Letter to the Editor

Periodic Manual Algorithm Updates and Generalizability: A Developer's Response. Comment on "Evaluation of Four Artificial Intelligence–Assisted Self-Diagnosis Apps on Three Diagnoses: Two-Year Follow-Up Study"

Stephen Gilbert, BSc, BVMS, MRes, PhD; Matthew Fenech, MD, PhD; Anisa Idris, LLM, MPH, MBA; Ewelina Türk, MD

Ada Health, Berlin, Germany

Corresponding Author:
Stephen Gilbert, BSc, BVMS, MRes, PhD
Ada Health
Karl-Liebknecht-Str 1
Berlin, 10178
Germany
Phone: 49 017680396015
Email: stephen.gilbert@ada.com

Related Articles:
Comment on: https://www.jmir.org/2020/12/e18097
Comment in: https://www.jmir.org/2021/6/e29336

(J Med Internet Res 2021;23(6):e26514) doi: 10.2196/26514

KEYWORDS

artificial intelligence; machine learning; mobile apps; medical diagnosis; mHealth; symptom assessment

We have several comments on the recent publication of Ćirković [1], in which repeated testing of four symptom assessment applications with clinical vignettes was carried out to look for "hints of 'non-locked learning algorithms'." As the developer of one of the symptom assessment applications studied by Ćirković [1], we are supportive of studies evaluating app performance; however, there are important limitations in the methodology of this study.

Most importantly, the methodology used in this study is not capable of addressing its main objective. The approach used to look for evidence of nonlocked algorithms was the quantification of differences in performance using 3 ophthalmology vignettes, first in 2018, then in 2020. This methodology, although highly limited due to the use of only 3 vignettes in one medical specialism, could be used to detect changes in app performance over time. It, however, cannot be used to distinguish between nonlocked algorithms and the manual updating of apps' medical intelligence, through the normal process of the manual release of updated app versions. Medical device regulations and quality system requirements provide standard mechanisms through which apps can be further developed, validated, and released as updated versions. The manual of medical knowledge in this manner has been acknowledged by the manufacturers of all the apps studied by Ćirković [1]. In response to previous independent vignettes studies [2,3], spokespeople for Your.MD and Babylon stated that they update their medical knowledge periodically, and this is also clear on Buoy's website. In Gilbert et al [4], the Ada app is described as having a knowledge base "built and reviewed by medical doctors in a curated process of knowledge integration from medical literature. It is being expanded continuously following this standardized process."

As is acknowledged in the limitations listed in Ćirković's work [1], the study used vignettes designed, entered, and with results adjudicated by a single clinician. This could result in bias and a narrow type of case. It is also acknowledged that 3 vignettes represent a small sample size for a vignettes study and that "standardized and transparent procedures" are needed for symptom assessment app–vignettes studies. We recently published a 200-vignette assessment of symptom assessment applications [4], including those studied by Ćirković [1], which used standardized and transparent procedures, including the separation of vignette design, entered and with results adjudication. It is our view that the effect of the limitations described by Ćirković [1], together with only including ophthalmological cases, is that the accuracy results reported have limited generalizability or repeatability. Our own internal validation testing shows an improvement in Ada's medical intelligence in all-condition top-3 suggestion accuracy (also known as M3, as defined by Miller et al [5]) of 4.8% between 2018 and 2020. We take account of all performance feedback

we receive, and incorporate this, when judged appropriate by our medical knowledge experts, into updates of our app, through periodic releases of locked versions of our app.

Conflicts of Interest

All authors are employees of Ada Health.

References

1. Ćirković A. Evaluation of Four Artificial Intelligence-Assisted Self-Diagnosis Apps on Three Diagnoses: Two-Year Follow-Up Study. J Med Internet Res 2020 Dec 04;22(12):e18097 [FREE Full text] [doi: 10.2196/18097] [Medline: 33275113]
2. Burgess M. Can you really trust the medical apps on your phone? Wired. 2017 Oct 1. URL: https://www.wired.co.uk/article/health-apps-test-ada-yourmd-babylon-accuracy [accessed 2020-12-15]
3. What happened when Pulse tested symptom checker apps. Pulse. 2019 Aug 30. URL: http://www.pulsetoday.co.uk/news/analysis/what-happened-when-pulse-tested-symptom-checker-apps/20039333.article [accessed 2020-12-15]
4. Gilbert S, Mehl A, Baluch A, Cawley C, Challiner J, Fraser H, et al. How accurate are digital symptom assessment apps for suggesting conditions and urgency advice? A clinical vignettes comparison to GPs. BMJ Open 2020 Dec 16;10(12):e040269 [FREE Full text] [doi: 10.1136/bmjopen-2020-040269] [Medline: 33328258]
5. Miller S, Gilbert S, Virani V, Wicks P. Patients' Utilization and Perception of an Artificial Intelligence-Based Symptom Assessment and Advice Technology in a British Primary Care Waiting Room: Exploratory Pilot Study. JMIR Hum Factors 2020 Jul 10;7(3):e19713 [FREE Full text] [doi: 10.2196/19713] [Medline: 32540836]

Edited by T Derrick; submitted 15.12.20; peer-reviewed by A Ćirković; accepted 13.05.21; published 16.06.21

Please cite as:
Gilbert S, Fenech M, Idris A, Türk E
Periodic Manual Algorithm Updates and Generalizability: A Developer's Response. Comment on "Evaluation of Four Artificial Intelligence–Assisted Self-Diagnosis Apps on Three Diagnoses: Two-Year Follow-Up Study"
J Med Internet Res 2021;23(6):e26514
URL: https://www.jmir.org/2021/6/e26514
doi: 10.2196/26514
PMID: