JMIR AI

Volume 1 (2022), Issue 1 ISSN 2817-1705 Editors-in-Chief: Khaled El Emam, Bradley Malin, PhD

Contents

Editorial

Original Papers

Editorial

Introducing JMIR AI

Khaled El Emam[1], BEng, PhD; Bradley Malin[2], BA, MSc, PhD

[1]School of Epidemiology and Public Health, University of Ottawa, Ottawa, ON, Canada

[2]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, United States

Corresponding Author:
Khaled El Emam, BEng, PhD
School of Epidemiology and Public Health
University of Ottawa
401 Smyth Road
Ottawa, ON, K1H 8L1
Canada
Phone: 1 613 737 7600
Email: kelemam@ehealthinformation.ca

Abstract

JMIR AI is a new journal with a focus on publishing applied artificial intelligence and machine learning research. This editorial provides an overview of the primary objectives, the focus areas of the journal, and the types of articles that are within scope.

(JMIR AI 2022;1(1):e42046) doi:10.2196/42046

KEYWORDS

artificial intelligence; AI; machine learning; methodology

The past decade has witnessed rapid growth in the development of artificial intelligence and machine learning (AI/ML) methods for biomedical research and clinical applications. At the same time, it has become clear that the translation of such methods into practice has met numerous challenges. This is perhaps best exemplified by the status of AI/ML work during the COVID-19 pandemic. The pandemic was one moment in time where powerful AI/ML-driven diagnostic and prognostic tools could have accelerated our understanding and development of effective treatments. With some notable exceptions [1], and despite many publications, the impact of AI/ML in practice has been limited [2,3]. The reasons are varied [4-6], such as a lack of representation in the populations to whom the data corresponds and poor quality in the data available, leading to a lack of generalizable methodologies and models and a lingering lack of trust in automated decision-making. In this respect, our main motivation for JMIR AI is to publish articles that focus on the practical issues involved in developing useful AI/ML solutions and implementing them in biomedical research and clinical settings.

At the same time, we are seeing the introduction of policies and statutes in disparate jurisdictions to regulate AI/ML systems [7,8]. These policies and statutes are being developed in anticipation of an AI-laden future. Yet, as with all policy-making, such activities are likely to impact data access, the definition of fit-for-use data, algorithmic explainability and transparency, patient access to data and decision justifications, and the need for continuous evaluation of models in clinical settings, to name a few.

JMIR AI aims to become a venue for identifying, discussing, and addressing such practical challenges, with a particular emphasis on applications. The journal will strive to publish technical articles, as well as those tackling societal aspects, including ethical, legal, policy, and regulatory issues. This will be accomplished through a mix of research, perspectives, tutorials, and articles describing benchmark data sets. By leveraging JMIR Publications' publishing processes and tools, we also expect to have a rigorous and rapid open access review and publication process.

To realize this vision, we are assembling a multidisciplinary editorial board covering a wide array of topics from academia and industry, as well as ensuring broad domain and regional representation. The founding members of the editorial board cover many years working in academic medical centers and with spin-off health technology companies, as well as working with and within the pharmaceutical and medical device industries. Given the continued rapid advancement of this multidisciplinary field, we intend to continue expanding the editorial board to cover relevant areas as they arise.

We also intend to use the journal as a platform to enable and facilitate code and data sharing. This will be achieved by providing authors with additional tools that address the many technical and regulatory obstacles to broader community sharing.

Conflicts of Interest

None declared.

References

1. Adams R, Henry KE, Sridharan A, Soleimani H, Zhan A, Rawat N, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nat Med 2022 Jul;28(7):1455-1460. [doi: 10.1038/s41591-022-01894-0] [Medline: 35864252]
2. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell 2021 Mar 15;3(3):199-217. [doi: 10.1038/s42256-021-00307-0]
3. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020 Apr 07;369:m1328 [FREE Full text] [doi: 10.1136/bmj.m1328] [Medline: 32265220]
4. Ghassemi M, Oakden-Rayner L, Beam A. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 2021 Nov;3(11):e745-e750. [doi: 10.1016/S2589-7500(21)00208-9]
5. Petersson L, Larsson I, Nygren JM, Nilsen P, Neher M, Reed JE, et al. Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in Sweden. BMC Health Serv Res 2022 Jul 01;22(1):850 [FREE Full text] [doi: 10.1186/s12913-022-08215-8] [Medline: 35778736]
6. Wadden JJ. Defining the undefinable: the black box problem in healthcare artificial intelligence. J Med Ethics 2021 Jul 21. [doi: 10.1136/medethics-2021-107529] [Medline: 34290113]
7. Good machine learning practice for medical device development: guiding principle. Food and Drug Administration. 2021 Oct. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles [accessed 2022-08-11]
8. Kop M. EU artificial intelligence act: the European approach to AI. Stanford - Vienna Transatlantic Technology Law Forum, Transatlantic Antitrust and IPR Developments, Stanford University. 2021 Oct 01. URL: https://law.stanford.edu/publications/eu-artificial-intelligence-act-the-european-approach-to-ai/ [accessed 2022-08-13]

Abbreviations

AI/ML: artificial intelligence and machine learning

Edited by T Leung; submitted 19.08.22; this is a non–peer-reviewed article;accepted 22.08.22; published 20.09.22.

Please cite as:
El Emam K, Malin B
Introducing JMIR AI
JMIR AI 2022;1(1):e42046
URL: https://ai.jmir.org/2022/1/e42046
doi:10.2196/42046
PMID:

Original Paper

Adolescents' Well-being While Using a Mobile Artificial Intelligence–Powered Acceptance Commitment Therapy Tool: Evidence From a Longitudinal Study

Dana Vertsberger[1,2], PhD; Navot Naor[2], PhD; Mirène Winsberg[1], PhD

[1]Stanford University, Stanford, CA, United States

[2]Kai.ai, Haifa, Israel

Corresponding Author:
Dana Vertsberger, PhD
Stanford University
450 Jane Stanford Way
Stanford, CA, 94305
United States
Phone: 1 6465468241
Email: dverts@stanford.edu

Abstract

Background: Adolescence is a critical developmental period to prevent and treat the emergence of mental health problems. Smartphone-based conversational agents can deliver psychologically driven intervention and support, thus increasing psychological well-being over time.

Objective: The objective of the study was to test the potential of an automated conversational agent named Kai.ai to deliver a self-help program based on Acceptance Commitment Therapy tools for adolescents, aimed to increase their well-being.

Methods: Participants were 10,387 adolescents, aged 14-18 years, who used Kai.ai on one of the top messaging apps (eg, iMessage and WhatsApp). Users' well-being levels were assessed between 2 and 5 times using the 5-item World Health Organization Well-being Index questionnaire over their engagement with the service.

Results: Users engaged with the conversational agent an average of 45.39 (SD 46.77) days. The average well-being score at time point 1 was 39.28 (SD 18.17), indicating that, on average, users experienced reduced well-being. Latent growth curve modeling indicated that participants' well-being significantly increased over time (β=2.49; P<.001) and reached a clinically acceptable well-being average score (above 50).

Conclusions: Mobile-based conversational agents have the potential to deliver engaging and effective Acceptance Commitment Therapy interventions.

(JMIR AI 2022;1(1):e38171) doi:10.2196/38171

KEYWORDS

well-being; adolescents; chatbots; conversational agents; mental health; mobile mental health; automated; support; self-management; self-help; smartphone; psychology; intervention; psychological; therapy; acceptance; commitment; engagement

Introduction

Adolescence is a developmental period that is filled with changes: changes to one's body, in one's social environment, and even to one's mind [1]. It is also a crucial period for mental, social, and emotional well-being, which is characterized with an increased risk to develop mental health problems, such as anxiety, depression, substance abuse, and eating disorders [1]. According to the Centers for Disease Control and Prevention, more than 10% of adolescents aged 12-17 years experience anxiety, almost 8% experience behavior disorders, and 6% experience depression. These problems tend to continue during adulthood, especially when left untreated, and impact not only those who experience it but also those around them, as well as society as a whole. For example, lost productivity due to anxiety and depression is estimated to cost the global economy US $1 trillion each year [2]. The increased risk in developing mental illnesses during adolescence marks it as a crucial period for prevention, as well as treatment, and emphasizes the need for accessible and customized mental health tools aimed at decreasing adolescents' ill-being and increasing their well-being.

In this study, we focused on adolescents' well-being and followed their well-being during a time in which they used a digital, artificial intelligence (AI)–powered, personal companion designed to promote well-being and mental health.

Well-being usually refers to (1) a desirable state of satisfaction; (2) the presence of positive affect (eg, happiness); and (3) the absence of negative affect [3,4]. Well-being does not refer to a specific moment but rather a continuous state [5]. Previous research that mainly focused on adult populations have shown that psychological well-being is associated with health and long-term adjustment [6]. Specifically, higher levels of well-being were found to be associated with fewer illnesses, increased life expectancy, and healthier behavior [6]. In comparison, lower levels of well-being were found to be associated with higher levels of depression, hopelessness, and suicidal intent [7], as well as actual suicide attempts [8,9].

The few longitudinal studies that did focus on adolescents' well-being tend to find that adolescents' well-being decreased over time. Specifically, it was found that well-being, which was often measured by satisfaction with school, family, friends, schoolwork, appearance, and life as a whole, showed weak stability over time [10] and that girls are at greater risk of a decrease in their well-being over time than boys [10-12]. Another study did not find a significant change over time in psychological well-being but did find that girls reported lower levels of well-being than boys [13]. Taken together, these studies suggest that adolescent girls are at a greater risk to experience reduced well-being than boys.

Studies on adult populations suggest that different interventions can improve well-being, compared to control groups who did not receive interventions or were in delayed intervention groups. For example, an intervention based on Acceptance and Commitment Therapy (ACT), which focuses on cognitive diffusion, psychological flexibility, mindfulness, and values clarification, was found to be associated with greater well-being posttreatment, as well as at the 3-month follow-up, compared to the control group who did not receive the intervention [14]. Another study tested the effectiveness of the "My Coping Plan" app in improving mental health and coping [15]. The "My Coping Plan" approach focuses on normalizing unpleasant emotions and coping as universal human experiences, both for those with and without mental illness, and encourages professional help–seeking as a healthy coping strategy when personal coping strategies are ineffective. It was found that participants in the intervention group reported improved well-being, compared to control group, over the 1-month period of using the app [15]. Finally, a study that tested the effect of mindfulness-based therapy versus a waitlist group among patients with breast cancer showed a significant increase in well-being in the experimental group at both 8- and 12-week assessments compared to the control group [16].

One promising avenue for delivering interventions, especially for adolescents, is through their smartphones. The use of smartphones and their different apps have been highly integrated in almost everyone's everyday lives. Almost 50% of 11-year-olds in the United States have a mobile phone, with this number reaching 85% among 14-year-olds [17]. On average, US adolescents aged 13-18 years engage with their mobile phone for more than 3 hours every day [18], making it a highly accessible and easy-to-use tool for presenting a mental health intervention. Indeed, recent years have seen an increase in the development of various mental mobile health (mHealth) interventions apps [19]. These mental mHealth interventions are either aimed at complementing traditional mental health treatments or providing mental health support to those who are unable to receive quality mental health services—for example, due to long waiting lists [19]—and were found to be beneficial to adult populations [19-21].

There are many possibilities for delivering mental mHealth interventions, such as web-based therapists, conversational interfaces (such as Amazon's Alexa), and delivering information. Another prospective approach, especially for adolescents, is by implementing a text-based approach. Using a text-based approach could be an easy and sustainable way to keep adolescents engaged with the process of the intervention. Thus, in this study, we tested whether adolescents' well-being improved while using a commercially developed text-based conversational agent named Kai.ai.

Kai.ai is an AI-powered, personal companion designed to promote well-being and mental health by initiating daily conversations and presenting short and simple exercises to users and is used within an instant messenger app (eg, iMessage and WhatsApp). The intervention delivered by Kai.ai is mainly based on ACT protocols and tools adapted from positive psychology theories. As previously described, ACT aims to improve cognitive flexibility, as well as coping with challenging experiences, by focusing on cognitive diffusion, practicing mindfulness, and reflection on one's values [22,23]. Kai.ai delivers the ACT intervention using an AI conversational bot that facilitates users in creating and enhancing habits for healthy living and resilience [24]. Kai.ai interacts with users throughout the day and leverages their responses to deliver a tailored ACT skilled coaching. The main advantages of Kai.ai are that (1) it integrates seamlessly with all the top messaging apps (iMessage, WhatsApp, Discord, Telegram, etc), making it accessible and easy to use; (2) it can reach adolescents who would otherwise not seek support; (3) it contacts the adolescents but also responds to adolescents when they initiate the conversation; (4) it is available 24/7; (5) it is free; and (6) it is anonymous.

In this study, we followed adolescents who voluntarily chose to join and interact with Kai.ai. At several time points during the period in which they interacted with Kai.ai, they were asked about their well-being. We hypothesized that their well-being would improve while using Kai.ai.

Methods

Participants

The initial sample included 43,237 adolescents from the United States, aged 14-18 years, who had a smartphone with either an iOS or Android operating system and freely chose to interact with Kai.ai through common messaging apps. We advertised Kai.ai in platforms that are frequently used by adolescents such as Instagram and Snapchat. Within these platforms, the

advertisements were specifically shown to adolescents aged 14-18 years, and especially girls, according to the information they entered during the registration to each platform. Users are presented with an ad inviting them to test how happy they are. When clicking the ad, users are taken to Kai.ai's landing page. The final sample included 10,387 participants who answered the questionnaires more than once. The onboarding process for the use of Kai.ai does not ask the user to report their actual age or gender; thus, it did not enable us to report this information.

Procedure

As part of the joining process to Kai.ai (in the onboarding process), all users were asked, but were not obligated, to complete different questions and questionnaires to assess their needs. Subsequently, users were prompted once every 6 weeks to answer these questionaries once again, to monitor their mental progress; however, they could have answered whenever it was convenient for them. These prompts were also optional, and users were not required to answer the questionnaires to continue using Kai.ai. Participants in this study completed the 5-item World Health Organization Well-being Index (WHO-5) questionnaire between 2 and 5 times, between February 2020 and January 2022. The study was based on anonymized data gathered during the engagement with the service.

Ethics Approval

This study was approved by the WCG Institutional Review board (approval #1-1504102-1) and determined as not human subject research.

Intervention

Kai.ai

During the onboarding process, the users are made clear that Kai—the name of AI in the program—is not a real person and has been built by clinical psychologists, coaches, and engineers. Users are being informed that Kai will reach out to them and will send them daily practices, techniques, and insights, but they are also encouraged to reach out to Kai whenever they need. Kai.ai initiates between 1 and 3 daily interactions with the users, usually in the morning, at noon, and in the evening.

Each daily interaction that is initiated by Kai.ai begins with a greeting (ie, "good morning" or "good evening") and an inspiration quote (eg, "Many people will walk in and out of your life, but only true friends will leave footprints in your heart," Eleanor Roosevelt), which are then followed by a short exercise (described below) related to ACT. Users can also initiate an interaction with Kai whenever they please, or when Kai initiates the interaction, they can direct it to whatever topic they wish. When Kai recognizes that participants are in some sort of distress, the bot switches off and a trained companion, who is practiced in giving support, goes on and encourages users to turn to a someone close to them for support or use available hotlines near them (ie, presents them with a list of available possibilities).

Process-Oriented Features

In addition to interacting with Kai, Kai.ai also presents users with different exercises to promote better mental health. These exercises are described below.

Gratitude

The aim of this exercise is to help users to develop flexible thinking patterns, balance negative biases, and develop a positive view over their lives. Each day, users are prompted to think about the things they are grateful for and share them with Kai. The system saves their responses, and they can view them whenever they want.

Learning

The aim of this exercise is to help users to adopt a routine of reflection and journaling to help them achieve a more centered, grounded, joyous, and purposeful state of mind. Users are prompted to focus on the lessons they can learn from their experiences. In addition, to reduce stress and anxiety, users are guided to focus on a single task they have instead of a long to-do list.

Breathing

The breathing exercises are meant to reduce stress and anxiety by ensuring a better flow of oxygen to the body through the operation of the parasympathetic nervous system [25]. Initially, users are taught in a relaxed state of mind with the aim that with continuous practice, the exercises will become a tool that can be enacted while the users feel distressed. The exercises teach users how to breathe through their noses, using their diaphragm, and note their posture.

Mindfulness

The mindfulness exercises help the users become aware of the present moment without being judgmental toward themselves. The mindfulness exercises are audio-based and help users practice the art of observing and visualizing thoughts, emotions, and body sensations as they arise. By practicing these exercises, users can benefit by letting go of any repeated unwanted thoughts, increase self-awareness and self-compassion, as well as reduce tension and stress.

ACT Training

The ACT training aims to develop psychological flexibility and diffusion of thoughts, emotions, and behaviors and accept the challenging moments in our lives that contain negative emotions, as well as unfold the users' values and help them commit to the values. It teaches users to treat pain and discomfort as facts of life that can be used for personal growth through a process of acceptance and validation [26]. ACT exercises are also presented as audio, in which their aim is to guide the users in observing and accepting their current and past thoughts and emotions, despite the discomfort they might elicit. These ACT exercises help build resilience, gain control over thoughts and emotions, and assist in building coping strategies when facing difficult moments.

Positive Psychology

These exercises aim to decrease negativity bias and increase positivity in users' lives and are also audio-based. They help connect positive intentions to the users themselves and assist in enhancing self-compassion.

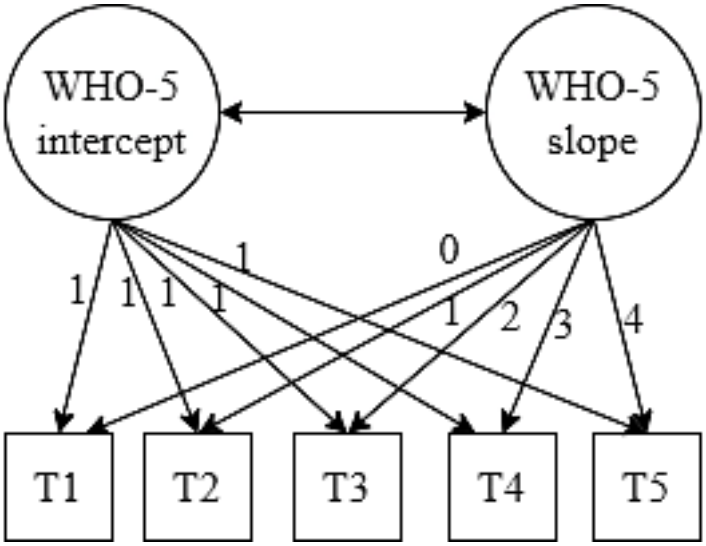
Measures: Well-being

Users' well-being was assessed using the WHO-5 Well-being Index, which is a brief 5-item self-reported measure [27]. Participants were asked to report their experiences in the past 2 weeks (eg, "I have felt cheerful and in good spirits") on a 6-point Likert scale, ranging from 0 (at no time) to 5 (all of the time). The raw well-being score theoretically ranges from 0 (absence of well-being) to 25 (maximal well-being). However, as scales measuring health-related quality of life are typically translated to a percentage-based scale, ranging from 0 to 100, it is recommended to multiply the raw score by 4. A score below 50% reflects poor well-being, and a 10-point change in the translated score is seen as clinically significant [28]. The WHO-5 has shown both clinical and psychometric validity (for a systematic review, see Topp et al [28]) and has been previously integrated in mental health– and physical health–related mobile apps [29-31]. The WHO-5 was also found to be reliable for assessing children and adolescents' well-being [32-34].

Statistical Approach

We first conducted a 1-way ANOVA to test for differences in the WHO-5 baseline scores between participants according to the number of times they have answered the questionnaire. In this analysis, we also considered those who answered only once, to test whether there were differences in well-being between those who chose to answer only once and those who continued to answer the questionnaire. Next, we described the descriptive statistics of the sample. Finally, to account for the structure of data, in which we had several time points of assessment for each user (between 2 and 5), we conducted latent growth curve modeling, in which we examined the change in the WHO-5 assessment over time. We first conducted an intercept-only model, in which the intercept variance was constrained to 0, so we only assessed the mean. In the second model, we conducted a random-intercept model and allowed users to differ in their starting point. In the third model, we added a random slope (ie, users are changing in different ways) but set the average slope to 0. Finally, in the fourth model, we estimated the real slope. After each step, we estimated the fit of the model to test whether adding each component improved the model. Figure 1 is an illustration for the final model. The analysis was conducted using the *lavaan* package in R statistical software (version 3.5.1; R Foundation for Statistical Computing) [35].

Figure 1. An illustration of the latent growth curve model. T: time point; WHO-5: 5-item World Health Organization Well-being Index.



Results

First, we tested for differences in WHO-5 baseline scores between all participants according to the number of times they have answered the questionnaire, using the *aov* function in R statistical software [36]. No significant difference was found between the groups ($F_{4,43,232}$=0.35; P=.84).

On average, users interacted with Kai.ai for 45.39 (SD 46.77; range 2-634) days. Table 1 describes the averages of the WHO-5 assessment at each time point (T). The average well-being score at T1 was below 50 (mean 39.28, SD 18.15), indicating that, on average, users experienced reduced well-being [28]. However, the average score increased over time, reaching an average of 53.63 (SD 21.32) in T5, 85 days after the first assessment (approximately 2.5 months). The average number of days that have passed between each assessment increased over time, with the average number of days being 25.80 days between T1 and T2 and 36.62 days between T4 and T5.

The results of the latent growth curve modeling, in which we assessed the difference in the WHO-5 assessments over time, are presented in Table 2. As can be seen in Table 2, the model fit improved from the first model to the last model, with the fourth model showing the best goodness of fit. As can be seen in the fourth model, participants' well-being significantly increased over time (β=2.49; P<.001). Figure 2 depicts the change in the WHO-5 over time. However, the negative covariance between the intercept and slope indicates that users who were lower in well-being show a bigger increase in well-being than users who were higher in well-being, as could be expected.

Table 1. Number of participants, mean, and SD of the WHO-5[a] assessments at each time point and average time elapsed between each time point.

T[b]	Participant, n	WHO-5 score, mean (SD)	Average time between $T_n - T_{n+1}$
1	10,387	39.28 (18.17)	
2	10,387	47.18 (19.68)	25.80
3	4801	49.85 (20.15)	25.70
4	2324	52.09 (20.45)	30.81
5	1072	53.64 (21.32)	36.62

[a]WHO-5: 5-item World Health Organization Well-being Index.

[b]T: time point.

Table 2. Latent growth curve modeling of the associations between time and users' engagement and the 5-item World Health Organization Well-being Index.

Model	Intercept (σ^2)	SE	Slope (σ^2)	Intercept-slope covariance	Goodness of fit			
					χ^2 (*df*)	RMSEA[a]	SRMR[b]	CFI[c]
Intercept only	50.25 (0)	393.2	N/A[d]	N/A	2884.86 (18)	0.39	0.43	0
Random intercept	50.25 (212.03)	181.70	N/A	N/A	794.52 (17)	0.21	0.14	0.7
Random slope	47.77 (188.68)	142.42	0 (14.22)	0	493.12 (16)	0.17	0.15	0.82
Real slope	45.28 (209.13)	138.13	2.49 (11.04)	–8.33	227.55 (14)	0.12	0.06	0.92

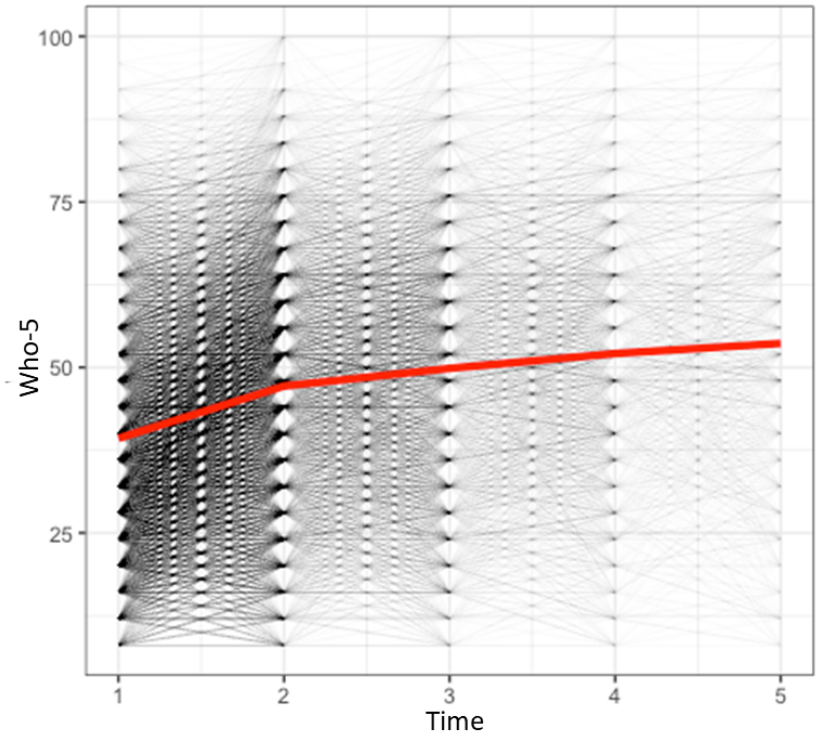
[a]RMSEA: root mean square error of approximation.

[b]SRMR: standardized root mean square residual.

[c]CFI: comparative fit index.

[d]N/A: not applicable.

Figure 2. Plot of the change in the WHO-5 scores across time. WHO-5: 5-item World Health Organization Well-being Index.



Discussion

Principal Findings

Adolescence is a crucial period for treating and, ideally, preventing mental health problems and increasing well-being: first, to improve the lives of adolescents and second, to decrease the risk for developing mental health problems in adulthood [1]. In this study, we followed the well-being of more than 1000 adolescents from the United States through their interaction with an AI-powered, personal companion named Kai.ai, over a period of 4 months. The results indicated that, on average, adolescents' well-being increased over time and went from, on average, a poor well-being score to an acceptable well-being score (above 50).

The importance of developing cost-effective, accessible, and engaging mental health interventions lies not only in the obvious benefits they have for adolescents and their families' well-being but also in the economic impact they may have. For example, the overall annual economic burden of depression among adults in the United States is estimated to be greater than US $326 billion [37], which were mainly accounted for by workplace costs (eg, missed days of work and reduced productivity while at work). Such studies among adolescents are scarce, but it was estimated that the annual societal cost of clinically referred adolescents ranged between US $42-66 million [38]. These costs were mainly attributed to their parents' loss of productivity and adolescents' school absence. When adding the economic impacts of other disorders, such as anxiety, these estimates are much higher. Thus, findings ways to efficiently treat, and more importantly prevent, mental health problems and increase well-being should be considered a priority for policy makers, health care providers, and entrepreneurs.

Limitations

The assessments were made through the Kai.ai platform, but we cannot infer that the improvement of users' well-being stemmed directly from the use of the service for several reasons. First, since the participants were all users who freely chose to use the service, there was no control group that was followed and studied for the same duration of the study. Therefore, users' improvement may represent a regression to the mean. Moreover, individuals who freely chose to use Kai.ai, a self-help tool, may have used other self-help tools or apps at the same time. Future studies should conduct a randomized control trial to better understand the effectiveness of the service compared to no intervention. Second, the use of common messaging apps for communicating with the users made it impossible to monitor whether they used the different process-oriented features such as breathing and ACT training. Therefore, we could not test their associations with users' well-being. Third, as the use of Kai.ai is anonymous, we did not have estimates regarding age, gender, and other demographic variables such as socioeconomic status, which limits the generalizability of the findings.

Conclusions

These initial results demonstrate the potential of a text-based conversational companion as a cost-effective and accessible tool to improve adolescents' well-being. Due to the great economic cost for poor well-being and mental health and the decrease in the accessibility of various support systems, partly due to the COVID-19 pandemic, developing efficient interventions should be considered a societal priority. Future studies should test if any of the process-oriented features of Kai.ai are beneficial for improving users' well-being or if the recognized increase in well-being represents a regression to the mean.

Conflicts of Interest

DV and NN are consultants for Kai.ai. MW declares no conflicts of interest.

References

1. Blakemore SJ. Adolescence and mental health. Lancet 2019 May 18;393(10185):2030-2031. [doi: 10.1016/S0140-6736(19)31013-X] [Medline: 31106741]
2. The Lancet Global Health. Mental health matters. Lancet Glob Health 2020 Nov;8(11):e1352 [FREE Full text] [doi: 10.1016/S2214-109X(20)30432-0] [Medline: 33069297]
3. González-Carrasco M, Casas F, Viñas F, Malo S, Gras ME, Bedin L. What leads subjective well-being to change throughout adolescence? an exploration of potential factors. Child Indic Res 2016 Jan 20;10(1):33-56. [doi: 10.1007/s12187-015-9359-6]
4. Deci E, Ryan RM. Hedonia, eudaimonia, and well-being: an introduction. J Happiness Stud 2008;9(1):1-11. [doi: 10.1007/s10902-006-9018-1]
5. Ben-Arieh A, Casas F, Frønes I, Korbin JE. Multifaceted concept of child well-being. In: Ben-Arieh A, Casas F, Frønes I, Korbin JE, editors. Handbook of Child Well-Being: Theories, Methods, and Policies in Global Perspective. Dordrecht, the Netherlands: Springer; 2014:1-27.
6. Ryff CD. Eudaimonic well-being, inequality, and health: recent findings and future directions. Int Rev Econ 2017 Jun;64(2):159-178 [FREE Full text] [doi: 10.1007/s12232-017-0277-4] [Medline: 29057014]
7. Sisask M, Värnik A, Kõlves K, Konstabel K, Wasserman D. Subjective psychological well-being (WHO-5) in assessment of the severity of suicide attempt. Nord J Psychiatry 2008 Jul 12;62(6):431-435. [doi: 10.1080/08039480801959273] [Medline: 18846444]
8. Stefanello S, Cais CFDS, Mauro M, Freitas G, Botega N. Gender differences in suicide attempts: preliminary results of the multisite intervention study on suicidal behavior (SUPRE-MISS) from Campinas, Brazil. Braz J Psychiatry 2008 Jun;30(2):139-143. [doi: 10.1590/s1516-44462006005000063] [Medline: 18176725]

9. Vijayakumar L, Ali ZSS, Umamaheswari C. Socio cultural and clinical factors in repetition of suicide attempts: a study from India. Int J Cult Ment Health 2008 Jun;1(1):3-9 [FREE Full text] [doi: 10.1080/17542860802102141]
10. Patalay P, Fitzsimons E. Development and predictors of mental ill-health and wellbeing from childhood to adolescence. Soc Psychiatry Psychiatr Epidemiol 2018 Dec 26;53(12):1311-1323. [doi: 10.1007/s00127-018-1604-0] [Medline: 30259056]
11. Booker CL, Kelly YJ, Sacker A. Gender differences in the associations between age trends of social media interaction and well-being among 10-15 year olds in the UK. BMC Public Health 2018 Mar 20;18(1):321 [FREE Full text] [doi: 10.1186/s12889-018-5220-4] [Medline: 29554883]
12. World Health Organization, Regional Office for Europe. Growing up unequal: gender and socioeconomic differences in young people's health and well-being. Health Behaviour in School-aged Children (HBSC) study: international report from the 2013/2014 survey. World Health Organization. 2016. URL: https://www.euro.who.int/__data/assets/pdf_file/0003/303438/HSBC-No.7-Growing-up-unequal-Full-Report.pdf [accessed 2022-11-03]
13. Meade T, Dowswell E. Adolescents' health-related quality of life (HRQoL) changes over time: a three year longitudinal study. Health Qual Life Outcomes 2016 Jan 25;14(1):1-8 [FREE Full text] [doi: 10.1186/s12955-016-0415-9] [Medline: 26810328]
14. Tol WA, Leku MR, Lakin DP, Carswell K, Augustinavicius J, Adaku A, et al. Guided self-help to reduce psychological distress in South Sudanese female refugees in Uganda: a cluster randomised trial. Lancet Glob Health 2020 Feb;8(2):e254-e263 [FREE Full text] [doi: 10.1016/S2214-109X(19)30504-2] [Medline: 31981556]
15. Stallman HM. Efficacy of the My Coping Plan mobile application in reducing distress: A randomised controlled trial. Clin Psychol 2020 Nov 09;23(3):206-212. [doi: 10.1111/cp.12185]
16. Hoffman CJ, Ersser SJ, Hopkinson JB, Nicholls PG, Harrington JE, Thomas PW. Effectiveness of mindfulness-based stress reduction in mood, breast- and endocrine-related quality of life, and well-being in stage 0 to III breast cancer: a randomized, controlled trial. J Clin Oncol 2012 Apr 20;30(12):1335-1342. [doi: 10.1200/JCO.2010.34.0331] [Medline: 22430268]
17. Odgers C. Smartphones are bad for some teens, not all. Nature 2018 Feb 22;554(7693):432-434 [FREE Full text] [doi: 10.1038/d41586-018-02109-8] [Medline: 29469108]
18. Rideout V. The Common Sense Census: media use by tweens and teens. Common Sense. 2015. URL: https://www.commonsensemedia.org/sites/default/files/research/report/census_researchreport.pdf [accessed 2022-11-03]
19. Bakker D, Kazantzis N, Rickwood D, Rickard N. Mental health smartphone apps: review and evidence-based recommendations for future developments. JMIR Ment Health 2016 Mar 01;3(1):1-31 [FREE Full text] [doi: 10.2196/mental.4984] [Medline: 26932350]
20. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017 Jun 06;4(2):1-11 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
21. Weisel KK, Fuhrmann LM, Berking M, Baumeister H, Cuijpers P, Ebert DD. Standalone smartphone apps for mental health-a systematic review and meta-analysis. NPJ Digit Med 2019 Dec 2;2(1):1-10 [FREE Full text] [doi: 10.1038/s41746-019-0188-8] [Medline: 31815193]
22. Krafft J, Potts S, Schoendorff B, Levin ME. A randomized controlled trial of multiple versions of an acceptance and commitment therapy matrix app for well-being. Behav Modif 2019 Mar;43(2):246-272. [doi: 10.1177/0145445517748561] [Medline: 29262693]
23. Landy LN, Schneider RL, Arch JJ. Acceptance and commitment therapy for the treatment of anxiety disorders: A concise review. Curr Opin Psychol 2015 Apr;2:70-74. [doi: 10.1016/j.copsyc.2014.11.004]
24. Naor N, Frenkel A, Winsberg M. Improving well-being with a mobile artificial intelligence-powered acceptance commitment therapy tool: pragmatic retrospective study. JMIR Form Res 2022 Jul 12;6(7):e36018 [FREE Full text] [doi: 10.2196/36018] [Medline: 35598216]
25. Jerath R, Edry JW, Barnes VA, Jerath V. Physiology of long pranayamic breathing: neural respiratory elements may provide a mechanism that explains how slow deep breathing shifts the autonomic nervous system. Med Hypotheses 2006;67(3):566-571. [doi: 10.1016/j.mehy.2006.02.042] [Medline: 16624497]
26. Hayes SC. Acceptance and Commitment Therapy, Relational Frame Theory, and the third wave of behavioral and cognitive therapies - republished article. Behav Ther 2016 Nov;47(6):869-885. [doi: 10.1016/j.beth.2016.11.006] [Medline: 27993338]
27. Wellbeing measures in primary health care/the DepCare Project: report on a WHO meeting, Stockholm, Sweden, 12-13 February 1998. World Health Organization. 1998. URL: https://www.euro.who.int/__data/assets/pdf_file/0016/130750/E60246.pdf [accessed 2022-11-03]
28. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 Well-Being Index: a systematic review of the literature. Psychother Psychosom 2015 Apr;84(3):167-176 [FREE Full text] [doi: 10.1159/000376585] [Medline: 25831962]
29. Andreasson K, Krogh J, Bech P, Frandsen H, Buus N, Stanley B, et al. MYPLAN -mobile phone application to manage crisis of persons at risk of suicide: study protocol for a randomized controlled trial. Trials 2017 Apr 11;18(1):1-7 [FREE Full text] [doi: 10.1186/s13063-017-1876-9] [Medline: 28399909]
30. Coelhoso CC, Tobo PR, Lacerda SS, Lima AH, Barrichello CRC, Amaro E, et al. A new mental health mobile app for well-being and stress reduction in working women: randomized controlled trial. J Med Internet Res 2019 Nov 07;21(11):e14269 [FREE Full text] [doi: 10.2196/14269] [Medline: 31697244]

31. Laird EA, Ryan A, McCauley C, Bond RB, Mulvenna MD, Curran KJ, et al. Using mobile technology to provide personalized reminiscence for people living with dementia and their carers: appraisal of outcomes from a quasi-experimental study. JMIR Ment Health 2018 Sep 11;5(3):e57 [FREE Full text] [doi: 10.2196/mental.9684] [Medline: 30206053]
32. Allgaier AK, Pietsch K, Frühe B, Prast E, Sigl-Glöckner J, Schulte-Körne G. Depression in pediatric care: is the WHO-Five Well-Being Index a valid screening instrument for children and adolescents? Gen Hosp Psychiatry 2012 May;34(3):234-241. [doi: 10.1016/j.genhosppsych.2012.01.007] [Medline: 22325631]
33. Blom EH, Bech P, Högberg G, Larsson JO, Serlachius E. Screening for depressed mood in an adolescent psychiatric context by brief self-assessment scales--testing psychometric validity of WHO-5 and BDI-6 indices by latent trait analyses. Health Qual Life Outcomes 2012 Dec 11;10:1-6 [FREE Full text] [doi: 10.1186/1477-7525-10-149] [Medline: 23227908]
34. de Wit M, Pouwer F, Gemke R, Delemarre-van de Waal HA, Snoek F. Validation of the WHO-5 Well-Being Index in adolescents with type 1 diabetes. Diabetes Care 2007 Aug;30(8):2003-2006. [doi: 10.2337/dc07-0447] [Medline: 17475940]
35. Rosseel Y. Iavaan: an R package for structural equation modeling. J Stat Soft 2012 May 24;48(2):1-36. [doi: 10.18637/jss.v048.i02]
36. R Core Team. R: a language and environment for statistical computing, reference index version 3.5.1. R Foundation for Statistical Computing. 2018. URL: https://www.r-project.org/ [accessed 2022-11-03]
37. Greenberg PE, Fournier AA, Sisitsky T, Simes M, Berman R, Koenigsberg SH, et al. The economic burden of adults with major depressive disorder in the United States (2010 and 2018). Pharmacoeconomics 2021 Jun 05;39(6):653-665 [FREE Full text] [doi: 10.1007/s40273-021-01019-4] [Medline: 33950419]
38. Bodden DHM, Stikkelbroek Y, Dirksen CD. Societal burden of adolescent depression, an overview and cost-of-illness study. J Affect Disord 2018 Dec 01;241:256-262. [doi: 10.1016/j.jad.2018.06.015] [Medline: 30138810]

Abbreviations

ACT: Acceptance Commitment Therapy
AI: artificial intelligence
mHealth: mobile health
T: time point
WHO-5: 5-item World Health Organization Well-being Index

Edited by K El Emam, B Malin; submitted 21.03.22; peer-reviewed by S Fudickar, Z Dai; comments to author 13.07.22; revised version received 26.10.22; accepted 29.10.22; published 29.11.22.

Please cite as:
Vertsberger D, Naor N, Winsberg M
Adolescents' Well-being While Using a Mobile Artificial Intelligence–Powered Acceptance Commitment Therapy Tool: Evidence From a Longitudinal Study
JMIR AI 2022;1(1):e38171
URL: https://ai.jmir.org/2022/1/e38171
doi:10.2196/38171
PMID:

Original Paper

Artificial Intelligence–Assisted Diagnosis of Anterior Cruciate Ligament Tears From Magnetic Resonance Images: Algorithm Development and Validation Study

Kun-Hui Chen[1,2,3], MD; Chih-Yu Yang[1,4], MD, PhD; Hsin-Yi Wang[2,5], MD; Hsiao-Li Ma[2,3], MD; Oscar Kuang-Sheng Lee[1,6], MD, PhD

[1]Institute of Clinical Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan
[2]Department of Surgery, School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan
[3]Department of Orthopedics and Traumatology, Taipei Veterans General Hospital, Taipei, Taiwan
[4]Division of Nephrology, Department of Medicine, Taipei Veterans General Hospital, Taipei, Taiwan
[5]Department of Anaesthesiology, Taipei Veterans General Hospital, Taipei, Taiwan
[6]China Medical University Hospital, Taichung, Taiwan

Corresponding Author:
Oscar Kuang-Sheng Lee, MD, PhD
Institute of Clinical Medicine
National Yang Ming Chiao Tung University
No 155, Sec 2, Linong Street
Taipei, 11221
Taiwan
Phone: 886 2 28757391
Email: oscarlee9203@gmail.com

Abstract

Background: Anterior cruciate ligament (ACL) injuries are common in sports and are critical knee injuries that require prompt diagnosis. Magnetic resonance imaging (MRI) is a strong, noninvasive tool for detecting ACL tears, which requires training to read accurately. Clinicians with different experiences in reading MR images require different information for the diagnosis of ACL tears. Artificial intelligence (AI) image processing could be a promising approach in the diagnosis of ACL tears.

Objective: This study sought to use AI to (1) diagnose ACL tears from complete MR images, (2) identify torn-ACL images from complete MR images with a diagnosis of ACL tears, and (3) differentiate intact-ACL and torn-ACL MR images from the selected MR images.

Methods: The sagittal MR images of torn ACL (n=1205) and intact ACL (n=1018) from 800 cases and the complete knee MR images of 200 cases (100 torn ACL and 100 intact ACL) from patients aged 20-40 years were retrospectively collected. An AI approach using a convolutional neural network was applied to build models for the objective. The MR images of 200 independent cases (100 torn ACL and 100 intact ACL) were used as the test set for the models. The MR images of 40 randomly selected cases from the test set were used to compare the reading accuracy of ACL tears between the trained model and clinicians with different levels of experience.

Results: The first model differentiated between torn-ACL, intact-ACL, and other images from complete MR images with an accuracy of 0.9946, and the sensitivity, specificity, precision, and F1-score were 0.9344, 0.9743, 0.8659, and 0.8980, respectively. The final accuracy for ACL-tear diagnosis was 0.96. The model showed a significantly higher reading accuracy than less experienced clinicians. The second model identified torn-ACL images from complete MR images with a diagnosis of ACL tear with an accuracy of 0.9943, and the sensitivity, specificity, precision, and F1-score were 0.9154, 0.9660, 0.8167, and 0.8632, respectively. The third model differentiated torn- and intact-ACL images with an accuracy of 0.9691, and the sensitivity, specificity, precision, and F1-score were 0.9827, 0.9519, 0.9632, and 0.9728, respectively.

Conclusions: This study demonstrates the feasibility of using an AI approach to provide information to clinicians who need different information from MRI to diagnose ACL tears.

(J AI 2022;1(1):e37508) doi:10.2196/37508

KEYWORDS

artificial intelligence; convolutional neural network; magnetic resonance imaging; MRI; deep learning; anterior cruciate ligament; sports medicine; machine learning; ligament; sport; diagnosis; tear; damage; imaging; development; validation; algorithm

Introduction

The anterior cruciate ligament (ACL), an important ligament of the knee joint, is a common and devastating sports injury that affects more than 200,000 people in the United States annually [1,2]. The early and proper diagnosis of ACL tears is crucial and can lead to early intervention to prevent subsequent chondral or meniscal damage and early osteoarthritis [3]. A neglected diagnosis can cause longer chronicity of ACL tears at the time of surgery and is positively correlated with the development of osteoarthritis [4]. Arthroscopy can directly visualize the intra-articular lesions of the knee and is the most accurate diagnostic tool for ACL tears [5]. However, this is an invasive procedure with potential surgical risks.

Magnetic resonance imaging (MRI) is a strong, noninvasive tool for detecting ACL tears with high sensitivity and specificity if interpreted by an experienced musculoskeletal radiologist [6,7]. However, reading MR images and making an accurate diagnosis of ACL tears are challenging for less experienced medical personnel.

Graphic identification using deep learning is an important and integral part of artificial intelligence (AI). Using a convolutional neural network (CNN) with repeated input and output data, established algorithms can learn layers of features and repeatedly adjust their neural network and thereby model the complex relationships between medical images and their interpretations [8]. CNNs may be useful in medical imaging tasks; thus, the development of a computer-assisted tool to detect ACL tears from MR images may be helpful in reducing doctor workload, increasing education, reducing misdiagnosis, and enhancing the quality of health care in resource-limited areas [9].

In this study, we aimed to use AI to (1) diagnose ACL tears from complete MR images (for those who were not trained to read knee MRI but nevertheless wanted to diagnose it); (2) identify torn-ACL images from complete MR images that have a diagnosis of an ACL tear (for those who need advanced information after they obtain the result of an ACL tear from the first model); and (3) differentiate torn-ACL and intact-ACL images from the selected MR images (for those who were able to identify the images containing ACL but do not have sufficient confidence in making the diagnosis).

Methods

Ethics Approval

This retrospective study was approved by the institutional review board of Taipei Veterans General Hospital (2018-11-005CC).

Patient Selection and Database

The sagittal MR images of torn ACL (n=1205) and intact ACL (n=1018) from 800 cases and the complete knee MR images of 200 cases (100 torn ACL and 100 intact ACL; torn- and intact-ACL images were extracted, n=335,742) of patients who underwent knee MRI examinations between January 2013 and December 2017 were retrospectively collected for training purposes (training set). The complete MR images of 200 independent cases (100 torn ACL and 100 intact ACL; n=34,914) were used for testing purpose (testing set). The mean age of these patients was 28.1 years and 66.4% (664/1000) were male. The patient population was similar to previous reports on the group with the higher prevalence of ACL tears [10]. We believe these models have routine applications in a majority of patient groups.

Knee MR images excluded patients with the following knee conditions: tumor around the knee, previous knee surgery, multiple ligament injuries, osteoarthritis (Kellgren-Lawrence classification grades 2 to 4), and previous fractures around the knee. MRI examinations were performed on the knee, either in our hospital or in other hospitals, and were then uploaded to our system for a second opinion. There were 6 different MRI scanners used to perform knee examination in our hospital, and we did not restrict the scanner from which we obtained the images. Moreover, we did not identify the scanners in the uploaded images. In this database, for the torn-ACL MRIs, 76.8% (384/500) were performed in our hospital and 23.2% (116/500) were from other hospital; and for the intact-ACL MRIs, 84.6% (423/500) were performed in our hospital and 15.4% (77/500) were from other hospital. Hence, the images used in this study were not restricted to one hospital or a specific MRI scanner.

The determination of a torn-ACL or intact-ACL case was formulated independently by 2 orthopedic doctors and 1 musculoskeletal radiologist who reviewed the MR images and issued the report officially. In addition, torn-ACL cases were also confirmed through arthroscopic examination as all patients with ACL tears underwent arthroscopic ACL reconstruction surgery. All 3 doctors had consistent opinions on the sagittal torn-ACL (Figure 1) and intact-ACL (Figure 2) images.

Figure 1. MR images of different torn-ACL patterns. Sagittal proton density images from 6 different patients show variations in the patterns of torn ACL on their respective images: (A) proximal third tear; (B) mid-substance tear; (C) distal third tear; (D) chronic tear with complete ligament resorption, such as ligament disappearance; (E) tear with folded ligament, which may cause extension difficulty; and (F) tear with cyst formation. White arrow: lesion site. ACL: anterior cruciate ligament; MR: magnetic resonance.

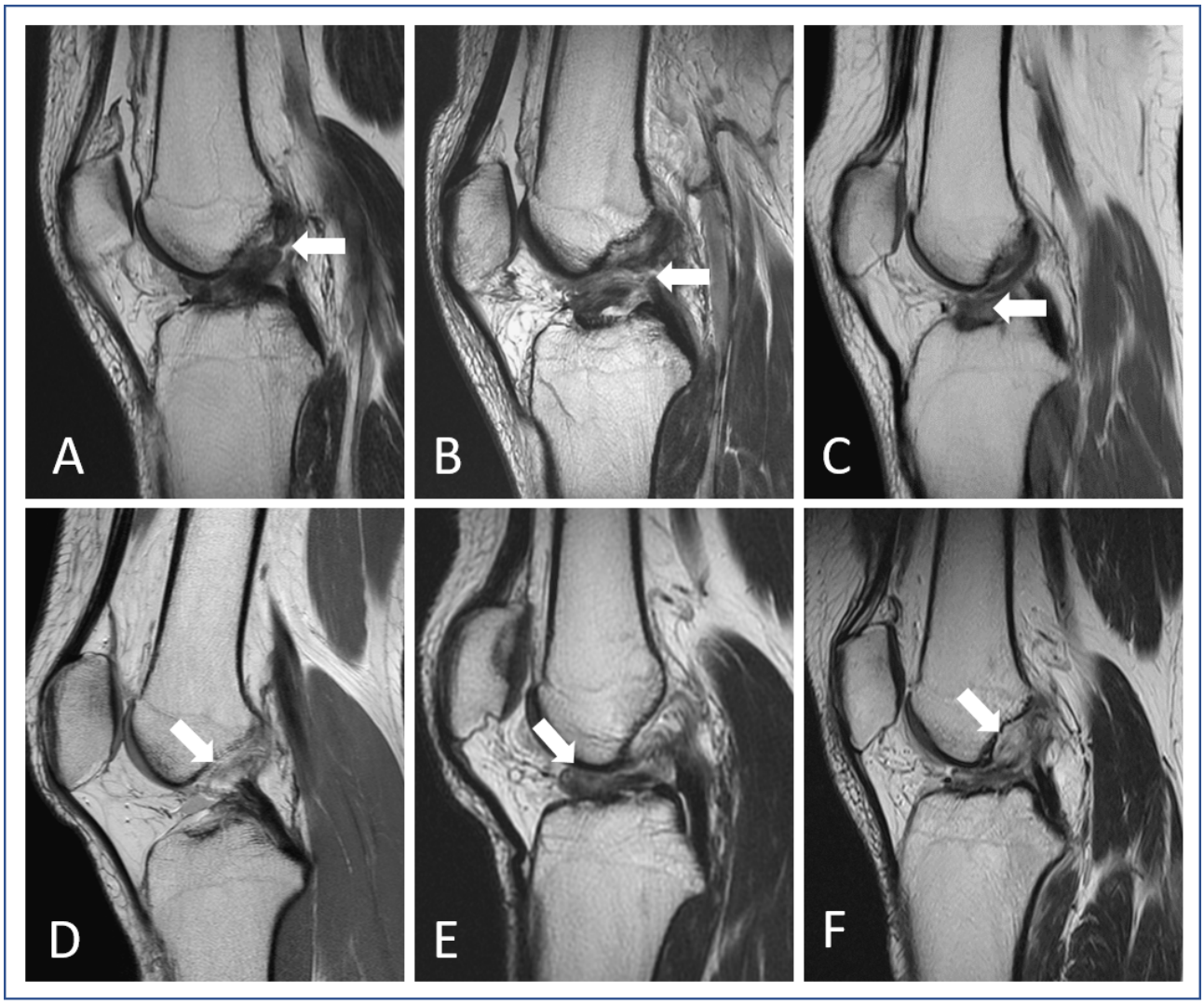
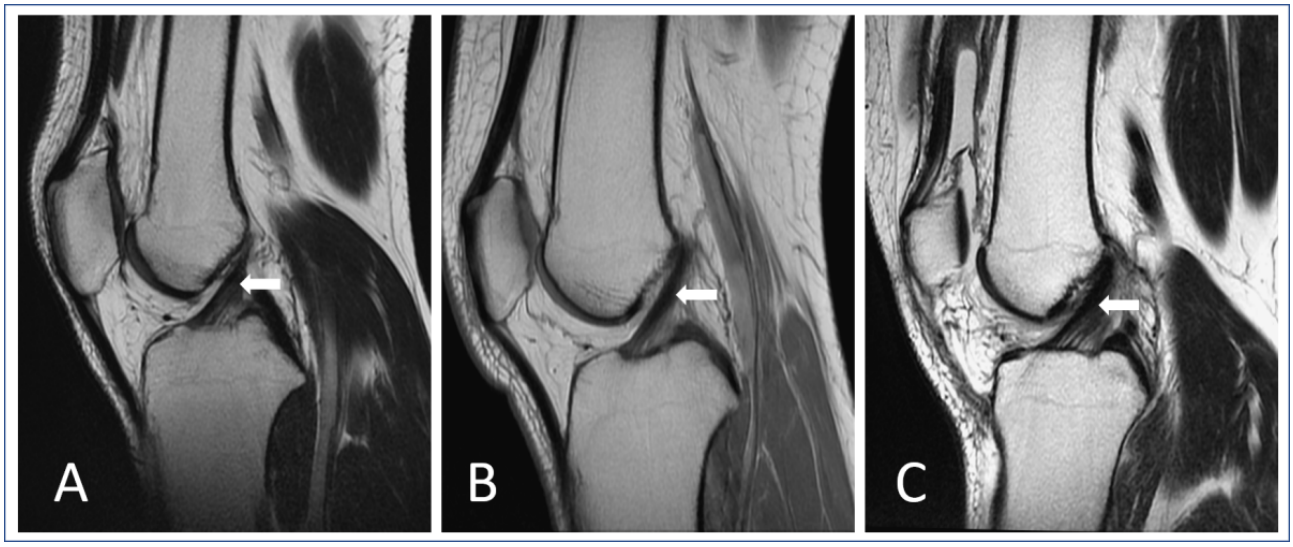


Figure 2. MR images of intact ACL. Sagittal proton density images of 3 different patients are shown. All images show the taut and straight bands parallel to the intercondylar roof with low signal intensity patterns of the intact ACL (white arrow). ACL: anterior cruciate ligament; MR: magnetic resonance.



The first model was for clinicians who were not trained to read knee MR images but wanted to know if the ACL was torn. For this purpose, we first trained a CNN model to differentiate between torn-ACL, intact-ACL, and other images from complete MR images of the knee. The sagittal MR images of torn and intact ACL from 800 cases and the images from 200 complete knee MR images (the torn- and intact-ACL images were extracted), regarded as other images, were used to train and validate the model (Table 1). Cases containing intact-ACL images or both intact- and torn-ACL images were regarded as intact-ACL cases, and cases containing torn-ACL images only were regarded as tear cases. This is similar to the strategy often used by some readers; if an intact-ACL image could be identified among complete MR images, then it might indicate that there is less probability of a torn ACL. Instead, if an intact ACL could not be found when examining the knee MRI of a patient, it would be indicative of a torn ACL.

As the first model did not provide information for identifying torn-ACL images, a second model was developed to identify them from complete MR images that had been diagnosed as ACL-tear case from the first model. Thus, the second model was intended for personnel who needed advanced information on torn-ACL images after obtaining the ACL-tear results. For this purpose, torn-ACL images and other images from 100 ACL-tear cases in the training set were used for training and validation (Table 2).

The third model was used to differentiate between torn-ACL and intact-ACL images from the selected MR images. This model was used by more experienced readers who were able to identify the sagittal images that contained ACLs but needed assistance in making the correct diagnosis. For this purpose, the sagittal MR images of torn and intact ACLs were included for training purposes (Table 3).

Table 1. Number of images used for training, validating, and testing the model to differentiate intact-ACL, torn-ACL, and other images from the complete magnetic resonance images.

Classification	Training and validation, n	Test, n
Intact-ACL[a] images	1018	270
Torn-ACL images	1205	346
Other images	335,742[b]	34,298[c]

[a]ACL: anterior cruciate ligament.

[b]Including sagittal, coronal, and axial images (torn- and intact-ACL images were extracted) from the training set (200 cases).

[c]Including sagittal, coronal, and axial images (torn- and intact-ACL images were extracted) from the test set (200 cases).

Table 2. Number of images used for training, validating, and testing the model to identify torn-ACL images from ACL-tear cases.

Classification	Training and validation, n	Test, n
Torn-ACL[a] images	1205	346
Other images	15,969[b]	16,800[c]

[a]ACL: anterior cruciate ligament.

[b]Including sagittal, coronal, and axial images (torn-ACL images were extracted) from 100 ACL-tear cases in the training set

[c]Including sagittal, coronal, and axial images (torn-ACL images were extracted) from 100 ACL-tear cases in the testing set.

Table 3. Number of images used for training, validating, and testing to differentiate between torn- and intact-ACL images.

Classification	Training and validation, n	Test, n
Intact-ACL[a] images	1018	270
Torn-ACL images	1205	346

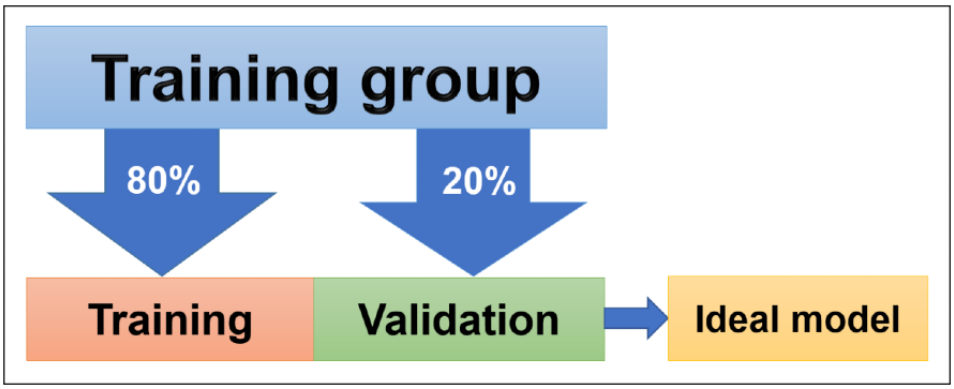
[a]ACL: anterior cruciate ligament.

Image Preprocessing and CNN Model Training by an Automatic Deep-Learning Software

All images were downloaded from the imaging system as a 256 × 256-pixel image in a portable network graphics format and subsequently grouped, as previously mentioned, for training the 3 different CNN models. The AI approach used MAIA automatic deep learning software for medical imaging analyses (version 1.2.0; Muen Biomedical and Optoelectronic Technologies Inc), which was used in a previous study [11]. The CNN model of MAIA was based on EfficientNet-B0, pretrained with ImageNet [12,13]. After inputting the MR images of the training group, 80% of the images were distributed to train and 20% were distributed to validate and find the most ideal CNN model (Figure 3). The MR images were then augmented with horizontal flipping and Gaussian noise [14]. The dropout function and different data augmentation methods were added to prevent the model from overfitting in the data set [15,16]. For hyperparameters in training, the number of epochs was set as 100, the batch size was selected automatically based on memory consumption, and the learning rate was dynamically scheduled through cosine annealing and a 1-cycle

policy [17,18]. The network was trained end-to-end using the Adam optimization algorithm, which optimized the cross-entropy as a loss function [19]. For classification, the softmax or sigmoid layer was applied as the output layer in multiclass or binary classification, respectively. The MAIA analysis was performed with Python (version 3.x; Python Software Foundation) and PyTorch (version 1.1.x; Meta AI) on a Windows 10 laptop with GeForce RTX2070 graphic cards (8 GB GDDR6 RAM, GT63 Titan 8SF; MSI).

Figure 3. Data organization for model training.



CNN Models Performance Evaluation

To evaluate how the model differentiated between torn-ACL, intact-ACL, and other images, the 200 independent cases were used to test the model. To evaluate the accuracy of an ACL-tear diagnosis, cases that were identified as containing intact-ACL images were regarded as intact-ACL cases, and the rest were diagnosed as ACL tears (Figure 4). To evaluate the secondary model of identifying torn images from cases diagnosed with ACL tears, 100 ACL-tear cases from the independent test set were used for testing purposes (Figure 5). To evaluate the third model of differentiating intact-ACL and torn-ACL images from the selected MR images, sagittal MR images labeled as torn and intact ACL from the independent test set were used (Figure 6). Finally, we compared the performance of the first model to diagnose ACL tears with those of orthopedic residents and medical students. For this purpose, 40 randomly selected cases (20 torn and 20 intact) from the test set were used to test differently experienced readers (ie, orthopedic residents and medical students). Complete images were provided to the readers after the removal of personal, clinical, surgical, and institutional information to focus on the reading of the MRI. The residents were split into 3 groups: Group 1 (chief residents and sports fellows), Group 2 (third- and fourth-year residents), and Group 3 (first- and second-year residents). There were 5 participants in each group. We excluded the highest and lowest accuracy results for each group, and the accuracy of each group is the mean accuracy of the 3 readers. The resultant accuracies of the machine and differently experienced readers were compared.

Figure 4. Flowchart of diagnosing ACL tears using the AI approach. ACL: anterior cruciate ligament; AI: artificial intelligence; MRI: magnetic resonance imaging.

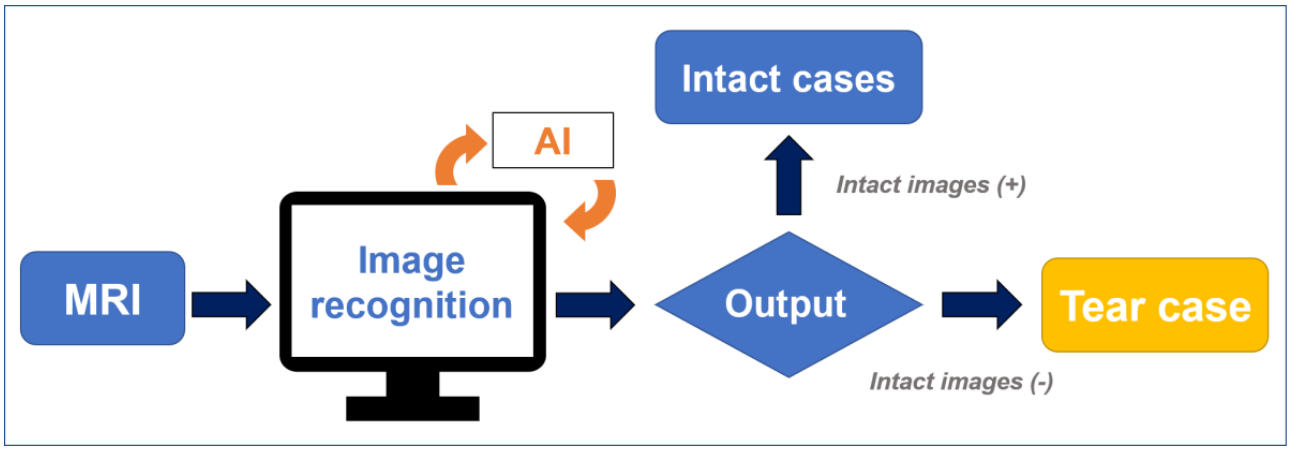


Figure 5. Flowchart of identifying torn-ACL images using the AI approach. ACL: anterior cruciate ligament; AI: artificial intelligence.

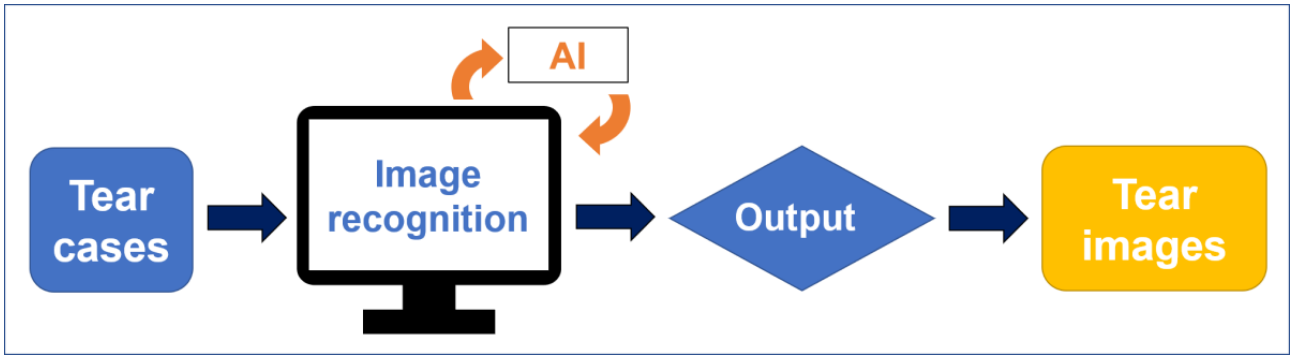
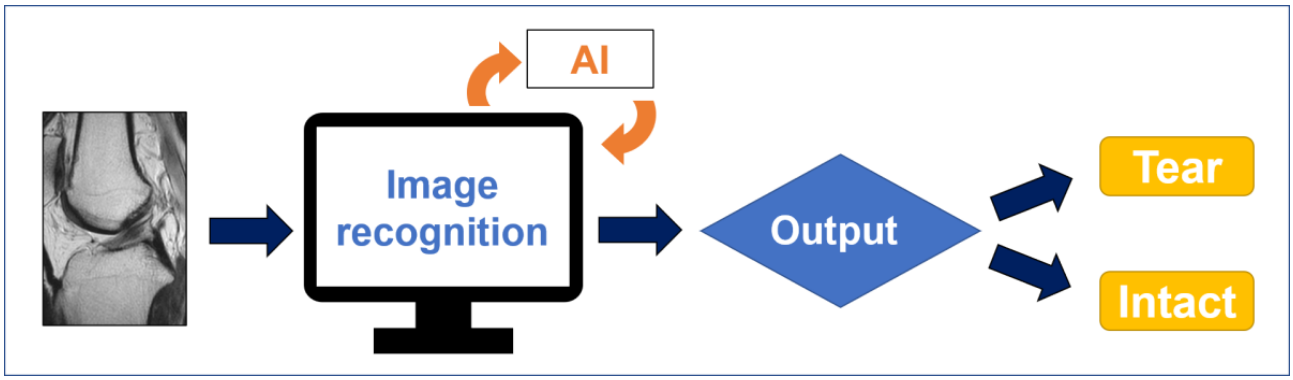


Figure 6. Flowchart of differentiating intact-ACL and torn-ACL images using the AI approach. ACL: anterior cruciate ligament; AI: artificial intelligence.



Statistical Analysis

The effectiveness of the 3 models was evaluated using several metrics, including the accuracy, sensitivity, specificity, F1-score, receiver operating characteristic curve, and the area under the curve, which were calculated using Python. The comparison of the models and doctors with different degrees was performed using SPSS software package (version 22; IBM Corp). Statistical significance was set at $P<.05$, with a 95% CI.

Results

The accuracy of the model that differentiated between torn-ACL, intact-ACL, and other images was 0.9946. The sensitivity, specificity, precision, and F1-scores were 0.9344, 0.9743, 0.8659, and 0.9980, respectively (Table 4 and Figure 7). The accuracy of ACL diagnosis was 0.96 (Figure 8). The accuracy of the model identifying torn-ACL images from the complete images of ACL-tear cases was 0.9943. The sensitivity, specificity, precision, and F1-scores were 0.9154, 0.9660, 0.8167, and 0.8632, respectively. (Table 4 and Figure 9). The accuracy of the model that differentiated torn- and intact-ACL images was 0.9691. The sensitivity, specificity, precision, and F1-scores were 0.9827, 0.9519, 0.9632, and 0.9782, respectively (Table 4 and Figure 10).

The accuracy of the first model and the differently experienced orthopedic residents and medical students for the diagnosis of ACL tears is shown in Table 5. When using the 40 randomly selected cases from the test set for reading comparison, the results showed a significantly higher reading accuracy for the model than those of the less experienced residents and medical students.

Table 4. Validation and test results for the 3 models.

Model	Torn-ACL[a], intact-ACL, and other images differentiation		ACL-tear image identification		ACL-tear or intact images differentiation	
	Validation	Test	Validation	Test	Validation	Test
Accuracy	0.9947	0.9946	0.9959	0.9943	1.0000	0.9691
Sensitivity	0.9702	0.9344	0.9834	0.9154	1.0000	0.9827
Specificity	0.9884	0.9743	0.9969	0.9660	1.0000	0.9519
Precision	0.9647	0.8659	0.9595	0.8167	1.0000	0.9632
F1-score	0.9674	0.8980	0.9713	0.8632	1.0000	0.9728

[a]ACL: anterior cruciate ligament.

Figure 7. Performance of the model in differentiating torn-ACL, intact-ACL, and other images. (A) Confusion matrix; (B) ROC curve of the model; (C) Precision recall curve for identifying torn-ACL images; (D) Precision recall curve for identifying intact-ACL images; and (E) Precision recall curve for identifying other images (images without torn or intact ACL). ACL: anterior cruciate ligament; ROC: receiver operating characteristic.

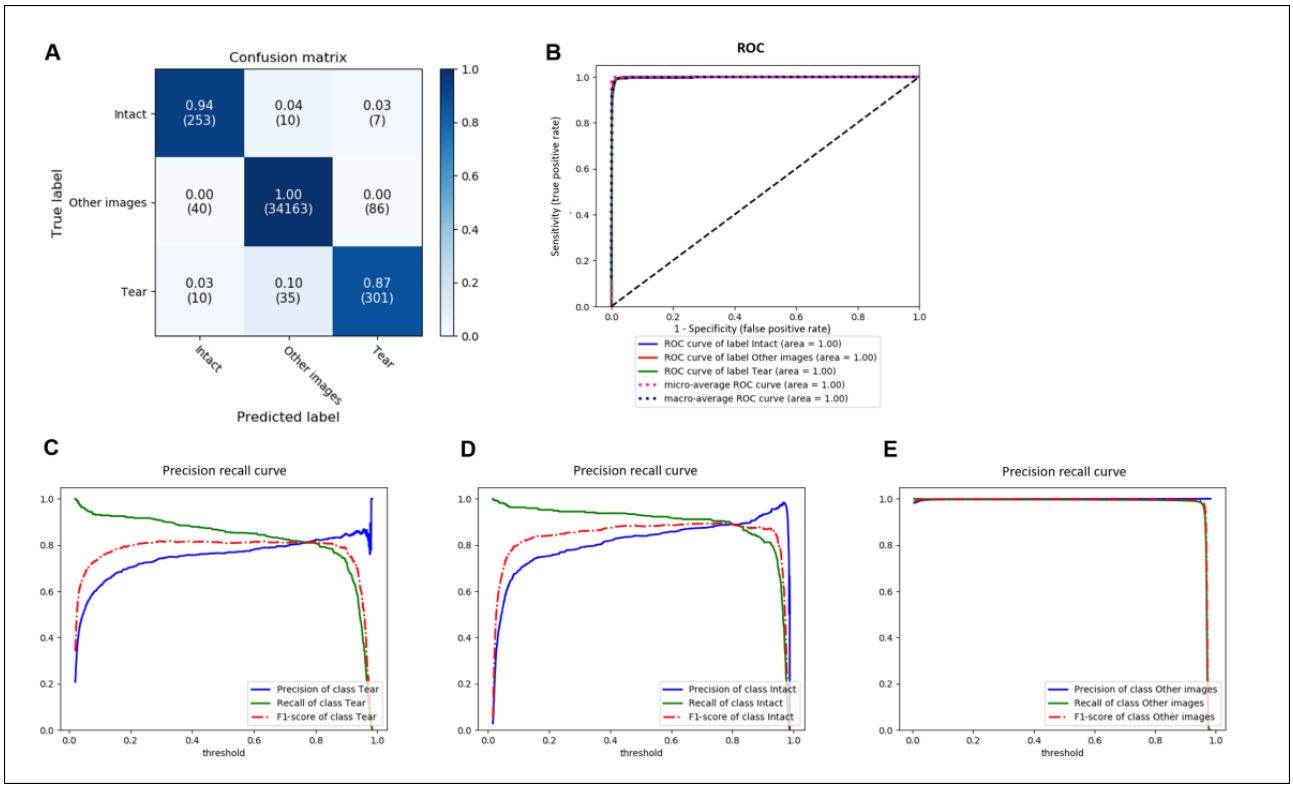
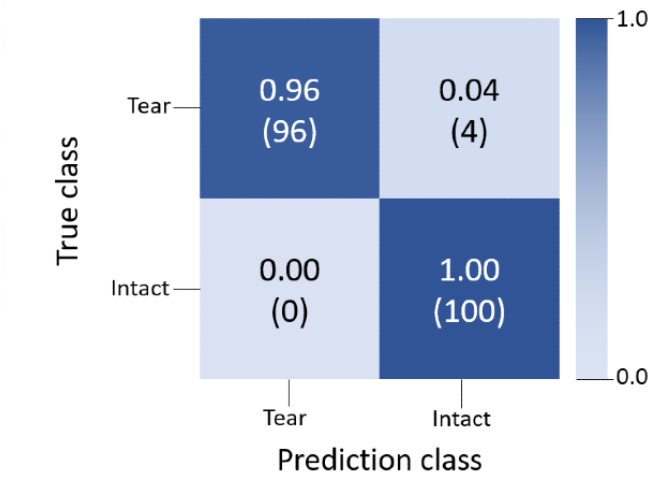


Figure 8. Classification matrix for diagnosing ACL-tear cases. ACL: anterior cruciate ligament.



Accuracy	Sensitivity	Specificity
0.98	0.96	1.00

Figure 9. Performance of the model in identifying torn-ACL images from complete MRI images with an ACL-tear diagnosis. (A) Confusion matrix; (B) ROC curve of the model; and (C) Precision recall curve. ACL: anterior cruciate ligament; ROC: receiver operating characteristic.

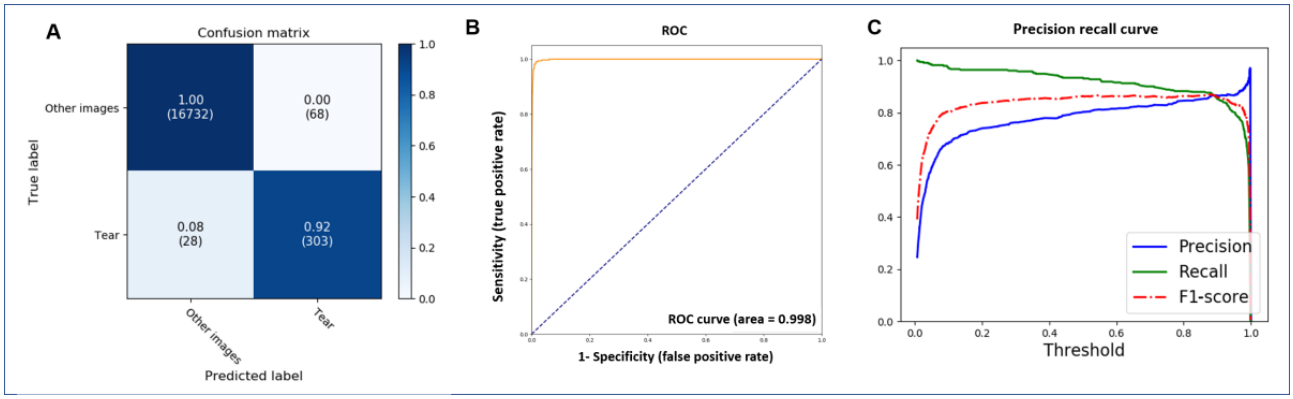


Figure 10. Performance of the model in differentiating between intact-ACL and torn-ACL images. (A) Confusion matrix; (B) ROC curve of the model; (C) Precision recall curve; and (D) torn-ACL image identified (left) and its representative heat map (right). ACL: anterior cruciate ligament; ROC: receiver operating characteristic.

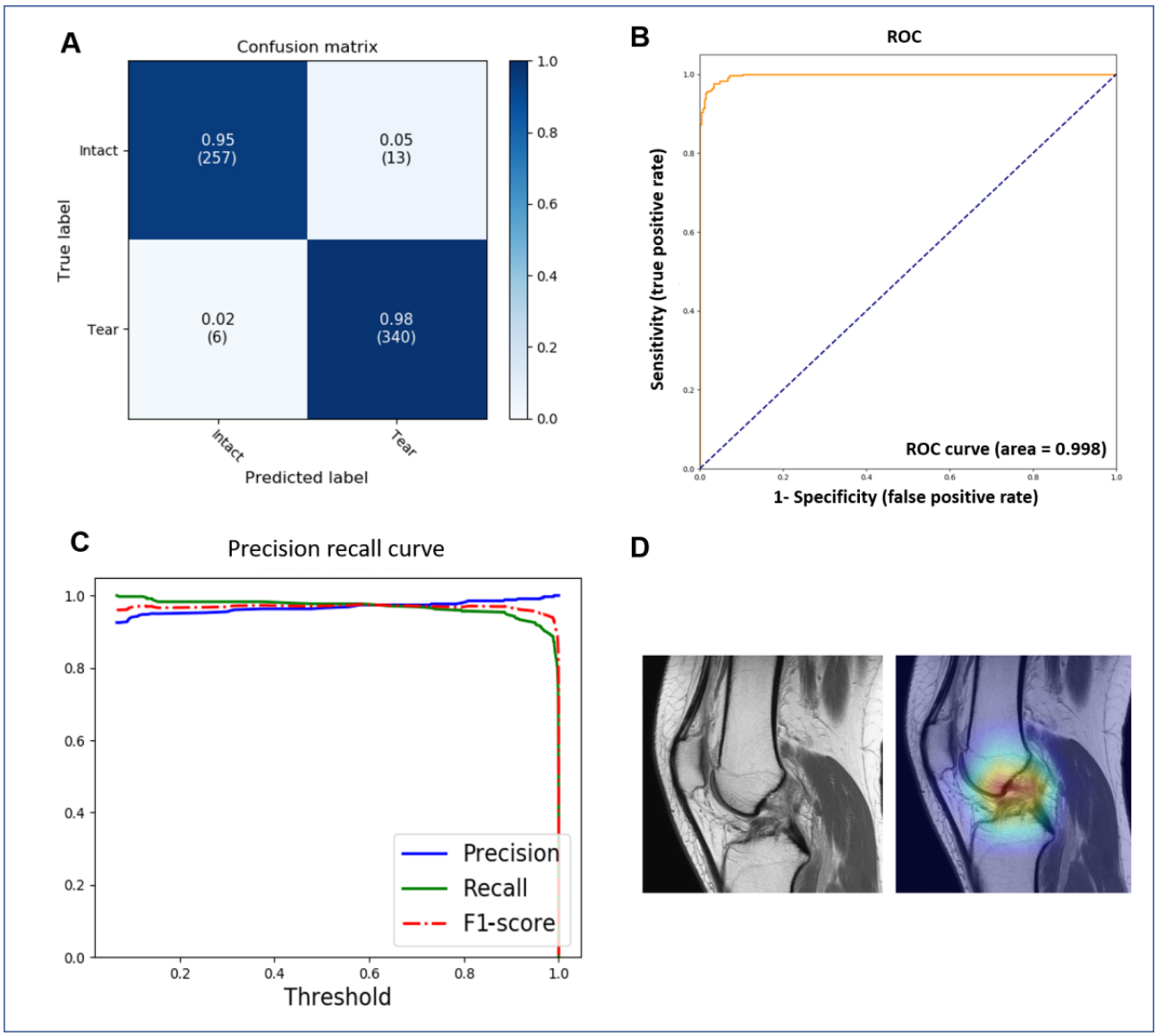


Table 5. Accuracy of the model and the differently experienced orthopedic residents and medical students in the diagnosis of anterior cruciate ligament tears in 40 randomly selected magnetic resonance imaging cases.

Reader	Accuracy, mean	*P* value[a]
Machine	0.975	Reference[b]
Group 1: chief residents and sports fellows (n=3)	0.888	.13
Group 2: third- and fourth-year residents (n=3)	0.817	.02
Group 3: first- and second-year residents (n=3)	0.742	.003
Medical students (n=3)	0.708	.001

[a]*P* values were based on statistical analyses using the chi-squared test. Statistical significance was set at $P<.05$.

[b]The accuracy of machine reading was used as a reference.

Discussion

Principal Findings

This study demonstrates the feasibility of using an AI approach to diagnose ACL tears from complete MR images with 96% accuracy, identify torn-ACL images from ACL tear cases with 99.4% accuracy, and differentiate intact-ACL and torn-ACL images from the selected MR images with 96.9% accuracy. The model also demonstrated a significantly higher diagnostic accuracy than orthopedic residents in training and medical students.

MRI is a highly accurate tool for evaluating ACL tears, with an accuracy, sensitivity, and specificity of more than 90% [20,21]. In a complete MR scan, the knee should ideally be imaged in 3 orthogonal planes: sagittal, coronal, and axial slices. During the examination, the patient was positioned supine in the scanner, with the knee relaxed in mild flexion and slight external rotation (5°-10°). This position enables the ACL to be orthogonal to the sagittal plane of imaging [22]. Therefore, of all 3 planes, sagittal plane images show the ACL most clearly, especially with T2-weighted sequences [23]. When reading knee MR images in clinical practice, sagittal images are more commonly used to evaluate the condition of the ACL than the other planes. For this reason, we chose to use the sagittal images of the intact or torn ACL as the target for the AI approach to develop the 3 models.

In a normal knee, the ACL is between the lateral femoral condyle and the anterior midportion of the tibia and attaches the anterior to the tibial spine. Sagittal MR images appear as a taut and straight band parallel to the intercondylar roof (Blumensaat line) and have low signal intensity on T1- and T2-weighted images (Figure 2). However, compared to intact-ACL images, there are many variations in the torn-ACL sign on the MR images. These variations include discontinuity in the different parts of the ligament (proximal, midsubstance, or distal) [24], abnormally increased signal intensity, and abnormal morphology, such as a wave, fold, or angulation. In chronic tears, the ACL can even be nonvisualized owing to the resorption of the torn ligament (Figure 1). Thus, the variable appearance of torn-ACL images makes them more complicated to read than intact-ACL images. In the first model, the results showed that the model had less accuracy in identifying torn-ACL images than intact-ACL images (0.87 vs 0.94). There was more misprediction of other images as torn-ACL images, and many of these mispredictions occurred in the intact-ACL cases, identifying both intact-ACL and torn-ACL images as intact-ACL cases (19 cases). However, there was less misprediction of other images as intact-ACL images in ACL-tear cases (4 cases). All the results reflected the variations in torn-ACL images. Accordingly, for the purpose of diagnosing ACL-tears, cases containing intact-ACL images were regarded as intact cases because the model identified them with a higher accuracy. The other cases without intact-ACL images were regarded as tear cases. By using this principle to exclude ACL-tear cases, the accuracy of the diagnosis of ACL tears could reach 96%, which is comparable to many studies using different AI approaches [25-27]. This method can be helpful for personnel who are not trained to read the knee MRI but want to know if the ACL is torn. In addition to diagnosing ACL tears, this study also demonstrates the feasibility of identifying ACL images from complete MR images of ACL-tear cases and differentiating intact- and torn-ACL images with a good accuracy and F1-score. These models can be useful for various user needs.

A total of 40 cases were randomly selected from the test set for the reading of the model and from differently experienced residents and medical students. The images provided for each case were complete MRI examinations, which included all planes and sequences. The results showed that the accuracy of the model in diagnosing ACL-tear cases was significantly higher than that of medical students and orthopedic residents in training. Reading MR images to identify ACL tears is relatively routine for attending orthopedic surgeons or radiologists. However, for less experienced readers, the model may provide a useful reference when they are uncertain of the diagnosis.

In this study, we did not extract images from only 1 specific MR scanner. This is because, in daily practice, a hospital may have multiple scanners, and sometimes a physician may need to read MR images from an unknown scanner from another hospital. The MR images for this study were obtained using 6 different MR knee scanners in our institute, which were obtained from 2 different companies and purchased in different years. In addition to MR images that were obtained in our hospital, images were also taken from other hospitals and uploaded to our image system when the patients came for a second opinion or asked for surgery. Therefore, our data set comprised images from different scanners, and it was less likely that the model would learn some artifacts from the scanners that are not related to the ACL condition. We demonstrated that the models can

perform well for an independent test set that contains MR images from different scanners.

Comparison With Prior Work

Using a deep-learning approach to detect ACL tears has been reported with an accuracy exceeding 95% in many studies using different AI approaches [25-28]. Nonetheless, there were some novelties in this study that we consider to be comparable for their use in daily practice. First, we extracted images from heterogeneous MR scanners. In previous studies, only 1 or 2 scanners were used; however, it is uncommon that there are only 1 or 2 MRI scanners in an institution. Thus, developing a deep-learning algorithm that is trained with images from different MR scanners may better represent real-world situations in many hospitals. For the independent test set, we used the complete images of the MRI examination, and there was no restriction on the protocol used by the scanner, which is different from previous studies. Second, we used a different approach to diagnose the ACL injuries. We excluded the cases containing the intact-ACL images, which were identified by the AI approach, to diagnose ACL-tear cases with an accuracy of 96%. Third, we developed 3 different models for users with different purposes: (1) to diagnose ACL tears from complete MR images; (2) to identify torn-ACL images from complete MR images with a diagnosis of ACL tears; and (3) to differentiate intact-ACL and torn-ACL MR images from the selected images. Users with different experiences require different types of help. These 3 models are tailored to assist users with different needs by providing them with relevant information using an AI approach, which has not been previously reported.

Limitations

Our study has several limitations. First, we did not label the partially torn–ACL images. Partial tears of the ACL are more difficult to diagnose than complete tears, and the accuracy of these diagnoses is poor on MR images [29]. Thus, we did not use the images of partial tears for training or testing in this study. However, should a partial tear case be input into the model, the model could diagnose the case as an ACL tear because this model cannot identify an intact-ACL image. This finding may alert the user that the case is a torn-ACL case, and the case may need to be double-checked by an orthopedic specialist. Second, we used only sagittal torn-ACL and intact-ACL images for the diagnosis of ACL tears. Considering that the images of other planes can also assist in the diagnosis, adding the other planes of images into the training might increase the reading accuracy. Third, we did not record the details of the MR scanners, because the information of the scanners of the images taken from other hospitals could not be identified.

Conclusions

This study demonstrates the feasibility of using an AI approach to diagnose ACL tears from a complete MR image (with 96.0% accuracy), identify torn-ACL images from ACL-tear cases, and differentiate intact-ACL and torn-ACL images from the selected MR images. These models may serve as clinical decision support systems for diagnosing ACL injuries for clinicians with different experiences and purposes in reading knee MRIs.

Acknowledgments

The authors acknowledge financial support from the Ministry of Science and Technology (MOST; MOST 109-2926-I-010-501, MOST 107-2314-B-010-015-MY3, MOST 109-2926-I-010-502, MOST 109-2321-B-010-005, MOST 108-2923-B-010-002-MY3, MOST 109-2823-8-010-003-CV, MOST 109-2622-B-010-006, and MOST 109-2321-B-010-006). This work is particularly supported by “Development and Construction Plan” of the School of Medicine, National Yang-Ming University, now known as National Yang Ming Chiao Tung University (107F-M01-0504); and Aiming for the Top University Plan, a grant from the Ministry of Education.

Conflicts of Interest

None declared.

References

1. Kaeding CC, Léger-St-Jean B, Magnussen RA. Epidemiology and diagnosis of anterior cruciate ligament injuries. Clin Sports Med 2017 Jan;36(1):1-8. [doi: 10.1016/j.csm.2016.08.001] [Medline: 27871652]
2. Salzler M, Nwachukwu BU, Rosas S, Nguyen C, Law TY, Eberle T, et al. State-of-the-art anterior cruciate ligament tears: a primer for primary care physicians. Phys Sportsmed 2015 May;43(2):169-177. [doi: 10.1080/00913847.2015.1016865] [Medline: 25703144]
3. Chen KH, Chiang ER, Wang HY, Ma HL. Correlation of meniscal tear with timing of anterior cruciate ligament reconstruction in patients without initially concurrent meniscal tear. J Knee Surg 2019 Nov;32(11):1128-1132. [doi: 10.1055/s-0038-1675783] [Medline: 30449021]
4. Cinque ME, Dornan GJ, Chahla J, Moatshe G, LaPrade RF. High rates of osteoarthritis develop after anterior cruciate ligament surgery: an analysis of 4108 patients. Am J Sports Med 2018 Jul;46(8):2011-2019. [doi: 10.1177/0363546517730072] [Medline: 28982255]
5. Roßbach BP, Pietschmann MF, Gülecyüz MF, Niethammer TR, Ficklscherer A, Wild S, et al. Indications requiring preoperative magnetic resonance imaging before knee arthroscopy. Arch Med Sci 2014 Dec 22;10(6):1147-1152 [FREE Full text] [doi: 10.5114/aoms.2014.47825] [Medline: 25624852]

6. Benjaminse A, Gokeler A, van der Schans CP. Clinical diagnosis of an anterior cruciate ligament rupture: a meta-analysis. J Orthop Sports Phys Ther 2006 May;36(5):267-288. [doi: 10.2519/jospt.2006.2011] [Medline: 16715828]
7. Sri-Ram K, Salmon LJ, Pinczewski LA, Roe JP. The incidence of secondary pathology after anterior cruciate ligament rupture in 5086 patients requiring ligament reconstruction. Bone Joint J 2013 Jan;95-B(1):59-64. [doi: 10.1302/0301-620X.95B1.29636] [Medline: 23307674]
8. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
9. Ogura T, Sato M, Ishida Y, Hayashi N, Doi K. Development of a novel method for manipulation of angiographic images by use of a motion sensor in operating rooms. Radiol Phys Technol 2014 Jul;7(2):228-234. [doi: 10.1007/s12194-014-0259-0] [Medline: 24609904]
10. Sanders TL, Maradit Kremers H, Bryan AJ, Larson DR, Dahm DL, Levy BA, et al. Incidence of anterior cruciate ligament tears and reconstruction: a 21-year population-based study. Am J Sports Med 2016 Jun;44(6):1502-1507. [doi: 10.1177/0363546516629944] [Medline: 26920430]
11. Chen HC, Tzeng SS, Hsiao YC, Chen RF, Hung EC, Lee OK. Smartphone-based artificial intelligence-assisted prediction for eyelid measurements: algorithm development and observational validation study. JMIR mHealth uHealth 2021 Oct 08;9(10):e32444 [FREE Full text] [doi: 10.2196/32444] [Medline: 34538776]
12. Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. In: Proc Mach Learn Res. 2019 Presented at: Proceedings of the 36th International Conference on Machine Learning, vol 97; June 9-15, 2019; Long Beach, CA p. 6105-6114.
13. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-Excitation Networks. IEEE Trans Pattern Anal Mach Intell 2020 Aug 1;42(8):2011-2023. [doi: 10.1109/tpami.2019.2913372]
14. Buslaev A, Iglovikov VI, Khvedchenya E, Parinov A, Druzhinin M, Kalinin AA. Albumentations: fast and flexible image augmentations. Information 2020 Feb 24;11(2):125. [doi: 10.3390/info11020125]
15. Wu H, Gu X. Towards dropout training for convolutional neural networks. Neural Netw 2015 Nov;71:1-10. [doi: 10.1016/j.neunet.2015.07.007] [Medline: 26277608]
16. Perez L, Wang J. The effectiveness of data augmentation in image classification using deep learning. arXiv 2017 Dec 13:1-8. [doi: 10.48550/arXiv.1712.04621]
17. Smith LN. A disciplined approach to neural network hyper-parameters: part 1 -- learning rate, batch size, momentum, and weight decay. arXiv 2018 Apr 24:1-21. [doi: 10.48550/arXiv.1803.09820]
18. Huang G, Li Y, Pleiss G, Liu Z, Hopcroft J, Weinberger K. Snapshot ensembles: train 1, get M for free. arXiv 2017 Apr 01:1-14. [doi: 10.48550/arXiv.1704.00109]
19. Kingma D, Ba J. Adam: a method for stochastic optimization. arXiv 2014 Dec 22:1-15. [doi: 10.48550/arXiv.1412.6980]
20. Ng WHA, Griffith JF, Hung EHY, Paunipagar B, Law BKY, Yung PSH. Imaging of the anterior cruciate ligament. World J Orthop 2011 Aug 18;2(8):75-84 [FREE Full text] [doi: 10.5312/wjo.v2.i8.75] [Medline: 22474639]
21. Ha TP, Li KC, Beaulieu CF, Bergman G, Ch'en IY, Eller DJ, et al. Anterior cruciate ligament injury: fast spin-echo MR imaging with arthroscopic correlation in 217 examinations. AJR Am J Roentgenol 1998 May;170(5):1215-1219. [doi: 10.2214/ajr.170.5.9574587] [Medline: 9574587]
22. Kam CW, Chee DWY, Peh WCG. Magnetic resonance imaging of cruciate ligament injuries of the knee. Can Assoc Radiol J 2010 Apr 01;61(2):80-89 [FREE Full text] [doi: 10.1016/j.carj.2009.11.003] [Medline: 20110155]
23. Lee JK, Yao L, Phelps CT, Wirth CR, Czajka J, Lozman J. Anterior cruciate ligament tears: MR imaging compared with arthroscopy and clinical tests. Radiology 1988 Mar;166(3):861-864. [doi: 10.1148/radiology.166.3.3340785] [Medline: 3340785]
24. van der List JP, Mintz DN, DiFelice GS. The location of anterior cruciate ligament tears: a prevalence study using magnetic resonance imaging. Orthop J Sports Med 2017 Jun 22;5(6):2325967117709966 [FREE Full text] [doi: 10.1177/2325967117709966] [Medline: 28680889]
25. Liu F, Guan B, Zhou Z, Samsonov A, Rosas H, Lian K, et al. Fully automated diagnosis of anterior cruciate ligament tears on knee MR images by using deep learning. Radiol Artif Intell 2019 May 08;1(3):180091 [FREE Full text] [doi: 10.1148/ryai.2019180091] [Medline: 32076658]
26. Chang PD, Wong TT, Rasiej MJ. Deep learning for detection of complete anterior cruciate ligament tear. J Digit Imaging 2019 Dec;32(6):980-986 [FREE Full text] [doi: 10.1007/s10278-019-00193-4] [Medline: 30859341]
27. Germann C, Marbach G, Civardi F, Fucentese SF, Fritz J, Sutter R, et al. Deep convolutional neural network-based diagnosis of anterior cruciate ligament tears: performance comparison of homogenous versus heterogeneous knee MRI cohorts with different pulse sequence protocols and 1.5-T and 3-T magnetic field strengths. Invest Radiol 2020 Aug;55(8):499-506 [FREE Full text] [doi: 10.1097/RLI.0000000000000664] [Medline: 32168039]
28. Zhang L, Li M, Zhou Y, Lu G, Zhou Q. Deep learning approach for anterior cruciate ligament lesion detection: evaluation of diagnostic performance using arthroscopy as the reference standard. J Magn Reson Imaging 2020 Dec 26;52(6):1745-1752. [doi: 10.1002/jmri.27266] [Medline: 32715584]
29. Yao L, Gentili A, Petrus L, Lee JK. Partial ACL rupture: an MR diagnosis? Skeletal Radiol 1995 May;24(4):247-251. [doi: 10.1007/BF00198407] [Medline: 7644934]

Abbreviations

AI: artificial intelligence
ACL: anterior cruciate ligament
CNN: convolutional neural network
MOST: Ministry of Science and Technology
MRI: magnetic resonance imaging

Edited by K El Emam, B Malin; submitted 24.02.22; peer-reviewed by CC Lin, JF Rajotte; comments to author 28.05.22; revised version received 15.06.22; accepted 05.07.22; published 26.07.22.

Please cite as:
Chen KH, Yang CY, Wang HY, Ma HL, Lee OKS
Artificial Intelligence–Assisted Diagnosis of Anterior Cruciate Ligament Tears From Magnetic Resonance Images: Algorithm Development and Validation Study
J AI 2022;1(1):e37508
URL: https://ai.jmir.org/2022/1/e37508
doi:10.2196/37508
PMID:38875555

Original Paper

Provider Perspectives on Artificial Intelligence–Guided Screening for Low Ejection Fraction in Primary Care: Qualitative Study

Barbara Barry[1,2], PhD; Xuan Zhu[2], PhD; Emma Behnken[3], BA; Jonathan Inselman[2], MS; Karen Schaepe[2], PhD; Rozalina McCoy[4], MS, MD; David Rushlow[5], MD; Peter Noseworthy[6], MD; Jordan Richardson[7], BS; Susan Curtis[7], MLIS; Richard Sharp[7], PhD; Artika Misra[8], MD; Abdulla Akfaly[9], MD; Paul Molling[10], MD; Matthew Bernard[5], MD; Xiaoxi Yao[1,2], PhD

[1]Division of Health Care Delivery Research, Mayo Clinic, Rochester, MN, United States
[2]Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN, United States
[3]Knowledge and Evaluation Research Unit, Mayo Clinic, Rochester, MN, United States
[4]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, United States
[5]Department of Family Medicine, Mayo Clinic, Rochester, MN, United States
[6]Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN, United States
[7]Biomedical Ethics Research Program, Mayo Clinic, Rochester, MN, United States
[8]Department of Family Medicine, Mayo Clinic Health System, Mankato, MN, United States
[9]Department of Community Internal Medicine, Mayo Clinic Health System, Eau Claire, WI, United States
[10]Department of Family Medicine, Mayo Clinic Health System, Onalaska, WI, United States

Corresponding Author:
Barbara Barry, PhD
Division of Health Care Delivery Research
Mayo Clinic
200 First Street SW
Rochester, MN, 55905
United States
Phone: 1 507 255 5123
Email: barry.barbara@mayo.edu

Abstract

Background: The promise of artificial intelligence (AI) to transform health care is threatened by a tangle of challenges that emerge as new AI tools are introduced into clinical practice. AI tools with high accuracy, especially those that detect asymptomatic cases, may be hindered by barriers to adoption. Understanding provider needs and concerns is critical to inform implementation strategies that improve provider buy-in and adoption of AI tools in medicine.

Objective: This study aimed to describe provider perspectives on the adoption of an AI-enabled screening tool in primary care to inform effective integration and sustained use.

Methods: A qualitative study was conducted between December 2019 and February 2020 as part of a pragmatic randomized controlled trial at a large academic medical center in the United States. In all, 29 primary care providers were purposively sampled using a positive deviance approach for participation in semistructured focus groups after their use of the AI tool in the randomized controlled trial was complete. Focus group data were analyzed using a grounded theory approach; iterative analysis was conducted to identify codes and themes, which were synthesized into findings.

Results: Our findings revealed that providers understood the purpose and functionality of the AI tool and saw potential value for more accurate and faster diagnoses. However, successful adoption into routine patient care requires the smooth integration of the tool with clinical decision-making and existing workflow to address provider needs and preferences during implementation. To fulfill the AI tool's promise of clinical value, providers identified areas for improvement including integration with clinical decision-making, cost-effectiveness and resource allocation, provider training, workflow integration, care pathway coordination, and provider-patient communication.

Conclusions: The implementation of AI-enabled tools in medicine can benefit from sensitivity to the nuanced context of care and provider needs to enable the useful adoption of AI tools at the point of care.

Trial Registration: ClinicalTrials.gov NCT04000087; https://clinicaltrials.gov/ct2/show/NCT04000087

(JMIR AI 2022;1(1):e41940) doi:10.2196/41940

KEYWORDS

artificial intelligence; AI; machine learning; human-AI interaction; health informatics; primary care; cardiology; pragmatic clinical trial; AI-enabled clinical decision support; human-computer interaction; health care delivery; clinical decision support; health care; AI tools

Introduction

Advances in artificial intelligence (AI) that are poised to transform health care are hindered by implementation challenges [1,2] that call for attention to provider needs and user-centeredness [3,4]. As AI models are increasingly pushed to the point of care, front-line care teams are often left to solve the challenges of AI integration on their own [5]. Research is needed to ensure the clinical value of AI tools is preserved through successful adoption at the point of care. To inform this knowledge gap, we present a case study of a pragmatic trial in which an AI-enabled screening tool was introduced in primary care to help identify patients with a high likelihood of unrecognized left ventricular low ejection fraction (EF) [6]. Low EF is often underdiagnosed but treatable; early diagnosis and treatment could prevent the progression of heart failure and reduce future hospitalization and mortality. We offer a qualitative analysis of provider reflections on the use of the AI screening tool and suggestions for the effective clinical adoption of AI-enabled tools.

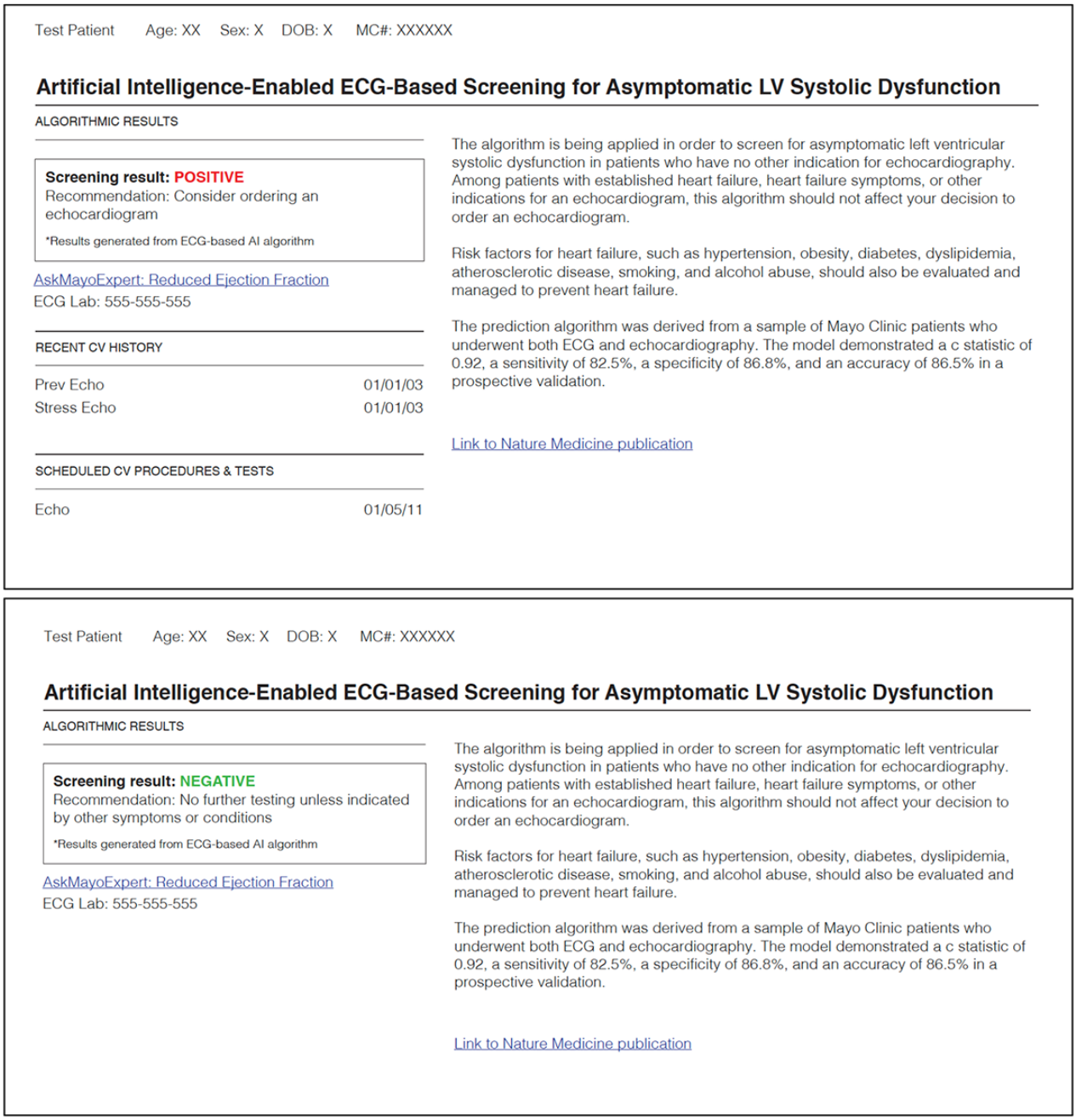
Methods

Overall Study Design

A pragmatic cluster randomized controlled trial (NCT04000087) was conducted to evaluate whether an electrocardiogram (ECG) AI-guided screening tool (ECG AI-Guided Screening for Low Ejection Fraction; EAGLE) improves the diagnosis of left ventricular EF in clinical practice [7,8]. Details on the trial design are reported elsewhere [6]. The intervention is a provider-facing action-recommendation report (Figure 1) that contains a screening result generated by the application of a deep learning algorithm to a patient's ECG [9].

Positive screening results were delivered to providers via an email alert that suggests a transthoracic echocardiogram (TTE) should be considered and remind them that the report was available in the electronic health record (EHR). The report included a brief description of the AI algorithm and a phone number to call for additional information. Follow-up emails were sent if no TTE was ordered or no rationale was provided for rejecting the TTE recommendation.

Figure 1. Example AI result report. AI reports are generated by the AI tool and embedded into the electronic health record. Note that only positive results would generate an email to a provider, and both positive and negative results could be accessed in the patient's health record. AI: artificial intelligence; ECG: electrocardiogram; LV: left ventricular.



Focus Group Study Design and Procedures

Semistructured focus groups were conducted with 10 primary care teams. We used a positive deviance approach to select the care teams [10,11]. Specifically, we selected the 5 care teams with the lowest TTE recommendation adherence and the 5 teams with highest adherence, with adherence defined as acting on the AI recommendation by ordering a TTE. Each focus group was conducted with providers from the same care team. Individual interviews were conducted to accommodate provider schedules when necessary. Discussion topics included provider experiences with the AI tool and their attitudes toward AI in medicine. Between December 2019 and February 2020, a total of 7 focus groups and 5 individual interviews were conducted, involving 29 providers consisting of physicians, physician assistants, and nurse practitioners. Participant characteristics are summarized in Table 1. The 2 interviewers and all interviewees were blinded to the adherence status of the care team to enable candid, nondefensive conversation as well as to avoid biasing the interviewers [12]. All focus groups were audio recorded, transcribed verbatim, deidentified, and reviewed for accuracy.

Table 1. Characteristics of focus group participants in high- and low-adherence care teams. Note that characteristic information from 2 participants is missing.

Characteristic	High adherence (N=17)	Low adherence (N=10)
Age (years)		
n	15	10
Mean (SD)	44.8 (9.07)	41.4 (3.92)
Median	46.0	41.0
Range	32.0-61.0	36.0-50.0
Gender (self-reported), n (%)		
Male	8 (47)	6 (60)
Female	9 (53)	4 (40)
Race, n (%)		
White	16 (94)	8 (80)
Other[a]	0 (0)	1 (10)
Prefer not to say	1 (6)	1 (10)
Position, n (%)		
Physician	12 (71)	8 (80)
Physician assistant	0 (0)	2 (20)
Nurse practitioner	5 (29)	0 (0)
Specialty, n (%)		
Family medicine	11 (65)	5 (50)
Internal medicine	6 (35)	5 (50)
Years in practice		
n	15	10
Mean (SD)	13.5 (9.04)	8.3 (5.19)
Median	11.0	5.5
Range	1.0-31.0	3.0-20.0
Years in current care team		
n	15	10
Mean (SD)	11.5 (9.08)	7.2 (6.00)
Median	11.0	5.0
Range	0.5-31.0	1.0-20.0

[a]Racial categories measured included American Indian or Alaskan Native, Asian, Black or African American, and Native Hawaiian or Other Pacific Islander. None of our participants identified as being in these categories.

Data Analysis

Thematic analysis was used to identify predominant themes regarding clinicians' experiences and perspectives regarding using the AI screening tool [13-15]. Two researchers (BB and XZ) open-coded transcripts and then categorized open codes into themes. The relationships between the themes were then articulated in a hierarchical structure of main themes and subthemes. The thematic structure was revised when new categories and themes were identified. Analytic memos were used to summarize the findings. NVivo software (version 12; QSR International) was used to facilitate analysis. Researchers were unaware of the adherence status of the care teams during coding. Adherence status was revealed to the researchers after all transcripts were coded to assess differences between groups.

Ethics Approval

The methods were performed in accordance with the relevant guidelines and regulations and approved by the Mayo Clinic Institutional Review Board (IRB #19-003137). The trial was registered on ClinicalTrials.gov (NCT04000087) on June 27, 2019.

Results

Perspectives and Themes

All providers received at least one positive AI screening result and were able to correctly describe the AI tool's functionality and purpose. Providers had polarized perspectives on the value of the AI tool: some expressed that the tool could improve patient care, whereas others thought it was unnecessary or costly. Dissatisfied providers agreed that honing the tool and its delivery would increase value, whereas a small number of providers disagreed with the need for the tool.

We did not observe prominent differences in themes between care teams with high and low adherence. We identified 7 dominant themes of provider reflections on AI tool use: (1) promising clinical value, (2) integration with clinical reasoning, (3) cost-effectiveness and resource allocation, (4) provider training, (5) workflow integration, (6) care pathway coordination, and (7) provider-patient communication.

Promising Clinical Value

Providers believed in the AI tool's capability to identify asymptomatic patients at risk for heart failure. Providers saw an opportunity to accelerate care for patients who might otherwise fail to report symptoms of low EF and saw value in implementing early management to save patients from acute cardiac events. Providers also noted the ability of the AI tool to make care more efficient by assessing the ECG more quickly than a provider could.

> *[The result] was definitely abnormal, and I was able to talk with this patient about lifestyle changes and actually have something coming behind me within that.* [Focus group #12]

> *I'm still pushing the button on the order cuz I agree with it but, you know, doing all the nuts and bolts behind it, if that's done for me, then I can focus my time on doing what only I can do.* [Focus group #9]

Integration With Clinical Reasoning

Providers expressed apprehension about the utility and long-term patient benefit of the tool based on how it fit into clinical practice during the trial. They were concerned about the increased burden, especially when a screening was not clinically useful in patient contexts such as preexisting cardiovascular conditions, and noted that for certain patients, other medical priorities (eg, cancer treatment) might take precedence over initiating a TTE and treatment for low EF. Providers expressed concern regarding the lack of clear guidelines about when to order a new TTE if there were prior TTEs in a patient's medical record. A few providers were unsure to what extent the AI tool could improve patients' long-term health outcomes and noted barriers treating a patient who may be at risk for heart failure but has not yet shown any symptoms, revealing a potential lack of knowledge about evidence-based recommendations for asymptomatic low EF treatment.

> *They had known heart disease. I was like, "Well, that doesn't make any sense." After the first couple of doing that, I started almost disregarding.* [Focus group #10]

> *I only had three, and I know them. I knew them very well, so the minute I got the one with the end-stage liver failure, cirrhosis, paracentesis, I knew that immediately that that wasn't gonna be valid, or not necessarily not valid, but is it correct?* [Focus group #12]

Providers gauged the AI's capability relative to their own. Some providers believed that the AI tool was superior in recognizing patterns to identify asymptomatic cases. A few others preferred face-to-face visits for physical examination and continuity of care. Some providers were also concerned a bias might occur if the AI algorithm was trained on data misaligned with their patient panels.

> *Good I got a notification. I woulda missed it [the diagnosis].* [Focus group #10]

> *Uh, we can do a lot by remote monitoring, but I need to touch you, and I need to listen to you, and I need to listen to your heart. And so if something just got triggered, I don't care.* [Focus group #7]

> *The struggle with AI so far has been that the breakdowns have come because of the data that's been input and a lot of that has been because of our geographic or our social or our...* [Interviewee #5]

> *There's bias.* [Interviewee #3; focus group #1]

Cost-effectiveness and Resource Allocation

Some providers questioned the cost-effectiveness of the recommended TTE follow-up given the current lack of outcome data and noted that the cost is especially concerning when the screening result is a false positive. A few clinicians were concerned about insurance coverage. Some clinicians noted potential cumulative cost savings from optimized treatment plans and the prevention of heart failure hospitalizations. However, providers noted that increased TTE order volume due to positive AI screenings could delay care for patients with more urgent TTE needs.

> *You can tell them, thankfully, it's normal. Obviously, the EKG picked up something that showed potential for concern. We have good news that everything is normal. We're gonna continue to optimize your treatment. That being said, it's several thousand dollars.* [Focus group #8]

Provider Training

Providers remembered being introduced to the AI tool and trial protocol by department leadership in meetings and via email yet did not recall the information when they received the AI result. Providers reported agreeing, sometimes enthusiastically, with the objectives of the trial but found it difficult to translate the instructions (eg, ordering TTE based on AI result) into their context of care. Championship by leadership set unintended high expectations for the AI tool and caused disappointment when the number of positive screenings was lower than expected. Providers also remarked that they could not remember

how to find the AI report in the EHR and did not have time to read the training packet.

> *Right. I mean, the first email inviting us, I read two paragraphs. "Boy, that sounds like a good idea." Then all of a sudden, we get this book, and then all of a sudden, they [AI results] start coming. You just get lost.* [Focus group #10]

> *No. We didn't see the video.* [Interviewee #3]

> *Right there, if it requires a video and a half hour lecture to figure out where [to find the report], then it's probably not well-placed.* [Interviewee #1; focus group #1]

Workflow Integration

Providers were unaware that the delivery of the AI result via email notification outside of the EHR was due to system security issues, Food and Drug Administration regulations, and IT barriers, which fragmented the digital workflow. The AI report in the EHR was rarely accessed by providers after the receipt of the AI result email notification. Repeated reminder emails urging providers to act on unresolved AI results were irritating and created confusion about which alerts had been completed and which needed attention. The AI result delivery was not timed to be part of a scheduled visit, which caused extra clerical and cognitive burden, and took time away from providers' already busy schedules.

> *There was an ECG that suggests you might do an echo and if I'm with the patient right there, done, but to [Interviewee's] point, it was when it was noncontiguous, non-need. It was an extra half an hour phone call in the day that I just simply don't have time for.* [Focus group #1]

> *I got emails, which made it very difficult because it's not linked to the chart.* [Focus group #3]

Care Pathway Coordination

The AI report was always routed to the patient's primary care provider regardless of why and by whom the ECG was ordered, causing confusion among primary care providers about care coordination and the chain of custody. In cases where the AI result email alerts were from ECGs ordered outside of primary care (eg, in the emergency department), the primary care providers questioned whether these AI results were within the scope of their responsibility. Some providers felt they were caught between the care pathway already underway (eg, for surgery) and a potentially new or redundant care pathway suggested by the AI result. They felt that they were stuck in an awkward position, either ignoring the alert or communicating a result to a colleague who would have already been aware. In these cases, the AI tool was seen as not being logically coordinated within the care pathway.

> *Right now he's in the hospital, and Cardiology's definitely onboard, and so I just gave them that heads-up.* [Focus group #2]

> *The EAGLE thing triggers to us. We don't know whether we are supposed to follow up and do everything. I don't know whether I ordered [the ECG].* [Focus group #6]

> *I'm a minutiae guy, so if someone's got an abnormal EKG, I look at their EKG. I look at their echo. It puts a fair amount of burden back to the PCP because no matter where it's ordered, it comes back to me as PCP.* [Focus group #8]

Provider-Patient Communication

Some providers stated that the unexpected nature of a result generated outside the context of a visit, the lack of explainability of deep learning, and the lack of reporting guidelines regarding false positives make communicating the results to patients challenging and time-consuming. In a worst-case scenario for patient-provider alliance and provider morale, a patient perceived that AI corrected an error made by the provider. Providers disagreed on whether patients can understand and cope with the AI results if the results are automatically delivered to patients without provider oversight and communication of the results. Some providers considered the AI tool new and complex, whereas others considered it similar to screenings that patients already view as routine (eg, a blood test).

> *You have to explain to the patient what you're gonna do when you get the low EF; "how come you didn't figure it out already, Dr. [Name]; if you're such a great clinician, how could you miss this?"* [Focus group #1]

> *But in terms of just calling somebody up out of the...on just like a cold call and saying, "I think you might have heart failure" because a computer said so, um, that's where my caveats come from.* [Focus group #7]

Provider Suggestions to Improve Future AI Tool Adoption

Providers articulated the following suggestions to improve future AI tool uptake and use: (1) setting appropriate expectations for how, when, and how often the AI tool would deliver a recommendation; (2) attuning the application of the AI tool to patient populations; (3) having reliable data that show positive clinical outcomes due to the tool; (4) having a demonstration of cost-effectiveness, (5) streamlining integration into clinical workflow, (6) clarifying provider responsibility, and (7) having support for the communication of results to patients.

Discussion

Principal Findings

We found that most providers saw the potential value of the AI tool for more accurate and faster diagnoses. They were willing to adopt such tools and collaborate with researchers to validate tools in clinical practice. However, during use in the clinical trial, providers identified challenges that should be taken into consideration as AI tools are introduced widely in primary care. Provider recommendations encompassed increased sensitivity to clinical decision-making, addressing digital implementation issues, and the awareness of system-wide impact.

AI tools that predict asymptomatic health conditions convoke a set of issues in medical decision-making that providers are

asked to resolve on a case-by-case basis and, in doing so, are confronted with a change in the scope of their clinical decision-making [16]. Although AI tools provide guidance, they rarely apply to all patients and often add a new dimension to already complex decision-making [17,18]. In practical use, providers have different ways of weighing evidence to inform the best next step in patient care. For example, although the confirmation or rejection of an AI screening result through a follow-up testing may seem low-risk and easy, the clinical action in an individual patient's case could shift focus from a more immediate threat to health and increase cost. In primary care, providers are positioned to see the entire context of care and together with patients navigate multiple risk-benefit decisions within complex situations that do not lend themselves to rapid, binary decisions for next steps [19]. Consequently, the incorporation of AI tools that support new diagnoses can further complicate the issues of distinguishing between the art and science of medicine in complex primary care decision-making [20]. Our research reaffirms that providers may find AI-enabled tools capable of delivering helpful information but that communication and actions taken by the care team in response to AI tools are complex and demand a balance between structured guidance and freedom to adapt information to a clinical case [21,22].

Providers offered suggestions for improving the applicability of the AI model, digital workflow, and patient communication. These suggestions can enhance AI tool use but may be difficult to achieve during the initial translation in a pragmatic clinical trial. Provider feedback to hone the AI model and digital workflow are necessary to ensure the best diagnostic performance over time, safety, and adherence to regulatory requirements. Additional burden on providers during the initial translation may exacerbate clerical burden, which can dampen interest in AI tool adoption. It is important to set expectations with providers that clumsy workarounds and added burden during initial translation in the clinical trial are temporary and that fine-tuning AI implementation to meet various clinical contexts and provider needs is a long-term, collaborative process [23]. Additionally, the silent testing of the AI tool before broader launch in a randomized controlled trial and more spontaneous, passive modes of collecting provider feedback (versus repeatedly requesting active input from providers) may be of value. Moreover, AI tools may illuminate existing issues in care delivery or cause new problems in new contexts, which prompts the need for real-time observations and auditing of AI models and tools to improve the design of the full implementation and enable effective use [24,25].

Study Limitations

Focus groups were conducted after trial completion, and thus, provider experience was communicated retrospectively. Future research could make use of more spontaneous data collection methods (eg, ecological momentary assessment) to capture provider experiences and perspectives at the point of care in real time. Our findings were based on the perspectives of 29 providers from 10 care teams that may not be representative of the primary care provider population and thus cannot capture the full scope of diverse perspectives among primary care providers. Additionally, it is unclear how our results will generalize beyond AI tools that use a deep learning algorithm and leverage knowledge from cardiology within a primary care setting. Future research with a broader range of AI tools in different clinical settings and specialties with more diverse provider samples is needed to triangulate our findings and uncover additional important themes.

Conclusion

Our work identified specific issues that providers faced when AI-enabled tools are introduced into primary care during a clinical trial as well as relevant techniques across algorithm development, point-of-care use, and broader systems that can drive the provider-centered adoption of AI tools. These findings corroborate the challenges of implementing AI-enabled tools in medicine: successful implementation must be sensitive to the nuanced context of care and provider sensibilities to enable the useful adoption of AI tools at the point of care.

Acknowledgments

This study was funded by the Mayo Clinic Robert D. And Patricia E. Kern Center for the Science of Health Care Delivery. We thank Monica Looze for supporting focus group planning and data collection and Cheri Rollo and Annie Farbisz for data formatting.

Authors' Contributions

BB contributed to supervision, conceptualization, methodology, formal analysis, writing the original draft, writing review, and editing. XZ contributed to conceptualization, methodology, formal analysis, writing the original draft, writing review, and editing. EB contributed to project administration, writing the original draft, writing review, and editing. JI contributed to formal analysis, writing review, and editing. KS contributed to methodology, investigation, writing review, and editing. RM, DR, PN, JR, SC, RS, AM, AA, PM, and MB contributed to writing review and editing. XY contributed to conceptualization, methodology, writing review, and editing.

Conflicts of Interest

None declared.

References

1. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019 Jan;25(1):30-36 [FREE Full text] [doi: 10.1038/s41591-018-0307-0] [Medline: 30617336]
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan 7;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
3. Amershi S, Weld D, Vorvoreanu M, Fourney A, Nushi B, Collisson P, et al. Guidelines for human-AI interaction. 2019 May 02 Presented at: CHI '19: CHI Conference on Human Factors in Computing Systems; May 4-9, 2019; Glasgow, United Kingdom p. 1-13. [doi: 10.1145/3290605.3300233]
4. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019 Oct 29;17(1):195 [FREE Full text] [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
5. Angehrn Z, Haldna L, Zandvliet AS, Gil Berglund E, Zeeuw J, Amzal B, et al. Artificial intelligence and machine learning applied at the point of care. Front Pharmacol 2020 Jun 18;11:759 [FREE Full text] [doi: 10.3389/fphar.2020.00759] [Medline: 32625083]
6. Yao X, Rushlow DR, Inselman JW, McCoy RG, Thacher TD, Behnken EM, et al. Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. Nat Med 2021 May;27(5):815-819. [doi: 10.1038/s41591-021-01335-4] [Medline: 33958795]
7. Yao X, McCoy RG, Friedman PA, Shah ND, Barry BA, Behnken EM, et al. ECG AI-Guided Screening for Low Ejection Fraction (EAGLE): rationale and design of a pragmatic cluster randomized trial. Am Heart J 2020 Jan;219:31-36. [doi: 10.1016/j.ahj.2019.10.007] [Medline: 31710842]
8. Yao X, McCoy RG, Friedman PA, Shah ND, Barry BA, Behnken EM, et al. Clinical trial design data for electrocardiogram artificial intelligence-guided screening for low ejection fraction (EAGLE). Data Brief 2020 Feb;28:104894 [FREE Full text] [doi: 10.1016/j.dib.2019.104894] [Medline: 31867424]
9. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. Lancet 2019 Sep 07;394(10201):861-867. [doi: 10.1016/S0140-6736(19)31721-0] [Medline: 31378392]
10. Krumholz HM, Curry LA, Bradley EH. Survival after acute myocardial infarction (SAMI) study: the design and implementation of a positive deviance study. Am Heart J 2011 Dec;162(6):981-987.e9 [FREE Full text] [doi: 10.1016/j.ahj.2011.09.004] [Medline: 22137070]
11. Curry LA, Spatz E, Cherlin E, Thompson JW, Berg D, Ting HH, et al. What distinguishes top-performing hospitals in acute myocardial infarction mortality rates? a qualitative study. Ann Intern Med 2011 Mar 15;154(6):384-390 [FREE Full text] [doi: 10.7326/0003-4819-154-6-201103150-00003] [Medline: 21403074]
12. Rose AJ, McCullough MB. A practical guide to using the positive deviance method in health services research. Health Serv Res 2017 Jun 28;52(3):1207-1222 [FREE Full text] [doi: 10.1111/1475-6773.12524] [Medline: 27349472]
13. Glaser BG. The constant comparative method of qualitative analysis. Soc Probl 1965 Apr;12(4):436-445. [doi: 10.2307/798843]
14. Patton MQ. Qualitative Evaluation and Research Methods. 2nd ed. Thousand Oaks, CA: SAGE Publications; 1990.
15. Boyatzis RE. Transforming Qualitative Information: Thematic Analysis and Code Development. Thousand Oaks, CA: SAGE Publications; 1998.
16. Wang TJ, Levy D, Benjamin EJ, Vasan RS. The epidemiology of "asymptomatic" left ventricular systolic dysfunction: implications for screening. Ann Intern Med 2003 Jun 03;138(11):907-916. [doi: 10.7326/0003-4819-138-11-200306030-00012] [Medline: 12779301]
17. Patel VL, Kaufman DR, Arocha JF. Emerging paradigms of cognition in medical decision-making. J Biomed Inform 2002 Feb;35(1):52-75 [FREE Full text] [doi: 10.1016/s1532-0464(02)00009-6] [Medline: 12415726]
18. Rushlow DR, Croghan IT, Inselman JW, Thacher TD, Friedman PA, Yao X, et al. Comparing the characteristics and effectiveness of early and late adopters utilizing an artificial intelligence tool to detect low cardiac ejection fraction. Mayo Clin Proc (in press) 2022.
19. Yancy CW, Januzzi JL, Allen LA, Butler J, Davis LL, Fonarow GC, et al. 2017 ACC Expert Consensus Decision Pathway for optimization of heart failure treatment: answers to 10 pivotal issues about heart failure with reduced ejection fraction: a report of the American College of Cardiology Task Force on Expert Consensus Decision Pathways. J Am Coll Cardiol 2018 Jan 16;71(2):201-230 [FREE Full text] [doi: 10.1016/j.jacc.2017.11.025] [Medline: 29277252]
20. Kantarjian H, Yu PP. Artificial intelligence, big data, and cancer. JAMA Oncol 2015 Aug;1(5):573-574. [doi: 10.1001/jamaoncol.2015.1203] [Medline: 26181906]
21. Romero-Brufau S, Wyatt KD, Boyum P, Mickelson M, Moore M, Cognetta-Rieke C. A lesson in implementation: a pre-post study of providers' experience with artificial intelligence-based clinical decision support. Int J Med Inform 2020 May;137:104072 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.104072] [Medline: 32200295]
22. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan 20;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
23. Matthiesen S, Diederichsen SZ, Hansen MKH, Villumsen C, Lassen MCH, Jacobsen PK, et al. Clinician preimplementation perspectives of a decision-support tool for the prediction of cardiac arrhythmia based on machine learning: near-live

feasibility and qualitative study. JMIR Hum Factors 2021 Nov 26;8(4):e26964 [FREE Full text] [doi: 10.2196/26964] [Medline: 34842528]
24. Schulam P, Saria S. Can you trust this prediction? auditing pointwise reliability after learning. In: Proc Mach Learn Res, vol 89.: Proceedings of Machine Learning Research; 2019 Presented at: 22nd International Conference on Artificial Intelligence and Statistics; April 16-18, 2019; Naha, Okinawa, Japan p. 1022-1031 URL: https://proceedings.mlr.press/v89/schulam19a.html
25. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, SPIRIT-AICONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ 2020 Sep 09;370:m3164 [FREE Full text] [doi: 10.1136/bmj.m3164] [Medline: 32909959]

Abbreviations

AI: artificial intelligence
EAGLE: ECG AI-Guided Screening for Low Ejection Fraction
ECG: electrocardiogram
EF: ejection fraction
EHR: electronic health record
TTE: transthoracic echocardiogram

Edited by G Eysenbach, K El Emam, B Malin; submitted 16.08.22; peer-reviewed by M Hofford, V Ochs; comments to author 11.09.22; revised version received 13.09.22; accepted 17.09.22; published 14.10.22.

Please cite as:
Barry B, Zhu X, Behnken E, Inselman J, Schaepe K, McCoy R, Rushlow D, Noseworthy P, Richardson J, Curtis S, Sharp R, Misra A, Akfaly A, Molling P, Bernard M, Yao X
Provider Perspectives on Artificial Intelligence–Guided Screening for Low Ejection Fraction in Primary Care: Qualitative Study
JMIR AI 2022;1(1):e41940
URL: https://ai.jmir.org/2022/1/e41940
doi:10.2196/41940
PMID:38875550

Original Paper

Visualizing the Interpretation of a Criteria-Driven System That Automatically Evaluates the Quality of Health News: Exploratory Study of 2 Approaches

Xiaoyu Liu[1,2], MBA, PhD; Hiba Alsghaier[1], MSc; Ling Tong[3], BSc; Amna Ataullah[1]; Susan McRoy[1], PhD

[1]Department of Computer Science, University of Wisconsin Milwaukee, Milwaukee, WI, United States

[2]School of Health Sciences, Southern Illinois University Carbondale, Carbondale, IL, United States

[3]Department of Health Informatics and Administration, University of Wisconsin Milwaukee, Milwaukee, WI, United States

Corresponding Author:
Susan McRoy, PhD
Department of Computer Science
University of Wisconsin Milwaukee
Engineering and Mathematical Sciences Bldg 1275
3200 N Cramer St
Milwaukee, WI, 53211
United States
Phone: 1 414 229 6695
Email: mcroy@uwm.edu

Abstract

Background: Machine learning techniques have been shown to be efficient in identifying health misinformation, but the results may not be trusted unless they can be justified in a way that is understandable.

Objective: This study aimed to provide a new criteria-based system to assess and justify health news quality. Using a subset of an existing set of criteria, this study compared the feasibility of 2 alternative methods for adding interpretability. Both methods used classification and highlighting to visualize sentence-level evidence.

Methods: A total of 3 out of 10 well-established criteria were chosen for experimentation, namely whether the health news discussed the costs of the intervention (the cost criterion), explained or quantified the harms of the intervention (the harm criterion), and identified the conflicts of interest (the conflict criterion). The first step of the experiment was to automate the evaluation of the 3 criteria by developing a sentence-level classifier. We tested Logistic Regression, Naive Bayes, Support Vector Machine, and Random Forest algorithms. Next, we compared the 2 visualization approaches. For the first approach, we calculated word feature weights, which explained how classification models distill keywords that contribute to the prediction; then, using the local interpretable model-agnostic explanation framework, we selected keywords associated with the classified criterion at the document level; and finally, the system selected and highlighted sentences with keywords. For the second approach, we extracted sentences that provided evidence to support the evaluation result from 100 health news articles; based on these results, we trained a typology classification model at the sentence level; and then, the system highlighted a positive sentence instance for the result justification. The number of sentences to highlight was determined by a preset threshold empirically determined using the average accuracy.

Results: The automatic evaluation of health news on the cost, harm, and conflict criteria achieved average area under the curve scores of 0.88, 0.76, and 0.73, respectively, after 50 repetitions of 10-fold cross-validation. We found that both approaches could successfully visualize the interpretation of the system but that the performance of the 2 approaches varied by criterion and highlighting the accuracy decreased as the number of highlighted sentences increased. When the threshold accuracy was ≥75%, this resulted in a visualization with a variable length ranging from 1 to 6 sentences.

Conclusions: We provided 2 approaches to interpret criteria-based health news evaluation models tested on 3 criteria. This method incorporated rule-based and statistical machine learning approaches. The results suggested that one might visually interpret an automatic criterion-based health news quality evaluation successfully using either approach; however, larger differences may arise when multiple quality-related criteria are considered. This study can increase public trust in computerized health information evaluation.

(JMIR AI 2022;1(1):e37751) doi:10.2196/37751

KEYWORDS

health misinformation; machine learning; local interpretable model-agnostic explanation; LIME; interpretable artificial intelligence; AI

Introduction

Background

The internet has grown in popularity as a source for the public to learn about their health and investigate potential treatments for their health conditions. It is estimated that 80% of internet users consult web-based health information before making decisions [1]. Web-based media outlets such as social media feeds, forum threads, blogs, and newspapers have made information access and sharing easier. However, this has also accelerated the propagation of misleading information. Misinformation about health has been detected on different social media sites, such as Twitter [2-5], Facebook [6-9], YouTube [10-13], Pinterest [14,15], and Weibo [16,17]. Waszak et al [18] found that 40% of the most frequently shared links on social media contained medical information related to the most common diseases and causes of death were classified as fake news. In addition, the spread of health-related misinformation is not confined by geography. A series of studies have reported and studied health misinformation in different geographic settings, such as in the United States [19-21], China [16,17,22,23], India [24], and Italy [25,26]. With the rise of seeking health information on the internet, the concerns and health-related harm cases regarding misinformation have increased [27-29].

Unlike other types of misinformation, health-related misleading information, especially claims of efficacy about health interventions, such as medical treatments, tests, products, or procedures, can cause immediate actual harm to real people. The public and patients may be misled into making bad decisions that could result in severe consequences regarding people's quality of life and even the risk of mortality. This negative influence has been observed in many countries worldwide, despite cultural, regulatory, and geographic variances [30]. When the COVID-19 pandemic started in 2019, health misinformation was further exacerbated globally as more people increasingly turned to social media to confirm possible symptoms and share treatment plans [31]. Misleading and erroneous information, information of low quality such as conspiracy theories, poorly sourced medical advice, and information trivializing the virus has not only contributed to widespread misconceptions about the novel coronavirus but also caused public panic, catastrophic consequences of public health, and even people's distrust in public health institutions at the global level [32,33].

To address this public health crisis, continuing efforts to counteract health misinformation are being made across a wide range of disciplines and organizations. Detection and fact-checking work that relies on human effort is limited in scope, considering the high volume of fake news generated on the internet. Many attempts have been made to leverage artificial intelligence (AI) to analyze enormous amounts of information generated daily on a scale that would be impossible for humans to handle [34]. AI-powered automated detection methods, in comparison with people, are faster, more efficient, and may be deployed on targeted platforms at a low cost and on larger scale, by replicating human intelligence using data-driven analysis by computers [35]. When combating misinformation, AI technology may distinguish between accurate and misleading information using terms or word patterns associated with misinformation as cues from a relatively small set of articles that have been previously annotated by experts. Therefore, AI techniques can automate the process of detection of misleading information, which is conventionally performed manually.

Related Work

In recent years, there has been an increasing trend in AI-based studies attempting to address health misinformation. The choice of health topic is a critical factor to consider, as it requires domain understanding and knowledge to assess the quality of health information and confirm the presence of misinformation. Health topics incorporated in past misinformation detection studies either focused on a specific topic, such as vaccination [36-38], Zika [39], autism [40], COVID-19 [41-44], or a collection of miscellaneous health conditions and lifestyle choices [45-50]. Health misinformation resides in various information outlets. Existing studies have proposed the detection of false, misleading health news on platforms such as Twitter [37,39,51,52], websites [36,45,46], and web-based forums [48,49].

Setting an appropriate benchmark for evaluating and annotating health information is unavoidable when developing detection systems. On the basis of the benchmark and objectives of this study, previous work on misinformation classification can be briefly categorized into a veracity-based approach or a criteria-based approach. Studies that follow a veracity-based approach involved training classifiers to assess the truth of each health-related claim using data that have been annotated to indicate whether the claim can be validated or refuted by finding a similar statement using a trusted source. These supporting sources might be experts from a third-party fact-checking organization (eg, Snopes [53]), medical and health-related professional organizations (eg, World Health Organization [54]), academic or research institutions (eg, John Hopkins Medicine [55]), and the federal government (eg, CDC [56]) which are typically considered as the officially sanctioned sources of bona fide accurate information and play an active role in myth debunking. For example, Ghenai and Mejova [39] proposed a novel pipeline that combines health experts, crowdsourcing, and machine learning (ML) to capture rumors on Twitter. The model was created using 13 million tweets concerning Zika infection between February 2016 and the Summer Olympics and rumors outlined by the World Health Organization and Snopes. The study found that rumor-related topics have a particularly burst behavior. The results demonstrated the feasibility of using automated techniques to remove rumor-bearing tweets when a questionable topic was detected.

In contrast, studies that followed the criteria-based approach looked at misinformation based on various quality-indicating criteria predefined by research. An example of such criteria might be the reliability or unreliability of the source; the rationale is that *intuitively, a news article published on an unreliable website and forwarded by unreliable users is more likely to be fake news than news posted by authoritative and credible users* [57]. For example, Liu et al [50] predefined a list of reliable and unreliable websites from which health-related articles from various sources on the Chinese Internet society were extracted for data set construction. Experiments were performed based on various ML classifiers using manually extracted features and text-classification modeling. The best performance among all models reached a precision of 0.8374. Other approaches were based on the idea that news that does not satisfy certain items on an assessment checklist for health information quality can be considered untrustworthy. For instance, Shah et al [37] used a 7-point checklist adapted from 2 validated tools, the DISCERN and Quality Index for health-related Media Reports checklists, to manually appraise the credibility of 474 web pages after sampling from 143,003 unique vaccine-related web pages shared on Twitter between January 2017 and March 2018. According to previous studies, the best-performing classifiers could distinguish between low, medium, and high credibility with an accuracy of 78% and labeled low-credibility web pages with a precision of >96%. Al-Jefri et al [58] and Afsana et al [59] both developed 10 classifiers to automatically evaluate the quality of health news based on the criteria developed by HealthNewsReview.org. However, the latter's models demonstrated better classification performance owing to the inclusion of more features. In summary, veracity-based studies examined the authenticity of the news. The criteria-based approach focused on the characteristics of the news content, but the results did not make claims about the veracity of information.

In addition to the wide range of themes and strategies in detecting misinformation identified in the literature, methodologically, current studies also show the effectiveness of AI-based algorithms in classifying misinformation and quality information. Traditional ML algorithms, including Logistic Regression [40,47,52,60], support vector machine [37,40,47,50], decision tree [52,61], and random forest (RF) [37,39,41,48,60] have been widely applied in these studies, yielding effective and accurate performance. More recent studies have shown improved performance on large data sets by incorporating deep learning techniques, including convolutional neural networks [49,61], bidirectional encoder representations from transformers [43], and long short term memory [42,44,61]. As part of the modeling process, feature engineering has also been a critical step in improving the performance of classifiers. Zhao et al [57] reviewed and summarized 12 features used in health misinformation detection models. These features were grouped into 4 subsets: linguistic, topic, sentiment, and behavioral features.

Compared with traditional human fact-checking, an AI-based model consists of an algorithm that can automatically learn latent patterns and relationships from the data. However, one of the major challenges is the lack of a human-understandable rationale to support the results of classification tasks. Approaches that attempt to address this concern are often called "interpretable ML," "explainable ML," or "explainable AI" [62]. Open-source software with implementations of various interpretable ML methods are also available, such as local interpretable model-agnostic explanation (LIME) [63], Shapley Additive Explanations [64], Eli5 [65], and InterpretML [66], etc. These tools have been applied to various tasks, including image classification and text classification. With interpretations or visualized cues, users can verify the model and determine whether it meets their expectations. In addition, users can discover knowledge, justify predictions, and improve the performance of models using interpretable ML methods. Therefore, interpretable AI improves the trust and usability of the classifiers.

However, to date, only a small body of research has incorporated explainable AI models to combat health misinformation [43,67]. All of these studies on health information classification were veracity-based. A knowledge gap remains regarding the effectiveness of constructing an interpretable, criteria-driven classification system to help users evaluate the quality of health information.

Objective of This Study

We aimed to address the aforementioned concerns and needs by creating an interpretable, criteria-driven system to assist the public in evaluating the quality of health news to mitigate the adverse consequences of health misinformation. Previous work using the HealthNewsReview.org data set and ML classifiers at the document level found that 3 criteria (cost, harm, and conflict) are more accurately classifiable among the 10 criteria, using linguistic features [58,68]; therefore, we selected these 3 criteria for this exploratory study. Our study, because it addressed interpretability, also focused on the use of features that are directly visualizable (linguistic features), excluding less visualizable features (such as average sentence length), which sometimes improved classification accuracy.

As an exploratory study, we opted to test 2 possible interpretation approaches, using 3 criteria. The evaluation results for the criteria will be visually explained with highlighted sentences as cues to enhance interpretability and reliability. As the number of highlighted sentences may affect the overall visual representation and effectiveness of the interpretation, we also attempted to determine the ideal range for the number of highlighted sentences.

Methods

Overview

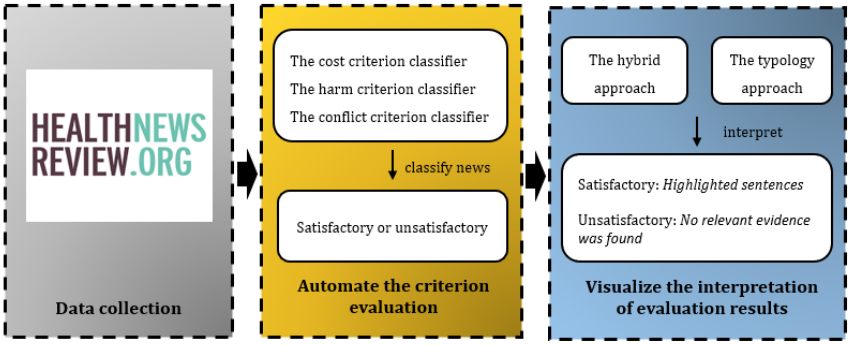
The experiment consisted of 3 components, as illustrated in Figure 1. In the first component, we collected reviewed health news from HealthNewsReview.org [69] to build the data set for modeling. Each criterion review result provided by HealthReview.org was treated as a classification target. The second component was a supervised document classification task that automated the criteria evaluation process. Each health news article was categorized automatically at the document level using established criteria and the output was binary

(satisfactory or unsatisfactory). “Satisfactory” meant the entire health news met the given criterion and “unsatisfactory” meant the opposite.

The last component visualized and interpreted the evaluation results provided by the health news quality-evaluation system. For example, for the criterion “Does the news adequately explain or quantify the harms of the intervention?” the method highlighted sentences that described the harms of intervention to help users quickly understand how well the criterion was met.

We examined 2 approaches to achieve this goal. The first was a hybrid approach (the hybrid approach). It was inspired by principles from rule-based systems, where patterns are cospecified by LIME and experts. The second approach (the typology approach) was a supervised sentence typology classification method, where hand-labeled training data are analyzed algorithmically to build models that can detect similar patterns when applied to unseen data.

Figure 1. Overview of the exploratory experiment.



Data Description and Collection

The data set that we used was adapted from an existing resource created by HealthNewsReview.org [69]. HealthNewsReview.org is a web-based project that reviewed articles from 2005 to 2018. Their team of experts rated the claims about health care interventions to improve the quality of health care information. Their rating instrument included 10 criteria used by the Australian and Canadian Media Doctor sites, and its interreviewer reliability was tested using a random sample of 30 stories [70]. HealthNewsReview.org included reviews of news stories from leading US media and news releases from institutes. The contents included efficacy claims about specific treatments, tests, products, or procedures. The news pieces were assessed using a standard rating system. At least 2 reviewers reviewed each news story. The reviewers were selected based on their years of experience in the health domain, spanning the fields of journalism, medicine, health services research, public health, or as patients, and each of them signed an industry-independent disclosure agreement. For each news story or news release reviewed, the criteria were scored as “satisfactory,” “unsatisfactory,” or “not applicable.” Total scores were posted for articles with ≤2 “not applicable” ratings and were expressed as proportions. It was acknowledged that increasing the diversity and independence of the reviewers could have reduced the potential for bias in the assessments. By the time the project ended, the website had accumulated 2616 health story reviews and 606 news release reviews.

For this study, we crawled health story news reviews and news release reviews, as archived by HealthNewsReviews.org, complying with the robots.txt. We scraped news contents that corresponded to the acquired reviews. Then, we visualized the results for the three selected criteria: (1) “Does the news adequately discuss the costs of the intervention?” (the cost criterion), (2) “Does the news adequately explain or quantify the harms of the intervention?” (the harm criterion), and (3) **“**Does the news identify conflicts of interest?” (the conflict criterion).

Automating the Criterion Evaluation

All 3 criteria applied to both news types, so we merged the 2 types of news content and treated them uniformly. We also combined health news that was scored as “unsatisfactory” or “not applicable” and named them as “unsatisfactory.” We preprocessed all news content via multiple text processing techniques, including removal of nonword elements (numbers, assented characters, and punctuation) and stop words, tokenization, stemming, and lemmatization. Then, we converted the textual representation into a vector space model using term frequency–inverse document frequency (TF-IDF).

We chose 4 representative algorithms: logistic regression, naive Bayes, support vector machine, and RF, from which we selected the best base algorithm that was suitable for automating the criterion evaluation. The 4 algorithms are commonly used in health misinformation classification tasks, as evident in previous studies [36,38,39,46,51,59], and were found to be effective. We applied RandomSearch to determine the optimal model hyperparameters for building the classifier. For our study, we defined the best classifier output from RandomSearch as the feature count, hyperparameter, and algorithm combination that produced the highest mean 5–cross-validated area under the curve (AUC) score. The performance of the classifier was further evaluated through 50-repeated 10–fold-validation.

Visualizing the Interpretation of Evaluation Result

We experimented with 2 approaches to visualize the interpretation of the evaluation results. The desired outcome was that all highlighted sentences were relevant to the examined criterion and provided evidence to assist end users in

comprehending and validating the evaluation results. To determine what qualified a sentence as evidence, we strictly adhered to the criteria definitions and review guidelines provided by HealthNewsReview.org [71-73]. For example, as per the explanation of the harm criterion provided by HealthNewsReview.org, satisfactory health news on the harm criterion should "include a discussion of harms and side effects, as well any measured 'adverse events' in a study" [71]. The measured "adverse events" can be addressed by a discussion of "both frequency of side effects and severity of side effects" and a discussion of "both major and minor side effects" [71].

The Hybrid Approach

The hybrid approach combined the interpretable AI technique, LIME, rule-based systems, and supervised document classification. LIME, proposed in 2016 by Ribeiro et al [74], belonged to a family of local model-agnostic methods, a type of interpretable AI method. It is used to explain the individual predictions of black-box ML based on a surrogate model, which is trained to approximate the predictions of the underlying black-box model [74,75]. The intuition of LIME is based on the idea that the behavior of a black-box model can be learned by perturbing the input. Specifically, a modified data set is generated by LIME through permutation by removing word features, corresponding to which predictions are obtained from the black-box model. Words with feature weights >0 indicate that the removal of such words affects the prediction result. For a negative case, no nonzero weight was estimated because regardless of which word was removed, the predicted evaluation result remained the same. Thus, an explanation can be generated by approximating the underlying model with a more interpretable model (such as a linear model or decision tree), learned locally on perturbations of the original instance [75]. Owing to the local fidelity nature of LIME, it does not guarantee a good global approximation [76]. A critique LIME often receives is that it lacks "stability" [77]. There are cases in which the surrogate model built by LIME can predict the instance correctly but provide incorrect reasons [75]. To address the instability of LIME, adding manually selected keywords can reduce the risk of obtaining incorrect keywords for highlighting. In this approach, we adopted the LIME method to facilitate the interpretable result of the predicted criterion evaluation. The Python packages used for implementing LIME algorithms were ELI5 [65] and LIME [63] application programming interface packages.

The explanation of the classification model for each criterion using the hybrid approach consisted of 3 steps. First, an ML classifier classifies health news as satisfactory or unsatisfactory based on the chosen criterion. Then, the classification model learned the difference of word distribution in satisfactory or unsatisfactory instances from the collection of health news document sets. LIME highlighted keywords in texts that contributed to the prediction. The keywords were also ranked using a weighted score, indicating their contribution to the prediction. Finally, we combined the keywords that contributed to a satisfactory prediction with a list of manually selected keywords, as shown in Table 1. The manual selection of the keywords was based on a consensus among the annotators who had taken part in the processes of evidence extraction for the typology approach.

We then extended the highlighting from the keyword to the sentence level to enhance the final visual representation. Sentences containing keywords with more weight were prioritized for highlighting. By default, manually selected keywords outweighed any keywords automatically picked by LIME.

Table 1. Lists of manually selected keywords for the cost, harm, and conflict criteria.

Criterion	Manually selected keywords
The cost criterion	Price, cost, charge, insurance, and pay
The harm criterion	Side effect, adverse reaction, adverse event, complication, and risk
The conflict criterion	Fund, sponsor, grant, spokesman, professor, and director

The Typology Approach

The typology approach was a sentence-level text-classification task. This approach was inspired by the study of persuasive communication and rhetoric. Reynolds and Reynolds [78] distinguished between statistical, testimonial, anecdotal, and analogical evidence. Hoeken and Hustinx [79] put forward 4 types of evidence in argumentation: individual examples, statistics, causal explanations, and expert opinions. Subsequent studies showed that machines can detect various types of evidence. For example, Fiok et al [80] built a classification model to automatically identify the evidence of respect in Twitter communication. There were 2 types of sentences in each health news item in our study. In the harm criterion, the first type of sentences was the evidence that supported the predicted evaluation result. Sentences of this type contained a description of side effects, including the symptoms, severity, and frequency of the symptoms. The second type of sentence referred to those that could not justify why a piece of certain health news satisfied a given criterion. Therefore, they were not characterized as evidence.

To implement the typology approach, for each criterion task, we designed and experimented with the typology approach in 2 stages. The first stage was to build an annotated data set of the sentence evidence. We extracted sentence evidence from health news that was evaluated as satisfactory by HealthNewsReview.org. A total of 3 people performed the sentence extraction tasks. The project investigator provided training and clarification to the other 2 extractors. The sentence extraction guideline fully adopted the criteria explained by HealthNewsReivew.org [71-73]. Two people performed most of the extraction work. Another individual worked as an independent reviewer to resolve disagreement. When combining the extracted sentences, sentences picked by the 2 extractors

were characterized as evidence. If a sentence was extracted by an extractor but not picked by the other, an independent third person was invited to resolve the disagreement. All approved sentences were considered positive. To build the negative class, we randomly selected the same number of sentences irrelevant to the evaluation of the pertinent criterion. An interannotator agreement was assessed using both simple counts and the percentage of the final quantity of the evidence in the total extracted items to address the relatively small sample size. The interrater agreement was also in line with the expectations of other studies [81]. The second stage involved building a supervised ML classifier. We followed the same steps as for automating the criterion evaluation.

For the final visual representation, the sentence classifier was applied to health news content to identify sentence evidence. Sentences with a higher probability of being categorized as evidence by the classifier were prioritized for highlighting purposes.

Evaluating and Optimizing the 2 Approaches

For each criterion's interpretation, we evaluated 2 visualization approaches to determine how accurately each scheme highlighted the sentences that supported the prediction result. The evaluation was conducted using 20 test cases. The selection of 20 test cases was based on the observation that the true positive health news counts in the test set (30% of the data set) ranged from 20 to 70, depending on the task criterion type. We measured the accuracy of 2 highlighting schemes by calculating the percentage of correctly highlighted evidence for all highlighted sentences. A total of 3 people evaluated the correctness of the highlighted sentence in accordance with each criterion's guideline. An independent reviewer was invited to handle any disputes.

As the number of highlighted sentences may affect the highlighting accuracy and thus the final visual representation, we calculated a spectrum of accuracies of both the highlighting approaches when the number of highlighted sentences increased from 1. A threshold was then selected with the lowest accuracy to determine the optimal range of sentence counts for highlighting.

Results

Classification Model Performance

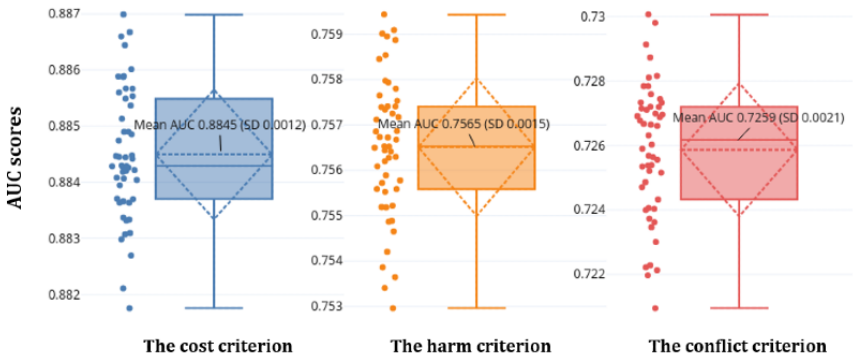
After removing dead links (to inaccessible news content), the acquired data set yielded 1453 stories and 579 news releases. Among the 2032 health news instances, the satisfactory or unsatisfactory instance ratios for the cost, harm, and conflict criteria were 25.03% (405/1618), 44.71% (625/1398), and 98.14% (1002/1021), respectively. Of the 4 experimental algorithms, RF was found to be the most effective in automating the evaluation of all the 3 criteria, as shown in Multimedia Appendix 1, despite the fact that the feature count varied according to the criterion. Table 2 shows the set of optimal hypermeters that RandomSearch selected for each criterion classifier.

For the cost, harm, and conflict criteria, Figure 2 shows that the average AUCs were 0.8845, 0.7565, and 0.7259, respectively, after 50 repeated 10-fold validations.

Table 2. Hyperparameters selected by RandomSearch for each criterion evaluation classifier.

Criteria	Base classifier	Word feature count, n	Hyperparameters
The cost criterion	Random forest	1000	("n_estimators": 600, "min_samples_split": 2, "min_samples_leaf": 4, "max_features": "sqrt," "max_depth": 10, and "bootstrap": false)
The harm criterion	Random forest	2000	("n_estimators": 1400, "min_samples_split": 10, "min_samples_leaf": 4, "max_features": "auto," "max_depth": 90, "bootstrap": false)
The conflict criterion	Random forest	1000	("n_estimators": 1200, "min_samples_split": 10, "min_samples_leaf": 1, "max_features": "auto," "max_depth": 20, and "bootstrap": true)

Figure 2. The performance of the cost, harm, and conflict criterion classifiers was measured with 10-fold cross-validated area under the curve (AUC) scores with a total of 50 repetitions.



Interpretable Model Performance

The Visual Interpretation by the Hybrid Approach

The LIME Text Explainer visualized how different word features contributed to the evaluation results for each classifier. Figure 3 illustrates the top 30 bigram or unigram word features that contributed to the classification learned from the entire data set related to a given criterion. For example, the binary word feature with the highest weight in the harm criterion classification was "side effect." Words that directly indicate the harm of intervention, such as "risk," "concern," "bleeding," and "harm," also ranked among the top features. Similarly, words that are commonly used to describe the intervention costs and insurance coverage such as "cost," "insurance," "expensive," and "pay" were also observed high in contribution to the evaluation for the cost criterion. For the conflicts criterion, the words were descriptive of one's affiliations such as "university," "dr," and "professor" stand out. The keyword "funded," which directly discloses funding information, also ranked high.

Figure 4 shows how LIME performed first-level visualization on a sample health news that was rated as satisfactory on the harm criterion. The classifier predicted the sample health news with a positive result of 65% probability. The words marked in orange were picked by LIME and explained as they contributed to the positive classification results of the model. Certain words were also highlighted in blue despite being scarce in number, indicating the likelihood of an unsatisfactory prediction. On the basis of the prediction result, the words "adverse," "reaction," "risk," "adverse," "serious," and "administration," were ranked among the most predictive words in the satisfactory classification result. A snapshot of the final visualized representation is shown in Figure 5, after highlighting sentences containing the keywords selected by LIME and the human expert. The 2-level visual interpretation cases for the cost and conflict criteria can be found in the Multimedia Appendix 2.

Figure 3. Top 30 word features with their feature weights in 3 criteria (the cost, harm, and conflict criteria) classifiers. The word feature weights signify how much discriminatory information each word contributes to the classification task by random forest algorithm.

Weight, mean(SD)	Feature
0.1070 (0.2493)	cost
0.0289 (0.0956)	price
0.0236 (0.0823)	insurance
0.0194 (0.0686)	expensive
0.0153 (0.0622)	pay
0.0147 (0.0532)	company
0.0132 (0.0409)	say
0.0121 (0.0441)	doctor
0.0112 (0.0367)	year
0.0088 (0.0443)	food drug
0.0085 (0.0361)	last
0.0083 (0.0317)	make
0.0078 (0.0372)	drug administration
0.0066 (0.0343)	said
0.0066 (0.0257)	approval
0.0074 (0.0246)	administration
0.0068 (0.0326)	would
0.0065 (0.0338)	last year
0.0057 (0.0214)	much
0.0050 (0.0218)	text
0.0049 (0.0249)	though
0.0048 (0.0280)	sale
0.0048 (0.0166)	one
0.0048 (0.0271)	mr
0.0046 (0.0229)	approved
0.0046 (0.0208)	several
0.0045 (0.0249)	fda
0.0044 (0.0209)	still
0.0044 (0.0178)	three
0.0044 (0.0199)	get
...970 more...	

Weight, mean(SD)	Feature
0.0178 (0.0545)	side
0.0166 (0.0512)	side effect
0.0082 (0.0224)	risk
0.0064 (0.0225)	concern
0.0063 (0.0179)	percent
0.0059 (0.0219)	serious
0.0059 (0.0215)	drug
0.0056 (0.0194)	expert
0.0053 (0.0160)	research
0.0051 (0.0179)	effect
0.0050 (0.0145)	may
0.0049 (0.0184)	review
0.0049 (0.0163)	american
0.0047 (0.0133)	study
0.0047 (0.0190)	safety
0.0047 (0.0183)	bleeding
0.0045 (0.0139)	one
0.0043 (0.0181)	adverse
0.0043 (0.0202)	food drug
0.0042 (0.0130)	say
0.0042 (0.0161)	percent patient
0.0040 (0.0130)	year
0.0040 (0.0177)	approved
0.0039 (0.0137)	many
0.0039 (0.0163)	severe
0.0039 (0.0116)	said
0.0039 (0.0160)	harm
0.0038 (0.0161)	safe
0.0037 (0.0159)	cancer institute
0.0036 (0.0159)	national cancer
...1970 more...	

Weight, mean(SD)	Feature
0.0106 (0.0330)	universitv
0.0080 (0.0247)	say
0.0073 (0.0244)	dr
0.0006 (0.0237)	funded
0.0054 (0.0214)	said dr
0.0054 (0.0213)	involved
0.0054 (0.0222)	national
0.0053 (0.0177)	study
0.0052 (0.0191)	one
0.0050 (0.0206)	professor
0.0048 (0.0168)	research
0.0045 (0.0196)	foundation
0.0040 (0.0175)	many
0.0037 (0.0142)	also
0.0036 (0.0142)	health
0.0036 (0.0149)	hospital
0.0036 (0.0152)	medical
0.0035 (0.0161)	would
0.0033 (0.0123)	said
0.0032 (0.0164)	grant
0.0032 (0.0133)	may
0.0031 (0.0140)	common
0.0031 (0.0133)	medicine
0.0030 (0.0135)	institute
0.0029 (0.0125)	cause
0.0029 (0.0151)	supported
0.0028 (0.0128)	show
0.0028 (0.0123)	center
0.0028 (0.0138)	dont
0.0026 (0.0139)	york
...970 more...	

Figure 4. Lime text explainer visualizes word's contribution to a satisfactory prediction on the harm criterion using random forest algorithm.

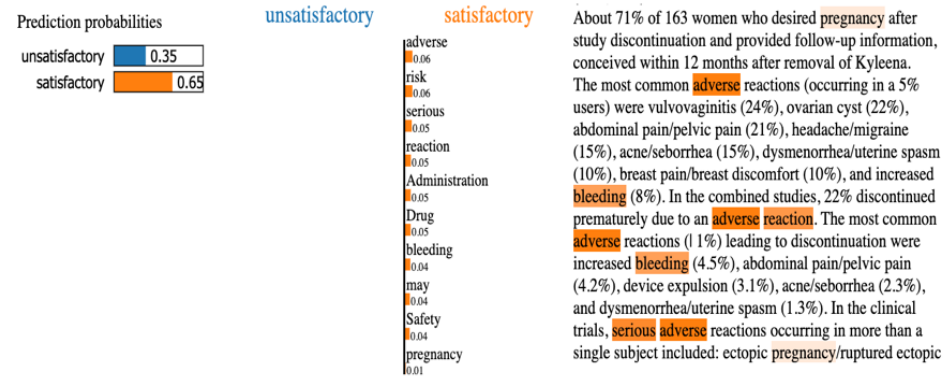
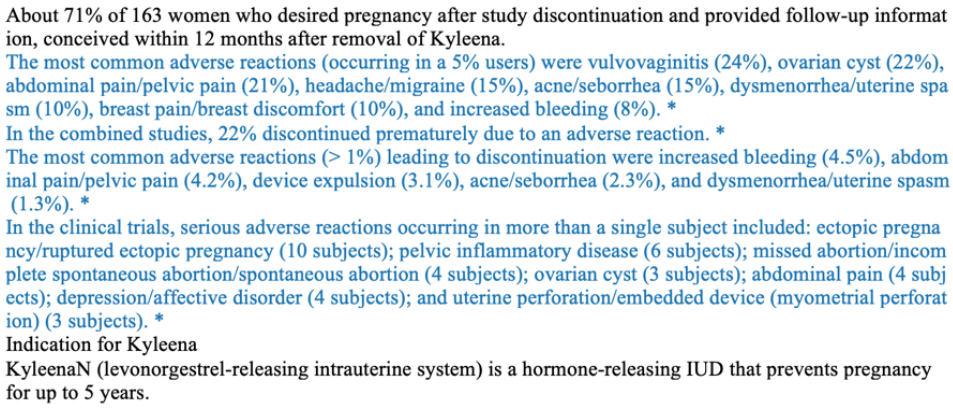


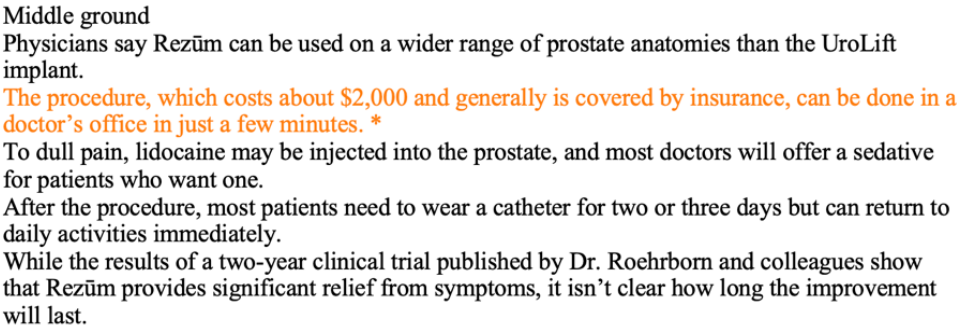
Figure 5. Example of a highlighting scheme for the harm criterion by the hybrid approach.



The Visual Interpretation by the Topology Approach

The interannotator agreement rates on evidence extraction for the cost, harm, and conflicts criteria were 72.04%, 72.24%, and 77.91%, respectively. The extraction task for each criterion yielded 201 (cost criterion), 318 (harm criterion), and 694 (conflict criterion) sentences in the positive class. We randomly selected the same number of sentences as the negative class to build the classification data sets. Following the same approach applied to the automation of criterion evaluation, which included base classifiers, word feature count selection, and hyperparameters tuning using RandomSearch, the classifiers of the 3 criteria attained an average AUC of 0.8791 (cost criterion), 0.7232 (harm criterion), and 0.8951 (conflict criterion) with 50 repetitions of 10–cross-fold validations. Figure 6 shows the result of applying the classifier to each sentence in the document and highlighting positive sentence instances that supported a cost criterion evaluation.

Figure 6. Example of a highlighting scheme for the cost criterion by the typology approach.



The Overall Performance and Optimization of the 2 Approaches

As the total number of highlighted sentences increased from 1, we calculated the varying rates of accurately highlighted sentences, as shown in Table 3. The numbers with footnotes suggest that the relevant approach could obtain a better result (accuracy >75%) within a certain number of sentences for highlighting.

According to Table 3, the accuracy of both approaches declined as the number of highlighted sentences increased. When both approaches highlighted the same number of sentences, the hybrid approach outperformed the typology approach in most scenarios. Typology, however, performed more accurately when the target was to pick <3 sentences to justify the harm criterion evaluation. When the threshold for highlighting accuracy was set at 75%, the optimal window size for the typology approach to achieve relatively better interpretation results was 2, 4, and 1 for the cost, harm, and conflict criteria, respectively. Comparatively, the hybrid approach still produced comparable outcomes when the window size for each criterion was extended by 2.

Table 3. The accuracy of both approaches for interpreting each criterion evaluation; maximum highlighting sentence count.

Number	Criterion and approach					
	Cost		Harm		Conflict	
	Typology (%)	Hybrid (%)	Typology (%)	Hybrid[a] (%)	Typology (%)	Hybrid (%)
1	80.00[a]	100.00[a]	100.00[a]	90.00	75.00[a]	90.00[a]
2	75.00[a]	92.50[a]	97.50[a]	90.00	72.50	87.50[a]
3	60.78	86.67[a]	86.67[a]	90.00	66.67	81.67[a]
4	66.67	76.25[a]	76.39[a]	85.00	61.11	68.75
5	60.00	67.00	72.94	81.00	55.29	59.00
6	54.55	59.65	66.67	75.83	56.41	50.93

[a]The relevant approach could obtain a better result (accuracy >75%) within a certain number of sentences for highlighting.

Discussion

Principal Findings

This study experimented with 2 AI-based approaches to visualize the interpretation of a criteria-based system designed to assist users in systematically evaluating the quality of health news.

The findings of our experiments were 3-fold. First, we found that both the hybrid and typology approaches could achieve the desired visualization result to justify the predicted evaluation result, despite the nature of the 2 approaches being differentiated. With 20 tests for each criterion, the performance of the hybrid approach was slightly better than that of the typology approach. Second, we were able to locate a window size to predetermine the sentences to be highlighted for a better visualization result for each criterion. The hybrid approach showed a higher capacity to reliably choose more sentences when the accuracy criterion was set at 75%. Third, the feasibility of the rule-based strategy to enhance LIME's interpretation work was supported by our observation during evidence extraction for the typology approach that specific words or phrases such as "adverse effect," "danger," "death," and "side effect" appeared repeatedly in the evaluation of the harm criterion; keywords such as "cost," "price," and "insurance" frequently appeared for the cost criterion evaluation; and "spokesman," "funding," and "sponsor" were typically used to disclose the conflicts of interests.

A Comparison of the 2 Approaches

The hybrid approach demonstrated both good accuracy and efficiency in visualizing the automatic model's interpretation for evaluating the 3 criteria. Compared with the typology approach, it was advantageous in saving manual effort because it did not require sentence extraction. We also observed that the hybrid approach tended to pick fewer sentences but with higher accuracy when not limiting the maximum number of sentences to be highlighted. By contrast, the typology approach selected more sentences, but only a few were relevant to the criterion.

However, the hybrid approach also had inherent weaknesses. The highlight scheme in the hybrid approach was to locate the sentence in which keywords were present. The drawback of this scheme was that it sometimes failed to discern the semantic differences between a sentence about the risk of the intervention and a sentence that described the benefits of the intervention by relieving or preventing adverse conditions. For example, in one of the test cases, the sentence, "Moreover, the study verified that long-term use of bisphosphonate drugs reduces the risk of typical osteoporosis fractures by 24 percent." was incorrectly highlighted. The sentence contained keywords, including "risk" and "fractures," which are relevant to adverse symptoms. However, it introduced how bisphosphonates are expected to benefit patients by decreasing the risk of negative outcomes. The other weakness associated with the hybrid approach was that it failed to distinguish between the intervention and stock prices. Both types of sentences typically shared many keywords that described the values associated with the intervention.

By contrast, the typology approach performed somewhat better at handling expressions with more lexical variations. For example, sentences, "Last fall the Food and Drug Administration issued a 'safety update' urging doctors and patients to be on the lookout for the problem." and "These medications are now linked to a growing number of complications, ranging in seriousness from nutrient deficiencies, joint pain and infections to bone fractures, heart attacks and dementia." were successfully picked by the typology approach; whereas they were missed by the hybrid approach, as keywords in those sentences were less commonly used to describe side effects. The typology approach distilled relevant information from text documents through sentence extraction by human experts. This information was key to building a knowledge base for the identification of sentences about side effects. We anticipated that the typology approach will be more robust and stable than the hybrid approach when visualizing the interpretation of criteria that are less keyword-reliant. For example, 1 of the 10 criteria, "Does the news compare the new approach with existing alternatives?" examined whether health news included a discussion on alternatives. Sentences that supported a satisfactory evaluation result may have been less likely to be observed with repetitive keywords than with the experimental criteria.

Limitations

This exploratory study had some limitations. The first limitation was that we only considered the TF-IDF values of words as

features for building both the document- and sentence-level classifiers. We acknowledged that the performance of our document-level classification model was lower compared with similar studies that adopted the same data set from HealthNewsReview.org. The performance of our doc-level classification models for the harm, cost, and conflict criteria were 0.71, 0.82, and 0.67, respectively, when measured by F_1 and 0.76, 0.88, and 0.72 when measured by AUC. The performance was better compared with a study by Al-Jefri et al [58] that focused on building health news quality classification models. The precision performance for classifying the harm, cost, and conflict criteria was reported to be 74.61, 77.61, and 70.89, respectively. The study incorporated more features, such as TF-IDF, comparative forms, and named-entity recognition tags and strategically changed the feature selections for different criterion classification tasks. In another study by Afsana et al [59], which also aimed to achieve the same research goal, the performance of their models for the harm, cost, and conflict measures by weighted F_1-score was reported 0.84, 0.899, and 0.835, respectively. However, superior performance was achieved through extensive work on feature engineering with 53,012 features applied. Considering that the key focus of this study was to experiment with 2 interpretation approaches, which both mentioned studies lack, we believed that the current performance of models was effective in serving the purpose of the study. In the future, we will incorporate some work on feature engineering for both document-level classification and especially the typology approach, which is embodied as a sentence-level classifier.

The second limitation is the simple rules of the hybrid approach. The hybrid approach takes advantage of both human knowledge and an autogenerated keyword list generated by the LIME. However, existing rules provided by human experts were keyword-based and did not contain complex rules for handling various expression variants. As part of the future plan, we will implement more complex rules for the hybrid to address the weak spots of the hybrid to enable it to distinguish different types of sentences when they share similar lexicons but different semantics.

A further limitation of the study was the absence of a user study to investigate how the final visual interpretation generated by the 2 interpretation approaches would increase user trust in a black-box model, particularly in the context of evaluating the quality of health news to mitigate misinformation. However, we have an ongoing user study to investigate whether a criteria-based system with visualized interpretation for evaluating health news quality will increase the trust of users compared with the system without interpretation. As of the completion of this study, the user study is still in the recruitment phase.

Comparison With Prior Work

Our study addressed the public's need to help evaluate the quality of health news and the typical opaqueness of an AI approach. The significance of this study is illustrated in 2 ways.

First, compared with previous interpretability work in suggested health-related misinformation detection systems, our work on adding the interpretability of a health misinformation system is innovative. To our knowledge, the current state of the art in explainable misinformation detection systems mostly looks to provide explanations for veracity predictions concerning inputs to the system. Our study fills a gap in the literature by explaining a criteria-based system for health misinformation. Moreover, developing an interpretable module on a criteria-based model is advantageous. The criteria-based approach inherently looks for the linguistic characteristics of health news, such as the presence or absence of crucial information, whereas a veracity-based system may face a challenge to be interpreted based on the linguistic features of text alone. In addition, we believe that our study exhibited a greater level of readability of the interpretation than the existing interpretation work on health misinformation, such as Alharbi et al [80] for fake news. The interpretation level achieved in the study by Alharbi et al remained at the word level, with both positive and negative words highlighted and dispersed throughout the articles; whereas our study presented 2 approaches to achieve sentence-level visualized interpretation, which demonstrated higher levels of readability to end users.

Second, this exploratory study demonstrated great potential for the development of a criteria-based system for evaluating the quality of health news as a way to counteract health misinformation. Compared with a veracity-based health misinformation detection system, a criteria-based system demonstrated high generalizability in handling health information on various topics. Most existing veracity-based fake news detectors are built on linguistic cues, leading to a lack of generalizability across topics, languages, and domains [82]. This weakness was also proven in a study by Gerts et al [41], as the team found a huge variation in the classifier performance (F_1-scores between 0.347 and 0.857) on 4 conspiracy topics and more narrowly defined topics could increase performance. In comparison, the idea of a criteria-driven system was to evaluate the quality of health news based on evidence for specified criteria. The evaluation procedure did not require a significant amount of domain knowledge. Thus, this type of system can be adapted to handle a variety of health news stories on various themes, as it did not rely on a data set with a strictly defined topic. In addition, an interpretable, criteria-based system may address the complexity and multidimensional attributes of the health information disorder [83-85]. Automatic tools for evaluating health misinformation have proven promising owing to their high accuracy and fast processing speed. However, existing studies are still predominantly binary classification tasks. This places a great challenge in identifying health misinformation, as the binary label is insufficient to represent the complicated evaluation process of health news in actual practice. This is especially the case with the veracity-based classification. Human-based fact-checking involves extensive knowledge understanding, inference, and source tracking, which remains a challenge, even in deep learning methods. This is because fabricated news is intended to mirror the truth to deceive readers; as a result, without cross-referencing and high-level inference, it might be impossible to determine the authenticity of news stories by text analysis alone [82]. Although it does not provide

veracity-level health news validation for users, it has the potential to provide another way of combating health misinformation by improving users' critical thinking about health news, as the slogan on HealthNewsReview.org indicates.

Conclusions

In this study, we described an interpretable, criteria-based strategy for evaluating the quality of health news. We explored 2 methods for visualizing the interpretation of the system. To aid in the exploration, an experiment was developed by comparing rule-based and statistical ML approaches. Our results suggested that either approach can successfully automate criterion-based health news quality ratings, with visual evidence supporting model explanation. This study has the potential to increase public trust in computer-assisted reviews of health information. We intend to expand on this study by applying 2 visualization approaches to more criteria and focusing on improving the performance of the classification model.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Performance of different base classifiers for automating 3 criteria evaluation.
[DOCX File , 27 KB - ai_v1i1e37751_app1.docx]

Multimedia Appendix 2
The 2-level visual interpretation cases for the cost and the conflict criteria.
[DOCX File , 11543 KB - ai_v1i1e37751_app2.docx]

References

1. Fox S. Health Topics. Pew Research Center. 2011 Feb 1. URL: https://www.pewresearch.org/internet/2011/02/01/health-topics-2/ [accessed 2021-09-21]
2. Becker BF, Larson HJ, Bonhoeffer J, van Mulligen EM, Kors JA, Sturkenboom MC. Evaluation of a multinational, multilingual vaccine debate on Twitter. Vaccine 2016 Dec 07;34(50):6166-6171. [doi: 10.1016/j.vaccine.2016.11.007] [Medline: 27840012]
3. Bonnevie E, Goldbarg J, Gallegos-Jeffrey AK, Rosenberg SD, Wartella E, Smyser J. Content themes and influential voices within vaccine opposition on Twitter, 2019. Am J Public Health 2020 Oct;110(S3):S326-S330. [doi: 10.2105/AJPH.2020.305901] [Medline: 33001733]
4. Hanson CL, Cannon B, Burton S, Giraud-Carrier C. An exploration of social circles and prescription drug abuse through Twitter. J Med Internet Res 2013 Sep 06;15(9):e189 [FREE Full text] [doi: 10.2196/jmir.2741] [Medline: 24014109]
5. Jamison A, Broniatowski DA, Smith MC, Parikh KS, Malik A, Dredze M, et al. Adapting and extending a typology to identify vaccine misinformation on Twitter. Am J Public Health 2020 Oct;110(S3):S331-S339. [doi: 10.2105/AJPH.2020.305940] [Medline: 33001737]
6. Buchanan R, Beckett RD. Assessment of vaccination-related information for consumers available on Facebook. Health Info Libr J 2014 Sep;31(3):227-234 [FREE Full text] [doi: 10.1111/hir.12073] [Medline: 25041499]
7. Faasse K, Chatman CJ, Martin LR. A comparison of language use in pro- and anti-vaccination comments in response to a high profile Facebook post. Vaccine 2016 Nov 11;34(47):5808-5814. [doi: 10.1016/j.vaccine.2016.09.029] [Medline: 27707558]
8. Johnson SB, Parsons M, Dorff T, Moran MS, Ward JH, Cohen SA, et al. Cancer misinformation and harmful information on Facebook and other social media: a brief report. J Natl Cancer Inst 2022 Jul 11;114(7):1036-1039 [FREE Full text] [doi: 10.1093/jnci/djab141] [Medline: 34291289]
9. Seymour B, Getman R, Saraf A, Zhang LH, Kalenderian E. When advocacy obscures accuracy online: digital pandemics of public health misinformation through an antifluoride case study. Am J Public Health 2015 Mar;105(3):517-523. [doi: 10.2105/AJPH.2014.302437] [Medline: 25602893]
10. Abukaraky A, Hamdan AA, Ameera MN, Nasief M, Hassona Y. Quality of YouTube TM videos on dental implants. Med Oral Patol Oral Cir Bucal 2018 Jul 01;23(4):e463-e468 [FREE Full text] [doi: 10.4317/medoral.22447] [Medline: 29924766]
11. Basch CH, Zybert P, Reeves R, Basch CE. What do popular YouTube videos say about vaccines? Child Care Health Dev 2017 Jul;43(4):499-503. [doi: 10.1111/cch.12442] [Medline: 28105642]
12. Biggs TC, Bird JH, Harries PG, Salib RJ. YouTube as a source of information on rhinosinusitis: the good, the bad and the ugly. J Laryngol Otol 2013 Aug;127(8):749-754. [doi: 10.1017/S0022215113001473] [Medline: 23866821]
13. Röchert D, Neubaum G, Stieglitz S. Identifying political sentiments on YouTube: a systematic comparison regarding the accuracy of recurrent neural network and machine learning models. In: Proceedings of the 2nd Multidisciplinary International Symposium on Disinformation in Open Online Media. 2020 Presented at: MISDOOM '20; October 26–27, 2020; Leiden, The Netherlands p. 107-121 URL: https://link.springer.com/chapter/10.1007/978-3-030-61841-4_8 [doi: 10.1007/978-3-030-61841-4_8]

14. Guidry J, Jin Y, Haddad L, Zhang Y, Smith J. How health risks are pinpointed (or not) on social media: the portrayal of waterpipe smoking on Pinterest. Health Commun 2016;31(6):659-667. [doi: 10.1080/10410236.2014.987468] [Medline: 26512916]
15. Guidry JP, Carlyle K, Messner M, Jin Y. On pins and needles: how vaccines are portrayed on Pinterest. Vaccine 2015 Sep 22;33(39):5051-5056. [doi: 10.1016/j.vaccine.2015.08.064] [Medline: 26319742]
16. Li A, Huang X, Jiao D, O'Dea B, Zhu T, Christensen H. An analysis of stigma and suicide literacy in responses to suicides broadcast on social media. Asia Pac Psychiatry 2018 Mar;10(1):e12314. [doi: 10.1111/appy.12314] [Medline: 29383880]
17. Xiao L, Chen S. Misinformation in the Chinese Weibo. In: Proceedings of the 12th International Conference on Social Computing and Social Media. Design, Ethics, User Behavior, and Social Network Analysis. 2020 Presented at: HCII '20; July 19–24, 2020; Copenhagen, Denmark p. 407-418 URL: https://link.springer.com/chapter/10.1007/978-3-030-49570-1_28
18. Waszak PM, Kasprzycka-Waszak W, Kubanek A. The spread of medical fake news in social media – the pilot quantitative study. Health Policy Technol 2018 Jun;7(2):115-118. [doi: 10.1016/j.hlpt.2018.03.002]
19. Chua AY, Banerjee S. Intentions to trust and share online health rumors: an experiment with medical professionals. Comput Human Behav 2018 Oct;87:1-9. [doi: 10.1016/j.chb.2018.05.021]
20. Radzikowski J, Stefanidis A, Jacobsen KH, Croitoru A, Crooks A, Delamater PL. The measles vaccination narrative in Twitter: a quantitative analysis. JMIR Public Health Surveill 2016 Jan 4;2(1):e1 [FREE Full text] [doi: 10.2196/publichealth.5059] [Medline: 27227144]
21. Vraga EK, Bode L. Using expert sources to correct health misinformation in social media. Sci Commun 2017 Sep 14;39(5):621-645. [doi: 10.1177/1075547017731776]
22. Chen B, Shao J, Liu K, Cai G, Jiang Z, Huang Y, et al. Does eating chicken feet with pickled peppers cause avian influenza? Observational case study on Chinese social media during the avian influenza a (H7N9) outbreak. JMIR Public Health Surveill 2018 Mar 29;4(1):e32 [FREE Full text] [doi: 10.2196/publichealth.8198] [Medline: 29599109]
23. Li Y, Zhang X, Wang S. Fake vs. real health information in social media in China. Proc Assoc Info Sci Tech 2017 Oct 24;54(1):742-743. [doi: 10.1002/pra2.2017.14505401139]
24. Leong AY, Sanghera R, Jhajj J, Desai N, Jammu BS, Makowsky MJ. Is YouTube useful as a source of health information for adults with type 2 diabetes? A South Asian perspective. Can J Diabetes 2018 Aug;42(4):395-403.e4. [doi: 10.1016/j.jcjd.2017.10.056] [Medline: 29282200]
25. Aquino F, Donzelli G, De Franco E, Privitera G, Lopalco PL, Carducci A. The web and public confidence in MMR vaccination in Italy. Vaccine 2017 Aug 16;35(35 Pt B):4494-4498. [doi: 10.1016/j.vaccine.2017.07.029] [Medline: 28736200]
26. Bessi A, Zollo F, Del Vicario M, Scala A, Caldarelli G, Quattrociocchi W. Trend of narratives in the age of misinformation. PLoS One 2015 Aug 14;10(8):e0134641 [FREE Full text] [doi: 10.1371/journal.pone.0134641] [Medline: 26275043]
27. Vraga EK, Bode L. Defining misinformation and understanding its bounded nature: using expertise and evidence for describing misinformation. Polit Commun 2020 Feb 06;37(1):136-144. [doi: 10.1080/10584609.2020.1716500]
28. Chou WS, Oh A, Klein WM. Addressing health-related misinformation on social media. JAMA 2018 Dec 18;320(23):2417-2418. [doi: 10.1001/jama.2018.16865] [Medline: 30428002]
29. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021 Jan 20;23(1):e17187 [FREE Full text] [doi: 10.2196/17187] [Medline: 33470931]
30. Bianco V. Countering Online Misinformation Resource Pack. UNICEF Regional Office for Europe and Central Asia. 2020 Aug. URL: https://www.unicef.org/eca/media/13636/file [accessed 2021-10-26]
31. Fighting misinformation in the time of COVID-19, one click at a time. World Health Organization. 2021 Apr 27. URL: https://www.who.int/news-room/feature-stories/detail/fighting-misinformation-in-the-time-of-covid-19-one-click-at-a-time [accessed 2022-09-21]
32. Bridgman A, Merkley E, Zhilin O, Loewen PJ, Owen T, Ruths D. Infodemic pathways: evaluating the role that traditional and social media play in cross-national information transfer. Front Polit Sci 2021 Mar 29;3:20. [doi: 10.3389/fpos.2021.648646]
33. Cui L, Lee D. CoAID: COVID-19 healthcare misinformation dataset. arXiv 2020 May 22 [FREE Full text]
34. Marr B. Fake News Is Rampant, Here Is How Artificial Intelligence Can Help. Forbes. 2021 Jan 25. URL: https://www.forbes.com/sites/bernardmarr/2021/01/25/fake-news-is-rampant-here-is-how-artificial-intelligence-can-help/ [accessed 2021-12-15]
35. Burns E, Laskowski N, Tucci L. What is Artificial Intelligence (AI)? - AI Definition and How it Works. SearchEnterpriseAI. URL: https://searchenterpriseai.techtarget.com/definition/AI-Artificial-Intelligence [accessed 2021-10-26]
36. Meppelink CS, Hendriks H, Trilling D, van Weert JC, Shao A, Smit ES. Reliable or not? An automated classification of webpages about early childhood vaccination using supervised machine learning. Patient Educ Couns 2021 Jun;104(6):1460-1466 [FREE Full text] [doi: 10.1016/j.pec.2020.11.013] [Medline: 33243581]
37. Shah Z, Surian D, Dyda A, Coiera E, Mandl KD, Dunn AG. Automatically appraising the credibility of vaccine-related web pages shared on social media: a Twitter surveillance study. J Med Internet Res 2019 Nov 04;21(11):e14007 [FREE Full text] [doi: 10.2196/14007] [Medline: 31682571]
38. Wang Z, Yin Z, Argyris YA. Detecting medical misinformation on social media using multimodal deep learning. IEEE J Biomed Health Inform 2021 Jun;25(6):2193-2203. [doi: 10.1109/JBHI.2020.3037027] [Medline: 33170786]

39. Ghenai A, Mejova Y. Catching Zika fever: application of crowdsourcing and machine learning for tracking health misinformation on Twitter. In: Proceedings of the 2017 IEEE International Conference on Healthcare Informatics. 2017 Jul 12 Presented at: ICHI '17; August 23-26, 2017; Park City, UT, USA p. 518. [doi: 10.1109/ichi.2017.58]
40. Zhao Y, Da J, Yan J. Detecting health misinformation in online health communities: incorporating behavioral features into machine learning based approaches. Inf Process Manag 2021 Jan;58(1):102390. [doi: 10.1016/j.ipm.2020.102390]
41. Gerts D, Shelley CD, Parikh N, Pitts T, Watson Ross C, Fairchild G, et al. "Thought I'd share first" and other conspiracy theory tweets from the COVID-19 infodemic: exploratory study. JMIR Public Health Surveill 2021 Apr 14;7(4):e26527 [FREE Full text] [doi: 10.2196/26527] [Medline: 33764882]
42. Abdelminaam DS, Ismail FH, Taha M, Taha A, Houssein EH, Nabil A. CoAID-DEEP: an optimized intelligent framework for automated detecting COVID-19 misleading information on Twitter. IEEE Access 2021 Feb 9;9:27840-27867 [FREE Full text] [doi: 10.1109/ACCESS.2021.3058066] [Medline: 34786308]
43. Ayoub J, Yang XJ, Zhou F. Combat COVID-19 infodemic using explainable natural language processing models. Inf Process Manag 2021 Jul;58(4):102569 [FREE Full text] [doi: 10.1016/j.ipm.2021.102569] [Medline: 33776192]
44. Kolluri NL, Murthy D. CoVerifi: a COVID-19 news verification system. Online Soc Netw Media 2021 Mar;22:100123 [FREE Full text] [doi: 10.1016/j.osnem.2021.100123] [Medline: 33521412]
45. Dhoju S, Main Uddin Rony M, Ashad Kabir M, Hassan N. Differences in health news from reliable and unreliable media. In: Companion Proceedings of The 2019 World Wide Web Conference. 2019 May Presented at: WWW '19; May 13-17, 2019; San Francisco, CA, USA p. 981-987. [doi: 10.1145/3308560.3316741]
46. Saengkhunthod C, Kerdnoonwong P, Atchariyachanvanich K. Detection of unreliable medical articles on Thai websites. In: Proceedings of the 13th International Conference on Knowledge and Smart Technology. 2021 Presented at: KST '21; January 21-24, 2021; Bangsaen, Thailand p. 102-107. [doi: 10.1109/kst51265.2021.9415756]
47. Dito FM, Alqadhi HA, Alasaadi A. Detecting medical rumors on Twitter using machine learning. In: Proceedings of the 2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies. 2020 Presented at: 3ICT '20; December 20-21, 2020; Sakheer, Bahrain p. 1-7. [doi: 10.1109/3ict51146.2020.9311957]
48. Kinsora A, Barron K, Mei Q, Vydiswaran VG. Creating a labeled dataset for medical misinformation in health forums. In: Proceedings of the 2017 IEEE International Conference on Healthcare Informatics. 2017 Presented at: ICHI '17; August 23-26, 2017; Park City, UT, USA p. 456-461. [doi: 10.1109/ichi.2017.93]
49. Parfenenko Y, Verbytska A, Bychko D, Shendryk V. Application for medical misinformation detection in online forums. In: Proceedings of the 2020 International Conference on e-Health and Bioengineering. 2020 Presented at: EHB '20; October 29-30, 2020; Iasi, Romania p. 1-4. [doi: 10.1109/ehb50910.2020.9280120]
50. Liu Y, Yu K, Wu X, Qing L, Peng Y. Analysis and detection of health-related misinformation on Chinese social media. IEEE Access 2019 Oct 14;7:154480-154489. [doi: 10.1109/ACCESS.2019.2946624]
51. Elhadad MK, Li KF, Gebali F. Detecting misleading information on COVID-19. IEEE Access 2020 Sep 9;8:165201-165215 [FREE Full text] [doi: 10.1109/ACCESS.2020.3022867] [Medline: 34786288]
52. Khanday A, Khan QR, Rabani ST. Identifying propaganda from online social networks during COVID-19 using machine learning techniques. Int J Inf Technol 2021;13(1):115-122 [FREE Full text] [doi: 10.1007/s41870-020-00550-5] [Medline: 33145473]
53. Snopes-Medical Archives. Snopes.com. URL: https://www.snopes.com/fact-check/category/medical/ [accessed 2022-02-15]
54. World Health Organization. URL: https://www.who.int [accessed 2022-02-15]
55. Johns Hopkins Medicine, based in Baltimore, Maryland. URL: https://www.hopkinsmedicine.org/ [accessed 2022-02-15]
56. CDC Works 24/7. Centers for Disease Control and Prevention. 2022. URL: https://www.cdc.gov/index.htm [accessed 2022-02-15]
57. Zhou X, Zafarani R. A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Comput Surv 2021 Sep;53(5):1-40. [doi: 10.1145/3395046]
58. Al-Jefri M, Evans R, Lee J, Ghezzi P. Automatic identification of information quality metrics in health news stories. Front Public Health 2020 Dec 18;8:515347 [FREE Full text] [doi: 10.3389/fpubh.2020.515347] [Medline: 33392124]
59. Afsana F, Kabir MA, Hassan N, Paul M. Automatically assessing quality of online health articles. IEEE J Biomed Health Inform 2021 Feb;25(2):591-601. [doi: 10.1109/JBHI.2020.3032479] [Medline: 33079686]
60. Hawa S, Lobo L, Dogra U, Kamble V. Combating misinformation dissemination through verification and content driven recommendation. In: Proceedings of the 3rd International Conference on Intelligent Communication Technologies and Virtual Mobile Networks. 2021 Presented at: ICICV '21; February 4-6, 2021; Tirunelveli, India p. 917-924. [doi: 10.1109/icicv50876.2021.9388406]
61. Choudrie J, Banerjee S, Kotecha K, Walambe R, Karende H, Ameta J. Machine learning techniques and older adults processing of online information and misinformation: a covid 19 study. Comput Human Behav 2021 Jun;119:106716 [FREE Full text] [doi: 10.1016/j.chb.2021.106716] [Medline: 34866770]
62. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. Entropy (Basel) 2020 Dec 25;23(1):18 [FREE Full text] [doi: 10.3390/e23010018] [Medline: 33375658]
63. Ribeiro MT. lime: Local Interpretable Model-Agnostic Explanations for machine learning classifiers. GitHub. 2021 Jul 30. URL: http://github.com/marcotcr/lime [accessed 2022-02-15]

64. Lundberg S. shap: A unified approach to explain the output of any machine learning model. GitHub. URL: https://github.com/slundberg/shap [accessed 2022-02-15]
65. Korobov M, Lopuhin K. eli5: Debug machine learning classifiers and explain their predictions. GitHub. URL: https://github.com/eli5-org/eli5 [accessed 2022-02-15]
66. InterpretML Team. interpret: Fit interpretable machine learning models - Explain blackbox machine learning. GitHub. URL: https://github.com/interpretml/interpret [accessed 2022-02-15]
67. Kotonya N, Toni F. Explainable automated fact-checking for public health claims. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020 Oct 19 Presented at: EMNLP '20; November 16-18, 2020; Virtual p. 7740-7754. [doi: 10.18653/v1/2020.emnlp-main.623]
68. Zuo C, Zhang Q, Banerjee R. An empirical assessment of the qualitative aspects of misinformation in health news. In: Proceedings of the 4th Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda. 2021 Presented at: NLP4IF '21; June 6, 2021; Virtual p. 76-81. [doi: 10.18653/v1/2021.nlp4if-1.11]
69. HealthNewsReview. URL: https://www.healthnewsreview.org/ [accessed 2022-02-15]
70. Schwitzer G. How do US journalists cover treatments, tests, products, and procedures? An evaluation of 500 stories. PLoS Med 2008 May 27;5(5):e95 [FREE Full text] [doi: 10.1371/journal.pmed.0050095] [Medline: 18507496]
71. Criterion #3 Does the story adequately explain/quantify the harms of the intervention? HealthNewsReview. URL: https://www.healthnewsreview.org/about-us/review-criteria/criterion-3/ [accessed 2022-02-15]
72. Criterion #1 Does the story adequately discuss the costs of the intervention? HealthNewsReview. URL: https://web.archive.org/web/20220629205927/http://www.healthnewsreview.org/about-us/review-criteria/criterion-1/ [accessed 2022-09-05]
73. Criterion #6 Does the story use independent sources and identify conflicts of interest? HealthNewsReview. URL: https://web.archive.org/web/20220629202851/https://www.healthnewsreview.org/about-us/review-criteria/criterion-6/ [accessed 2022-09-05]
74. Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Feb 16 Presented at: KDD '16; August 13-17, 2016; San Francisco, CA, USA p. 1135-1144.
75. Ribeiro MT, Singh S, Guestrin C. Local interpretable model-agnostic explanations (LIME): an introduction. O'Reilly Media. 2016 Aug 12. URL: https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/ [accessed 2022-02-15]
76. Molnar C. Interpretable Machine Learning. URL: https://christophm.github.io/interpretable-ml-book/ [accessed 2022-02-15]
77. Zafar MR, Khan N. Deterministic local interpretable model-agnostic explanations for stable explainability. Mach Learn Knowl Extr 2021 Jun 30;3(3):525-541. [doi: 10.3390/make3030027]
78. Reynolds RA, Reynolds JL. Evidence. In: Dillard JP, Pfau M, editors. The Persuasion Handbook: Developments in Theory and Practice. Thousand Oaks, CA, USA: Sage Publications; 2002:427-445.
79. Hoeken H, Hustinx L. The relative persuasiveness of anecdotal, statistical, causal, and expert evidence. In: Proceedings of the 5th Conference of the International Society for the Study of Argumentation. 2002 Presented at: ISSA '02; June 26-28, 2002; Amsterdam, The Netherlands p. 497-502 URL: https://repository.ubn.ru.nl/handle/2066/82921
80. Fiok K, Karwowski W, Gutierrez E, Liciaga T, Belmonte A, Capobianco R. Automated classification of evidence of respect in the communication through Twitter. Appl Sci 2021 Feb 01;11(3):1294. [doi: 10.3390/app11031294]
81. Freund AJ, Giabbanelli PJ. Are we modeling the evidence or our own biases? A comparison of conceptual models created from reports. In: Proceedings of the 2021 Annual Modeling and Simulation Conference. 2021 Presented at: ANNSIM '21; July 19-22, 2021; Fairfax, VA, USA p. 1-12. [doi: 10.23919/annsim52504.2021.9552054]
82. Sharma K, Qian F, Jiang H, Ruchansky N, Zhang M, Liu Y. Combating fake news: a survey on identification and mitigation techniques. ACM Trans Intell Syst Technol 2019 Apr;10(3):1-42. [doi: 10.1145/3305260]
83. Wardle C, Derakhshan H. Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making. Council of Europe. 2017. URL: https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html [accessed 2021-10-26]
84. Habgood-Coote J. Stop talking about fake news!. Inquiry 2018 Aug 11;62(9-10):1033-1065 [FREE Full text] [doi: 10.1080/0020174x.2018.1508363]
85. Wang Y, McKee M, Torbica A, Stuckler D. Systematic literature review on the spread of health-related misinformation on social media. Soc Sci Med 2019 Nov;240:112552 [FREE Full text] [doi: 10.1016/j.socscimed.2019.112552] [Medline: 31561111]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
LIME: local interpretable model-agnostic explanation
ML: machine learning
RF: random forest

TF-IDF: term frequency–inverse document frequency

Edited by K El Emam, B Malin; submitted 04.03.22; peer-reviewed by P Giabbanelli, L Bošnjak, B Puladi; comments to author 06.06.22; revised version received 22.09.22; accepted 11.11.22; published 20.12.22.

Please cite as:
Liu X, Alsghaier H, Tong L, Ataullah A, McRoy S
Visualizing the Interpretation of a Criteria-Driven System That Automatically Evaluates the Quality of Health News: Exploratory Study of 2 Approaches
JMIR AI 2022;1(1):e37751
URL: https://ai.jmir.org/2022/1/e37751
doi:10.2196/37751
PMID:38875559

Original Paper

Chronic Disease Prediction Using the Common Data Model: Development Study

Chanjung Lee[1], BA; Brian Jo[1], MD, PhD; Hyunki Woo[1], MSc; Yoori Im[1], MPH; Rae Woong Park[2], MD, PhD; ChulHyoung Park[2], MD

[1]Evidnet, Seongnam, Republic of Korea

[2]Department of Biomedical Informatics, Ajou University Hospital, Suwon, Republic of Korea

Corresponding Author:
Rae Woong Park, MD, PhD
Department of Biomedical Informatics
Ajou University Hospital
164, World cup-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do
Suwon, 16499
Republic of Korea
Phone: 82 01073375540
Email: veritas@ajou.ac.kr

Abstract

Background: Chronic disease management is a major health issue worldwide. With the paradigm shift to preventive medicine, disease prediction modeling using machine learning is gaining importance for precise and accurate medical judgement.

Objective: This study aimed to develop high-performance prediction models for 4 chronic diseases using the common data model (CDM) and machine learning and to confirm the possibility for the extension of the proposed models.

Methods: In this study, 4 major chronic diseases—namely, diabetes, hypertension, hyperlipidemia, and cardiovascular disease—were selected, and a model for predicting their occurrence within 10 years was developed. For model development, the Atlas analysis tool was used to define the chronic disease to be predicted, and data were extracted from the CDM according to the defined conditions. A model for predicting each disease was built with 4 algorithms verified in previous studies, and the performance was compared after applying a grid search.

Results: For the prediction of each disease, we applied 4 algorithms (logistic regression, gradient boosting, random forest, and extreme gradient boosting), and all models show greater than 80% accuracy. As compared to the optimized model's performance, extreme gradient boosting presented the highest predictive performance for the 4 diseases (diabetes, hypertension, hyperlipidemia, and cardiovascular disease) with 80% or greater and from 0.84 to 0.93 in area under the curve standards.

Conclusions: This study demonstrates the possibility for the preemptive management of chronic diseases by predicting the occurrence of chronic diseases using the CDM and machine learning. With these models, the risk of developing major chronic diseases within 10 years can be demonstrated by identifying health risk factors using our chronic disease prediction machine learning model developed with the real-world data–based CDM and National Health Insurance Corporation examination data that individuals can easily obtain.

(JMIR AI 2022;1(1):e41030) doi:10.2196/41030

KEYWORDS

common data model; chronic disease; prediction model; machine learning; disease management; data model; disease prediction; prediction; risk prediction; risk factors; health risk

Introduction

World Health Organization's Global Action Plan (2013-2020) for noninfectious diseases aims to reduce the premature death rate stemming from chronic diseases by 25% by 2025 [1]. The plan also urges the establishment of national policies and management of performance indicators.

Accordingly, the Ministry of Health and Welfare of South Korea has designated cardiovascular disease, diabetes, chronic respiratory disease, and cancer as chronic diseases to be managed by the government [2] and established a chronic

disease management system centered on local hospitals. In March 2014, a community primary care pilot project for high blood pressure and patients with diabetes was initiated. In September 2016, the chronic disease management pilot project was carried out. From January 2019 to the present, a primary medical chronic disease management pilot project was conducted. Nevertheless, chronic diseases remain the primary cause of mortality and increasing medical expenses. According to the Korea Centers for Disease Control and Prevention, in 2020, chronic diseases were responsible for 7 out of 10 deaths in the country, accounting for 83.7% of total medical expenses [3].

Chronic diseases develop from metabolic syndrome that are caused by lifestyle or individual genetic and environmental factors [4]. The development of chronic disease leads to various complications or requires long-term treatment [5]. Therefore, it is important to take preemptive measures along with the prevention of metabolic syndrome. In this respect, it is necessary to develop various disease prediction models to reduce the risk of complications and medical costs.

Fortunately, early-stage disease prediction is gaining momentum with the use of real-world data combined with machine learning technology. Lee et al [6] predicted the risk of metabolic syndrome (area under the curve [AUC]=0.879) using machine learning techniques, and Choi et al [7] predicted disease occurrence using recurrent neural networks (diagnosis up to 79%). Lipton et al [8] predicted the probability of chronic disease by applying long short-term memory (AUC=0.81-0.99). However, the disease occurrence prediction models developed in South Korea used traditional statistical techniques, and most international predictive models were developed for Western White populations and therefore have reduced applicability to other countries and racial groups.

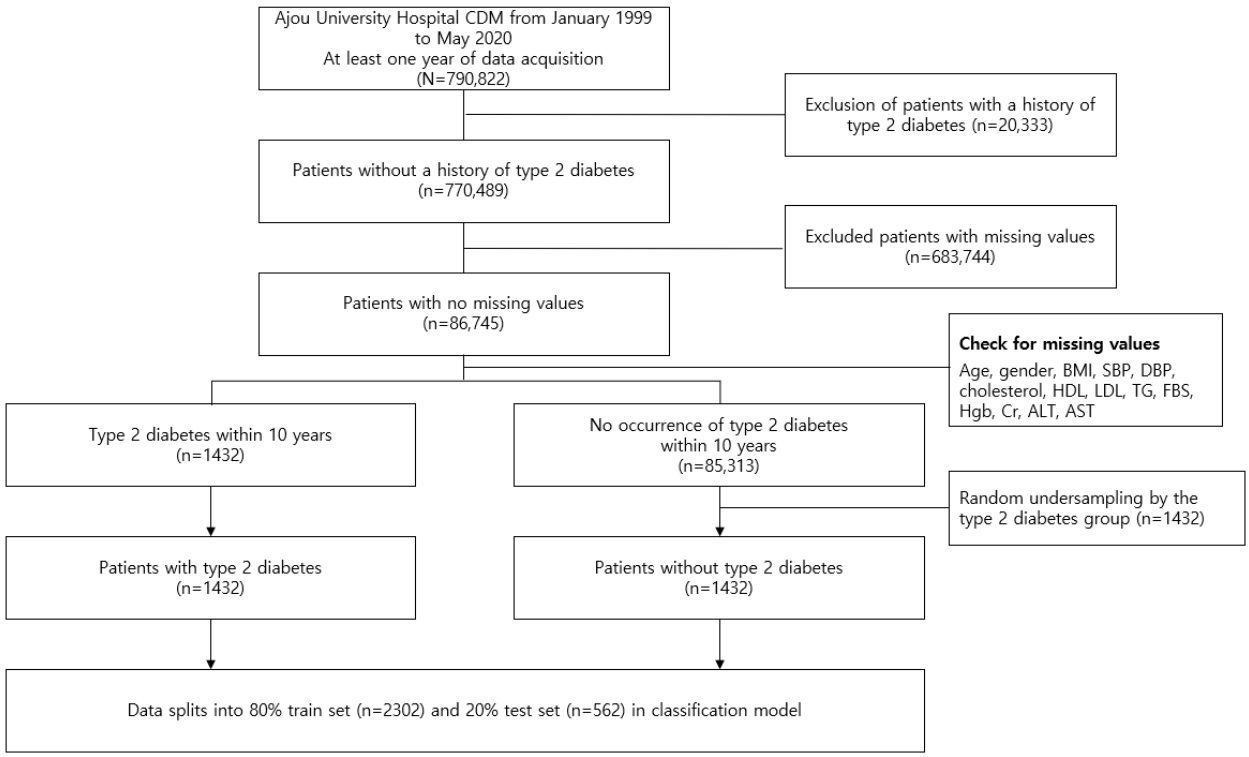
Although there are already many studies using electronic medical record (EMR) and machine learning, they have limitations in requiring a definition of medical terms or preprocessing for standardization in multicentered studies, entailing that these studies cannot be synchronized with other prediction models. There are relatively few papers on predictive model development using the common data model (CDM; although version 6.0 was release recently, version 5.4 of the Observational Medical Outcomes Partnership CDM is supported by the Observational Health Data Sciences and Informatics suite of tools and methods), which can overcome these limitations. In this paper, we aimed to develop a scalable chronic disease prediction model using the CDM.

Methods

Subjects

We used the data of 790,822 subjects with at least one year of hospital records among subjects aged ≥20 years who had also visited a tertiary hospital in South Korea (Ajou University Hospital in Suwon) from 1999 to 2020. To predict the risk of developing chronic diseases for the subjects as they age, patients with chronic diseases (type 2 diabetes, high blood pressure, hyperlipidemia, and cardiovascular disease) as the underlying disease were excluded (Figure 1).

Figure 1. The process of selecting subjects for the type 2 diabetes study. ALT: alanine aminotransferase; AST: aspartate aminotransferase; CDM: common data model; Cr: creatinine; DBP: diastolic blood pressure; FBS: fasting blood glucose; HDL: high-density lipoprotein; Hgb: hemoglobin; LDL: low-density lipoprotein; SBP: systolic blood pressure; TG: triglyceride.



Select Model Variables

The public health checkup is a test for adults aged >18 years in South Korea, and anyone can use it for free. Variables were selected based on the general examination of items from the National Health Insurance Service. A total of 19 variables were included, such as basic information, measurement information, lifestyle information, and history of diseases.

Data Extraction

The data used in the predictive model were extracted using the Atlas analysis tool (Observational Health Data Sciences and Informatics—a nonprofit consortium that allows researchers to perform design, characterization, and analysis). A cohort for chronic diseases was created through Atlas design for the variables used in the cohort. Concept IDs following the Systematized Nomenclature Of Medicine–Clinical Terms terminology were used, which are mapped to the International Classification of Diseases, 10th Revision code and currently used as a diagnostic name in clinical practice. Systematized Nomenclature Of Medicine–Clinical Terms were developed to meet the various needs and expectations of clinicians around the world, and it is an international standard terminology system used in more than 80 countries, helping to consistently express clinical contents in medical information records. Additionally, concept IDs following the Local Laboratory Result Code terminology, mapped with the managed local code, was used. Local Laboratory Result Code refers to international standard test terms, and medical terms are defined and standardized for the standardization of test codes. Table 1 shows the concept IDs used in the defined cohort group.

The defined cohort group was divided into a disease-occurring group and a nonoccurring group according to the presence or absence of a diagnosed chronic disease within 10 years from the index date (when the criteria for participation in the study were met). Cohort generation and data extraction were performed according to the design criteria shown in Textbox 1.

Table 1. Concept ID information.

Variables	Concept ID	Concept name	Type	Vocabulary
Type 2 diabetes	• 201826	Type 2 diabetes mellitus	Factor	SNOMED-CT[a]
Hypertension	• 316866	Hypertensive disorder	Factor	SNOMED-CT
Hyperlipidemia	• 432867	Hyperlipidemia	Factor	SNOMED-CT
Cardiovascular disease	• 134057	Disorder of cardiovascular system	Factor	SNOMED-CT
BMI	• 3038553	Body mass index (ratio)	Numeric	LOINC[b]
SBP[c]	• 3004249	Systolic blood pressure	Numeric	LOINC
DBP[d]	• 3012888	Diastolic blood pressure	Numeric	LOINC
Total cholesterol	• 3027114	Cholesterol (mass/volume) in serum or plasma	Numeric	LOINC
HDL[e]	• 3007070	Cholesterol in high-density lipoprotein (mass/volume) in serum or plasma	Numeric	LOINC
LDL[f]	• 3028437	Cholesterol in low-density lipoprotein (mass/volume) in serum or plasma	Numeric	LOINC
TG[g]	• 3022038 • 3022192	Triglyceride (mass/volume) in serum or plasma	Numeric	LOINC
FBS[h]	• 3040820 • 36303387 • 3037110	Fasting glucose (mass/volume) in serum or plasma	Numeric	LOINC
Hgb[i]	• 3000963 • 3027484	Hemoglobin (mass/volume) in blood	Numeric	LOINC
Cr[j]	• 3016723 • 3051825	Creatinine (mass/volume) in serum or plasma	Numeric	LOINC
AST[k]	• 3013721	Aspartate aminotransferase (enzymatic activity/volume) in serum or plasma	Numeric	LOINC
ALT[l]	• 3006923 • 46236949	Alanine aminotransferase (enzymatic activity/volume) in serum or plasma	Numeric	LOINC

[a]SNOMED-CT: Systematized Nomenclature Of Medicine–Clinical Terms.

[b]LONIC: Local Laboratory Result Code.

[c]SBP: systolic blood pressure.

[d]DBP: diastolic blood pressure.

[e]HDL: high-density lipoprotein.

[f]LDL: low-density lipoprotein.

[g]TG: triglyceride.

[h]FBS: fasting blood glucose.

[i]Hgb: hemoglobin.

[j]Cr: creatinine.

[k]AST: aspartate aminotransferase.

[l]ALT: alanine aminotransferase.

Textbox 1. Design criteria.

Target group

- Patients who visited the hospital from January 1, 1999, to May 31, 2020
- Patients with data for 180 days before and after the index date
- Patients diagnosed with chronic diseases (type 2 diabetes, hypertension, hyperlipidemia, or cardiovascular disease) within 10 years from the index date

Comparator group

- Patients who visited the hospital from January 1, 1999, to May 31, 2020
- Patients with data for 180 days before and after the index date
- Patients who have not been diagnosed with chronic diseases (type 2 diabetes, hypertension, hyperlipidemia, or cardiovascular disease) within 10 years from the index date.

Exclusion criteria

- A history of chronic diseases (diabetes, high blood pressure, hyperlipidemia, or cardiovascular disease) for any period before the selection duration
- Missing basic information, examination, and questionnaire items that were selected as essential items in the study for the development of the chronic disease prediction model

Data Preparation

We used the patient information, medical treatment, and examination data from a tertiary hospital in South Korea for the CDM. If the missing value was a numeric variable, it was replaced with the median of the matching gender for each age group (stratified into 5-year units), and in the case of a categorical variable, it was replaced with the mode of the matching gender for each age group. Since the number of samples between the 2 groups was unbalanced, random undersampling was performed in the nondiabetic group within 10 years to match the size of the diabetic group. Although data balancing can be hidden from the actual prevalence in practice, it ensures model performance for new data by preventing biased learning from highly imbalanced class problems.

Statistical Analysis

The descriptive statistics of each group (target group and comparator group) are shown in Table 2.

Table 2. Descriptive statistics.

Feature	Processed data	Target	Comparator
Sex, n			
Female	1157	586	571
Male	1691	838	853
Age (years), mean (SD)	47.56 (15.03)	54.94 (12.50)	40.17 (13.60)
BMI, mean (SD)	24.59 (5.68)	25.69 (7.05)	23.50 (3.55)
SBP[a], mean (SD)	128.1 (16.95)	132.5 (17.49)	123.6 (14.86)
DBP[b], mean (SD) , mean (SD)	79.29 (11.89)	81.56 (12.04)	77.03 (11.16)
Total cholesterol, mean (SD)	189.4 (39.84)	192.9 (42.40)	185.8 (36.76)
HDL[c], mean (SD)	51.45 (12.52)	47.55 (10.82)	55.36 (13.01)
LDL[d], mean (SD)	111.9 (29.57)	114 (28.95)	109.7 (30.13)
TG[e], mean (SD)	143.1 (55.0)	145.0 (59.06)	115.2 (43.63)
FBS[f], mean (SD)	116.4 (45.64)	136.7 (54.40)	96.04 (20.31)
Hgb[g], mean (SD)	14.27 (1.70)	14.18 (1.83)	14.36 (1.56)
Cr[h], mean (SD)	0.99 (0.68)	1.045 (0.92)	0.93 (0.26)
AST[i], mean (SD)	27.93 (9.64)	31.71 (8.91)	24.15 (11.20)
ALT[j], mean (SD)	31.16 (11.89)	37.46 (8.19)	24.87 (14.87)

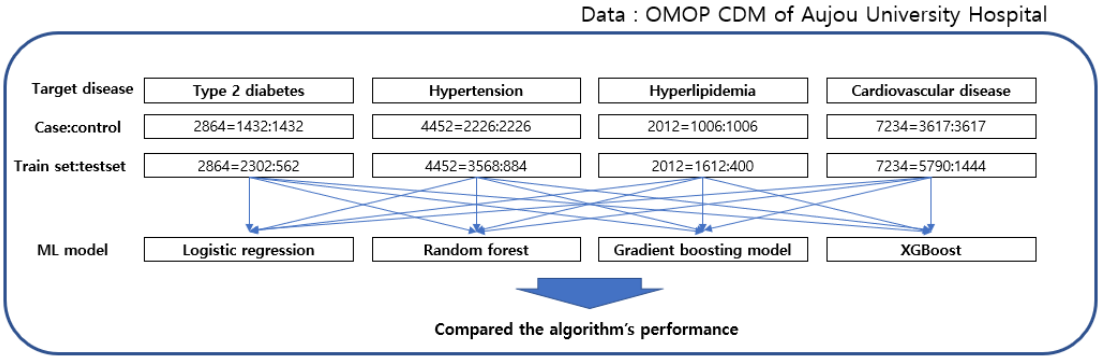
[a]SBP: systolic blood pressure.
[b]DBP: diastolic blood pressure.
[c]HDL: high-density lipoprotein.
[d]LDL: low-density lipoprotein.
[e]TG: triglyceride.
[f]FBS: fasting blood glucose.
[g]Hgb: hemoglobin.
[h]Cr: creatinine.
[i]AST: aspartate aminotransferase.
[j]ALT: alanine aminotransferase.

Models

Overview

In this study, to select the most suitable model for disease prediction, we used the following 4 algorithms: logistic regression (LR), random forest (RF), gradient boosting model (GBM), and extreme gradient boosting (XGBoost). LR using binary classification in the statistics field and the other 3 machine learning algorithms had shown better performance than similar prior research [9]. Afterward, the prediction performance was compared. Model validation was conducted with the same 80% training data and 20% validation data derived from the entire data set. Accuracy, sensitivity, specificity, and AUC were used as model performance indicators. The prediction model flow is shown in Figure 2.

Figure 2. Prediction model flow. CDM: common data model; ML: machine learning; OMOP: Observational Medical Outcomes Partnership; XGBoost: extreme gradient boosting.



LR Algorithm

LR was devised by Cox [10] as a regression model that predicts the probability of the occurrence of an event with respect to a binary dependent variable. Unlike general linear regression analysis, the range of LR is limited to 0-1 because the dependent variable is dichotomous, and the conditional probability of the occurrence of an event also follows a binomial distribution. That is, if the estimated value following the logistic function satisfying the above assumption is less than 0.5, the predicted value is classified as “nonoccurring,” and if it is greater than 0.5, then the predicted value is classified as “occurring.” Although LR was developed in 1970, it is still being used for statistical analysis and predictive research in various fields.

RF Algorithm

RF is a tree-based ensemble model capable of both classification and regression and selects the most appropriate forest model by collecting the results of randomly generated independent decision trees [11]. Bagging-based training data inputted to the tree provides model diversity, and the randomness of variable combinations constituting the tree can prevent model noise and the risk of overfitting. The fact that RF is less sensitive to missing values than other algorithms is also an advantage.

GBM Algorithm

GBM is a tree-based ensemble model similar to RF, but unlike RF, it creates a tree using a boosting method. The boosting method increases the performance of classification or prediction by sequentially combining several small models [12]. GBM reduces the errors generated by the previous model. Although GBM shows high performance in prediction, it may take a lot of time to fit the model because training requires extensive computation. In recent years, GBM-based algorithms such as LightGBM [13], CatBoost, and XGBoost have been developed to overcome the shortcomings of GBM.

XGBoost Algorithm

XGBoost is a representative tree-based ensemble model devised by Chen and Guestrin [14]. It is a machine learning algorithm actively used in prediction and classification research because of its powerful performance and has many advantages such as fast learning due to parallel processing, overfitting regulation, and linkage with other algorithms. Since XGBoost is based on GBM, it optimizes the model by assigning weights using a boosting method, reducing the residual error of the model created with classification and regression tree algorithm–based trees.

Grid Search

Unlike LR analysis, machine learning algorithms support various parameters (hyperparameters) so that users can optimize the model. Grid search is a technique to find the parameter value when the model has the highest performance by sequentially applying the parameter values set by the user. We optimized the model by applying grid search to the above 3 algorithms (RF, GBM, and XGBoost). Table S1 in Multimedia Appendix 1 presents the parameters and ranges used in the grid search for each algorithm.

Results

Model Results

Comparing model performance by chronic disease, the predictive model using XGBoost based on accuracy showed superior performance in all diseases compared to the other 3 models (Table 3).

Table 3. Performance comparison of disease prediction models.

Parameter, chronic disease	LR[a]	RF[b]	GBM[c]	XGBoost[d]
Accuracy				
Type 2 Diabetes	0.877	0.8743	0.8743	0.8824
Hypertension	0.7783	0.793	0.7896	0.8213
Hyperlipidemia	0.8125	0.82	0.8325	0.8325
Cardiovascular disease	0.7941	0.8162	0.8235	0.8429
Sensitivity				
Type 2 Diabetes	0.8852	0.8804	0.8684	0.8705
Hypertension	0.7758	0.7758	0.7783	0.7934
Hyperlipidemia	0.8141	0.8556	0.8077	0.8182
Cardiovascular disease	0.8143	0.8644	0.8030	0.8243
Specificity				
Type 2 Diabetes	0.8691	0.8684	0.8804	0.8950
Hypertension	0.7808	0.7808	0.8019	0.8550
Hyperlipidemia	0.8109	0.8122	0.8333	0.8482
Cardiovascular disease	0.8333	0.7792	0.7857	0.8636

[a]LR: logistic regression.

[b]RF: random forest.

[c]GBM: gradient boosting model.

[d]XGBoost: extreme gradient boosting.

Model Validation Results

Table 4 shows the parameter values of each disease model outputted by the XGBoost grid search.

The model evaluation indicators used were accuracy, sensitivity, specificity, and AUC. Over 80% prediction accuracy was achieved for all diseases, with AUC from 0.84 to 0.93. The XGBoost model performance by disease is shown in Table 5 and Figure 3.

Table 4. Extreme gradient boosting grid search result.

Target disease	Subsample[a]	Max depth[b]	Min child[c]	Eta[d]
Type 2 diabetes	0.7	7	2	0.1
Hypertension	0.9	3	2	0.01
Hyperlipidemia	1	3	2	0.01
Cardiovascular disease	0.9	3	1	0.01

[a]Subsample: sample's rate of each tree.

[b]Max depth: maximum depth of Tree.

[c]Min child: minimum sum of weights for all observations needed in the child.

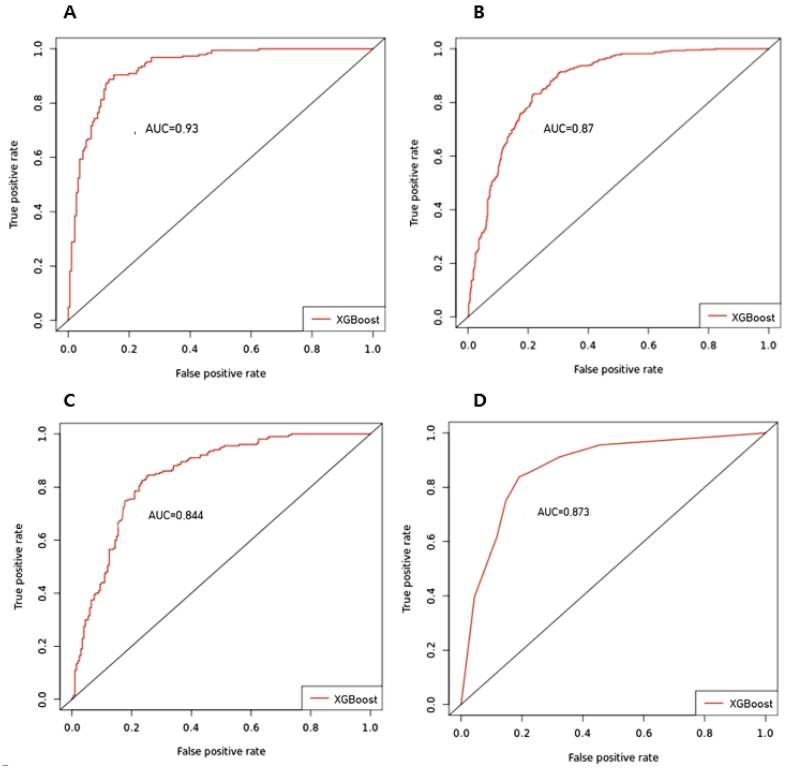
[d]Eta: learning rate.

Table 5. Predictive performance by model.

Target disease	Accuracy	Sensitivity	Specificity	AUC[a]
Type 2 diabetes	0.8824	0.8705	0.8950	0.9303
Hypertension	0.8213	0.7934	0.8550	0.8704
Hyperlipidemia	0.8325	0.8182	0.8432	0.8442
Cardiovascular disease	0.8429	0.8243	0.8636	0.8726

[a]AUC: area under the curve.

Figure 3. Receiver operating characteristic curves for XGBoost (A) type 2 diabetes model, (B) hypertension model, (C) hyperlipidemia model, and (D) cardiovascular disease model. AUC: area under the curve; XGBoost: extreme gradient boost.



Shapley Additive Explanations Model Variable Importance

In open-source program languages (eg, Python and R), the XGBoost package shows model feature importance using its own library. However, small models are more combined and complicated, and the feature importance of small models becomes inconsistent. Therefore, we used the Shapley additive explanations (SHAP) method to represent the model's feature importance, which had high consistency and accuracy [15]. SHAP's feature importance used the weighted average of marginal contribution for each feature (Shapley value). It gave the importance of the features and the positive or negative effect of each feature. The formula of Shapley value is a follows:

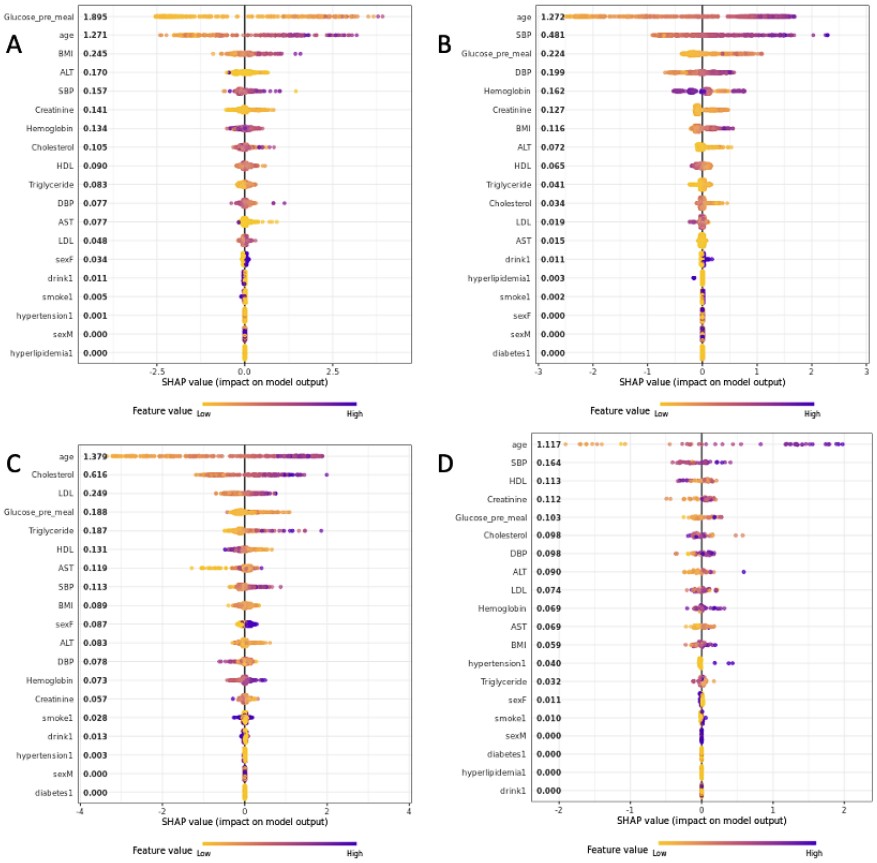
$$\text{Contribution of feature}_i = \beta_i x_i - E(\beta_i x_i) = \beta_i x_i - \beta_i E(x_i)$$

where $\emptyset_i$ is the Shaley value of data$_i$, F is the full set, S is the subsets in total set excluding data$_i$,  is the contribution of the full set including data$_i$, and $f_S(x_S)$ is the contribution of subsets excluding data$_i$.

The SHAP value graph of the fitted model for each disease is presented in Figure 4.

In the case of type 2 diabetes, fasting blood glucose (Shapley value=1.895), age (1.271), and BMI (0.245) influenced the occurrence of diabetes within 10 years [16]. For hypertension, hyperlipidemia (1.272), cardiovascular disease (1.379), and age (1.117) had the greatest influence on disease occurrence. Furthermore, in the case of hyperlipidemia, it was found that among the variables excluding age, total cholesterol (0.616) and low-density lipoprotein (0.249) influenced disease occurrence, in the order presented [17]. In case of cardiovascular disease, systolic blood pressure (0.164) and high-density lipoprotein (0.113) had the second highest influence on disease occurrence. These results are consistent with the results of previous studies that studied the risk factors of the 4 chronic diseases [16,17].

Figure 4. Shapley additive explanations (SHAP) value graph of the fitted model for the importance of (A) type 2 diabetes variables, (B) hypertension variables, (C) hyperlipidemia variables, and (D) cardiovascular disease variables. ALT: alanine aminotransferase; AST: aspartate aminotransferase; DBP: diastolic blood pressure; HDL: high-density lipoprotein; LDL: low-density lipoprotein; SBP: systolic blood pressure.



Discussion

Principal Findings

This study develops a disease prediction model with more than 80% accuracy by using the 16 National Health Insurance system test variables from the real-world data of a tertiary hospital in South Korea. Our study:

1. Presents the possibility of predicting diseases with universal and useful information on public health examinations,
2. Explains the ability of model prediction results, and
3. Presents the external verification and scalability using other organizations' CDM.

By observing recent research trends relating to disease prediction models using a CDM, the number of cases focusing on multicenter studies rather than single-center studies is increasing. Lee et al [18] established an artificial intelligence (AI) learning platform for multicenter clinical research focusing on CDM linkage. Using data from Gachon University Gil Hospital to develop a machine learning model that predicts 5-year risk in patients with inflammatory bowel disease who started biologics, Choi et al [19] externally validated the model with CDM data (Ministry of Food and Drug Safety). Johnston et al [20] developed a model to predict whether patients will stop taking antihyperglycemic drugs within 1 to 2 years after laparoscopic metabolic surgery. Using psychiatric patient notes at Ajou University Hospital, Lee et al [21] developed an NLP model that predicts the onset of psychosis in patients by learning, which is a representative case. As such, if the same cohort criterion is applied to multiple institutions in an expanded form along with disease prediction model construction and cross-validation, a more universal and robust model can be developed.

Research is being conducted globally to reduce medical costs by predicting disease occurrence using AI. As the AI industry has gone bigger, AI can reduce costs in providing care and increase the efficiency of medical jobs [22]. In 2019, researchers from the Boston Institute of Technology and Boston Health Center conducted joint research using electronic health records and lifelog big data with AI in an attempt to prevent disease outbreaks and medical fraud. The findings may help reduce hospitalization costs, which account for a substantial portion of US medical expenses [23]. The model developed through this study is expected to evolve into a similar system for South Koreans, by predicting the risk of future disease development and aiding self-health management.

Comparison With Prior Work

With the recent development of AI and data processing technology, research on disease prediction model development using National Health Insurance Service data [9] or single-institution EMR data [24] is steadily progressing. In this study, we developed chronic disease prediction models with relatively high performance compared to previous papers. A difference between this study and existing work is that the models have been developed using the CDM, so that we can

expect improved precision through variable expansion and by simultaneously using multiorganizational data.

Limitations

A limitation of this study is that it uses a single-institution CDM from a tertiary hospital. Therefore, it cannot ensure generalizability. Additionally, the demographic variables (educational level, residential area, marital status, etc.) are insufficient compared to the health insurance service examination items. They are limited due to the focus on South Korean public checkups; by using more features related with the disease (hemoglobin A1C and biopsy data), the model becomes more accurate. Lastly, the model was trained using the cross-sectional data of patients. If the model is trained using time-series data (eg, the cohort of patients' information that includes changes of laboratory results as time goes by), it could be much more comprehensive.

Conclusions

In this study, 4 metabolic chronic diseases were selected, and disease prediction models were developed using the Ajou University Hospital CDM. To obtain a model suitable for disease prediction, the predictive performance of each model for disease occurrence was compared using the LR, GBM, RF, and XGBoost algorithms. The XGBoost model shows the best performance for all diseases. The performance of the XGBoost model was calculated as 0.9303, 0.8704, 0.8442, and 0.8726 AUC standards for type 2 diabetes, hypertension, hyperlipidemia, and cardiovascular disease, respectively. In addition, the importance of the variables was calculated through modeling, and the results are in line with previous clinical studies. We have confirmed that chronic diseases can be predicted, not just using single-institution EMR or public clinical data, but using the CDM in each local hospital.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Model's hyperparameter range for grid search.
[DOCX File , 14 KB - ai_v1i1e41030_app1.docx]

References

1. World Health Organization, Regional Office for Europe. Action plan for the prevention and control of noncommunicable diseases in the WHO European Region. World Health Organization. 2016. URL: https://apps.who.int/iris/handle/10665/341522 [accessed 2022-12-09]
2. Song E, Kim YE, Ji S. Impact of a primary health care chronic diseases management pilot program. Article in Korean. Korean J Med 2021 Feb 1;96(1):7-12. [doi: 10.3904/kjm.2021.96.1.7]
3. Jeong EK. 2020 chronic disease fact sheet. Article in Korean. Korean Disease Control and Prevention Agency. 2021 Jan 26. URL: https://www.kdca.go.kr/gallery.es?mid=a20503020000&bid=0003&act=view&list_no=144928 [accessed 2022-12-14]
4. Jung C, Son JW, Kang S, Kim WJ, Kim H, Kim HS, et al. Diabetes fact sheets in Korea, 2020: an appraisal of current status. Diabetes Metab J 2021 Jan 13;45(1):1-10 [FREE Full text] [doi: 10.4093/dmj.2020.0254] [Medline: 33434426]
5. Jang JS, Lee MJ, Lee TR. Development of T2DM prediction model using RNN. Article in Korean. Journal of Digital Convergence 2019 Aug 28;17(8):249-255. [doi: 10.14400/JDC.2019.17.8.249]
6. Lee S, Lee H, Choi JR, Koh SB. Development and validation of prediction model for risk reduction of metabolic syndrome by body weight control: a prospective population-based study. Sci Rep 2020 Jun 19;10(1):10006 [FREE Full text] [doi: 10.1038/s41598-020-67238-5] [Medline: 32561810]
7. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. JMLR Workshop Conf Proc 2016 Aug;56:301-318 [FREE Full text] [Medline: 28286600]
8. Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to diagnose with LSTM recurrent neural networks. arXiv Preprint posted online on November 11, 2015. [doi: 10.48550/arXiv.1511.03677]
9. Kim J, Jeong Y, Kim JH, Lee J, Park D, Kim H. Machine learning-based cardiovascular disease prediction model: a cohort study on the Korean National Health Insurance Service Health Screening Database. Diagnostics (Basel) 2021 May 25;11(6):943 [FREE Full text] [doi: 10.3390/diagnostics11060943] [Medline: 34070504]
10. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) 2018 Dec 05;34(2):187-202. [doi: 10.1111/j.2517-6161.1972.tb00899.x]
11. Louppe G. Understanding random forests: from theory to practice. arXiv Preprint posted online on July 28, 2014. [doi: 10.48550/arXiv.1407.7502]
12. Bühlmann P, Hothorn T. Boosting algorithms: regularization, prediction and model fitting. Statist Sci 2007 Nov 1;22(4):477-505. [doi: 10.1214/07-STS242]
13. Wang C, Chang L, Liu T. Predicting student performance in online learning using a highly efficient gradient boosting decision tree. 2022 Presented at: IIP 2022: Intelligent Information Processing XI; May 27-30, 2022; Qingdao, China p. 508-521. [doi: 10.1007/978-3-031-03948-5_41]

14. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016 Aug 13 Presented at: KDD '16: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA p. 785-794. [doi: 10.1145/2939672.2939785]
15. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. 2017 Presented at: NIPS 2017: 31st Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA URL: https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
16. Kim SA, Yu SJ, Choi D. Prevalence and risk factors of type 2 diabetes according to gender among Korean employees. Article in Korean. Journal of the Korea Academia-Industrial cooperation Society 2015 Nov 30;16(11):7589-7598. [doi: 10.5762/KAIS.2015.16.11.7589]
17. Kim KY. Risk factors for hypertension in elderly people aged 65 and over, and adults under age 65. Article in Korean. Journal of the Korea Academia-Industrial cooperation Society 2019 Jan 31;20(1):162-169. [doi: 10.5762/KAIS.2019.20.1.162]
18. Lee CS, Kim JE, No SH, Kim TH, Yoon KH, Jeong CW. Construction of artificial intelligence training platform for multi-center clinical research. Article in Korean. KIPS Transactions on Computer and Communication Systems 2020 Oct 31;9(10):239-246. [doi: 10.3745/KTCCS.2020.9.10.239]
19. Choi YI, Park SJ, Chung JW, Kim KO, Cho JH, Kim YJ, et al. Development of machine learning model to predict the 5-year risk of starting biologic agents in patients with inflammatory bowel disease (IBD): K-CDM network study. J Clin Med 2020 Oct 26;9(11):3427 [FREE Full text] [doi: 10.3390/jcm9113427] [Medline: 33114505]
20. Johnston SS, Morton JM, Kalsekar I, Ammann EM, Hsiao C, Reps J. Using machine learning applied to real-world healthcare data for predictive analytics: an applied example in bariatric surgery. Value Health 2019 May;22(5):580-586 [FREE Full text] [doi: 10.1016/j.jval.2019.01.011] [Medline: 31104738]
21. Lee DY, Kim C, Lee S, Son SJ, Cho S, Cho YH, et al. Psychosis relapse prediction leveraging electronic health records data and natural language processing enrichment methods. Front Psychiatry 2022 Apr 5;13:844442 [FREE Full text] [doi: 10.3389/fpsyt.2022.844442] [Medline: 35479497]
22. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019 Jun 13;6(2):94-98 [FREE Full text] [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]
23. Raven MC, Doran KM, Kostrowski S, Gillespie CC, Elbel BD. An intervention to improve care and reduce costs for high-risk patients with frequent hospital admissions: a pilot study. BMC Health Serv Res 2011 Oct 13;11(1):270 [FREE Full text] [doi: 10.1186/1472-6963-11-270] [Medline: 21995329]
24. Twick I, Zahavi G, Benvenisti H, Rubinstein R, Woods MS, Berkenstadt H, et al. Towards interpretable, medically grounded, EMR-based risk prediction models. Sci Rep 2022 Jun 15;12(1):9990 [FREE Full text] [doi: 10.1038/s41598-022-13504-7] [Medline: 35705550]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
CDM: common data model
EMR: electronic medical record
GBM: gradient boosting model
LR: logistic regression
RF: random forest
SHAP: Shapley additive explanations
XGBoost: extreme gradient boosting

Edited by B Malin, K El Emam; submitted 13.07.22; peer-reviewed by A Finny, SJC Soerensen , V Kumar; comments to author 12.10.22; revised version received 21.11.22; accepted 26.11.22; published 22.12.22.

Please cite as:
Lee C, Jo B, Woo H, Im Y, Park RW, Park C
Chronic Disease Prediction Using the Common Data Model: Development Study
JMIR AI 2022;1(1):e41030
URL: https://ai.jmir.org/2022/1/e41030
doi:10.2196/41030
PMID:38875545

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org