bergman W, and KuKutsh NA (2017). Poor agreement between a mobile phone application for the analysis of skin lesions and the clinical diagnosis of the dermatologist, a pilot study. *Br J Dermatol* 177:583–584.

Ngoo A, Finnane A, McMeniman E, Tan JM, Janda M and Soyer HP (2017). Efficacy of smartphone applications in high-risk pigmented lesions. *Australas. J Dermatol* 59:e175-e182.

Plante TB. et al. (2018) User experience of instant blood pressure: exploring reasons for the popularity of an inaccurate mobile health app. *npj Digit. Med*. 1, 31.

Powell AC, Landman AB and Bates DW (2014). In search of a few good apps. *JAMA* 311:, 1851–1852.

Powley L, McIlroy G, Simons G and Raza K (2016). Are online symptoms checkers useful for patients with inflammatory arthritis? *BMC Musculoskelet Disord* 17:362.

Robillard JM, Illes J, Arcand M, Beattie BL, Hayden S, Lawrence P, et al (2015). Scientific and ethical features of English-language online tests for Alzheimer's disease. *Alzheimers Dement (Amst)* 1:281–8.

Ronquillo JG and Zuckermann DM (2017) Software-Related Recalls of Health Information Technology and Other Medical Devices: Implications for FDA Regulation of Digital Health. *The Milbank Quarterly*, 95: 535553.

Rowland SP, Fitzgerald JE, Holme T, Powell J and McGregor A (2020). What is the clinical value of mHealth for patients? *npj Digit. Med.* 3, 4 (2020). https://doi.org/10.1038/s41746-019-0206-x

Semigran HL, Linder JA, Gidengil C and Mehrotra A (2015) Evaluation of symptom checkers for self diagnosis and triage: Audit study. *BMJ* 351, h3480.

Semigran HL, Levine DM, Nundy S and Mehrotra A. (2016) Comparison of physician and computer diagnostic accuracy. *JAMA Intern Med* 176:1860–1861.

Singh K, Drouin K, Newmark LP, Lee J, Faxvaag A, Rozenblum R, Pabo E A, Landman A, Klinger E and Bates DW (2016). Many mobile health apps target high-need, high-cost populations, but gaps remain. *Health Affairs Millward* 35: 2310–2318.

Thissen M, Udrea A, Hacking M, von Braunmuehl T and Ruzicka T (2017). mHealth app for risk assessment of pigmented and nonpigmented skin lesions-a study on sensitivity and specificity in detecting malignancy. *Telemed J E Health* 23:948–54.

van Velthoven MH, Wyatt JC, Meinert E, Brindley D and Wells G (2018) How standards and user involvement can improve app quality: a lifecycle approach. *Int. J Med Inform*. 118:54–57.

Wang JV, Chapman LW, Irvine Keller M (2016) Challenges to smartphone applications for melanoma detection. Department of Dermatology and Cutaneous Biology Faculty Papers. *Paper 69*. https://jdc.jefferson.edu/dcbfp/69

Wolf JA, Moreau JF, Akilov O, Patton T, English JC III, Ho J and Ferris LK (2013). Diagnostic inaccuracy of smartphone applications for melanoma detection. *JAMA Dermatol*. 149:422–6.

Wyatt JC (2018) How can clinicians, specialty societies and others evaluate and improve the quality of apps for patient use? *BMC Med* 16, 225. https://doi.org/10.1186/s12916-018-1211-7.

Zhang Y, Masci P, Jones P, and Thimbleby H (2019) User Interface Errors in Medical Devices: Study of U.S. Recall Data. *Biomedical Instrumentation Technology* 53:182-194.

Version history

Version	Description of change	Author	Effective date
V1.0	Original publication	Devices Emerging Technology & Diagnostic	July 2020

Therapeutic Goods Administration

PO Box 100 Woden ACT 2606 Australia
Email: info@tga.gov.au Phone: 1800 020 653 Fax: 02 6203 1605
https://www.tga.gov.au

Reference/Publication #