CENTER FOR
THE GOVERNANCE
OF CHANGE

CASE STUDY 1: BABYLON HEALTH

Saif Ahmad, Cambridge University Hospitals.

Innovation, Sustainability, and the Future of Healthcare

Babylon Health is a health technology company which provides users with remote access to medical expertise. It is primarily an app-based service which allows users to interact with health professionals through text and video. The Babylon app also employs a Symptom Checker Chatbot which employs AI-based techniques to triage patients and help diagnose the underlying condition. In 2016, it was widely publicized that the Symptom Checker Chatbot outperformed a nurse and junior doctor in triaging patients, both in terms of speed and accuracy, and thus offers significant promise in making high-quality healthcare more accessible, especially in developing countries. However, limitations of the study design have been highlighted within the methodology of this research. Shortcomings within the regulatory landscape have also been identified as it is perceived that this and other similar technologies may require more robust scrutiny. This case study summarizes the exciting potential for this technology and, more specifically, highlights some of the challenges around validating clinical effectiveness.

BACKGROUND

Babylon Health was founded in January 2013 by Dr Ali Parsa – a healthcare entrepreneur with a PhD in engineering and background in investment banking. In its own words, "Babylon's mission is to put an accessible and affordable health service in the hands of every person on earth" [1]. Babylon Health offers a suite of different services based on its AI technology and app-based telemedicine interface. This allows patients to have their symptoms assessed by an AI-driven symptom checker chatbot and remotely engage with clinicians and allied health professionals.

Sissons' study, "Virtual Primary Care: Fragmentation or Integration?", examined the challenges facing traditional primary-care service providers as a result of telemedicine services, such as Babylon Health. In contrast, this case study focuses more on Babylon Health's symptom checker chatbot, specifically detailing research into how its effectiveness has been assessed. Moreover, this article examines how the introduction of this AI chatbot for clinical use has provided a test case for how regulatory systems appraise this increasingly significant and pervasive form of "medical device".

BABYLON HEALTH'S PROPOSITION FOR USERS

The global telemedicine industry has been valued at $29.6 billion in 2017 and is anticipated to grow at a compound annual growth rate of around 19% from 2017 to 2022 [2] . Factors driving the market's growth include an increasing acceptance by users to utilize telemedicine, a rising incidence of chronic diseases, a growing elderly population, government initiatives and a shortage of physicians [2].

Across the globe, a large number of companies have entered the telemedicine market. Within the United Kingdom, Babylon Health is regarded as the market leader. Since 2017 Babylon Health has operated an NHS contract under their "GP at Hand" label which offers access to NHS GP services through Babylon's technology.

Outside of the United Kingdom, Babylon Health's only other active market, at present, is in Rwanda, where since 2016 it has been offering the Babyl digital health service for Rwandans to access a similar service, customized to function within its more challenging healthcare environment. For instance, many of its functions (e.g. receiving a prescription) are performed using SMS text message codes rather than through apps which are reliant on smartphones. Babylon Health does report plans to grow its business in target markets such as the United States.

Sitting alongside these telemedicine services is an AI-based symptom checker chatbot which allows patients to input symptoms and answer questions enabling the chatbot to give advice and/or recommend a medical review.

EVIDENCE SUPPORTING THE USE OF THE SYMPTOM CHECKER CHATBOT

In June 2016, Babylon Health published a paper on arXiv, a moderated but non-peer reviewed online publication describing its "babylon check" automated triage system [3] . The triage system was developed using a computational model which was built on questions, answers, triggers and outcomes. Essentially, answers to questions would help to inform further questioning while feeding into recommending an outcome, for example whether a patient should attend their GP. This model was iterated multiple times based on review of triage flow by clinicians and medical experts.

The authors of this research, all of whom were part of Babylon Health, described how babylon check compared against 12 clinicians and 17 nurses recruited for the study. This assessment, which they describe as semi-naturalistic, involved 102 clinical vignettes simulated by actors. The vignettes were created in-house by a Babylon Health clinician, who was not directly involved in the development of babylon check. The vignettes were reviewed by independent experts although details of how this was done were not explained within the article. The simulated clinical vignettes were diagnosed and triaged, in terms of how urgently they required assessment (i.e. self-care, GP or accident and emergency) by the doctors and nurses within the study, and the actors also used the automated babylon check symptom checker.

The results from the study showed that babylon check was more accurate, safer and faster than the nurses and clinicians. Safety was assessed by levels of under-triaging of patients, for example patients who required accident and emergency referral not being referred. Across the 102 vignettes, babylon check had 90.2% accuracy and 100% safety compared to accuracy

of around 75% for the nurses and clinicians and safety of 97% and 98% for nurses and doctors, respectively.

In the paper, one of the limitations of the study cited by the authors was that the relative lack of accuracy of clinicians and nurses could "be attributed to actors deviating from the patient vignettes during mock consultations, an issue we are currently investigating in our continued validation and testing efforts". There are a number of additional limitations of this study including the lack of detailed methodology being presented (e.g. how clinical vignettes were reviewed by independent experts) and the absence of authors who were independent of Babylon Health.

The claims made within this paper were partly used to support a trial of the symptom checker within the NHS 111 service, a telephone service that patients can use to access non-immediate medical advice and assessment. Prior to the pilot, the NHS did internally validate the symptom checker. This validation was performed by using the NHS 111 app against all reported "serious incidents" in the pilot area, and showed that it triaged patients appropriately in all cases; however, any further details of this internal validation do not appear to be publicly available [4]. The research community within the UK raised concerns about the lack of high-quality evidence supporting the symptom checker's use [5]. In particular, it was argued that an independent study should be published, and more detail of the methodology should be presented. Moreover, rather than simulated vignettes, data from real world use would be regarded as more representative of the true effectiveness of the technology.

Robust, independent data to support the symptom checker has yet to be published. However, in June 2018, Babylon Health published a follow-up article in arXiv [6], which described a newly updated version of the symptom checker also known as the Babylon Triage and Diagnostic System. In this study, it was reported that the symptom checker scored higher on a selection of Membership of the Royal College of General Practitioners (MRCGP) exam questions when compared to in-house GPs and a historical average from doctors taking the exam (81% versus 72%).

Again, concerns were raised about the methodology of the research: data in the trials were entered by doctors rather than the intended lay users, and despite only small numbers of GPs' data being used for comparison, no statistical significance testing was performed [7]. However, Babylon Health were also commended in this article for publishing data and it was acknowledged that issues around clinical testing methodology were not limited to their platform.

One significant reason that rigorous clinical studies have not been performed to support the use of symptom checkers is that it is not currently required for regulatory approval (see Figure 1).

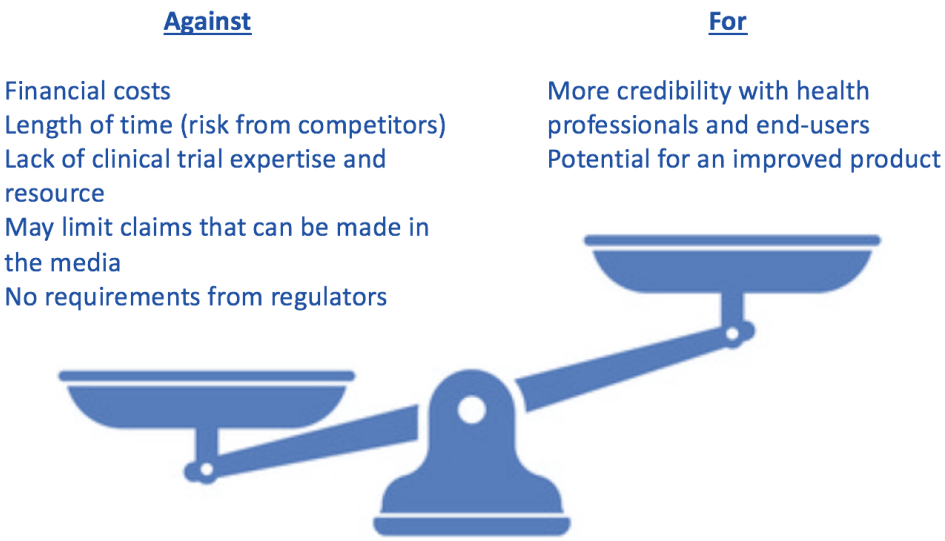

Figure 1. Reasons affecting decision-making for health app developers to perform rigorous clinical studies.

SYMPTOM CHECKERS AND REGULATION

Medicines and Healthcare products Regulatory Agency (MHRA) guidance classes symptom checkers as Class 1 devices, which means that they can be registered by the manufacturer against safety, reliability and usability standards. Post-market surveillance is also considered the responsibility of the manufacturer, although users can raise concerns directly with the MHRA.

The lack of regulatory hurdles for app deployment in healthcare contributes to a low barrier to entry for health apps in general, and governance is further complicated by the fluidity in app function as a result of iterative software modifications [8]. Current European medical device regulation (CE marking) presumes apps are clinically effective, and thereby attributes liability for use-errors, even if caused by software bugs, to end users. As apps such as the Babylon Health symptom checker gain increasing traction across healthcare, it is increasingly likely that regulators will act to increase governance of this sector. Yet, at the present time what form these changes will take remains uncertain.

CONCLUSIONS AND FUTURE DIRECTIONS

The Babylon Health symptom checker is an example of a potentially valuable clinical tool that complements the rise in telemedicine uptake. Several companies are developing similar tools, including Germany-based Ada. It is likely that these tools will not only be applied to triaging

care but will also have value in managing chronic conditions such as mental health disorders [9]. Moreover there are an increasing number of clinician-facing apps that are aimed at helping clinical decision making [10]. Babylon Health's success in terms of clinical uptake within the NHS demonstrates how attractive clinical triage tools can be to healthcare providers as well as the public. Importantly, one of the key challenges in appraising triage tools is defining what a good triage tool looks like. Recently, the UK Care Quality Commission (CQC) published a report on this topic building on a regulatory sandbox pilot which involved building a consensus with multiple stakeholders. This exercise identified key criteria that a "good" triage tool should satisfy including safety, efficacy, caring, responsiveness and being well-led [11] . In general, robust data has yet to clearly emerge supporting these criteria and the economic benefits of symptom checkers or triage tools, partly as a result of a lack of randomized studies [12]. However, as the usage of these apps increases the development of these datasets will be of critical importance.

REFERENCES

[1] *Babylon Health*. 20 November 2019. [Online] https://www.babylonhealth.com/about

[2] Global Telemedicine Market Outlook 2022. (2019, 11 22). Retrieved from https://www.researchandmarkets.com/reports/3766749/global-telemedicine-market-outlook-2022

[3] K. Middleton et al. (2016, 6 7) [Online] https://arxiv.org/pdf/1606.02041.pdf

[4] Babylon. (2020, January 29). [Online] https://assets.babylonhealth.com/nhs/NHS-111-Evaluation-of-outcomes.pdf

[5] M. McCartney. "Innovation without sufficient evidence is a disservice to all," *BMJ*, j3980. 2017.

[6] S. Razzaki et al. (2018, 6 27). [Online] https://arxiv.org/pdf/1806.10698.pdf

[7] H. Fraser, E. Coeira & D. Wong. "Safety of patient-facing digital symptom checkers," *Lancet*, pp. 2263-2264. 2018.

[8] F. Magrabi et al. "Why is it so difficult to govern mobile apps in healthcare?," *BMJ Health and Care Informatics*, e100006. 2019.

[9] A.A. Abd-alrazaq et al. "An overview of the features of chatbots in mental health: A scoping review," *International Journal of Medical Informatics*, 103978. 2019.

[10] C.L. Ventola, "Mobile Devices and Apps for Health Care Professionals: Uses and Benefits," *P&T, 39* (5), pp. 356-364, 2014.

[11] CQC. *Getting to the right care in the right way – digital triage in health services.* 2020.

[12] CCG. (2019, 11 26). [Online] https://www.hammersmithfulhamccg.nhs.uk/media/156123/Evaluation-of-Babylon-GP-at-Hand-Final-Report.pdf