Viewpoint

What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask

Isaac S Kohane[1], MD, PhD; Bruce J Aronow[2], PhD; Paul Avillach[1], MD, PhD; Brett K Beaulieu-Jones[1], PhD; Riccardo Bellazzi[3,4], PhD; Robert L Bradford[5], BSc; Gabriel A Brat[1], MD, MPH; Mario Cannataro[6,7], MS; James J Cimino[8], MD; Noelia García-Barrio[9], MS; Nils Gehlenborg[1], PhD; Marzyeh Ghassemi[10], PhD; Alba Gutiérrez-Sacristán[1], PhD; David A Hanauer[11], MS, MD; John H Holmes[12], PhD; Chuan Hong[1], PhD; Jeffrey G Klann[13,14], PhD; Ne Hooi Will Loh[15], MBBS, FRCA, FFICM, EDIC; Yuan Luo[16], PhD; Kenneth D Mandl[17], MPH, MD; Mohamad Daniar[18], MSIS; Jason H Moore[19], PhD; Shawn N Murphy[1,20], MD, PhD; Antoine Neuraz[21,22], MD; Kee Yuan Ngiam[15], MBBS, MRCS, MMed; Gilbert S Omenn[23], MD, PhD; Nathan Palmer[1], PhD; Lav P Patel[24], MS; Miguel Pedrera-Jiménez[9], MS; Piotr Sliz[17], PhD; Andrew M South[25], MS, MD; Amelia Li Min Tan[1,26], BSc, PhD; Deanne M Taylor[27,28], PhD; Bradley W Taylor[29], MS; Carlo Torti[7], MD; Andrew K Vallejos[29], MS; Kavishwar B Wagholikar[13,14], MBBS, PhD; The Consortium For Clinical Characterization Of COVID-19 By EHR (4CE)[30]; Griffin M Weber[1*], MD, PhD; Tianxi Cai[1*], SCD

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, United States

[2]Biomedical Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati, Cincinnati, OH, United States

[3]Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy

[4]ICS Maugeri, Pavia, Italy

[5]North Carolina Translational and Clinical Sciences Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

[6]Data Analytics Research Center, University Magna Graecia of Catanzaro, Catanzaro, Italy

[7]Department of Medical and Surgical Sciences, University Magna Graecia of Catanzaro, Catanzaro, Italy

[8]Informatics Institute, University of Alabama at Birmingham, Birmingham, AL, United States

[9]Department of Informatics, 12 de Octubre University Hospital, Madrid, Spain

[10]Department of Computer Science and Medicine, University of Toronto, Toronto, ON, Canada

[11]Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, MI, United States

[12]Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

[13]Department of Medicine, Harvard Medical School, Boston, MA, United States

[14]Laboratory of Computer Science, Massachusetts General Hospital, Boston, MA, United States

[15]National University Health Systems, Singapore, Singapore

[16]Department of Preventive Medicine, Northwestern University, Chicago, IL, United States

[17]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, United States

[18]Clinical Research Informatics, Boston Children's Hospital, Boston, MA, United States

[19]Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, United States

[20]Department of Neurology, Massachusetts General Hospital, Boston, MA, United States

[21]Department of Biomedical Informatics, Necker-Enfant Malades Hospital, Assistance Publique - Hôpitaux de Paris, Paris, France

[22]Centre de Recherche des Cordeliers, INSERM UMRS 1138 Team 22, Université de Paris, Paris, France

[23]Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI, United States

[24]Department of Internal Medicine, Division of Medical Informatics, University of Kansas Medical Center, Kansas City, KS, United States

[25]Section of Nephrology, Department of Pediatrics, Brenner Children's Hospital, Wake Forest School of Medicine, Winston Salem, NC, United States

[26]Department of Biomedical Informatics, National University of Singapore, Singapore, Singapore

[27]Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia, Philadelphia, PA, United States

[28]Department of Pediatrics, Perelman School of Medicine, The University of Pennsylvania, Philadelphia, PA, United States

[29]Clinical and Translational Science Institute, Medical College of Wisconsin, Milwaukee, WI, United States

[30]See Acknowledgments

[*]these authors contributed equally

Corresponding Author:
Isaac S Kohane, MD, PhD

Department of Biomedical Informatics
Harvard Medical School
10 Shattuck Street
Boston, MA, 02115
United States
Phone: 1 617 432 3226
Email: isaac_kohane@harvard.edu

Abstract

Coincident with the tsunami of COVID-19–related publications, there has been a surge of studies using real-world data, including those obtained from the electronic health record (EHR). Unfortunately, several of these high-profile publications were retracted because of concerns regarding the soundness and quality of the studies and the EHR data they purported to analyze. These retractions highlight that although a small community of EHR informatics experts can readily identify strengths and flaws in EHR-derived studies, many medical editorial teams and otherwise sophisticated medical readers lack the framework to fully critically appraise these studies. In addition, conventional statistical analyses cannot overcome the need for an understanding of the opportunities and limitations of EHR-derived studies. We distill here from the broader informatics literature six key considerations that are crucial for appraising studies utilizing EHR data: data completeness, data collection and handling (eg, transformation), data type (ie, codified, textual), robustness of methods against EHR variability (within and across institutions, countries, and time), transparency of data and analytic code, and the multidisciplinary approach. These considerations will inform researchers, clinicians, and other stakeholders as to the recommended best practices in reviewing manuscripts, grants, and other outputs from EHR-data derived studies, and thereby promote and foster rigor, quality, and reliability of this rapidly growing field.

(J Med Internet Res 2021;23(3):e22219) doi: 10.2196/22219

KEYWORDS

COVID-19; electronic health records; real-world data; literature; publishing; quality; data quality; reporting standards; reporting checklist; review; statistics

Introduction

What should researchers and clinicians conclude about the recent high-profile retractions of COVID-19 studies based on electronic health record (EHR) data? It is impressive that two publications involving patients with COVID-19, one in The Lancet [1] and the other in the New England Journal of Medicine [2], were determined to be unsound and were retracted in less than 2 months from publication, as these journals' review processes and quality checks are among the most rigorous in the world. Yet, upon closer inspection by those of us familiar with EHR-based research, there were many flaws to these studies involving data quality issues and a lack of transparency that should have been more readily identified during the peer and editorial review process. This is not to say that in-depth statistical analysis might not have eventually uncovered concerns but rather to point out incongruities and anomalies unique to EHR-based studies that should immediately raise concerns to experienced biomedical informaticians, much like an experienced contractor explaining to a homeowner why a competing bid is too good to be true.

In this viewpoint, we present six key questions that are necessary to consider when appraising EHR-based research, especially for research studies investigating the pandemic:

1. How complete are the data?
2. How were the data collected and handled?
3. What were the specific data types?
4. Did the analysis account for EHR variability?
5. Are the data and analytic code transparent?
6. Was the study appropriately multidisciplinary?

In particular, we focus on general aspects of these questions that are crucial to study and data quality and validity of and interpretability of the results and that are broadly applicable to many stakeholders, including researchers and clinicians, in order to optimize the review of submitted manuscripts, published studies, and grant applications containing preliminary data. These desiderata were compiled by the 96 members of the Consortium for Clinical Characterization of COVID-19 by EHR (4CE)—a self-assembled group of collaborating hospitals focused specifically on studying the clinical course of patients with COVID-19 using EHR-based data—most of whom are biomedical informaticians—across 7 countries. 4CE members were invited to contribute their specific key concerns to a shared checklist. This list was then pared down into a less technical list for a more general audience. We excluded those items that are generally considered to be good biostatistical practices (eg, manual review of sample data sets, detecting and understanding outliers [3,4]) to present EHR-specific concerns to a broad biomedical audience. We also excluded recommendations that are contained within the Reporting of Studies Conducted Using Observational Routinely Collected Health Data (RECORD) statement [5,6], which are not specific to EHR-derived data. Finally, we did not focus on the specific limitations of EHR-derived studies, which have been amply documented [7,8], or on the methods to minimize the impact of these limitations,

as this viewpoint is not focused on reviewing specific methodological options for investigators using EHR-derived data, which has been reviewed in detail previously [9-11]. We acknowledge that there are many other criteria that can inform evaluations of EHR-based studies, but we have purposefully limited this discussion to those issues that are most relevant to a general audience, centered on studies investigating the pandemic.

Data Completeness

There are several statistical tests to query data completeness and methods for incorporating missing data [12,13], but here we describe the reasonable expectations for such completeness with knowledge of current, state-of-the-art EHR usage. A publication that is specific about which data were obtained from the EHR (eg, specific laboratory tests or billing codes) is more credible than a study that simply claims it obtained 100% of the EHR data (as did the two recently retracted publications [1,2]). The range of data types from EHRs is extensive and highly varied; each data type requires its own specific quality control and transformations to standard terminologies. For example, laboratory measurements alone can have as many as hundreds of thousands of local codes at a large health care system such as the Veterans Health Administration. In many cases, these data require some level of manual record review to assure data quality and completeness.

Similarly, if a study reports a deidentification procedure, it must describe the details of said procedure. The goals of the deidentification process determine the nature of the deidentification process and the associated regulatory requirements. For example, US hospitals can meet HIPAA (Health Insurance Portability and Accountability Act) standards [14] if they require obfuscation of the counts of patients with rare clinical presentations below a specified prevalence threshold and if they employ date shifting. Knowledge of these methods is essential to analyzing and interpreting the derived data.

Some data types are represented theoretically in the EHR but in practice are only recorded occasionally. For example, standardized codes for smoking history or a family history of specific diseases exist but their underuse is well known. Thus, one cannot assume that the lack of smoking history codes equates to the patient being a nonsmoker. In such scenarios, one must provide an explicit description of the management of missing/null values. Many data elements, such as a complete pulmonary function test, exist in a fragmented form, scattered across different fields in the EHR, and are difficult to extract reliably. In addition, clinical notes allow clinicians greater qualitative expressivity on some of the above values, like smoking history, where they are documented more frequently but not consistently. The quality criteria for reporting narrative content from clinical notes are further addressed below.

Many clinical states are not represented explicitly in the EHR but can be inferred (often referred to as computational phenotypes). When a publication refers to hyperlipidemia, readers should ask themselves whether the hyperlipidemic phenotype is assessed from one or more lipid laboratory tests, billing diagnostic codes, prescription of lipid-lowering medication, or a combination of the above. It is important to document if only structured codes were used or if the phenotype was defined based on information extracted from clinical notes by using natural language processing (NLP) or manual chart review. Either a table describing these phenotypic methods or a reference to a public set of definitions (eg, Phenotype Knowledgebase, PheKB [15]) or a published algorithm with reported accuracy (as seen, for example, in Zhang et al [16] and Ananthakrishnan et al [17]) can provide transparency and precision to these EHR-driven computational phenotypes. The lack of this transparency should be a warning sign. If onset time or temporal trends of clinical events are used as outcomes, it is important to provide sufficient details on how the data were used to derive these outcomes, how granular time was incorporated (eg, by day, 24-hour period, or hour/minute), and to comment on their accuracy, since EHR data are particularly noisy with regards to capturing the timing of events [18,19].

If one uses EHR data to obtain population estimates (eg, prevalence of a complication per 100,000 patients), then additional information should be provided so that readers can determine which subset of patients from that population a given hospital's EHR can capture. For example, if the EHR captures a patient's hospitalization for heart failure, will the EHR also capture the preceding or subsequent outpatient clinic visits related to that hospitalization? With health maintenance organizations, such as Kaiser Permanente, that is much less of a concern, but many hospitals operate in a patchwork system where the patient's data are spread across multiple heterogeneous EHRs that do not necessarily communicate. In our recent COVID-19 study [20], we found many instances in which patients with COVID-19 were transferred from another hospital; unless that other hospital was part of our consortium, it was impossible to have a complete record of their COVID-19 clinical course. It is also important to recognize that a given EHR may not fully capture the clinical course of certain patients, such as those infected with SAR-CoV-2 who have mild symptoms and are discharged home from the emergency room. In these instances, integration of EHR data with data from other sources (eg, primary care providers' offices or nursing homes) may increase the reliability of analysis, although in practice this is rare and such integration methods have to be well documented. EHR systems may also fail to capture acute events that occur outside of the system, especially in the coded data. Leveraging NLP data from the clinical notes can potentially recover partial information if the patient has follow-up visits within that particular system.

Data Collection and Handling

Often the units of measurement and the codes used for data elements like laboratory tests, medications, and diagnoses are not the same across hospitals and may even differ within the same health care system or change over time. Single analytic concepts (eg, the troponin T test) can balloon into dozens of local codes at each hospital, since these tests may be performed at different diagnostic laboratories, each with its own distinct codes or with different technologies over time. Therefore, they have to be "harmonized," or mapped, to agreed-upon standard terminologies and scales [21]. Even when they are the same,

their meaning can differ based on population or practice differences (eg, which sensitive troponin test is used or which reference range defines a test result being normal, or in children rather than in adults, whose normative values often change across the age range) [7]. In both instances, readers should expect that the specific procedures for harmonization or site-specific semantic alignment are described adequately in the Methods section (or via supplementary materials). A summary of this process can become increasingly complex within the usual confines of a Methods section for multisite and international studies where, by necessity, the site-by-site variability is high.

Data Type

There are large methodological divides and divergent ethical challenges between codified data (eg, discrete laboratory values such as serum glucose) and narrative text (eg, discharge summary) from which characterizations are obtained using NLP. While both data types have their own limitations, methods that incorporate both can greatly improve the sensitivity and/or specificity of the clinical characterizations and phenotyping of a group of patients. For example, signs and symptoms are often not codified discreetly or consistently (eg, not entered into the EHR's Problem List) but are written in the clinical notes. Similarly, outpatient medication documentation in clinical notes does not necessarily represent accurately the medications that the patient is actually taking, but prescriptions entered into the EHR may. Combining both codified and NLP data can substantially improve sensitivity and/or specificity and ideally one should always use this complementarity [22-24]. For example, only about 10% of pregnant women with suicide ideation have related codes and vast majority of the cases are only documented in the notes [25]. However, the ability to extract NLP data and the accuracy of those data may be limited by each institution's informatics infrastructure and expertise as well as local institutional review board (IRB) constraints. Furthermore, NLP application to clinical narrative text is relatively new and more prone to large variability in the quality of the obtained characterizations. Particularly in countries with different languages, the NLP techniques and their performance may vary widely. For this reason, readers should expect a reference to the specific NLP methods used and their performance characteristics on data of the sort that the study collected and analyzed. For example, if someone describes the use of an NLP approach on discharge summaries in intensive care units in Italy, but the provided citation was validated only for use in outpatient notes written in English, readers can be legitimately concerned about the accuracy and validity of the patient characterizations in that study. Furthermore, if a study claims very high accuracy, readers should expect a report (or citation of a report) that shows an expert review of the NLP method validated against a representative sample confirming the claimed performance.

Robustness Against EHR Variability

Beyond any variation in human biology across countries and continents, different styles of practice, and how different reimbursement schemes influence styles of practice and use of EHRs, have a very large impact on the nature of EHR data. Therefore, a multinational study should at least acknowledge these differences as a limitation or explicitly attempt to account for them in the analyses. For example, in COVID-19–related research, it has become increasingly apparent that there is an association between patient race/ethnicity and their risk for acquisition of and complications from COVID-19. However, this association is much less detectable in EHR data, as, for example, it is mostly invisible in data from Europe because several countries forbid collecting self-reported race in the EHR. Even in the United States, the coding of different ethnicities or multiracial identification is not standardized. In addition, some countries have far more comprehensive primary care EHR data sharing, whereas others (like the United States) cannot aggregate data systematically and consistently across major health care centers.

Transparency

In order to ensure patients' rights to privacy, patient-level data can rarely be shared outside an institution. In many EHR-driven studies, the code to extract data from a source EHR can be protected by confidentiality agreements with the EHR vendor and is thus difficult to share. Nonetheless, the code or algorithm for creating the variables used for analyses should be provided even if the detailed data extraction procedures are not shared because of commercial restrictions. Running the code on synthetic data sets that follow a standard data model can demonstrate code functionality and facilitate code reuse [26]. The code used to conduct statistical analyses and create visualizations—after data extraction—should also be shared in public repositories to enable other researchers to follow each step of the analysis and provide further transparency. While there are significant challenges to sharing patient-level data, one can share intermediate results and aggregate distributions to increase transparency and understand between-institution differences [27]. One should archive the data used for analyses, along with the associated data extraction codes, at the local institution to ensure reproducibility. Authors should also make the deidentified data available—either publicly in a repository or by request. While only a small fraction of readers typically look at the code, whether referenced on a file server or shared as supplementary methods, the availability of the code provides reassurance and validation that the study utilized proper methodologies.

Multidisciplinary Approach

There may come a time when data can be aggregated automatically from multiple EHR environments to answer a particular question without relying on a human to understand the particular idiosyncrasies of each institution's data and EHR system. Until that day, effective EHR data set analysis requires collaboration with clinicians and scientists who have knowledge of the diseases being studied and the practices of their particular health care systems; informaticians with experience in the underlying structures of biomedical record repositories at their own institutions and the characteristics of their data; data

harmonization experts to help with data transformation, standardization, integration, and computability; statisticians and epidemiologists well versed in the limitations and opportunities of EHR data sets and related sources of potential bias; machine learning experts; and at least one expert in regulatory and ethical standards. Data provenance records should already exist to ensure compliance with privacy standards, so that authors can readily point to these processes and reference institutional officials who grant data access similarly to IRBs. In our experience, we often have an interdisciplinary team participate in the process of establishing the research question and study design, defining the data elements, and determining what analyses can be performed given the available data. It is also important that people with complementary skills work together to review and interpret the results [28]. Each of these steps is a major contribution deserving of authorship. Just as a population genetics study reporting across countries often has dozens of authors, so do we expect multihospital EHR-driven studies to acknowledge and name the individuals as authors and in doing so provide accountability for the dozens of procedures, checks, and balances necessary for the reliable extraction of EHR patient data. Consequently, contribution statements should list explicitly the responsibilities of each author with regard to study conceptualization and design, data extraction, data harmonization, data integration, data analysis, results interpretation, and regulatory and ethical oversight. Additionally, although reputation is sometimes overvalued, having *no* reputation or at least a track record of appropriate success should trigger greater attention to documenting the process to reach the same level of trust. Unlike a mathematical proof, simple inspection of the data may be insufficient and will become increasingly so in the era of data generated by machine learning algorithms purposefully built for the task of conditioning data to appear real. Trust and accountability become essential companions to transparency and clarity during the EHR analytic process.

Conclusion

Similar to publications from the early days of the genomic revolution, which initially included extensive sections on DNA sequencing validation, methods, reagents, and conditions that became progressively briefer as trust was built and the methods commoditized, comprehensively and transparently reported methods of EHR data extraction and transformation are at least as important as subsequent statistical analysis and interpretation. We need to be open and transparent about the inherent limitations of the data and the analyses. We should also acknowledge alternative interpretations of the results (eg, outlier prescribing practices in one country that confound the apparent effects of that drug in that country). Extra caution is also needed in how we draw causal inferences from EHR data, especially given the noisiness and incompleteness of the data in addition to several sources of bias, though application of a causal model framework and specific causal inference methods may help mitigate some of these concerns. The recommendations we have outlined here (see Table 1 for our 12-item checklist) do not substitute for a durable research infrastructure that would enable tracking EHR data provenance along explicit source, ownership, and data protocols, which would allow for rigorous and routine quality assurance in the use of EHR data [29].

Table 1. 12-item checklist to assess electronic health record (EHR) data–driven studies.

Item	Reassuring	Concerning
Defining study cohort/data extraction	Reporting the precise definition of the domains and/or subsets of EHR data extracted for the study cohort and the information system sources	100% of the EHR said to be extracted or no specification of which subsets of the EHR data were obtained
Deidentification	Specific deidentification algorithm documented with acknowledgment of analytic consequences/limitations	Only a statement that deidentification was performed
Defining clinical variables/data type–specific omissions/limitations	For data types represented poorly in EHR codified data, either NLP[a] is deployed on the EHR clinical notes or additional data sources (eg, self-reported questionnaires) are used. Procedures to deal with missing values should also be made explicit	Referencing data types like family/social history without explaining how they are obtained through NLP or exceptional codified data practice
Phenotypic transparency	Computational phenotypes that are more than just a specific native EHR variable (eg, hyperlipidemia vs a specific LDL[b] measurement) are either defined in the study or a citation is given to algorithmic phenotype definitions	Clinical phenotypes are used in the study without specifying how they were derived from the EHR data
Generalizing EHR findings to the population/population denominator	Study heavily cautions on using prevalence/incidence estimates from the EHR data or refers to empirical estimates on how much of a patient's entire health care is captured in that particular EHR	Direct estimates of prevalence or incidence from EHR frequencies without justifying that generalization
Data collection	Clinical forms or data models implemented in health care information systems are shared or clearly described. This includes the coding systems used	Mention structured data without specifying the clinical forms or data models. Mention coded data without mentioning coding systems
Data transformation/harmonization	Data transformation process shared or clear description of which methods were used to harmonize data to a standardized terminology, scale units, and account for different local usage	Mention of harmonization methods without specifying which ones and what problems were identified and addressed/overcome
Textual vs codified data	If textual data are used in the study, then specification of which clinical notes, in what language, with which NLP algorithm with either an explanation of or a citation to that algorithm's validation, sensitivity, and specificity for comparable data	Harmonization efforts for codified and textual data treated as if they are the same process. Lack of specificity in describing the NLP algorithm and performance
Manual coding of data	Qualifications of coders described, formal coding criteria described or at least mentioned, intercoder reliability measured and reported	No description of process for turning text or nonstandard coded data into standard coded data; use of crowd-sourced coders (eg, graduate students or Mechanical Turk) without mention of quality assurance processes
Regional and global variation	A study describes how they adjust for (or exclude) differences that are due to variation in practice, regulation, and clinical documentation through the EHR from site to site	A study says they adjusted for regional or country differences in practice or EHR documentation but do not describe how they do it
Sharing analytic code	Analytic code is deposited in a public repository or study-specific public website	Code is not shared or only "shared on demand"
Acknowledge a multidisciplinary team	Authorships for all parts of the extraction-through-analysis pipeline with precision as to each contribution	Health care system sources not named or local health care system site collaborators not named

[a]NLP: natural language processing.

[b]LDL: low-density lipoprotein.

Finally, in crises such as the COVID-19 pandemic, we need to recognize that many studies can contribute to our understanding of what is happening to our patients and how our practices might affect patient outcomes. Overly generalized conclusions will likely strain the boundaries of what can be reasonably inferred from the kinds of data currently obtained through EHRs. Recommendations that flow from overly broad claims may irreversibly harm stakeholders, including patients and clinicians. Increased reader awareness of EHR-derived data quality indicators is crucial in critically appraising EHR-driven studies and to prevent harm from misleading studies, which will ensure sustainable quality in this rapidly growing field.

Acknowledgments

The members of the Consortium for Clinical Characterization of COVID-19 By EHR (4CE) are as follows: Adem Albayrak, Danilo F Amendola, Li LLJ Anthony, Bruce J Aronow, Andrew Atz, Paul Avillach, Brett K Beaulieu-Jones, Douglas S Bell, Antonio Bellasi, Riccardo Bellazzi, Vincent Benoit, Michele Beraghi, José Luis Bernal Sobrino, Mélodie Bernaux, Romain Bey,

Alvar Blanco Martínez, Martin Boeker, Clara-Lea Bonzel, John Booth, Silvano Bosari, Florence T Bourgeois, Robert L Bradford, Gabriel A Brat, Stéphane Bréant, Mauro Bucalo, Anita Burgun, Tianxi Cai, Mario Cannataro, Aize Cao, Charlotte Caucheteux, Julien Champ, Luca Chiovato, James J Cimino, Tiago K Colicchio, Sylvie Cormont, Sébastien Cossin, Jean Craig, Juan Luis Cruz Bermúdez, Arianna Dagliati, Mohamad Daniar, Christel Daniel, Anahita Davoudi, Batsal Devkota, Julien Dubiel, Scott L DuVall, Loic Esteve, Shirley Fan, Robert W Follett, Paula SA Gaiolla, Thomas Ganslandt, Noelia García Barrio, Nils Gehlenborg, Alon Geva, Tobias Gradinger, Alexandre Gramfort, Romain Griffier, Nicolas Griffon, Olivier Grisel, Alba Gutiérrez-Sacristán, David A Hanauer, Christian Haverkamp, Martin Hilka, John H Holmes, Chuan Hong, Petar Horki, Meghan R Hutch, Richard Issitt, Anne Sophie Jannot, Vianney Jouhet, Mark S Keller, Katie Kirchoff, Jeffrey G Klann, Isaac S Kohane, Ian D Krantz, Detlef Kraska, Ashok K Krishnamurthy, Sehi L'Yi, Trang T Le, Judith Leblanc, Guillaume Lemaitre, Leslie Lenert, Damien Leprovost, Molei Liu, Ne Hooi Will Loh, Yuan Luo, Kristine E Lynch, Sadiqa Mahmood, Sarah Maidlow, Alberto Malovini, Kenneth D Mandl, Chengsheng Mao, Patricia Martel, Aaron J Masino, Michael E Matheny, Thomas Maulhardt, Michael T McDuffie, Arthur Mensch, Marcos F Minicucci, Bertrand Moal, Jason H Moore, Jeffrey S Morris, Michele Morris, Karyn L Moshal, Sajad Mousavi, Danielle L Mowery, Douglas A Murad, Shawn N Murphy, Kee Yuan Ngiam, Jihad Obeid, Marina P Okoshi, Karen L Olson, Gilbert S Omenn, Nina Orlova, Brian D Ostasiewski, Nathan P Palmer, Nicolas Paris, Lav P Patel, Miguel Pedrera Jimenez, Hans U Prokosch, Robson A Prudente, Rachel B Ramoni, Maryna Raskin, Siegbert Rieg, Gustavo Roig Domínguez, Elisa Salamanca, Malarkodi J Samayamuthu, Arnaud Sandrin, Emily Schiver, Juergen Schuettler, Luigia Scudeller, Neil Sebire, Pablo Serrano Balazote, Patricia Serre, Arnaud Serret-Larmande, Domenick Silvio, Piotr Sliz, Jiyeon Son, Andrew M South, Anastasia Spiridou, Amelia LM Tan, Bryce WQ Tan, Byorn WL Tan, Suzana E Tanni, Deanne M Taylor, Valentina Tibollo, Patric Tippmann, Andrew K Vallejos, Gael Varoquaux, Jill-Jênn Vie, Shyam Visweswaran, Kavishwar B Wagholikar, Lemuel R Waitman, Demian Wassermann, Griffin M Weber, Yuan William, Zongqi Xia, Alberto Zambelli, Aldo Carmona, Charles Sonday, and James Balshi.

Authors' Contributions

ISK led the 4CE international consortium, conceived and designed the study, and drafted the manuscript. TC led 4CE analytics strategies and made contributions to the study design and drafting of the manuscript. JJC contributed a validation strategy and made edits to the manuscript. NG-B was responsible for data extraction and transformation to 4CE format and quality control of the results and made internal contributions. NG led 4CE visualization strategies and made contributions/edits to the manuscript. JGK contributed to the 4CE validation strategy and data submission strategies and made edits to the manuscript. KDM made contributions to the text and framework and made edits to the manuscript. DM was involved in data extraction and transformation to 4CE format. SNM led 4CE data validation strategies and made contributions/edits to the manuscript. GSO made contributions to strategy and edits to the manuscript. NP contributed to 4CE data analysis, aggregation, and quality control. KBW contributed to validation strategies and made edits to the manuscript. BJA, PA, BKB-J, RB, RLB, GAB, MC, MG, AG-S, DAH, JHH, CH, NHW, YL, JHM, AN, KYN, LPP, MP-J, PS, AMS, ALMT, DMT, BMT, CT, AKV, and GMW made contributions/edits to the manuscript.

Conflicts of Interest

RB and AM are shareholders of Biomeris srl. GSO is affiliated with BoD, Galectin Therapeutics, Angion Biomedica, and Amesite, Inc. DMT consulted on a legal matter for AstraZeneca last year.

References

1. Mehra MR, Desai SS, Ruschitzka F, Patel AN. RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. The Lancet 2020 May [FREE Full text] [doi: 10.1016/S0140-6736(20)31180-6]
2. Mehra MR, Desai SS, Kuy S, Henry TD, Patel AN. Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19. N Engl J Med 2020 Jun 18;382(25):e102. [doi: 10.1056/nejmoa2007621]
3. Cox D, Donnelly C. Principles of Applied Statistics. Cambridge, UK: Cambridge University Press; 2011.
4. Eriksson L, Byrne T, Johansson E, Trygg J, Vikström C. Multi- and Megavariate Data Analysis Basic Principles and Applications. Malmo, Sweden: Umetrics Academy; 2013.
5. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med 2015 Oct 6;12(10):e1001885 [FREE Full text] [doi: 10.1371/journal.pmed.1001885] [Medline: 26440803]
6. Langan SM, Schmidt SA, Wing K, Ehrenstein V, Nicholls SG, Filion KB, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). BMJ 2018 Nov 14;363:k3532 [FREE Full text] [doi: 10.1136/bmj.k3532] [Medline: 30429167]
7. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. Medical Care 2013;51:S30-S37. [doi: 10.1097/mlr.0b013e31829b1dbd]

8. Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. J Med Internet Res 2018 May 29;20(5):e185 [FREE Full text] [doi: 10.2196/jmir.9134] [Medline: 29844010]
9. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC) 2016 Sep 11;4(1):1244 [FREE Full text] [doi: 10.13063/2327-9214.1244] [Medline: 27713905]
10. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 01;20(1):144-151 [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
11. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. Annu Rev Public Health 2016 Mar 18;37(1):61-81 [FREE Full text] [doi: 10.1146/annurev-publhealth-032315-021353] [Medline: 26667605]
12. Capocaccia R, De Angelis R. Estimating the completeness of prevalence based on cancer registry data. Statist Med 1997 Feb 28;16(4):425-440. [doi: 10.1002/(sici)1097-0258(19970228)16:4<425::aid-sim414>3.0.co;2-z]
13. Smirnov VB. Earthquake catalogs: Evaluation of data completeness. Volc Seis 1998;19:497-510 [FREE Full text]
14. Methods for De-identification of PHI. Office for Civil Rights. 2015 Nov 6. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html [accessed 2020-06-16]
15. Kirby J, Speltz P, Rasmussen L, Basford M, Gottesman O, Peissig P, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc 2016 Nov;23(6):1046-1052 [FREE Full text] [doi: 10.1093/jamia/ocv202] [Medline: 27026615]
16. Zhang J, Can A, Lai PMR, Mukundan S, Castro VM, Dligach D, et al. Age and morphology of posterior communicating artery aneurysms. Sci Rep 2020 Jul 14;10(1):11545 [FREE Full text] [doi: 10.1038/s41598-020-68276-9] [Medline: 32665589]
17. Ananthakrishnan AN, Cagan A, Cai T, Gainer VS, Shaw SY, Churchill S, et al. Statin Use Is Associated With Reduced Risk of Colorectal Cancer in Patients With Inflammatory Bowel Diseases. Clin Gastroenterol Hepatol 2016 Jul;14(7):973-979 [FREE Full text] [doi: 10.1016/j.cgh.2016.02.017] [Medline: 26905907]
18. Uno H, Ritzwoller DP, Cronin AM, Carroll NM, Hornbrook MC, Hassett MJ. Determining the Time of Cancer Recurrence Using Claims or Electronic Medical Record Data. JCO Clinical Cancer Informatics 2018 Dec(2):1-10. [doi: 10.1200/cci.17.00163]
19. Liu C, Wang F, Hu J, Xiong H. Temporal Phenotyping from Longitudinal Electronic Health Records: A Graph Based Framework. New York, NY: Association for Computing Machinery; 2015 Presented at: KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data; August 2015; Sydney, NSW, Australia p. 705-714. [doi: 10.1145/2783258.2783352]
20. Brat G, Weber G, Gehlenborg N, Avillach P, Palmer N, Chiovato L, et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. NPJ Digit Med 2020;3:109 [FREE Full text] [doi: 10.1038/s41746-020-00308-0] [Medline: 32864472]
21. Klann J, Abend A, Raghavan V, Mandl K, Murphy S. Data interchange using i2b2. J Am Med Inform Assoc 2016 Sep;23(5):909-915 [FREE Full text] [doi: 10.1093/jamia/ocv188] [Medline: 26911824]
22. Ananthakrishnan AN, Cai T, Savova G, Cheng S, Chen P, Perez RG, et al. Improving Case Definition of Crohn s Disease and Ulcerative Colitis in Electronic Medical Records Using Natural Language Processing. Inflammatory Bowel Diseases 2013;19(7):1411-1420. [doi: 10.1097/mib.0b013e31828133fd]
23. Ning W, Chan S, Beam A, Yu M, Geva A, Liao K, et al. Feature extraction for phenotyping from semantic and knowledge resources. J Biomed Inform 2019 Mar;91:103122 [FREE Full text] [doi: 10.1016/j.jbi.2019.103122] [Medline: 30738949]
24. Zhang Y, Cai T, Yu S, Cho K, Hong C, Sun J, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). Nat Protoc 2019 Dec 20;14(12):3426-3444 [FREE Full text] [doi: 10.1038/s41596-019-0227-6] [Medline: 31748751]
25. Zhong Q, Karlson EW, Gelaye B, Finan S, Avillach P, Smoller JW, et al. Screening pregnant women for suicidal behavior in electronic medical records: diagnostic codes vs. clinical notes processed by natural language processing. BMC Med Inform Decis Mak 2018 May 29;18(1):30 [FREE Full text] [doi: 10.1186/s12911-018-0617-7] [Medline: 29843698]
26. Morin A, Urban J, Adams PD, Foster I, Sali A, Baker D, et al. Research priorities. Shining light into black boxes. Science 2012 Apr 13;336(6078):159-160 [FREE Full text] [doi: 10.1126/science.1218263] [Medline: 22499926]
27. Beaulieu-Jones BK, Greene CS. Reproducibility of computational workflows is automated using continuous analysis. Nat Biotechnol 2017 Apr 13;35(4):342-346 [FREE Full text] [doi: 10.1038/nbt.3780] [Medline: 28288103]
28. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ 2015 Apr 24;350(apr24 11):h1885-h1885 [FREE Full text] [doi: 10.1136/bmj.h1885] [Medline: 25911572]
29. Geissbuhler A, Safran C, Buchan I, Bellazzi R, Labkoff S, Eilenberg K, et al. Trustworthy reuse of health data: a transnational perspective. Int J Med Inform 2013 Jan;82(1):1-9. [doi: 10.1016/j.ijmedinf.2012.11.003] [Medline: 23182430]

Abbreviations

4CE: Consortium for Clinical Characterization of COVID-19 by EHR
EHR: electronic health record
HIPAA: Health Insurance Portability and Accountability Act
RECORD: Reporting of Studies Conducted Using Observational Routinely Collected Health Data
NLP: natural language processing
IRB: institutional review board
PheKB: Phenotype Knowledgebase

Edited by R Kukafka; submitted 13.07.20; peer-reviewed by N Delvaux, M Adly, P Harris, A Adly, A Adly, J Li, L Genaro; comments to author 04.08.20; revised version received 14.09.20; accepted 10.01.21; published 02.03.21

Please cite as:
Kohane IS, Aronow BJ, Avillach P, Beaulieu-Jones BK, Bellazzi R, Bradford RL, Brat GA, Cannataro M, Cimino JJ, García-Barrio N, Gehlenborg N, Ghassemi M, Gutiérrez-Sacristán A, Hanauer DA, Holmes JH, Hong C, Klann JG, Loh NHW, Luo Y, Mandl KD, Daniar M, Moore JH, Murphy SN, Neuraz A, Ngiam KY, Omenn GS, Palmer N, Patel LP, Pedrera-Jiménez M, Sliz P, South AM, Tan ALM, Taylor DM, Taylor BW, Torti C, Vallejos AK, Wagholikar KB, The Consortium For Clinical Characterization Of COVID-19 By EHR (4CE), Weber GM, Cai T
What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask
J Med Internet Res 2021;23(3):e22219
URL: https://www.jmir.org/2021/3/e22219
doi: 10.2196/22219
PMID: 33600347

Original Paper

Exploring Usage of COVID Coach, a Public Mental Health App Designed for the COVID-19 Pandemic: Evaluation of Analytics Data

Beth K Jaworski[1*], PhD; Katherine Taylor[1*], MPH, PsyD; Kelly M Ramsey[1*]; Adrienne Heinz[1,2*], PhD; Sarah Steinmetz[1*], PhD; Ian Pagano[3*], PhD; Giovanni Moraja[4*]; Jason E Owen[1*], MPH, PhD

[1]National Center for PTSD, Dissemination & Training Division, US Department of Veterans Affairs, Menlo Park, CA, United States

[2]School of Medicine, Stanford University, Stanford, CA, United States

[3]University of Hawaii Cancer Center, Honolulu, HI, United States

[4]Vertical Design, LLC, Berkeley, CA, United States

[*]all authors contributed equally

Corresponding Author:
Beth K Jaworski, PhD
National Center for PTSD, Dissemination & Training Division
US Department of Veterans Affairs
795 Willow Road
Menlo Park, CA, 94025
United States
Phone: 1 650 308 9437
Email: beth.jaworski@va.gov

Abstract

Background: The COVID-19 pandemic has significantly impacted mental health and well-being. Mobile mental health apps can be scalable and useful tools in large-scale disaster responses and are particularly promising for reaching vulnerable populations. COVID Coach is a free, evidence-informed mobile app designed specifically to provide tools and resources for addressing COVID-19–related stress.

Objective: The purpose of this study was to characterize the overall usage of COVID Coach, explore retention and return usage, and assess whether the app was reaching individuals who may benefit from mental health resources.

Methods: Anonymous usage data collected from COVID Coach between May 1, 2020, through October 31, 2020, were extracted and analyzed for this study. The sample included 49,287 unique user codes and 3,368,931 in-app events.

Results: Usage of interactive tools for coping and stress management comprised the majority of key app events (n=325,691, 70.4%), and the majority of app users tried a tool for managing stress (n=28,009, 58.8%). COVID Coach was utilized for ≤3 days by 80.9% (n=34,611) of the sample whose first day of app use occurred within the 6-month observation window. Usage of the key content in COVID Coach predicted returning to the app for a second day. Among those who tried at least one coping tool on their first day of app use, 57.2% (n=11,444) returned for a second visit; whereas only 46.3% (n=10,546) of those who did not try a tool returned (P<.001). Symptoms of anxiety, depression, and posttraumatic stress disorder (PTSD) were prevalent among app users. For example, among app users who completed an anxiety assessment on their first day of app use (n=4870, 11.4% of users), 55.1% (n=2680) reported levels of anxiety that were moderate to severe, and 29.9% (n=1455) of scores fell into the severe symptom range. On average, those with moderate levels of depression on their first day of app use returned to the app for a greater number of days (mean 3.72 days) than those with minimal symptoms (mean 3.08 days; t_1=3.01, P=.003). Individuals with significant PTSD symptoms on their first day of app use utilized the app for a significantly greater number of days (mean 3.79 days) than those with fewer symptoms (mean 3.13 days; t_1=2.29, P=.02).

Conclusions: As the mental health impacts of the pandemic continue to be widespread and increasing, digital health resources, such as apps like COVID Coach, are a scalable way to provide evidence-informed tools and resources. Future research is needed to better understand for whom and under what conditions the app is most helpful and how to increase and sustain engagement.

(J Med Internet Res 2021;23(3):e26559) doi: 10.2196/26559

KEYWORDS

COVID-19; coronavirus; mobile app; mHealth; digital health; mental health; public mental health; stress; coping; public health; app

Introduction

Impact of COVID-19 on Mental Health and Well-Being

In the United States, the COVID-19 pandemic has led to over 500,000 deaths, millions of job losses, and disruption of nearly every aspect of daily life. COVID-19 has also negatively impacted mental health and well-being globally [1-3]. One-third of American adults report a high level of psychological distress due to the pandemic [4].

Several studies now indicate that an unprecedented mental health crisis is underway. In a poll conducted by Harris [5] on behalf of the American Psychological Association, nearly 8 in 10 adults said the pandemic is a significant source of stress in their lives. The prevalence of depression symptoms among adults in the United States has risen from 8.5% of the population prior to the COVID-19 pandemic to 27.8% in the midst of the pandemic [6]. Researchers from the US Centers for Disease Control and Prevention found that 40% of respondents of a survey administered in June 2020 endorsed at least one adverse mental or behavioral health condition including symptoms of depression, anxiety, posttraumatic stress, or having started or increased substance use to cope with stress or emotions related to COVID-19. Over 10% of respondents reported seriously considering suicide in the previous 30 days [7]. Furthermore, there appears to be a bidirectional relationship between COVID-19 and psychiatric disorders, such that having a psychiatric disorder is associated with a greater likelihood of contracting COVID-19, and contracting COVID-19 is associated with an increased risk of receiving a psychiatric diagnosis [8].

Digital Mental Health as a Strategy for Addressing the Mental Health Impact of COVID-19

Digital mental health options are needed to help address the mental health effects of COVID-19 as well as the secondary impacts of the pandemic, such as fear of contracting the virus, financial stress related to job loss, loss of childcare, or the need to balance work with remote education. Mobile mental health apps are a promising strategy for addressing mental health impacts of the pandemic because of their potential scalability, reach, and utility, particularly during a time when in-person care may not be accessible due to social distancing and safety regulations. High-quality, accessible, and sustainable apps have been identified as part of an integrated "blueprint" for digital mental health services during the pandemic [9]. They may be a particularly useful tool for reaching a large number of individuals from highly impacted populations at risk for posttraumatic stress disorder (PTSD) or other mental health conditions, including those who have contracted COVID-19 and frontline health care workers [10].

Apps are a particularly appealing medium because of their potential reach. Individuals rarely turn off mobile devices [11], making apps available 24/7. Additionally, in the United States, 81% of adults own smartphones, with few differences among sociodemographic groups [12]. This reach is important because the pandemic has a disproportionate and complex impact on Black, Indigenous, people of color (BIPOC), people from low-income backgrounds, and women [13], and it is clear that vulnerable groups are at greater risk for behavioral and mental health consequences [6,7,14]. Systemic disadvantage with respect to social determinants of health, such as lack of internet access and reduced educational opportunities, has been associated with increased COVID-19 mortality rates [15]. Free, evidence-informed apps, such as COVID Coach, that are developed by government or not-for-profit entities and made specifically to address such systemic barriers, can contribute to a digital mental health safety net for vulnerable individuals. Beyond the ability to reach many people, apps have been shown to be useful adjunctive resources for a range of mental health concerns, including anxiety and depression [16] and PTSD [17].

Creation of the COVID Coach App

In response to the anticipated mental health impact of the COVID-19 pandemic, and as part of the Veterans Affairs' (VA) "Fourth Mission" to help during times of national emergencies and support public health, the National Center for PTSD created COVID Coach (Multimedia Appendix 1). COVID Coach is a free, publicly available mental health app designed to help people cope with stress, find resources, and track mental health over time. It is intended to be simple to use, does not require an internet connection or data plan to access primary content, and all recommended activities and resources are low in cost or free to users. COVID Coach is one of only a few public mental health apps available for specifically addressing mental health concerns stemming from or exacerbated by COVID-19, and it is the newest in a suite of free mental health apps designed to support mental health [18,19].

COVID Coach is based upon the model of the empirically supported PTSD Coach app [20], which has been identified as a potential approach for the behavioral and mental health impact of COVID-19 [21]. COVID Coach provides app users with many of the features of PTSD Coach, including tools for coping with challenging situations and managing stress, psychoeducation, tracking of mental health symptoms, and quick access to support networks and crisis resources. COVID Coach also provides symptom management tools adapted for life during the pandemic (eg, sleep struggles; isolation; stress; sadness; and indoor, socially distanced activities), goal-setting, and over 50 unique psychoeducational topics about managing COVID-19–related concerns (ie, staying well, staying balanced, staying together, staying safe, and staying healthy). The app was released at the end of April 2020 and has been promoted as part of the VA's response to the pandemic and highlighted as an important resource [22].

Evaluating COVID Coach in the Context of a Public Health Disaster

Mobile mental health apps can be useful tools in large-scale disaster responses [23], and their use has been indicated

specifically within the context of the COVID-19 pandemic (eg, [24,25]). However, the utility of standalone apps "in the wild" can be limited by poor engagement and high attrition (eg, [26,27].) A host of challenges renders it difficult to conduct formal research and evaluation on disaster mental health interventions and resources [28]. Accordingly, there is often insufficient data on when, how, and why individuals utilize disaster mental health resources to help guide policy and budgetary allocation. Although COVID Coach has been well received in the general population, usage of the app, particularly the key content areas, and retention have not yet been formally evaluated.

Objective

This study utilized anonymous mobile analytics data to characterize the overall usage of an app designed specifically to provide tools and resources for addressing COVID-19–related stress, explore retention and return usage, and assess whether the app was reaching individuals that may benefit from mental health resources. Three key aims guided the study: (1) describe general usage trends between May 1, 2020, and October 31, 2020 (a key period of time during the pandemic), and identify how frequently specific types of key app content were used (ie, coping tools, psychoeducation, self-assessments, and accessing resources); (2) explore usage patterns, with a particular focus on understanding how usage of key content on the first day of use may be related to return use and retention; and (3) characterize baseline mental health and well-being among COVID Coach users.

Methods

COVID Coach Mobile App Description

COVID Coach, available for Android [29] and iOS [30], is an app designed specifically for the COVID-19 pandemic to provide users with interactive, evidence-informed tools for coping with stress and anxiety, information about how to stay well, stay connected, and navigate challenges, self-monitoring mental health symptoms and goals, and resources to discover and connect with various types of verified and vetted support. The app can be used independently or in conjunction with professional mental health care but is not a replacement for therapy. Users are not required to create an account or log in to access any of the content, and the app is fully compatible with assistive software technologies (eg, VoiceOver or TalkBack).

Mobile Analytics Data

COVID Coach collects anonymous information about app use for the purposes of quality improvement. Fully nonidentifying, anonymous, and encrypted event sequences were stored using JavaScript Object Notation (JSON) format on a remote GovCloud server that meets VA security and privacy requirements. Data are accessible from VA App Connect software, which has been approved for use under the VA's Technical Reference Model [31]. Upon first launch of the app, a unique, randomly generated 32-character (256-bit) code is assigned to that particular app installation. Completely anonymous usage data, such as screens selected, button presses, and other nonidentifying patterns, are collected and associated with this install code. Install codes serve as a proxy for app users since the unique identity of each app user cannot be determined. Each in-app event contains a timestamp (in Coordinated Universal Time [UTC]) that corresponds to when the event actually occurred, but data are only transmitted to the server when the app is in use and connected to Wi-Fi or utilizing a data plan.

Procedures

For the purpose of this study, mobile analytics data with timestamps between May 1, 2020, and October 31, 2020, were extracted from the research server on November 4, 2020. Between May 1 and October 31, 3,368,931 in-app related events were captured (Android: n=847,260; iOS: n=2,521,612) across 49,297 unique install codes (Android: n=12,938; iOS: n=36,359).

Measures

App Use Metrics

Daily active users and monthly active users were measured by the total number of app users that used COVID Coach on a given day or at least once within a given month. Overall, frequencies for key content usage were computed for each of the four key sections in the app: *Manage Stress* (tried a tool), *Learn* (viewed a learn topic), *Mood Check* (created and rated a goal or completed an assessment), and *Find Resources* (viewed at least one specific subsection within *Find Resources*). These frequencies were computed for all key events and for all app users that had activity during the observation window (May 1, 2020, through October 31, 2020). Based on a rationale similar to Kwasny and colleagues [32], we decided a priori that frequency of use within the observation window would be measured in terms of unique days of use, rather than sessions or visits because of the variability in establishing the end of an app session, within and across platforms. Additionally, all app users were categorized according to whether their first day of app use occurred during the observation window (first-time users) or prior to the start of the observation window. Thus, all analyses related to distinct days of app use, return usage, and patterns of usage by day of use focused only on app events associated with first-time users. Among all first-time users, distinct days of app use within the observation window were calculated, as well as retention days (the number of days between the first day of use and the last day of use) and the number of days between the first day of use and the second day of use (for all individuals who used the app for at least 2 distinct days). For each first-time user, completion of tasks within each of the four key content areas were totaled, by each distinct day of use. First-time users who completed one or more assessments on their first day of app use were identified as "baseline" assessment completers.

In-App Assessments

Four assessments are available within the *Mood Check* section of COVID Coach. These assessments can be accessed and taken at any time by app users.

The Warwick-Edinburgh Mental Well-Being Scale (WEMWBS) [33] is a measure to assess the feelings and functional aspects

of positive mental health. COVID Coach contains the 14-item version of the scale, with each item measured on a Likert-type scale ranging from 1 (“none of the time”) to 5 (“all of the time”). For each item, respondents are asked to consider how they have been feeling over the past 2 weeks. Total score is obtained by summing all items. Scores of less than 42 are indicative of low well-being [34]. The scale was found to be a valid and reliable tool for measuring mental well-being in diverse populations and across project types, and has adequate internal reliability (α=.89) [35].

The Generalized Anxiety Disorder-7 (GAD-7) [36] is a measure to screen for GAD and assess severity of GAD symptoms. The scale consists of 7 items, each measured on a Likert-type scale ranging from 0 (“not at all”) to 3 (“nearly every day”), and total score is obtained by summing all items. For each item, respondents are asked to consider how they have been feeling over the past 2 weeks. Anxiety symptom severity is categorized as: minimal (total score=0-4), mild (total score=5-9), moderate (total score=10-14), and severe (total score=15 or higher). The scale has acceptable internal reliability and good psychometric properties, including among general population samples [37].

The Patient Health Questionnaire-9 (PHQ-9) [38] is a measure to assess the severity of depression symptoms. The scale consists of 9 items, each measured on a Likert-type scale ranging from 0 (“not at all”) to 3 (“nearly every day”), and total score is obtained by summing all items. For each item, respondents are asked to consider how they have been feeling over the past 2 weeks. Depression symptom severity is categorized as: minimal (total score=0-4), mild (total score=5-9), moderate (total score=10-14), moderately severe (total score=15-19), and severe (total score=20 or higher). The scale has acceptable internal reliability (α=.86-.89) and overall sound psychometric properties across settings [39].

The Posttraumatic Stress Disorder Checklist (PCL-5) [40] is a measure to assess symptoms of PTSD. The scale consists of 20 items, each measured on a Likert-type scale ranging from 0 (“not at all”) to 4 (“extremely”), and total score is obtained by summing all items. In COVID Coach, the PCL-5 is administered with only a brief introduction, followed by the assessment items. For each item, respondents are asked to consider how they have been feeling over the past month. Initial research suggests that total scores of 31 to 33 (or higher) are indicative of probable PTSD. For this study, we use 33 as the cut-off for significant PTSD symptoms. The PCL-5 was found to be reliable and valid in both veteran [41] and civilian populations [42].

Analyses

SQLPro Studio (Hankinsoft Development, Inc) was used for all data preprocessing and extraction. SAS University Edition (SAS Institute) software in conjunction with Oracle’s VirtualBox were used for all data analyses. We calculated descriptive statistics for key content usage, retention, and baseline levels of mental health symptom severity and levels of well-being. Chi-square analyses were conducted to understand differences in returning to the app for a second day of use based on key content usage on the first day of app use and baseline mental health symptoms. We ran separate chi-square analyses for each predictor. Independent samples *t* tests were conducted to examine differences in total unique days of app use and total manage stress tools utilized among app users who completed an assessment on their first day of app use compared to those who did not. An analysis of variance (ANOVA) was conducted with a Tukey test for post hoc analysis to examine differences in baseline WEMWBS scores, by month, among users who completed a well-being assessment on their first day of app use. Regression analyses were conducted to examine the relationship between baseline mental health symptoms and unique days of app usage.

Results

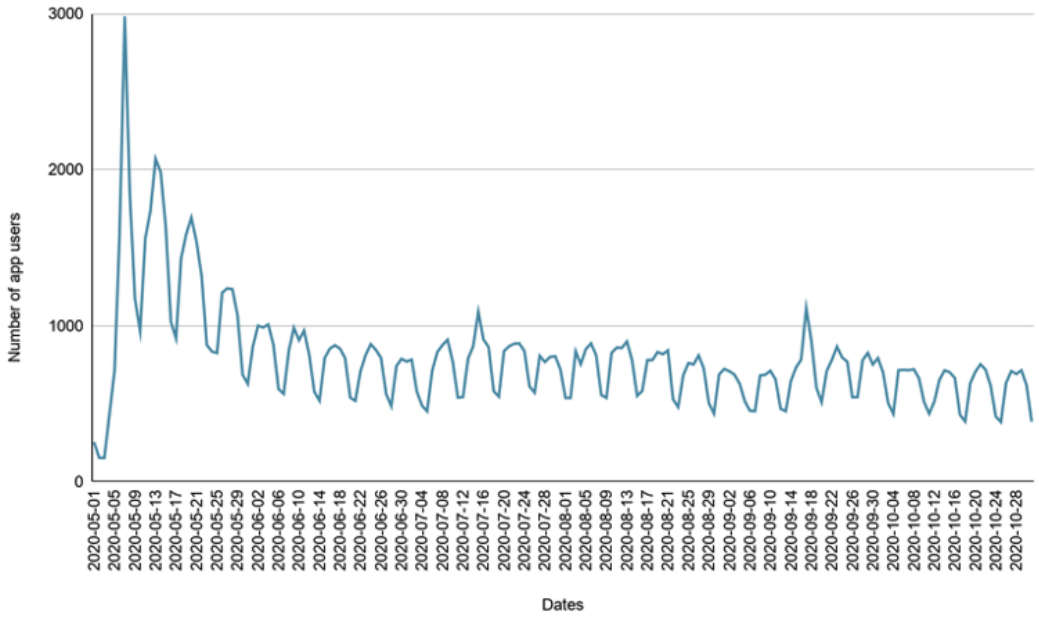
Reach and Reception

The app was released at the end of April 2020, and as of October 31, 2020, it has been downloaded 143,097 times. It is highly rated on both the Apple App Store (4.8 out 5 stars) and the Google Play Store (4.7 out of 5 stars). Users had the opportunity to provide written reviews along with star ratings. The majority of written reviews were overwhelmingly positive, with comments such as “Beautifully calming…,” “a necessity for our new normal,” “one of the best free apps I’ve found,” and “this is amazing… it has all you may need… mood trackers, resources, meditation-not too frilly, just important.” Notably, due to Google’s restrictions on mobile apps related to COVID-19 (including hiding certain results for apps among searches containing “COVID”), COVID Coach has been installed at a ratio of over 3:1 for iOS compared to Android mobile devices.

Daily and Monthly Active Users

The number of daily active users spiked in May 2020 (mean 1205.77, SD 615.70), shortly after the app’s release. The number of daily active users has leveled off but remained stable with average daily active users of 778.67 (SD 161.16), 752.03 (SD 152.07), 712.71 (SD 141.85), 682.83 (SD 150.83), and 611.35 (SD 128.60), respectively, during the months of June, July, August, September, and October 2020 (Figure 1). Although timestamp information is only captured in UTC, there appears to be a consistent, weekly pattern of usage such that the app is used more during the week than on weekends. The number of monthly active users followed a similar pattern as daily active users. The number of monthly active users peaked in May but remained steady through October 2020, at approximately 11,000 unique app users per month.

Figure 1. Daily active COVID Coach users, from May 1, 2020, through October 31, 2020.



Key Content Usage

Within the observation window (May 1, 2020, through October 31, 2020), there were 49,297 unique app users and 462,651 app events associated with the four key content areas (*Manage Stress*, *Learn*, *Mood Check*, and *Find Resources*). Table 1 provides an overview.

Of the four key sections of the app, the *Manage Stress* section, which contains tools for coping with stress and anxiety, was the most utilized. Across the observation window, there were 325,691 total tool use events (70.4% of all key events), among 28,009 unique install codes (56.82% of all unique install codes).

Within the *Manage Stress* section, app users can directly select individual tools from a list of all tools, or they can have a tool recommended to them by selecting from one of seven possible challenges related to the pandemic: (1) coping with stress, (2) feeling lonely, (3) creating space for myself, (4) feeling sad or hopeless, (5) handling anger and irritability, (6) navigating relationships, and (7) sleep struggles. Across all app users, 48.5% (n=23,885) selected at least one challenge. Among this group of app users, challenges related to coping with stress were the most commonly selected (n=12,696, 53.2%), followed by sleep struggles (n=9308, 39.0%) and feeling lonely (n=9153, 38.3%).

Table 1. Overall key content usage among all COVID Coach users within the observation window (between May 1, 2020, and October 31, 2020).

Key content area (specific in-app action)	Unique app users, n (%)[a]	Key events, n (%)[b]	Totals per app user, mean (SD); range
Manage Stress (tried at least one tool)	28,009 (56.8)	325,691 (70.4)	11.63 (30.33); 1-2124
Learn (viewed at least one topic)	10,124 (20.5)	52,123 (11.3)	5.15 (8.09); 1-267
Mood Check (entered and rated at least one goal or completed at least one assessment)	13,510 (27.4)	47,821 (10.3)	3.54 (12.52); 1-1008
Find Resources (viewed at least one specific subsection)	9418 (19.1)	37,016 (8.0)	3.93 (7.82); 1-329

[a]Total number of unique app users during the observation window=49,297. Percentage of total app users. Percentages in this column will not sum to 100% because app users could have completed actions across the four types of key content areas.

[b]Total key app events during the observation window=462,651. Percentage of total key app events.

Overall, the psychoeducation content within the *Learn* section of the app was consumed less frequently and by fewer users than the *Manage Stress* tools. Within the observation sample, there were 52,123 unique learn topic views (11.3% of all key events), among 20.5% of all app users (10,124/49,297). Four out of the five most viewed topics appeared in the first subsection within *Learn* (Staying Well).

In total, core activities within the *Mood Check* section comprised 10.3% (47,821/462,651) of all key events. Across the observation window, 27.4% of all app users (13,510/49,297) submitted at least one goal success rating or completed at least one of the four available assessments in the *Mood Check* section. There were 10,253 submitted goal success ratings across 2285 app users (4.6% of the total sample), and 37,568 completed assessments across 13,223 unique app users (26.8% of all users).

Across the eleven subsections within *Find Resources*, 19.1% (9418/49,297) of all app users viewed the resource pages 37,016 times across the observation window, representing 8% of all key events. Notably, although not the most frequently viewed subsection, *Crisis Resources* (which includes direct links to

phone lines, text support, and online chat for services such as the National Suicide Prevention Lifeline, Crisis Text Line, and Substance Abuse and Mental Health Services Administration's Helpline) was visited 3297 times (8.9% of all *Find Resources* visits) among 2131 unique users. Table 2 presents detailed information about key events within each of the four key sections.

Table 2. Detailed key content usage among all COVID Coach users within the observation window (between May 1, 2020, and October 31, 2020).

Key content area	Unique app users, n (%)[a]	Key app events, n (%)[b]
Manage Stress: top 5 most frequently used tools		
Ambient Sounds (an audio-only tool with no narration)	7041 (14.3)	18,493 (4.0)
Deep Breathing (an audio-guided exercise)	7870 (16.0)	16,011 (3.5)
Change Your Perspective (a tool with tips for how to replace negative thoughts with more helpful ones)	6721 (13.6)	12,480 (2.7)
Muscle Relaxation (an audio-guided exercise focused on relaxing distinct core body parts)	6037 (12.2)	11,599 (2.5)
Grounding (a tool with tips on how to stay connected to the present moment and surroundings)	5718 (11.6)	9767 (2.1)
Learn: top 5 most frequently viewed topics		
Prioritizing Yourself, Right Now	2131 (4.3)	2811 (0.6)
Managing Irritability	1856 (3.8)	2281 (0.5)
Finding Humor	1506 (3.1)	1817 (0.4)
Finding Calm	1442 (2.9)	1813 (0.4)
Sleep	1184 (2.4)	1511 (0.3)
Find Resources: top 3 most frequently viewed sections		
Finding Local Resources (for locating state-specific COVID-19 guidelines and information)	2927 (5.9)	6451 (1.4)
Meeting Your Needs (for basic needs support)	3176 (6.4)	6223 (1.3)
Mobile Apps to Support Mental Health (information about other free apps to support mental health)	2461 (5.0)	3748 (0.8)
Mood Check: completion of assessments, by type		
Track Mood (PHQ-9[c])	7698 (15.6)	11,732 (2.5)
Track Anxiety (GAD-7[d])	8115 (16.5)	11,649 (2.5)
Track Well-Being (WEMWBS[e])	6151 (12.5)	8860 (1.9)
Track PTSD[f] Symptoms (PCL-5[g])	3568 (7.2)	5327 (1.2)

[a]Total number of unique app users during the observation window=49,297. Percentage of total app users.

[b]Total key app events during the observation window=462,651. Percentage of total key app events.

[c]PHQ-9: Patient Health Questionnaire-9.

[d]GAD-7: Generalized Anxiety Disorder-7.

[e]WEMWBS: Warwick-Edinburgh Mental Well-Being Scale.

[f]PTSD: posttraumatic stress disorder.

[g]PCL-5: Posttraumatic Stress Disorder Checklist-5.

Return Usage and Retention

Among the 49,297 app user install codes present in the observation window, 86.8% (n=42,783) of COVID Coach users had their first day of app use occur within the observation window. Thus, for the analyses presented in this section, usage patterns will be restricted to only the app users and events associated with those whose first day of app use occurred during the observation window.

Nearly half of COVID Coach users used the app for a single day (n=20,793, 48.6%), and an additional 32.3% (n=13,818) used the app for 2 or 3 days in total. Less than 2% of the sample (n=709) used the app for 15 or more distinct days (Table 3). On average, across all app users with ≥2 distinct days of app use (n=21,990), the number of days retained was 42.44 (SD 44.40, median 25, range 1-179). On average, the number of days between the first day of app use and the second day of app use was 14.65 (SD 24.52, median 4, range 1-176). Although the majority of app users who returned to the app for at least a second day returned within 14 days, there was variability,

including users whose second day of use occurred over 90 days after the first (see Table 4 for a detailed analysis among users whose first month of use occurred in May, June, or July so that returns within a 90-day or longer window could be examined).

Table 3. Total number of distinct days of COVID Coach use, by month of first app use.

Month of first app use	Frequency of users per distinct day, n (%)					
	1 day only	2 days	3 days	4-6 days	7-14 days	≥15 days
May	6573 (42.67)	3406 (22.11)	1891 (12.28)	2137 (13.87)	1049 (6.81)	348 (2.26)
June	2533 (44.26)	1205 (21.10)	693 (12.14)	770 (13.49)	372 (6.51)	137 (2.40)
July	2939 (48.79)	1246 (20.68)	640 (10.62)	737 (12.23)	362 (6.01)	100 (1.66)
August	3165 (51.31)	1267 (20.54)	648 (10.51)	698 (11.32)	314 (5.09)	76 (1.23)
September	2783 (53.66)	1119 (21.58)	548 (10.57)	503 (9.70)	190 (3.66)	43 (0.83)
October	2800 (65.25)	849 (19.79)	306 (7.13)	273 (6.36)	58 (1.35)	5 (0.12)
All	20,793 (48.6)	9092 (21.25)	4726 (11.05)	5118 (11.96)	2345 (5.48)	709 (1.66)

Table 4. Analysis of days between first and second app use among COVID Coach users with at least 2 distinct days of app use, by month of first use.

First month of app use	Users that returned at least once, n	First return within 7 days, n (%)	First return within 8-14 days, n (%)	First return within 15-30 days, n (%)	First return within 31-60 days, n (%)	First return within 61-90 days, n (%)	First return after more than 90 days, n (%)
May	8831	4956 (56.12)	1071 (12.13)	1107 (12.54)	870 (9.85)	391 (4.43)	436 (4.94)
June	3177	1786 (56)	403 (12.68)	457 (14.38)	298 (9.38)	118 (3.71)	115 (3.62)
July	3085	1922 (62.30)	385 (12.48)	366 (11.86)	232 (7.52)	130 (4.21)	50 (1.62)

Differential Day 2 Return Rates Based on Day 1 Key Content Usage

On both the first and second days of app use (see Table 5 for an overview of usage), many app users tried at least one tool within the *Manage Stress* section (46.80% [n=20,222] on the first day, 41.85% [n=9202] among individuals who returned to the app for a second day). Usage of the key content in COVID Coach predicted returning to the app for a second day.

Of those who tried at least one *Manage Stress* tool on their first day of app use, 57.2% (n=11,444) returned for a second visit; whereas only 46.3% (n=10,546) of those who did not try a tool returned (P<.001). Among those who viewed at least one *Learn* topic on their first day of app use, 58.8% (n=3292) returned for a second day of use; whereas only 50.3% (n=18,698) who did not view a learn topic returned (P<.001). With respect to the *Mood Check* section, 57.2% (n=4892) of app users that completed at least one goal rating or one assessment activity returned for a second day of use, compared to 50.0% (n=17,098) of users who did not complete any *Mood Check* activities (P<.001). Lastly, among app users who viewed at least one specific *Find Resources* subsection, 57.4% (n=3014) returned for a second day of app use, compared to only 50.6% (n=18,976) of users who did not view any resources returned (P<.001).

Additionally, usage patterns among individuals who completed an assessment on the first day of app use were significantly different than those who did not complete an assessment on their first day. On average, individuals who completed at least one assessment on their first day of app use utilized COVID Coach for more unique days within the observation window (mean 3.29 days, SD 5.44) compared to individuals who did not complete an assessment on the first day (mean 2.66 days, SD 4.37; P<.001). Similarly, individuals who completed at least one assessment on their first day of app use utilized, on average, significantly more *Manage Stress* tools within the observation window (mean 9.2 tools, SD 24.6) compared to individuals who did not complete an assessment on the first day (mean 5.8 tools, SD 19.3; P<.001).

Table 5. Comparison of key content area usage, by first and second day of app use.

Number of key content areas accessed	App users	
	First day of app use (n=42,783), n (%)	Second day of app use (n=21,990), n (%)
All four key areas		
Completed at least one action within all four key content areas	650 (1.5)	192 (0.9)
Two to three key areas		
Manage Stress (with one or two other key areas; tried at least one tool and completed another action within one or two other key areas)	7953 (18.6)	3143 (14.3)
Two or three key areas (excluding Manage Stress; completed at least one action within two or more of the Learn, Mood Check, or Find Resources sections)	1129 (2.6)	405 (1.8)
One key area		
Manage Stress only (only tried at least one tool)	11,419 (26.7)	5867 (26.7)
Mood Check only (only completed at least one goal rating or assessment)	2677 (6.3)	1355 (6.2)
Find Resources only (only viewed at least one resource subsection)	1196 (2.8)	660 (3.0)
Learn only (only viewed at least one learn topic)	805 (1.9)	485 (2.2)
No key area actions		
Did not complete an action within any of the four key areas	16,954 (39.6)	9883 (44.9)

Characterizing Baseline Mental Health Among COVID Coach Users

Baseline well-being among individuals using COVID Coach appeared to be relatively low and decreased over time. Among app users who completed a WEMWBS assessment on their first day of app use (n=3558, 8.32% of all users whose first day of app use occurred during the observation window), average well-being scores, by month, were all less than 42, which has been used as a cut-off to identify low well-being [34]. These average baseline scores decreased over time, with app users who completed their first WEMWBS on their first day of app use in September 2020 (n=416; mean 38.7, SD 0.04) or October 2020 (n=341; mean 38.1, SD 9.50) demonstrating significantly lower average well-being scores than app users who completed their first WEMWBS on their first day of using the app in May 2020 (n=1361; mean 41.2, SD 9.65).

Symptoms of anxiety, depression, and PTSD were prevalent among app users. For all app users who completed a GAD-7 assessment on their first day of app use (n=4870; 11.4% of users), 12.8% (n=625) had scores suggesting minimal anxiety (total score=0-4), 32.1% (n=1565) endorsed mild levels of anxiety (total score=5-9), 25.2% (n=1225) indicated moderate levels of anxiety (total score=10-14), and 29.9% (n=1455) of scores fell into the severe symptom range (total score=15 or higher).

Among app users who completed a PHQ-9 on their first day of app use (n=4548, 10.6% of users), 16.5% (n=749) had scores suggesting minimal depression (total score=0-4), 28.9% (n=1312) endorsed mild levels of depression (total score=5-9), 25.0% (n=1136) indicated moderate levels of depression (total score=10-14), 17.5% (n=795) endorsed moderately severe levels of depression (total score=15-19), and 12.2% (n=556) of scores fell into the severe symptom range (total score=20 or higher).

Unlike the GAD-7 and the PHQ-9, the PCL-5 does not have symptom severity categorizations. However, among app users who completed a PCL-5 on their first day of app use (n=2064, 4.8% of users), the majority of individuals who completed the assessment (n=1234, 59.8%) had a total score ≥33, which is consistent with significant PTSD symptoms.

Baseline Mental Health Characteristics and Return Usage

Baseline PTSD symptoms predicted returning to the app for a second day. Among individuals with a baseline PCL-5 score of 33 or greater, 62.5% returned to the app for a second day of use, compared to only 56.4% of individuals with scores below 33 (P=.006). Neither symptom severity for anxiety or depression nor levels of well-being were predictive of return usage.

We conducted regression analyses to examine the relationship between baseline mental health symptoms and unique days of app usage. Depression and PTSD symptoms were predictive of the total number of unique days of app use. With respect to depression symptoms, we utilized the group with minimal symptoms as the reference group in comparison to those with mild, moderate, moderately severe, and severe symptoms. On average, those with moderate levels of depression on their first day of app use returned to the app for a greater number of days (mean 3.72 days) than those with minimal symptoms of depression (mean 3.08 days; t_1=3.01, P=.003). Individuals with mild, moderately severe, and severe depression did not significantly differ from the reference group. Although the difference in usage between moderately severe and minimal

symptom severity categories was not statistically significant, it was trending in the predicted direction. With respect to PTSD symptoms, individuals with baseline PCL-5 scores indicating significant PTSD symptoms utilized the app for a significantly greater number of days (mean 3.79 days) than those with subthreshold symptom levels (mean 3.13 days; t_1=2.29, P=.02).

Discussion

Principal Findings

This exploration of COVID Coach usage among the general population suggests that mobile apps may have the reach and accessibility necessary to be a useful medium for disseminating mental health information and resources to individuals experiencing stress related to the COVID-19 pandemic.

Between May 1, 2020, and October 31, 2020, the app was used by nearly 50,000 individuals, and daily active usage has remained steady over time. In addition to the total number of individuals reached, the key content within the app was utilized in over 450,000 instances. The stress management tools were most frequently used with over 28,000 users utilizing individual tools over 300,000 times. Further, each of the other three key content areas in the app were accessed tens of thousands of times by tens of thousands of users. This reach and scalability of COVID Coach across the general population is an example of how digital mental health tools can become successfully integrated into disaster response strategies. From a public mental health perspective (eg, [43]), being able to rapidly deploy evidence-informed tools and reliable health information via a free, accessible, and secure app is a way for the federal government to contribute to a digital mental health safety net and reduce barriers to accessing mental health resources.

Importantly, COVID Coach appears to be reaching individuals in need of mental health resources. On average, among app users who completed assessments during their first day of use, well-being was low, and the majority of individuals were indicating greater than minimal symptoms of anxiety, depression, and PTSD. Additionally, among app users who identified challenges they are facing, the majority reported difficulties with managing stress, troubles with sleep, and feelings of loneliness. We cannot determine if individuals utilizing COVID Coach are representative of the general population, but elevated levels of anxiety, depression, and posttraumatic stress are consistent with other research conducted during the pandemic [6,7,44]. Individuals with significant PTSD symptoms at baseline were more likely to return to the app for a second day of app use. On average, individuals with significant PTSD symptoms used the app for a greater number of days than those with subthreshold symptoms, and individuals with moderate depression used the app for more days than those with minimal symptoms. Greater usage among individuals with moderate depression symptoms is consistent with previous research [45].

Although overall app utilization data suggested considerable reach, engagement proved to be less consistent. Our analyses revealed that the majority of COVID Coach users (80.9%) utilized the app on ≤3 days. This finding is consistent with research indicating that self-management apps for mental health are often not used over extended periods of time [26,27,46]. However, as noted by Ng and colleagues [47], there is a need for more standardized reporting of measures related to user engagement and retention. The average number of retention days, as well as the number of days between days of use, suggest that the app may not be something that individuals use on a daily basis, but rather during moments of distress or need. This type of usage is consistent with the overall design of COVID Coach as a self-management tool, which does not provide any guidance on how often or when to use the app.

This research also provides some guidance on how engagement might be encouraged in future app versions. In general, app users that completed actions within the key content areas on the first day of app use were more likely to return for a second day of app use. More specifically, users that completed an assessment on the first day of app use were significantly more likely to use the app for a greater number of days and to use a greater number of stress management tools than app users who did not complete an assessment on the first day of app use. These findings suggest that finding ways to motivate users to complete actions within key areas on their first day of app use, particularly tools and assessments, may be one way to enhance engagement and retention. For example, having recommendations for a tool or assessment to try, easily accessible from the app home screen, may encourage users to try a specific in-app activity. Additionally, the onboarding sequence could include a few brief questions to help tailor in-app recommendations to the user's intentions and preferences, and guide them through the process of setting customized goals for using the features within the app most relevant to them. Lastly, finding ways to regularly disseminate and highlight new app content (eg, managing stress around prolonged distance learning, vaccine information) may encourage users to return to the app more frequently.

Limitations

Because COVID Coach does not collect any identifying information, we cannot say anything about the populations that we have reached, other than what we can characterize based upon in-app actions. Future research that permits collection of identifying information is needed, particularly given the disproportionate impact the pandemic has had on vulnerable groups of people. A Spanish version of COVID Coach has recently been released, and plans for data collection on app usage within Spanish-speaking populations are underway.

Additionally, we utilize the unique install codes as a proxy for an individual user. We assume that most individuals do not delete and reinstall the app multiple times. However, if an individual were to download COVID Coach on more than one mobile device, or delete it and reinstall, each of those installations would be assigned a unique install code, and would appear as a new user.

Although the app includes assessments for individuals to self-monitor well-being and symptoms of anxiety, depression, and PTSD, it is difficult to reliably measure change in these constructs via the app, due to the naturalistic nature of this study and the changing landscape of the pandemic over time. It is

important to highlight that even though a score of 33 or higher on the PCL-5 is suggestive of PTSD, the assessment questions in the app do not ask app users to respond to the questions while focusing on a particular traumatic incident, so caution in interpreting the meaning of these scores is warranted. Because the PCL-5 refers to "the stressful experience" in each item, in the context of COVID Coach, the PCL-5 may be capturing overall levels of distress. While desirable, we also did not have a way to measure other potential proxy variables of interest such as coping self-efficacy, perceived helpfulness of the app, improved opinions about mental health care, or reduction in stress related to enhanced support access, as these cannot be determined solely by in-app usage data.

Future research is needed to better understand who is interested in public mental health apps like COVID Coach, what their primary goals are for using the app, which outcomes are most useful in understanding engagement patterns, and how successful usage is defined. For example, someone may use the app only once, find the exact resource they need, and not use the app again, whereas someone else may be experiencing significant stress, use tools in moments of distress, and track mental health symptoms on a weekly basis. Findings from this type of research could be used to advance the science of mobile mental health and also be directly applied to a suite of publicly available apps that have been downloaded over 4 million times and are in widespread use across the VA, the largest health care organization in the United States.

Conclusions

As the mental health impacts of the pandemic continue to be widespread and increasing, digital health resources, such as apps like COVID Coach, are a scalable way to provide evidence-informed tools and resources. We believe that this is the first evaluation of a mobile mental health app designed specifically for use during the COVID-19 pandemic. This work shows that tens of thousands of people are accessing the app, with a particular focus on the tools for stress and coping. Such rapid uptake of a public mobile mental health app is unprecedented and signals perceived value. Specially, the findings from this evaluation suggest that apps may play a helpful role in providing mental health resources in the context of a public health disaster.

Future research should attempt to elucidate for whom and under what conditions the app is most helpful, and how to increase and sustain engagement. Additional areas of focus should include how to optimize the app for populations impacted by disparities related to mental health literacy, digital literacy, and stigma around mental health care. As noted by many mHealth (mobile health) scholars [48-50], there is no reason to believe that digital mental health care and blended options will disappear after the pandemic, so it is important to find strategies for increasing reach and optimizing for engagement within self-management tools. These strategies must also attend to issues of health inequities [48,49,51]. Due to the scale of the crisis, the pandemic may have opened the door to conversations about mental health, and apps may be a helpful first step in providing tools, accurate information, and connecting people with reliable resources. Those in government and nonprofit organizations may be able to provide these kinds of tools as a way to contribute to a digital mental health safety net and help alleviate mental health disparities.

Acknowledgments

This paper was not sponsored by any funder. The views expressed in this work are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.

Authors' Contributions

All authors contributed to the conceptualization, writing, and editing of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

COVID Coach screenshots.
[PDF File (Adobe PDF File), 2624 KB-Multimedia Appendix 1]

References

1. Liu CH, Zhang E, Wong GTF, Hyun S, Hahm HC. Factors associated with depression, anxiety, and PTSD symptomatology during the COVID-19 pandemic: Clinical implications for U.S. young adult mental health. Psychiatry Res 2020 Aug;290:113172 [FREE Full text] [doi: 10.1016/j.psychres.2020.113172] [Medline: 32512357]
2. Rodríguez-Rey R, Garrido-Hernansaiz H, Collado S. Psychological Impact and Associated Factors During the Initial Stage of the Coronavirus (COVID-19) Pandemic Among the General Population in Spain. Front Psychol 2020;11:1540 [FREE Full text] [doi: 10.3389/fpsyg.2020.01540] [Medline: 32655463]
3. Torjesen I. Covid-19: Mental health services must be boosted to deal with "tsunami" of cases after lockdown. BMJ 2020 May 15;369:m1994. [doi: 10.1136/bmj.m1994] [Medline: 32417756]
4. Keeter S. A third of Americans experienced high levels of psychological distress during the coronavirus outbreak. Pew Research Center. 2020 May 7. URL: https://www.pewresearch.org/fact-tank/2020/05/07/

a-third-of-americans-experienced-high-levels-of-psychological-distress-during-the-coronavirus-outbreak/ [accessed 2021-02-24]
5. Stress in America 2020: A national mental health crisis. American Psychological Association. 2020 Oct. URL: https://www.apa.org/news/press/releases/stress/2020/sia-mental-health-crisis.pdf [accessed 2021-02-24]
6. Ettman CK, Abdalla SM, Cohen GH, Sampson L, Vivier PM, Galea S. Prevalence of Depression Symptoms in US Adults Before and During the COVID-19 Pandemic. JAMA Netw Open 2020 Sep 01;3(9):e2019686 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.19686] [Medline: 32876685]
7. Czeisler ME, Lane RI, Petrosky E, Wiley JF, Christensen A, Njai R, et al. Mental Health, Substance Use, and Suicidal Ideation During the COVID-19 Pandemic - United States, June 24-30, 2020. MMWR Morb Mortal Wkly Rep 2020 Aug 14;69(32):1049-1057 [FREE Full text] [doi: 10.15585/mmwr.mm6932a1] [Medline: 32790653]
8. Taquet M, Luciano S, Geddes JR, Harrison PJ. Bidirectional associations between COVID-19 and psychiatric disorder: retrospective cohort studies of 62 354 COVID-19 cases in the USA. The Lancet Psychiatry 2021 Feb 09;8(2):130-140 [FREE Full text] [doi: 10.1016/S2215-0366(20)30462-4] [Medline: 33181098]
9. Balcombe L, De Leo D. An Integrated Blueprint for Digital Mental Health Services Amidst COVID-19. JMIR Ment Health 2020 Jul 22;7(7):e21718 [FREE Full text] [doi: 10.2196/21718] [Medline: 32668402]
10. Alexopoulos AR, Hudson JG, Otenigbagbe O. The Use of Digital Applications and COVID-19. Community Ment Health J 2020 Oct 30;56(7):1202-1203 [FREE Full text] [doi: 10.1007/s10597-020-00689-2] [Medline: 32734311]
11. Rainie L, Zickhur K. Americans' views on mobile etiquette. Pew Research Center. 2015 Jul 26. URL: https://www.pewresearch.org/internet/2015/08/26/americans-views-on-mobile-etiquette/#:~:text=Some%2092%25%20of%20U.S.%20adults,they%20rarely%20turn%20it%20off [accessed 2021-02-24]
12. Mobile fact sheet. Pew Research Center. 2019 Jun 12. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2021-02-24]
13. Kantamneni N. The impact of the COVID-19 pandemic on marginalized populations in the United States: A research agenda. J Vocat Behav 2020 Jun;119:103439 [FREE Full text] [doi: 10.1016/j.jvb.2020.103439] [Medline: 32390658]
14. Double Jeopardy: COVID-19 and Behavioral Health Disparities for Black and Latino Communities in the U.S. Substance Abuse and Mental Health Services Administration (SAHMSA). URL: https://www.samhsa.gov/sites/default/files/covid19-behavioral-health-disparities-black-latino-communities.pdf [accessed 2021-02-24]
15. Dalsania AK, Fastiggi MJ, Kahlam A, Shah R, Patel K, Shiau S, et al. The Relationship Between Social Determinants of Health and Racial Disparities in COVID-19 Mortality. J Racial Ethn Health Disparities 2021 Jan 05 [FREE Full text] [doi: 10.1007/s40615-020-00952-y] [Medline: 33403652]
16. Mohr DC, Schueller SM, Tomasino KN, Kaiser SM, Alam N, Karr C, et al. Comparison of the Effects of Coaching and Receipt of App Recommendations on Depression, Anxiety, and Engagement in the IntelliCare Platform: Factorial Randomized Controlled Trial. J Med Internet Res 2019 Aug 28;21(8):e13609 [FREE Full text] [doi: 10.2196/13609] [Medline: 31464192]
17. Miner A, Kuhn E, Hoffman JE, Owen JE, Ruzek JI, Taylor CB. Feasibility, acceptability, and potential efficacy of the PTSD Coach app: A pilot randomized controlled trial with community trauma survivors. Psychol Trauma 2016 May 25;8(3):384-392. [doi: 10.1037/tra0000092] [Medline: 27046668]
18. Mobile Apps. VA National Center for PTSD. URL: https://www.ptsd.va.gov/appvid/mobile/index.asp [accessed 2021-02-24]
19. Mobile Mental Health. VA National Center for PTSD, Dissemination & Training Division. URL: https://www.myvaapps.com/ [accessed 2021-02-24]
20. Kuhn E, Kanuri N, Hoffman JE, Garvert DW, Ruzek JI, Taylor CB. A randomized controlled trial of a smartphone app for posttraumatic stress disorder symptoms. J Consult Clin Psychol 2017 Mar;85(3):267-273. [doi: 10.1037/ccp0000163] [Medline: 28221061]
21. Meshick AB, Ryan L, Cullen T. Beyond PPE: Protecting Health Care Workers To Prevent A Behavioral Health Disaster. Health Affairs. 2020 Jun 4. URL: https://www.healthaffairs.org/do/10.1377/hblog20200603.842660/full/ [accessed 2021-02-24]
22. Huang L. Applications as Tools Amid the COVID-19 Pandemic. Journal of Hospital Librarianship 2020 Oct 09;20(4):376-390. [doi: 10.1080/15323269.2020.1823304]
23. Ruzek JI, Kuhn E, Jaworski BK, Owen JE, Ramsey KM. Mobile mental health interventions following war and disaster. Mhealth 2016;2:37 [FREE Full text] [doi: 10.21037/mhealth.2016.08.06] [Medline: 28293610]
24. Marshall JM, Dunstan DA, Bartik W. Treating Psychological Trauma in the Midst of COVID-19: The Role of Smartphone Apps. Front Public Health 2020;8:402 [FREE Full text] [doi: 10.3389/fpubh.2020.00402] [Medline: 33014955]
25. Wong AH, Pacella-LaBarbara ML, Ray JM, Ranney ML, Chang BP. Healing the Healer: Protecting Emergency Health Care Workers' Mental Health During COVID-19. Ann Emerg Med 2020 Oct;76(4):379-384 [FREE Full text] [doi: 10.1016/j.annemergmed.2020.04.041] [Medline: 32534830]
26. Baumel A, Muench F, Edan S, Kane JM. Objective User Engagement With Mental Health Apps: Systematic Search and Panel-Based Usage Analysis. J Med Internet Res 2019 Sep 25;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]

27. Owen JE, Jaworski BK, Kuhn E, Makin-Byrd KN, Ramsey KM, Hoffman JE. mHealth in the Wild: Using Novel Data to Examine the Reach, Use, and Impact of PTSD Coach. JMIR Ment Health 2015;2(1):e7 [FREE Full text] [doi: 10.2196/mental.3935] [Medline: 26543913]
28. Grolnick WS, Schonfeld DJ, Schreiber M, Cohen J, Cole V, Jaycox L, et al. Improving adjustment and resilience in children following a disaster: Addressing research challenges. Am Psychol 2018 Apr;73(3):215-229. [doi: 10.1037/amp0000181] [Medline: 29446960]
29. US Department of Veterans Affairs. COVID Coach (Version 1.0). Google Play Store. 2020. URL: https://play.google.com/store/apps/details?id=gov.va.mobilehealth.ncptsd.covid&hl=en_US&gl=US [accessed 2021-02-24]
30. US Department of Veterans Affairs. COVID Coach (Version 1.0). App Store. 2020. URL: https://apps.apple.com/us/app/covid-coach/id1504705038 [accessed 2021-02-24]
31. VA Technical Reference Model. URL: https://www.oit.va.gov/Services/TRM/TRMHomePage.aspx [accessed 2020-11-20]
32. Kwasny MJ, Schueller SM, Lattie E, Gray EL, Mohr DC. Exploring the Use of Multiple Mental Health Apps Within a Platform: Secondary Analysis of the IntelliCare Field Trial. JMIR Ment Health 2019 Mar 21;6(3):e11572 [FREE Full text] [doi: 10.2196/11572] [Medline: 30896433]
33. Tennant R, Hiller L, Fishwick R, Platt S, Joseph S, Weich S, et al. The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. Health Qual Life Outcomes 2007 Nov 27;5:63 [FREE Full text] [doi: 10.1186/1477-7525-5-63] [Medline: 18042300]
34. Stewart-Brown S, Samaraweera PC, Taggart F, Kandala N, Stranges S. Socioeconomic gradients and mental health: implications for public health. Br J Psychiatry 2015 Jun 02;206(6):461-465. [doi: 10.1192/bjp.bp.114.147280] [Medline: 25792696]
35. Stewart-Brown S, Platt S, Tennant A. The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): a valid and reliable tool for measuring mental well-being in diverse populations and projects. J Epidemiol Community Health 2011;65:A38-A39 [FREE Full text] [doi: 10.1136/jech.2011.143586.86]
36. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
37. Löwe B, Decker O, Müller S, Brähler E, Schellberg D, Herzog W, et al. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. Med Care 2008 Mar;46(3):266-274. [doi: 10.1097/MLR.0b013e318160d093] [Medline: 18388841]
38. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
39. Gilbody S, Richards D, Brealey S, Hewitt C. Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. J Gen Intern Med 2007 Nov;22(11):1596-1602 [FREE Full text] [doi: 10.1007/s11606-007-0333-y] [Medline: 17874169]
40. Weathers FW, Litz BT, Keane TM, Palmieri PA, Marx BP, Schnurr PP. The PTSD Checklist for DSM-5 (PCL-5). The National Center for PTSD. 2013. URL: https://www.ptsd.va.gov/professional/assessment/adult-sr/ptsd-checklist.asp [accessed 2021-02-24]
41. Bovin MJ, Marx BP, Weathers FW, Gallagher MW, Rodriguez P, Schnurr PP, et al. Psychometric properties of the PTSD Checklist for Diagnostic and Statistical Manual of Mental Disorders-Fifth Edition (PCL-5) in veterans. Psychol Assess 2016 Nov;28(11):1379-1391. [doi: 10.1037/pas0000254] [Medline: 26653052]
42. Blevins CA, Weathers FW, Davis MT, Witte TK, Domino JL. The Posttraumatic Stress Disorder Checklist for DSM-5 (PCL-5): Development and Initial Psychometric Evaluation. J Trauma Stress 2015 Dec;28(6):489-498. [doi: 10.1002/jts.22059] [Medline: 26606250]
43. Owen JE, Kuhn E, Jaworski BK, McGee-Vincent P, Juhasz K, Hoffman JE, et al. VA mobile apps for PTSD and related problems: public health resources for veterans and those who care for them. Mhealth 2018;4:28 [FREE Full text] [doi: 10.21037/mhealth.2018.05.07] [Medline: 30148141]
44. Rajkumar RP. COVID-19 and mental health: A review of the existing literature. Asian J Psychiatr 2020 Aug 10;52:102066 [FREE Full text] [doi: 10.1016/j.ajp.2020.102066] [Medline: 32302935]
45. Firth J, Torous J, Nicholas J, Carney R, Pratap A, Rosenbaum S, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry 2017 Oct;16(3):287-298 [FREE Full text] [doi: 10.1002/wps.20472] [Medline: 28941113]
46. Kozlov E, Bantum E, Pagano I, Walser R, Ramsey K, Taylor K, et al. The Reach, Use, and Impact of a Free mHealth Mindfulness App in the General Population: Mobile Data Analysis. JMIR Ment Health 2020 Nov 27;7(11):e23377 [FREE Full text] [doi: 10.2196/23377] [Medline: 33245289]
47. Ng MM, Firth J, Minen M, Torous J. User Engagement in Mental Health Apps: A Review of Measurement, Reporting, and Validity. Psychiatr Serv 2019 Jul 01;70(7):538-544 [FREE Full text] [doi: 10.1176/appi.ps.201800519] [Medline: 30914003]
48. Figueroa CA, Aguilera A. The Need for a Mental Health Technology Revolution in the COVID-19 Pandemic. Front Psychiatry 2020 Jun 3;11:523 [FREE Full text] [doi: 10.3389/fpsyt.2020.00523] [Medline: 32581891]

49. Torous J, Jän Myrick K, Rauseo-Ricupero N, Firth J. Digital Mental Health and COVID-19: Using Technology Today to Accelerate the Curve on Access and Quality Tomorrow. JMIR Ment Health 2020 Mar 26;7(3):e18848 [FREE Full text] [doi: 10.2196/18848] [Medline: 32213476]
50. Wind TR, Rijkeboer M, Andersson G, Riper H. The COVID-19 pandemic: The 'black swan' for mental health care and a turning point for e-health. Internet Interv 2020 Apr;20:100317 [FREE Full text] [doi: 10.1016/j.invent.2020.100317] [Medline: 32289019]
51. Smith JA, Judd J. COVID-19: Vulnerability and the power of privilege in a pandemic. Health Promot J Austr 2020 Apr;31(2):158-160 [FREE Full text] [doi: 10.1002/hpja.333] [Medline: 32197274]

Abbreviations

ANOVA: analysis of variance
BIPOC: Black, Indigenous, people of color
GAD-7: Generalized Anxiety Disorder-7
JSON: JavaScript Object Notation
mHealth: mobile health
PCL-5: Posttraumatic Stress Disorder Checklist-5
PHQ-9: Patient Health Questionnaire-9
PTSD: posttraumatic stress disorder
UTC: Coordinated Universal Time
VA: Veterans Affairs
WEMWBS: Warwick-Edinburgh Mental Well-Being Scale

Edited by C Basch; submitted 16.12.20; peer-reviewed by J Ruzek, J Ray; comments to author 04.01.21; revised version received 14.01.21; accepted 17.02.21; published 01.03.21

Please cite as:
Jaworski BK, Taylor K, Ramsey KM, Heinz A, Steinmetz S, Pagano I, Moraja G, Owen JE
Exploring Usage of COVID Coach, a Public Mental Health App Designed for the COVID-19 Pandemic: Evaluation of Analytics Data
J Med Internet Res 2021;23(3):e26559
URL: https://www.jmir.org/2021/3/e26559
doi: 10.2196/26559
PMID: 33606656

Original Paper

Cost-effectiveness of a Telemonitoring Program for Patients With Heart Failure During the COVID-19 Pandemic in Hong Kong: Model Development and Data Analysis

Xinchan Jiang, BSc, MPhil; Jiaqi Yao, BSc, MPhil; Joyce Hoi-Sze You, PharmD

School of Pharmacy, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

Corresponding Author:
Joyce Hoi-Sze You, PharmD
School of Pharmacy
Faculty of Medicine
The Chinese University of Hong Kong
8th Floor, Lo Kwee-Seong Integrated Biomedical Sciences Building
Shatin, NT
Hong Kong
China (Hong Kong)
Phone: 852 39436830
Email: joyceyou@cuhk.edu.hk

Abstract

Background: The COVID-19 pandemic has caused patients to avoid seeking medical care. Provision of telemonitoring programs in addition to usual care has demonstrated improved effectiveness in managing patients with heart failure (HF).

Objective: We aimed to examine the potential clinical and health economic outcomes of a telemonitoring program for management of patients with HF during the COVID-19 pandemic from the perspective of health care providers in Hong Kong.

Methods: A Markov model was designed to compare the outcomes of a care under COVID-19 (CUC) group and a telemonitoring plus CUC group (telemonitoring group) in a hypothetical cohort of older patients with HF in Hong Kong. The model outcome measures were direct medical cost, quality-adjusted life-years (QALYs), and incremental cost-effectiveness ratio. Sensitivity analyses were performed to examine the model assumptions and the robustness of the base-case results.

Results: In the base-case analysis, the telemonitoring group showed a higher QALY gain (1.9007) at a higher cost (US $15,888) compared to the CUC group (1.8345 QALYs at US $15,603). Adopting US $48,937/QALY (1 × the gross domestic product per capita of Hong Kong) as the willingness-to-pay threshold, telemonitoring was accepted as a highly cost-effective strategy, with an incremental cost-effective ratio of US $4292/QALY. No threshold value was identified in the deterministic sensitivity analysis. In the probabilistic sensitivity analysis, telemonitoring was accepted as cost-effective in 99.22% of 10,000 Monte Carlo simulations.

Conclusions: Compared to the current outpatient care alone under the COVID-19 pandemic, the addition of telemonitoring-mediated management to the current care for patients with HF appears to be a highly cost-effective strategy from the perspective of health care providers in Hong Kong.

(J Med Internet Res 2021;23(3):e26516) doi: 10.2196/26516

KEYWORDS

telemonitoring; mobile health; smartphone; heart failure; COVID-19; health care avoidance; cost-effectiveness

Introduction

Heart failure (HF) is a chronic disease affecting 38 million patients worldwide, with high in-hospital mortality (6.4%), 1-year readmission rate (24%-30%), and 1-year postdischarge mortality (20%) [1-5]. This chronic cardiac disease imposes a substantial global economic burden of US $108 billion per annum (approximated in 2012) [6], which is expected to increase considerably with the aging of the population [7]. Hong Kong is a developed city with an aging population, and the local epidemiological findings on outcomes of patients with HF were consistent with those of western countries [8,9].

The COVID-19 pandemic has imposed major burdens and barriers on the operation of health care systems worldwide. COVID-19 has not only disrupted the provision of routine

medical care but has also caused patients to delay and avoid seeking medical care [10]. COVID-19 was reported to be a factor associated with avoiding medical consultation in Hong Kong [11]. Patients with chronic conditions such as HF are therefore at risk of suboptimal care during the COVID-19 pandemic as a result of disruption or avoidance of routine medical care. The treatment outcomes of HF under current care during the COVID-19 pandemic are expected to be compromised.

Telehealth is a potential timely alternative to minimize the risk of COVID-19 transmission by reducing direct physical contact and to sustain continuous medical care to patients with HF during the COVID-19 pandemic [12]. The benefits of telemonitoring programs have been examined in clinical studies for the management of patients with HF. A meta-analysis reported that the application of telemonitoring program was associated with reduced risk of all-cause mortality and HF-related mortality [13].

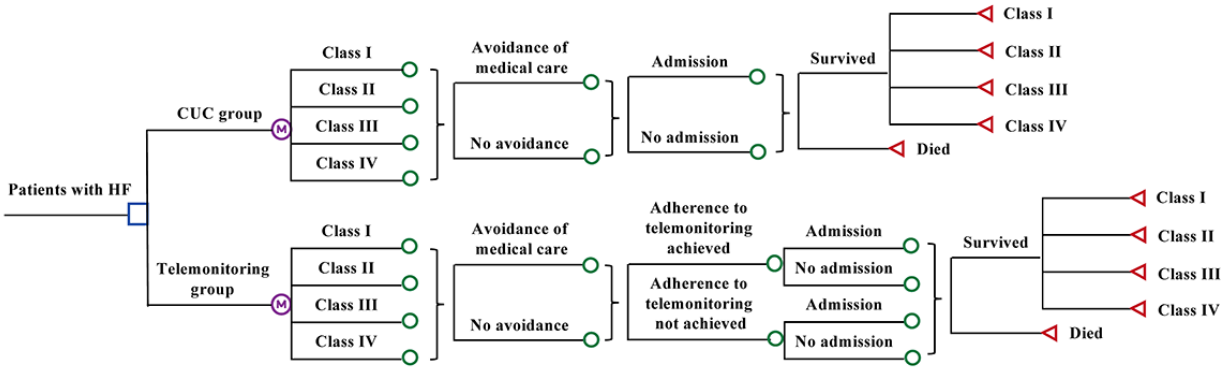
The Markov model is a well-established decision-analytic model for simulation of expected treatment costs and health-related outcomes by incorporating relevant clinical probabilities, costs, and utility inputs. In a Markov model, hypothetical subjects proceed through health states (Markov states) in the next model cycle according to transition probabilities. Markov modeling is recommended for evaluating the outcomes of diseases that might progress, improve, or relapse through transition over a series of health states [14]. The cost-effective application of telemonitoring for the management of HF was demonstrated by Markov model–based analyses prior to the era of COVID-19 [15,16], and the patients' medical avoidance was therefore not evaluated as an influential factor. In this study, COVID-related medical avoidance was considered in the model-based analysis. The aim of our study was to examine the potential clinical and health economic outcomes of adding telemonitoring programs to current medical care during the COVID-19 pandemic for the management of patients with HF from the perspective of health care providers in Hong Kong.

Methods

Model Design

A Markov decision-analytic model was designed to estimate the potential outcomes of current care under COVID-19 (CUC) with and without telemonitoring in a hypothetical cohort of older patients with HF (age 65 years or above) in Hong Kong (Figure 1). The outcomes were simulated from the entry of the model for a time frame of 10 years or until death, whichever occurred first. The two strategies examined in this study were (1) CUC plus telemonitoring (telemonitoring group) and (2) CUC alone (CUC group). The hypothetical cohort entered the model at one of the New York Heart Association (NYHA) classes I-IV and proceeded to another health status by the corresponding probability in each monthly cycle. The model outcome measures were direct medical cost, quality-adjusted life-years (QALYs), and incremental cost-effectiveness ratio (ICER).

Figure 1. Simplified Markov model of telemonitoring for patients with HF. CUC: care under COVID-19; HF: heart failure.



Multidisciplinary care is the standard management approach in usual care for patients with HF in Hong Kong, as recommended by the American College of Cardiology Foundation/American Health Association Guideline for the Management of Heart Failure [17]. Patients in the CUC and telemonitoring groups therefore all received multidisciplinary care, while patients in telemonitoring group received telemonitoring-mediated HF management in addition to multidisciplinary care. The telemonitoring-mediated management approach evaluated in a clinical outcome study was adopted in this model [18]. The patients in the telemonitoring group transmitted cardiac measures (heart rate, blood pressure, and weight) daily to the HF management team and answered a short series of questions pertinent to their HF symptoms via an app downloaded to a smartphone. A clinically validated algorithm that was embedded in the app stratified patients into different states and further identified patients with urgent needs. The patients with urgent needs would receive an alert message and an automated call suggesting emergent services. The on-call clinician would also be alerted to provide timely intervention at the onset of symptom exacerbations. Patients who were classified as nonurgent cases would receive self-instruction on administration of medications and when to contact a care provider.

Because of patients' concerns about the risk of acquiring COVID-19 at health care facilities during the pandemic, patients in both arms might or might not have avoided attending the in-person medical care clinic. The telemonitoring-mediated care also required daily transmission of cardiac measures via a smartphone app, and patients in the telemonitoring group might or might not have achieved adherence to the telemonitoring requirements. Patients in both arms might have experienced

HF-related hospitalization. For the patients who survived (with or without hospitalization) in each cycle, they might have remained in the same NYHA classification or improved/progressed to another NYHA classification.

Model Inputs

All the model inputs are shown in Table 1. The clinical inputs were retrieved from published reports written in English, identified from a literature search on MEDLINE over the period of 2000-2020. Epidemiology or disease burden studies in the Chinese population, randomized clinical trials, and meta-analyses were the preferred sources for clinical model inputs.

Table 1. Model parameters.

Parameters	Base case value	Range of sensitivity analysis	Distribution	Reference
Clinical inputs				
Proportion of NYHA[a] classification (%)			Dirichlet	[19]
Class I	9	8.1-9.9		
Class II	44	39.6-48.4		
Class III	34	30.6-37.4		
Class IV	13	8.6-17.4		
Transition probability (monthly)			Dirichlet	[20]
I to I	0.9597	0.9538-0.9678		
I to II	0.0394	0.0315-0.0473		
I to III	0.0009	0.0007-0.0011		
I to IV	0	0-0.0011		
II to I	0.0073	0.0058-0.0088		
II to II	0.9877	0.9852-0.9902		
II to III	0.0039	0.0031-0.0047		
II to IV	0.0011	0.0009-0.0013		
III to I	0.001	0.0008-0.0012		
III to II	0.0443	0.0354-0.0532		
III to III	0.8843	0.8612-0.9074		
III to IV	0.0704	0.0563-0.0845		
VI to I	0.0010	0.0008-0.0012		
VI to II	0.0443	0.0354-0.0532		
VI to III	0.8515	0.8612-0.9074		
VI to IV	0.1032	0.0563-0.0845		
Probability of HF[b]-related hospitalization in multidisciplinary care (monthly)	0.0296	0.0237-0.15	Beta	[9]
Probability of all-cause mortality in multidisciplinary care (monthly)	0.0279	0.0076-0.0383	Beta	[9]
Risk ratio of event with versus without multidisciplinary care				
HF-related hospitalization	0.74	0.64-0.87	Lognormal	[21]
All-cause mortality	0.75	0.59-0.96	Lognormal	[21]
Risk ratio of event with versus without telemonitoring				
HF-related hospitalization	0.5	0.36-0.64	Lognormal	[18]
All-cause mortality	0.81	0.70-0.94	Lognormal	[13]
Adherence to telemonitoring-guided management (%)	80	64-96	Triangular	[22]
COVID 19–related health care avoidance (%)	26.1	21-31.5	Triangular	[11]
Duration of COVID 19–related health care avoidance (years)	1.5	0.5-2	Triangular	[23]
Utility inputs				
Utilities			Uniform	[24]
NYHA class I	0.82	0.78-0.85		
NYHA class II	0.74	0.69-0.75		

Parameters	Base case value	Range of sensitivity analysis	Distribution	Reference
NYHA class III	0.64	0.55-0.77		
NYHA class IV	0.46	0.41-0.61		
Disutilities of hospitalization			Uniform	[24]
NYHA class I	0.04	0.03-0.05		
NYHA class II	0.07	0.06-0.08		
NYHA class III	0.10	0.08-0.12		
NYHA class IV	0.29	0.23-0.35		
Cost inputs				
Daily cost of hospitalization (US $)	654	523-785	Gamma	[25]
Length of hospitalization for HF (days)	8	6-10	Triangular	[26]
Monthly outpatient cost for HF (US $)	197	158-236	Gamma	[27]
Telemonitoring-mediated care (US $)				
Site implementation cost per patient	80	64-96	Gamma	[16]
Monthly cost of telemonitoring	50	40-60	Gamma	[16]

[a]NYHA: New York Heart Association.

[b]HF: heart failure.

At the entry of the model, the distribution of patients among the four statuses (NYHA class I: 9%, NYHA class II: 44%, NYHA class III: 34%, and NYHA class IV: 13%) adopted the baseline characteristics of patients with HF in Northeast Asia [19]. The yearly transition rates between NYHA classes were retrieved from the Eplerenone in Mild Patients Hospitalization And Survival Study in Heart Failure [20], and MATLAB (MathWorks) was used to generate the monthly transition matrix. HF-related hospitalization (2.96%) and all-cause mortality for patients aged ≥65 years (2.79%) with multidisciplinary care were approximated from the Hong Kong Heart Failure Registry. In this study, a total of 1940 new-onset HF cases were identified in the Hong Kong Chinese population between 2005 and 2012. Both of the above estimates were retrieved from patients followed in the outpatient setting, with a prior history of hospitalization for decompensated HF [9]. The clinical impacts of multidisciplinary care (vs without multidisciplinary care) on HF-related admission (risk ratio [RR] 0.74; 95% CI 0.63-0.87) and all-cause mortality (RR 0.75; 95% CI 0.59-0.96) were retrieved from a systematic review of 29 trials (5039 patients) on multidisciplinary strategies for management of patients with HF [21]. The probabilities of HF-related hospitalization and all-cause mortality in patients who avoided medical care during the COVID-19 pandemic were approximated using the risks of events without multidisciplinary care. The relative change of hospitalization rate associated with telemonitoring-medicated care (RR 0.5, 95% CI 0.36-0.64) was obtained from an outcome study of a smartphone-based telemonitoring system in 315 patients with HF [18]. The relative impact of telemonitoring on all-cause mortality (RR 0.81, 95% CI 0.70-0.94) was estimated from a meta-analysis of 37 trials that evaluated the comparative effectiveness of telemonitoring versus no telemonitoring for HF management [13]. The adherence of telemonitoring was defined as achieving 70% of scheduled daily data transmission and HF symptom reporting. The percentage of achieved adherence was assumed to be 80% based on a study investigating the patient adherence of a smartphone-based telemonitoring system for HF [22]. The percentage of medical avoidance among patients with HF (26.1%) was approximated from a public survey of 765 subjects on the use of health services during the COVID-19 pandemic in Hong Kong [11]. The base-case value of health care avoidance duration was estimated to be 1.5 years with a range of 0.5-2 years, based upon the epidemiologic projections of the COVID-19 pandemic [23].

Both the utility scores of the NYHA classes and disutilities due to hospitalization were retrieved from the predicted utilities of patients with HF in the Systolic Heart Failure Treatment with the I_f Inhibitor Ivabradine Trial (n=5313) [24]. The expected QALY gain in each group was calculated by the time spent in the health statuses and the corresponding utility scores. The QALY gain was discounted at an annual rate of 3%.

The cost analysis in this model was conducted using direct medical costs in the year 2020 from the perspective of public health care providers in Hong Kong. The costs of telemonitoring-medicated care (in the telemonitoring group) and the costs of HF-related inpatient and outpatient care (in both groups) were included. The cost of HF-related hospitalization was estimated by the daily cost of inpatient care and the length of stay of the patients. The daily cost of inpatient care was approximated from the fees and charges of public hospital services provided by the Hospital Authority in Hong Kong [25]. The length of hospital stay was estimated from a review on the burden of HF in 9 countries or regions (including Hong Kong) in Asia [26]. The monthly outpatient cost was estimated from the findings of a retrospective observational study on the total management cost (including hospitalization cost and ambulatory care cost) of patients with HF (n=73) recruited from a public hospital in Hong Kong [27]. The implementation cost of telemonitoring per capita (US $80) and

monthly cost of telemonitoring (US $50) were approximated from the reported costs of a smartphone-based telemonitoring system [16], including a smartphone, blood pressure monitor, weight scale, and licensing fee. The implementation cost was a one-time charge, while the monthly cost of telemonitoring was a recurrent cost for maintenance of the app. Hong Kong is a developed city with a high smartphone penetration rate of 85.5% in the overall population [28]. In this study, the monthly cost of telemonitoring was estimated at US $50 (US $1=HK $7.8), assuming the patients used their smartphones and installed the telemonitoring app. All costs were discounted annually by 3%.

Cost-effectiveness Analysis and Sensitivity Analysis

Expected costs and QALY gains were simulated for the two strategies in the base-case analysis. The ICERs were calculated using the equation (total $cost_{telemonitoring\ group}$–total $cost_{CUC\ group}$)/($QALY_{telemonitoring\ group}$–$QALY_{CUC\ group}$). As recommended by the World Health Organization in 2002, an ICER less than 1 × the gross domestic product per capita was considered to be highly cost‐effective [29]. The gross domestic product per capita of Hong Kong was US $48,937 in 2019 and was adopted as the willingness‐to‐pay (WTP) threshold [30]. A treatment alternative was preferred if (1) it was effective in saving QALYs at lower cost or (2) it was effective in saving QALYs at a higher cost with an acceptable ICER (< the WTP threshold).

Deterministic and probabilistic sensitivity analyses using Monte Carlo simulations were performed to examine the robustness of the base-case results. In the deterministic sensitivity analysis, each model input was evaluated over the range reported in the retrieved studies. If no range was specified, the parameter was examined over a range of ±20% of the base-case value. In the probabilistic analysis, 10,000 Monte Carlo simulations of each model outcome measure were generated by randomly drawing the value of all model inputs simultaneously from the distribution specified in Table 1. The probabilities of each strategy to be accepted as cost-effective in the 10,000 Monte Carlo simulations were determined against the variation of the WTP threshold (from US $0-100,000/QALY) in the acceptability curve. All analyses were performed using TreeAge Pro 2020 (TreeAge Software, Inc).

Results

Changes of Outcomes With Versus Without COVID-19–Related Health Care Avoidance

Over a time frame of 1.5 years (base-case value of health care avoidance duration), the expected direct medical cost and QALYs of the CUC group (with COVID-19–related health care avoidance) were US $7114 and 0.7960 QALYs, respectively. The expected cost and QALYs of usual care (without COVID-19–related health care avoidance) over a period of 1.5 years were US $6888 and 0.8135 QALYs, correspondingly. Compared with usual care (without COVID-19–related health care avoidance), CUC (with COVID-19–related health care avoidance) increased the cost by US $226 with a loss of 0.0175 QALYs.

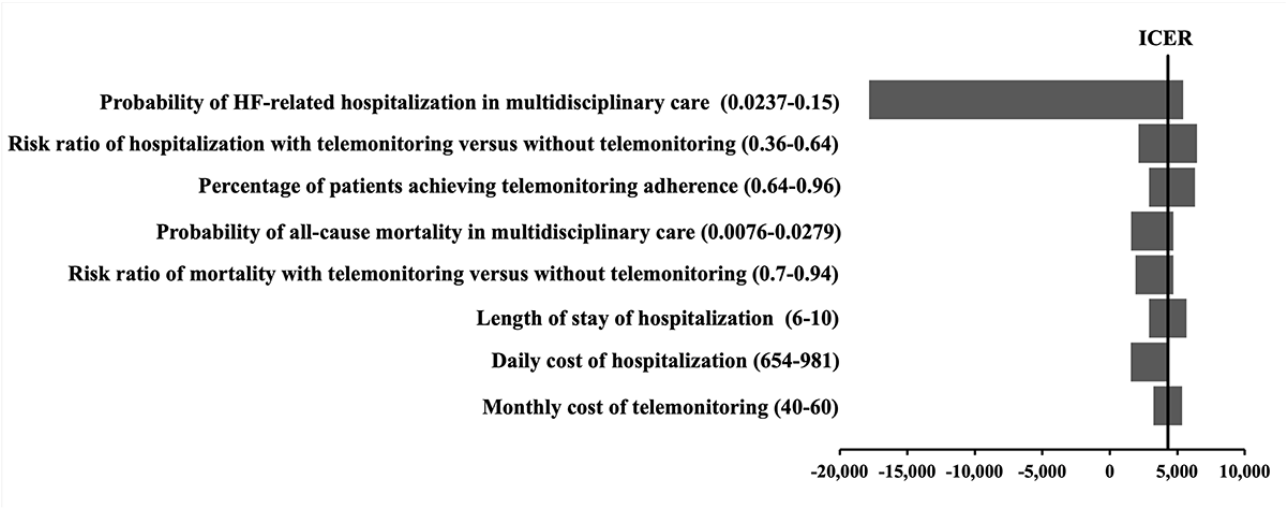
Base-Case Analysis

The expected QALY gains and total costs of the telemonitoring group and the CUC group were compared. The direct medical cost for the CUC group was US $15,603 and the QALYs were 1.8345, while these values for the telemonitoring group were US $15,888 and 1.9007, respectively. The incremental QALYs saved by the telemonitoring group (versus the CUC group) were 0.0662, with an additional cost of US $284. The ICER for the telemonitoring group versus the CUC group was US $4292/QALY, which is below the WTP threshold of 48,937 USD/QALY (1× gross domestic product per capita in Hong Kong). Telemonitoring was therefore a highly cost-effective strategy in the base-case analysis.

Sensitivity Analyses

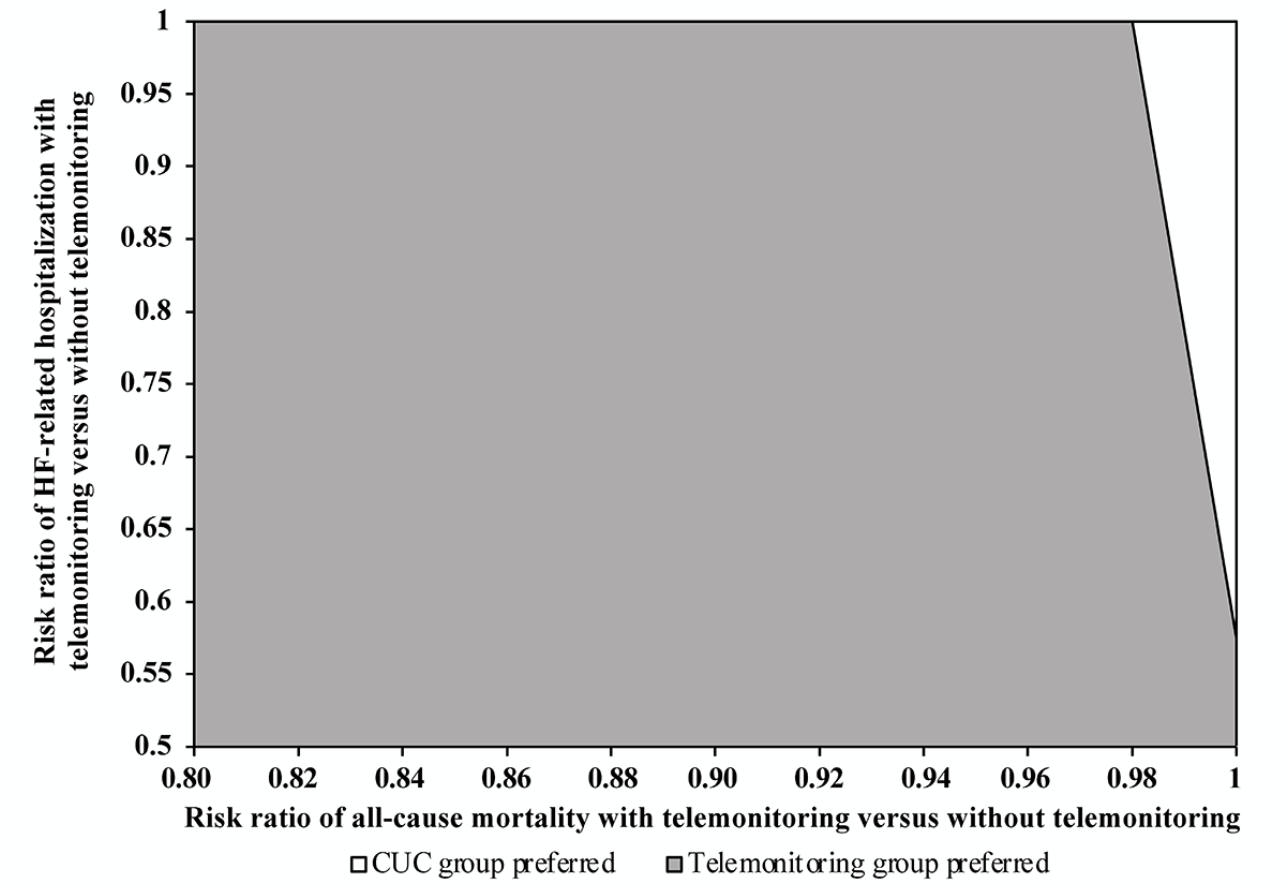
One-way deterministic sensitivity analyses were conducted for all model inputs. The ICERs of the telemonitoring group remained below the WTP threshold in the one-way variation of all parameters. No influential factor with the threshold value was found. For eight critical parameters, the ICERs varied by more than 20% (Figure 2): probability of HF-related hospitalization in multidisciplinary care, risk ratio of hospitalization with telemonitoring versus without telemonitoring, percentage of patients achieving telemonitoring adherence, probability of all-cause mortality in multidisciplinary care, risk ratio of mortality with telemonitoring versus without telemonitoring, length of stay of hospitalization, daily cost of hospitalization, and monthly cost of telemonitoring. Of these eight critical parameters, the probability of HF-related hospitalization in multidisciplinary care had the highest impact on the total cost. When the monthly probability of HF-related hospitalization in multidisciplinary care increased from the base-case value of 0.0296 to >0.0515, the telemonitoring group gained higher QALYs at a lower cost than the CUC group.

Figure 2. One-way sensitivity analysis of the ICER of the telemonitoring group versus the CUC group. CUC: care under COVID-19; ICER: incremental cost-effectiveness ratio.



The risk ratios of telemonitoring versus no telemonitoring for HF-related hospitalization and all-cause mortality were two parameters representing the relative effectiveness of the telemonitoring-mediated care. To further investigate the interaction of these two parameters with the cost-effective acceptance of telemonitoring, a two-way deterministic sensitivity analysis was conducted with the risk ratios of telemonitoring versus without telemonitoring for HF-related hospitalization (range 0.5-1) and all-cause mortality (range: 0.81-1). The gray area in Figure 3 indicates the combinations of these two variables for telemonitoring to be acceptable as the preferred option (higher QALY gained at lower cost or at higher cost with an ICER< the WTP threshold).

Figure 3. Two-way variation of the risk ratios with telemonitoring versus without telemonitoring on HF-related hospitalization and all-cause mortality. CUC: care under COVID-19; HF: heart failure.



The incremental costs versus incremental QALYs gained by telemonitoring (when compared with the CUC group) in 10,000 Monte Carlo simulations are shown in a scatter plot in Figure 4. The telemonitoring group gained an average QALY of 0.0688 (95% CI 0.0681-0.0695, *P*<.001), with a mean additional cost of US $319 (95% CI US $306-US $333, *P*<.001). In 10,000 Monte Carlo simulations, the probability of the telemonitoring group to be more effective in QALY gain and cost-saving was 23.5%. The telemonitoring group gained a higher QALY at a higher cost, with ICER<WTP (US $48,937/QALY) 75.7% of the time.

The probabilities of each strategy to be accepted as cost-effective are shown in the acceptability curve over a wide WTP range of US $0-100,000/QALY (Figure 5). The probabilities of the telemonitoring and CUC groups were the same (50%) at a WTP threshold of US $4700/QALY. The telemonitoring group was accepted to be cost-effective 99.2% of the time at the WTP threshold of US $48,937/QALY.

Figure 4. Scatter plot of the incremental cost-effectiveness ratios for the telemonitoring group versus the care under COVID-19 group. QALY: quality-adjusted life-year; WTP: willingness-to-pay.

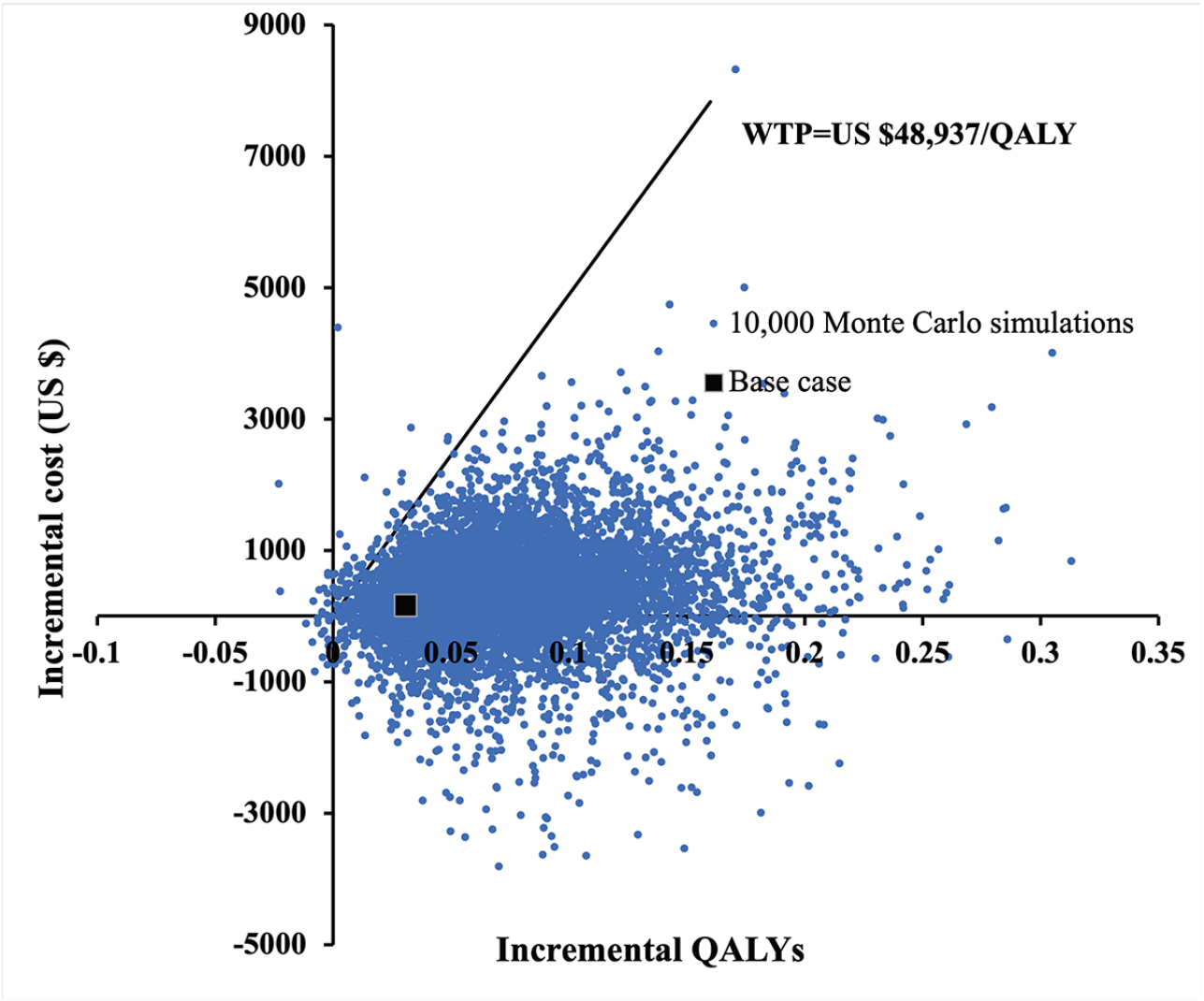
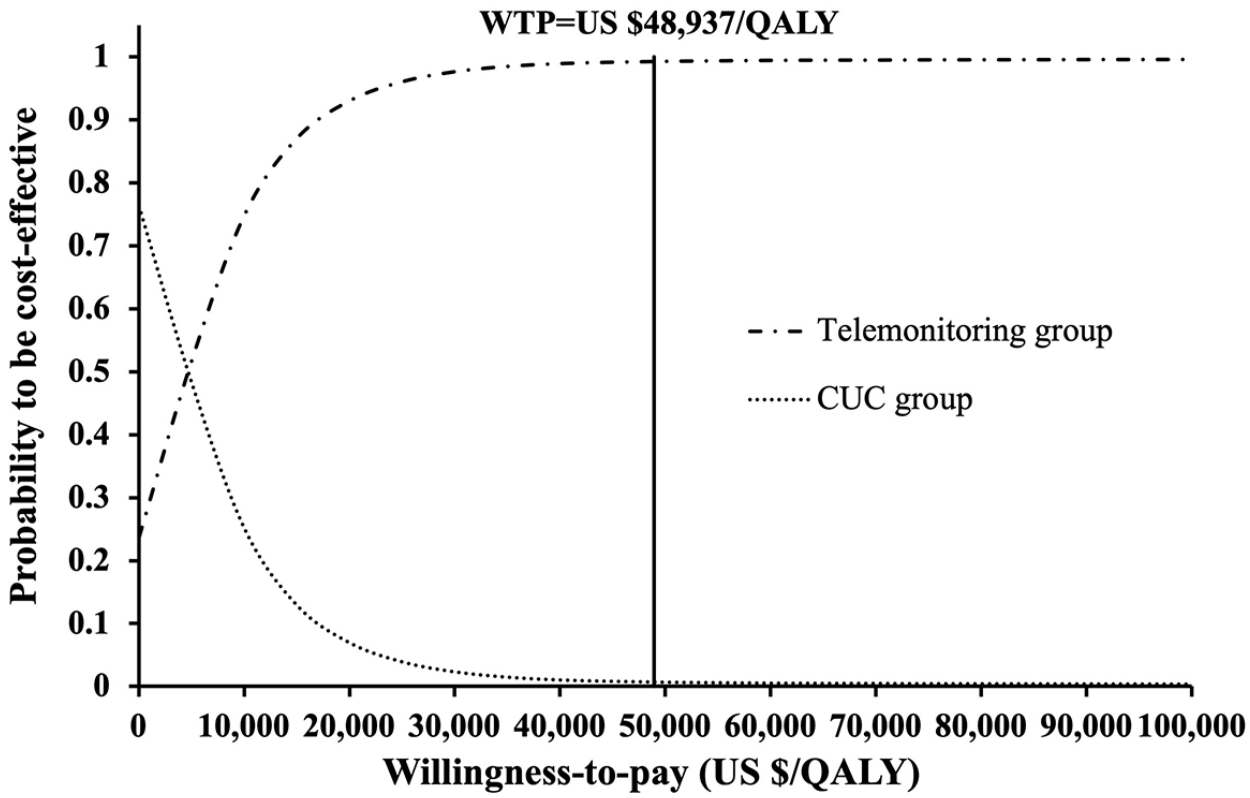


Figure 5. Cost-effectiveness acceptability curve for each strategy to be the preferred strategy against the WTP threshold. CUC: care under COVID-19; QALY: quality-adjusted life year; WTP: willingness-to-pay.



Discussion

Principal Results

This is the first analysis of the potential cost-effectiveness of smartphone-based telemonitoring systems for HF management during the COVID-19 pandemic. Our model results indicated that adding telemonitoring to current CUC for the management of patients with HF is a cost-effective strategy in the base-case analysis, with an ICER (US $4292/QALY) 10-fold below the WTP threshold (US $48,937/QALY). One-way sensitivity analysis supported the robustness of the base-case findings in that no influential parameter with a threshold value was identified. The high probability of the telemonitoring group to be accepted as the preferred strategy throughout a wide WTP range in the probabilistic sensitivity analysis further supported that adding telemonitoring to HF management is a highly cost-effective strategy.

The implementation cost is a modifiable factor when introducing a new technology in a health care system. In this study, telemonitoring was assumed to have a monthly cost of US $50 based on the estimated cost of a currently available smartphone-based telemonitoring system in Canada [16,18]. We further examined the impact of the monthly cost of the telemonitoring system in an extended one-way sensitivity analysis, and we found that telemonitoring-mediated care remained highly cost-effective if the monthly cost of telemonitoring was below US $467. Our findings were consistent with a cost-utility study of a telemonitoring-mediated HF care system in Canada in that the telemonitoring strategy was highly acceptable to be cost-effective, with an ICER of US $6701/QALY (WTP threshold=US $37,718/QALY) [16]. Our study further evaluated the interacting impact of two key parameters (risk ratios of events with telemonitoring vs without telemonitoring), which represented the relative effectiveness of telemonitoring in lowering HF-related hospitalization and all-cause mortality, on the cost-effective acceptance of the telemonitoring strategy. The combinations of these two parameters, as indicated in the two-way sensitivity analysis (Figure 3), provided the effectiveness thresholds required for the telemonitoring program to be accepted as cost-effective.

Health care systems in many countries worldwide are facing unprecedented challenges to maintaining routine medical care. This is particularly difficult when the target patients are older people with chronic cardiac diseases, who also belong to the high-risk group for life-threatening complications if they acquire COVID-19. In Hong Kong, the public health care system has struggled to provide care to patients with COVID-19 and protection against the disease to staff and other patients. Under these circumstances, public health care providers deferred some nonurgent care, and older patients also avoided attending their scheduled routine care appointments. As a result of fewer in-person clinic follow-ups, the risks of unplanned HF-related hospitalization and subsequently mortality inevitably increased.

The benefits of providing telemonitoring programs for HF management were recognized long before the COVID-19 pandemic. The pandemic has highlighted the urgency of adding telemonitoring-mediated care to in-person routine care for patients with HF [31]. Hong Kong is a developed city with a

high smartphone penetration rate [28]. An effective smartphone-based telemonitoring system with a clinician-approved algorithm is a feasible and practical option for patients with HF in Hong Kong. In light of social distancing measures in the landscape of the COVID-19 pandemic, the acceptance of applying telemonitoring-mediated care is expected to highly increase at the levels of policy decision-makers, health care providers, and patients. The COVID-19 pandemic will surely catalyze the application of telemonitoring-mediated health care services in the very near future. Cost-effectiveness evaluation of telemonitoring-based medical care is therefore highly warranted to assist policy makers in the decision-making process of resource allocation.

Limitations

There are limitations to this analysis. The cohort-based Markov model simplified real-life HF events with a limited number of health states. Other factors can impact the cost-effectiveness of HF management. For instance, influenza infection is associated with increased morbidity and mortality of patients with HF [32], and the influenza infection rate has dramatically decreased since the COVID-19 outbreak in Hong Kong [33]. Further evaluation of the impact of reduced influenza infections on HF outcome measures is highly warranted. The impact of telemonitoring on HF hospitalization and all-cause mortality varied among different types of telemonitoring, as indicated by the findings of a comprehensive network meta-analysis [13]. The cost-effectiveness of telemonitoring may therefore vary subject to the specific type of telemonitoring. Some model inputs were retrieved from overseas trials, which may affect the applicability of the model results for patients with HF in Hong Kong. Vigorous sensitivity analysis was therefore conducted on all model inputs over a broad range. The base-case results were found to be robust over the variation of all model inputs in both the deterministic and probabilistic sensitivity analyses. Additionally, the adherence of telemonitoring is not a parameter ready to be transferred between different health care systems. Health care practitioners should therefore examine the adherence of local patients when implementing a telemonitoring program for patients with HF.

Conclusion

Compared to the current CUC during the pandemic alone, the addition of telemonitoring-mediated management to current care for patients with HF appears to be a highly cost-effective strategy from the perspective of health care providers in Hong Kong. Our findings provide evidence to inform decision makers on the application of telemonitoring amid the COVID-19 pandemic. Telemonitoring has long been considered as a future model of care, and the COVID-19 pandemic has fast-forwarded the application timeline of telemonitoring in clinical settings worldwide. It is expected that a mixed mode of disease management with in-person and telemonitoring-mediated care is likely to be sustained beyond the pandemic era. Further cost-effectiveness evaluations of mixed modes of care for the management of high-burden chronic diseases, such as diabetes mellitus, are highly warranted.

Conflicts of Interest

None declared.

References

1. Braunwald E. The war against heart failure: the Lancet lecture. Lancet 2015 Feb 28;385(9970):812-824. [doi: 10.1016/S0140-6736(14)61889-4] [Medline: 25467564]
2. Lloyd-Jones DM, Larson MG, Leip EP, Beiser A, D'Agostino RB, Kannel WB, Framingham Heart Study. Lifetime risk for developing congestive heart failure: the Framingham Heart Study. Circulation 2002 Dec 10;106(24):3068-3072. [doi: 10.1161/01.cir.0000039105.49749.6f] [Medline: 12473553]
3. Nieminen MS, Brutsaert D, Dickstein K, Drexler H, Follath F, Harjola V, EuroHeart Survey Investigators, Heart Failure Association, European Society of Cardiology. EuroHeart Failure Survey II (EHFS II): a survey on hospitalized acute heart failure patients: description of population. Eur Heart J 2006 Nov;27(22):2725-2736. [doi: 10.1093/eurheartj/ehl193] [Medline: 17000631]
4. Cheng RK, Cox M, Neely ML, Heidenreich PA, Bhatt DL, Eapen ZJ, et al. Outcomes in patients with heart failure with preserved, borderline, and reduced ejection fraction in the Medicare population. Am Heart J 2014 Nov;168(5):721-730. [doi: 10.1016/j.ahj.2014.07.008] [Medline: 25440801]
5. Tromp J, Bamadhaj S, Cleland JGF, Angermann CE, Dahlstrom U, Ouwerkerk W, et al. Post-discharge prognosis of patients admitted to hospital for heart failure by world region, and national level of income and income disparity (REPORT-HF): a cohort study. Lancet Glob Health 2020 Mar;8(3):e411-e422 [FREE Full text] [doi: 10.1016/S2214-109X(20)30004-8] [Medline: 32710857]
6. Cook C, Cole G, Asaria P, Jabbour R, Francis DP. The annual global economic burden of heart failure. Int J Cardiol 2014 Feb 15;171(3):368-376. [doi: 10.1016/j.ijcard.2013.12.028] [Medline: 24398230]
7. Heidenreich PA, Albert NM, Allen LA, Bluemke DA, Butler J, Fonarow GC, American Heart Association Advocacy Coordinating Committee, Council on Arteriosclerosis, Thrombosis and Vascular Biology, Council on Cardiovascular Radiology and Intervention, Council on Clinical Cardiology, Council on Epidemiology and Prevention, Stroke Council. Forecasting the impact of heart failure in the United States: a policy statement from the American Heart Association. Circ Heart Fail 2013 May;6(3):606-619 [FREE Full text] [doi: 10.1161/HHF.0b013e318291329a] [Medline: 23616602]

8. Kelvin W, Yeung M. Population ageing trend of Hong Kong. Office of the Government Economist, The Government of the Hong Kong Special Administrative Region. 2019 Jan. URL: https://www.hkeconomy.gov.hk/en/pdf/el/el-2019-02.pdf [accessed 2020-11-24]
9. Hai J, Chan P, Huang D, Ho M, Ho C, Cheung E, et al. Clinical characteristics, management, and outcomes of hospitalized heart failure in a Chinese population-The Hong Kong Heart Failure Registry. J Card Fail 2016 Aug;22(8):600-608. [doi: 10.1016/j.cardfail.2016.03.007] [Medline: 27002944]
10. Czeisler M, Marynak K, Clarke KEN, Salah Z, Shakya I, Thierry JM, et al. Delay or avoidance of medical care because of COVID-19-related concerns - United States, June 2020. MMWR Morb Mortal Wkly Rep 2020 Sep 11;69(36):1250-1257 [FREE Full text] [doi: 10.15585/mmwr.mm6936a4] [Medline: 32915166]
11. Hung KK, Walline JH, Chan EYY, Huang Z, Lo ESK, Yeoh EK, et al. Health service utilization in Hong Kong during the COVID-19 pandemic - a cross-sectional public survey. Int J Health Policy Manag. Online ahead of print 2020 Oct 19. [doi: 10.34172/ijhpm.2020.183] [Medline: 33105965]
12. Monaghesh E, Hajizadeh A. The role of telehealth during COVID-19 outbreak: a systematic review based on current evidence. BMC Public Health 2020 Aug 01;20(1):1193 [FREE Full text] [doi: 10.1186/s12889-020-09301-4] [Medline: 32738884]
13. Yun JE, Park J, Park H, Lee H, Park D. Comparative effectiveness of telemonitoring versus usual care for heart failure: a systematic review and meta-analysis. J Card Fail 2018 Jan;24(1):19-28. [doi: 10.1016/j.cardfail.2017.09.006] [Medline: 28939459]
14. Caro JJ, Briggs AH, Siebert U, Kuntz KM. Modeling good research practices—overview. Med Decis Making 2012 Sep 18;32(5):667-677. [doi: 10.1177/0272989x12454577]
15. Thokala P, Baalbaki H, Brennan A, Pandor A, Stevens JW, Gomersall T, et al. Telemonitoring after discharge from hospital with heart failure: cost-effectiveness modelling of alternative service designs. BMJ Open 2013 Sep 18;3(9):e003250 [FREE Full text] [doi: 10.1136/bmjopen-2013-003250] [Medline: 24048626]
16. Boodoo C, Zhang Q, Ross HJ, Alba AC, Laporte A, Seto E. Evaluation of a heart failure telemonitoring program through a microsimulation model: cost-utility analysis. J Med Internet Res 2020 Oct 06;22(10):e18917 [FREE Full text] [doi: 10.2196/18917] [Medline: 33021485]
17. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE, Colvin MM, et al. 2017 ACC/AHA/HFSA focused update of the 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Failure Society of America. J Card Fail 2017 Aug;23(8):628-651. [doi: 10.1016/j.cardfail.2017.04.014] [Medline: 28461259]
18. Ware P, Ross HJ, Cafazzo JA, Boodoo C, Munnery M, Seto E. Outcomes of a heart failure telemonitoring program implemented as the standard of care in an outpatient heart function clinic: pretest-posttest pragmatic study. J Med Internet Res 2020 Feb 08;22(2):e16538 [FREE Full text] [doi: 10.2196/16538] [Medline: 32027309]
19. Lam CSP, Teng TK, Tay WT, Anand I, Zhang S, Shimizu W, et al. Regional and ethnic differences among patients with heart failure in Asia: the Asian sudden cardiac death in heart failure registry. Eur Heart J 2016 Nov 01;37(41):3141-3153. [doi: 10.1093/eurheartj/ehw331] [Medline: 27502121]
20. Ademi Z, Pasupathi K, Liew D. Cost-effectiveness of eplerenone compared to usual care in patients with chronic heart failure and NYHA class II symptoms, an Australian perspective. Medicine (Baltimore) 2016 May;95(18):e3531 [FREE Full text] [doi: 10.1097/MD.0000000000003531] [Medline: 27149456]
21. McAlister FA, Stewart S, Ferrua S, McMurray JJJV. Multidisciplinary strategies for the management of heart failure patients at high risk for admission: a systematic review of randomized trials. J Am Coll Cardiol 2004 Aug 18;44(4):810-819 [FREE Full text] [doi: 10.1016/j.jacc.2004.05.055] [Medline: 15312864]
22. Ware P, Dorai M, Ross HJ, Cafazzo JA, Laporte A, Boodoo C, et al. Patient adherence to a mobile phone-based heart failure telemonitoring program: a longitudinal mixed-methods study. JMIR mHealth uHealth 2019 Feb 26;7(2):e13259 [FREE Full text] [doi: 10.2196/13259] [Medline: 30806625]
23. Moore K, Lipsitch M, Barry J, Osterholm M. Part 1: The future of the COVID-19 pandemic: lessons learned from pandemic influenza. COVID-19: The CIDRAP Viewpoint. 2020 Apr 30. URL: https://www.cidrap.umn.edu/sites/default/files/public/downloads/cidrap-covid19-viewpoint-part1_0.pdf [accessed 2020-11-24]
24. Griffiths A, Paracha N, Davies A, Branscombe N, Cowie MR, Sculpher M. Analyzing health-related quality of life data to estimate parameters for cost-effectiveness models: an example using longitudinal EQ-5D data from the SHIFT randomized controlled trial. Adv Ther 2017 Mar;34(3):753-764 [FREE Full text] [doi: 10.1007/s12325-016-0471-x] [Medline: 28205056]
25. Fees and Charges. Hong Kong Hospital Authority. URL: https://www.ha.org.hk/visitor/ha_visitor_index.asp?Content_ID=10045&Lang=ENG&Dimension=100&Parent_ID=10044&Ver=HTML [accessed 2020-11-13]
26. Reyes EB, Ha J, Firdaus I, Ghazi AM, Phrommintikul A, Sim D, et al. Heart failure across Asia: same healthcare burden but differences in organization of care. Int J Cardiol 2016 Nov 15;223:163-167 [FREE Full text] [doi: 10.1016/j.ijcard.2016.07.256] [Medline: 27541646]
27. Leung AW, Chan CY, Yan BP, Yu CM, Lam YY, Lee VW. Management of heart failure with preserved ejection fraction in a local public hospital in Hong Kong. BMC Cardiovasc Disord 2015 Feb 25;15:12 [FREE Full text] [doi: 10.1186/s12872-015-0002-8] [Medline: 25887230]

28. Usage of information technology and the internet by Hong Kong residents, 2000 to 2016. Hong Kong Census and Statistics Department. 2000. URL: https://www.statistics.gov.hk/pub/B71711FB2017XXXXB0100.pdf [accessed 2020-11-28]
29. The World Health Report 2002: reducing risks, promoting healthy life. World Health Organization. 2002. URL: https://www.who.int/whr/2002/en/whr02_en.pdf?ua=1 [accessed 2020-11-24]
30. National Income. Hong Kong Census and Statistics Department. URL: https://www.censtatd.gov.hk/hkstat/sub/sp250.jsp?tableID=030&ID=0&productType=8 [accessed 2020-11-24]
31. Abraham WT, Fiuzat M, Psotka MA, O'Connor CM. Heart failure collaboratory statement on remote monitoring and social distancing in the landscape of COVID-19. JACC Heart Fail 2020 Aug;8(8):692-694 [FREE Full text] [doi: 10.1016/j.jchf.2020.06.006] [Medline: 32731947]
32. Panhwar MS, Kalra A, Gupta T, Kolte D, Khera S, Bhatt DL, et al. Effect of influenza on outcomes in patients with heart failure. JACC Heart Fail 2019 Feb;7(2):112-117 [FREE Full text] [doi: 10.1016/j.jchf.2018.10.011] [Medline: 30611718]
33. Flu Express (volume 18), number 5 (week 5). Surveillance Division of the Communicable Disease Branch of the Hong Kong Centre for Health Protection. 2021 Feb 03. URL: https://www.chp.gov.hk/files/pdf/fluexpress_week5_04_02_2021_eng.pdf [accessed 2021-01-30]

Abbreviations

CUC: care under COVID-19
HF: heart failure
ICER: incremental cost-effective ratio
NYHA: New York Heart Association
QALY: quality-adjusted life-year
RR: risk ratio
WTP: willingness-to-pay

Edited by R Kukafka, C Basch; submitted 15.12.20; peer-reviewed by Q Zhang, CP Lau; comments to author 23.01.21; revised version received 08.02.21; accepted 19.02.21; published 03.03.21

Please cite as:
Jiang X, Yao J, You JHS
Cost-effectiveness of a Telemonitoring Program for Patients With Heart Failure During the COVID-19 Pandemic in Hong Kong: Model Development and Data Analysis
J Med Internet Res 2021;23(3):e26516
URL: https://www.jmir.org/2021/3/e26516
doi: 10.2196/26516
PMID: 33656440

Review

Compliance With Mobile Ecological Momentary Assessment of Self-Reported Health-Related Behaviors and Psychological Constructs in Adults: Systematic Review and Meta-analysis

Marie T Williams[1*], BAppSc, PhD; Hayley Lewthwaite[1,2*], BAppSc, BHlthScHons, PhD; François Fraysse[3], PhD; Alexandra Gajewska[3], BClinExPhys, BAHons, Dip Ed; Jordan Ignatavicius[3], BPhysio; Katia Ferrar[3*], BAppSc, BHlthScHons, PhD

[1]Innovation, Implementation And Clinical Translation in Health, Allied Health and Human Performance, University of South Australia, Adelaide, Australia

[2]Department of Kinesiology and Physical Education, Faculty of Education, McGill University, Montreal, QC, Canada

[3]Alliance for Research in Exercise, Nutrition and Activity, Allied Health and Human Performance, University of South Australia, Adelaide, Australia

[*]these authors contributed equally

Corresponding Author:
Marie T Williams, BAppSc, PhD
Innovation, Implementation And Clinical Translation in Health
Allied Health and Human Performance
University of South Australia
City East Campus, North Terrace
Adelaide, 5000
Australia
Phone: 61 8 8302 1153
Email: Marie.Williams@unisa.edu.au

Abstract

Background: Mobile ecological momentary assessment (mEMA) permits real-time capture of self-reported participant behaviors and perceptual experiences. Reporting of mEMA protocols and compliance has been identified as problematic within systematic reviews of children, youth, and specific clinical populations of adults.

Objective: This study aimed to describe the use of mEMA for self-reported behaviors and psychological constructs, mEMA protocol and compliance reporting, and associations between key components of mEMA protocols and compliance in studies of nonclinical and clinical samples of adults.

Methods: In total, 9 electronic databases were searched (2006-2016) for observational studies reporting compliance to mEMA for health-related data from adults (>18 years) in nonclinical and clinical settings. Screening and data extraction were undertaken by independent reviewers, with discrepancies resolved by consensus. Narrative synthesis described participants, mEMA target, protocol, and compliance. Random effects meta-analysis explored factors associated with cohort compliance (monitoring duration, daily prompt frequency or schedule, device type, training, incentives, and burden score). Random effects analysis of variance ($P \leq .05$) assessed differences between nonclinical and clinical data sets.

Results: Of the 168 eligible studies, 97/105 (57.7%) reported compliance in unique data sets (nonclinical=64/105 [61%], clinical=41/105 [39%]). The most common self-reported mEMA target was affect (primary target: 31/105, 29.5% data sets; secondary target: 50/105, 47.6% data sets). The median duration of the mEMA protocol was 7 days (nonclinical=7, clinical=12). Most protocols used a single time-based (random or interval) prompt type (69/105, 65.7%); median prompt frequency was 5 per day. The median number of items per prompt was similar for nonclinical (8) and clinical data sets (10). More than half of the data sets reported mEMA training (84/105, 80%) and provision of participant incentives (66/105, 62.9%). Less than half of the data sets reported number of prompts delivered (22/105, 21%), answered (43/105, 41%), criterion for valid mEMA data (37/105, 35.2%), or response latency (38/105, 36.2%). Meta-analysis (nonclinical=41, clinical=27) estimated an overall compliance of 81.9% (95% CI 79.1-84.4), with no significant difference between nonclinical and clinical data sets or estimates before or after data exclusions. Compliance was associated with prompts per day and items per prompt for nonclinical data sets. Although

widespread heterogeneity existed across analysis (I^2>90%), no compelling relationship was identified between key features of mEMA protocols representing burden and mEMA compliance.

Conclusions: In this 10-year sample of studies using the mEMA of self-reported health-related behaviors and psychological constructs in adult nonclinical and clinical populations, mEMA was applied across contexts and health conditions and to collect a range of health-related data. There was inconsistent reporting of compliance and key features within protocols, which limited the ability to confidently identify components of mEMA schedules likely to have a specific impact on compliance.

(J Med Internet Res 2021;23(3):e17023) doi: 10.2196/17023

KEYWORDS

mobile momentary ecological assessment; adult; compliance; systematic review; meta-analysis; mobile phone

Introduction

Background

Ecological momentary assessment (EMA) is a survey method that allows collection of data on participant behaviors, affect, and perceptual experiences in real-time (momentary) and real-life environments (ecological) [1]. In its original form, EMA required pen and paper diaries or logs to be completed on random (signal) or fixed (interval) time-based schedules or in response to a specific target behavior, psychological or social event (event-based). With the advent of handheld technologies, mobile EMA (mEMA) and increasingly mobile ecological momentary interventions (mEMIs) can be completed through automated schedules via handheld devices such as tablets and mobile phones.

As mEMA or mEMI have the potential to capture data in real time, the level of recall bias is potentially reduced. In addition, contextual (where and who the respondent is with) and antecedents to the specific target behavior or psychological construct can be obtained [1,2]. As a survey approach, mEMA or mEMI has undeniable utility, but data are dependent on participants consistently responding to the mEMA or mEMI schedule (compliance) [3]. Although electronically delivered surveys to personal mobile devices provide a means of time or date stamping and limit the possibility of hoarding, back and forward filling [4], concerns have been raised about protocol burden, missing data (especially if systematic), mindless answering, and survey habituation when lengthier questionnaires can be circumvented by a no response to initial questions [2]. EMA data with low compliance rates are unlikely to be ecologically valid; however, it is also possible to have good individual compliance with data of questionable accuracy [5,6].

In the last 5 years, there have been at least 10 systematic reviews focused on EMA and/or reporting aspects of compliance to EMA schedules in youth (<18 years [7-9]; <22 years [10]), mixed youth and adult cohorts [11-13], or specific adult populations [5,14-16]. Compliance with EMA in youth (nonclinical and clinical samples) has been reported to range between 44% and 96% [8-10] and in mixed youth and adult cohorts, between 23% and 94% [11-14]. Reports of compliance in specific adult clinical populations range from 21% to 99% (chronic pain, 21%-99% [15]; psychotic disorders, 78%-86% [16]; substance use, 75%, (95% CI 72.37-77.65) [5].

Although Stone and Shiffman [17] have highlighted the need for explicit reporting of compliance in their original reporting guidelines for EMA, recurring issues relating to the reporting of compliance include (1) missing, incomplete, or ambiguous data; (2) heterogeneity in reporting; (3) impact of data exclusions; and (4) combining traditional (paper-based) and mEMA data [5]. Participant compliance with mEMA or mEMI—in theory—is related to the total protocol burden, which is a function of monitoring duration, frequency and complexity of prompts, and familiarity with the technology. However, as Jones et al [5] note, to date, there is little compelling, systematic evidence to support an association between EMA burden and compliance rates. These issues make it difficult to determine which, if any, features of EMA protocols positively or negatively influence compliance to EMA schedules.

The purpose of this systematic review is to guide the development of an mEMA protocol, which could be used for future studies of health-related behaviors and psychological constructs (including symptoms) in adults with and without chronic disease. The primary question for this systematic review is as follows: In adult nonclinical and clinical populations, which factors are associated with increased compliance to mEMA protocols for collection of health-related behaviors and psychological constructs (including symptoms)?

Objectives

The objectives of this systematic review were to describe:

1. Health-related behaviors and psychological constructs assessed using mEMA
2. mEMA protocol and compliance reporting
3. Associations between key components of mEMA protocols and participant compliance

Methods

Search Registration

The search strategy and review protocol were registered prospectively with the International Prospective Register of Systematic Reviews (PROSPERO 2016: CRD42016051726).

Eligibility

Observational studies (cohort, cross-sectional) of mEMA in adults (>18 years of age) were eligible for inclusion in this review if these (1) reported participant compliance with mEMA; (2) were a primary study published in English between 2006 and 2016 inclusive; (3) included adults (≥18 years) either apparently healthy (nonclinical population) or with health conditions (clinical population); and (4) collected mEMA data

using mobile devices as a primary or secondary outcome. References were excluded if these were (1) experimental designs investigating intervention efficacy; (2) duplicate publications or secondary analysis of the same data set; or (3) conference abstracts, protocols, commentaries (editorials or letters), or systematic or narrative reviews.

Information Sources and Search Strategy

A range of electronic databases were searched to identify eligible studies: AMED (Allied and Complementary Medicine), CINAHL, Cochrane Library and CENTRAL (Cochrane Central Register of Controlled Trials), Embase, MEDLINE (including epub ahead of print), PsycINFO, Scopus, and Web of Science. An academic librarian (Carole Gibbs, University of South Australia) assisted with the development of the search strategy regarding conceptualization, operators (operational terms), and limiters [18] with the final search undertaken during a single week. Search terms and associated MeSH (Medical Subject Heading) alternatives, which were adapted for use in all databases, related to the population (adults), assessment (mEMA), and outcomes of interest (health behaviors, perceptual experiences including symptoms, affect or mood). Key search terms included "ecological momentary assessment," "EMA," "mobile ecological momentary assessment," "mEMA," "electronic diary," "SMS or short message service," "prompting," "text messaging," "health behaviour," "symptom," and "adult." Reference lists of included studies and systematic reviews identified during the search were reviewed to identify additional potentially relevant studies.

Study Selection

The titles and abstracts of studies identified from the search process were screened against a priori eligibility criteria and full-text versions imported into Covidence (Covidence systematic review software, Veritas Health Innovation). Both screening steps were undertaken by individual members of the research team working in pairs (AG and MW, HL and FF) with each person completing the task independently, before meeting with their partner to compare results and resolve disagreements (consensus).

Data Collection

A data extraction template was prospectively developed; it was guided by the Checklist for Reporting EMA studies proposed by Liao et al [10] and pilot-tested on 5 randomly selected eligible studies. Working in pairs (AG and MW, JI and KF, HL and FF), individual members of the research team extracted all data before meeting with their partner to compare results and resolve disagreements by discussion. As this review aims to describe the features of mEMA schedules associated with increased mEMA protocol adherence, assessment of methodological bias was not planned.

Data Items

Data were extracted across 4 domains:

Publication demographics: title, authors, year of publication.

Participants: recruitment source, medical condition or diagnosis (clinical populations), sample size (enrolled, attrition or withdrawn and included in analysis), and age (mean/median, SD).

mEMA protocol: target behavior or psychological construct, mobile device type (PDA, palmtop computer, electronic diary, mobile or smartphone, tablet, other), participant training (yes/no), provision of incentives (course credit, financial, other, or none), incentive thresholds (yes/no) monitoring duration (days), prompt type (random signal, interval, event-based), frequency per day, number of questions/items per prompt type (reported or estimated from information reported in studies), strategy to deal with unanswered prompts, and time allowed for survey response. Where authors did not report the number of items per prompt type, but rather included descriptions of standardized instruments which were converted to mEMA survey items, a full version of the standardized instrument was accessed, and number of items calculated.

mEMA compliance: verbatim (or where possible calculated from reported data), participant completion (number included in analysis, data exclusions), criteria/thresholds for mEMA data, number of prompts delivered/answered per person/cohort (planned, actual, average, range), and response latency as time (mean, SD) [8,10].

Data Management

Data were tabulated to provide descriptive summaries. The mEMA surveys commonly included multiple questions reflecting behavioral or psychological constructs. Although the authors of mEMA studies did not always specify the primary outcome for these observational studies, most studies explicitly reported the key variable of interest for mEMA, which we interpreted to be the primary mEMA target. Where other data were also collected by the same mEMA survey, we denoted those as secondary mEMA targets. The primary mEMA target of studies was identified, and studies were grouped and reported according to two broad domains: (1) behavior (eg, dietary, physical activity, and smoking) and (2) psychological construct (eg, affect, cognition, and sensations/symptoms). For each domain, a narrative synthesis was used to summarize participants, mEMA protocol, and compliance data for nonclinical and clinical data sets.

With the exception of device type, where possible, we adopted the operationalization of variables common to Wen et al [9] or Jones et al [5] unless the distribution of our data resulted in very unbalanced cells or our data could provide greater resolution. Potential mEMA protocol factors related to compliance were categorized for analysis. *Monitoring duration* was categorized as follows: <7 days, >7 days to <14 days, or >14 days. *Prompt frequency* was grouped as follows: 1-3 prompts per day; 4-5 prompts per day; or ≥6 prompts per day. Minimum *items per prompt* were categorized as follows: ≤5, >5 to ≤9.5, >9.5 to ≤26, and >26. *Device type* was categorized as mobile phone, PalmPilot/PDA, or other. The reporting of *training or familiarization sessions or provision of incentives* were dichotomized as yes/no or labeled as not reported.

Given ongoing concerns about the burden imposed by EMA schedules and compliance, in addition to these individual factors, we explored a novel composite metric to reflect aspects

previously identified as possible contributing factors (monitoring duration, frequency, type, and complexity of prompts).

Where possible, a mEMA *burden score* was calculated for each study by multiplying:

- the total monitoring duration in days (d; all days included in all waves)
- by the maximum frequency of time-based prompts (random and interval) per day (f)
- by the minimum number of compulsory questions/items within all prompts per day (i) and
- by a weighting reflecting the number of prompt types scheduled per day (w; eg, time-based [signal or interval] and/or event-based) with each prompt type weighted as 1 (min weight=1, max=3).

For example, the mEMA burden score for a 14-day monitoring schedule *(d)*, where 5 random signal prompts were delivered per day (*f*), with each prompt requiring responses to a minimum of 12 items/questions (*I*; 60 items in total per day), would be 840. If event-based prompts (irrespective of the number of items within the prompt) were added to this schedule (*w*), the burden score would rise to 1680. *Burden scores* were calculated and reported in quartiles: 0 to 283.5, 284 to 810, 811 to 1806, or ≥1807.

Meta-analysis

Random effects restricted maximum likelihood estimator meta-analyses were undertaken using the approach reported by Jones et al [5] and Wen et al [9], with both authors advising to assist in accurate replication. All statistical analyses were conducted using JASP (Jeffreys's Amazing Statistics Program, version 0.9.2; 2019). Studies were included in the meta-analysis if they reported all data necessary for the meta-analysis procedure and cohort compliance (%) could be extracted before data exclusions when possible. Sensitivity analysis was conducted to explore the impact of compliance rates reported before and after data exclusion. The effect sizes (ESs) were calculated by logit transforming the proportion of completed prompts (ie, compliance rates; proportion/[1−proportion]). SEs were then estimated using the following equation:

$$\sqrt{([1/np]+[1/n\{1-p\}])}$$

Where, *n* is the sample size and *p* is the proportion.

To adjust for clustering within participants, the SE was adjusted by the effective sample size (ESS). The ESS equation is as follows:

$$kn/(1+[k-1]\ ICC)$$

Where, k is the number of study prompts, n is the participant number, ICC is either the reported intraclass correlation coefficient (ICC) or the SD of reported compliance, and p is the proportion of completed prompts.

For studies that did not report SD data, sensitivity analyses were conducted by computing the SEs using the 25 and 75 percentiles of available SDs. The sensitivity analyses did not show any differences. Therefore, analysis used imputed median SD (where the original SD was not reported). To aid interpretation, inverse logit transformation was conducted to enable reporting of proportions. The I^2 statistic was used to quantify heterogeneity across the ES. Pooled compliance rates were initially explored for combined nonclinical and clinical data sets and then compared between nonclinical and clinical studies.

To explore the relationships between the pooled compliance rates (nonclinical and clinical data sets) and EMA protocol factors (ie, monitoring duration, prompt frequency, device type, training, incentives, and burden score), random effects analysis of variance was conducted as part of the meta-analysis program. Moderator analyses were conducted separately for nonclinical and clinical pooled compliance.

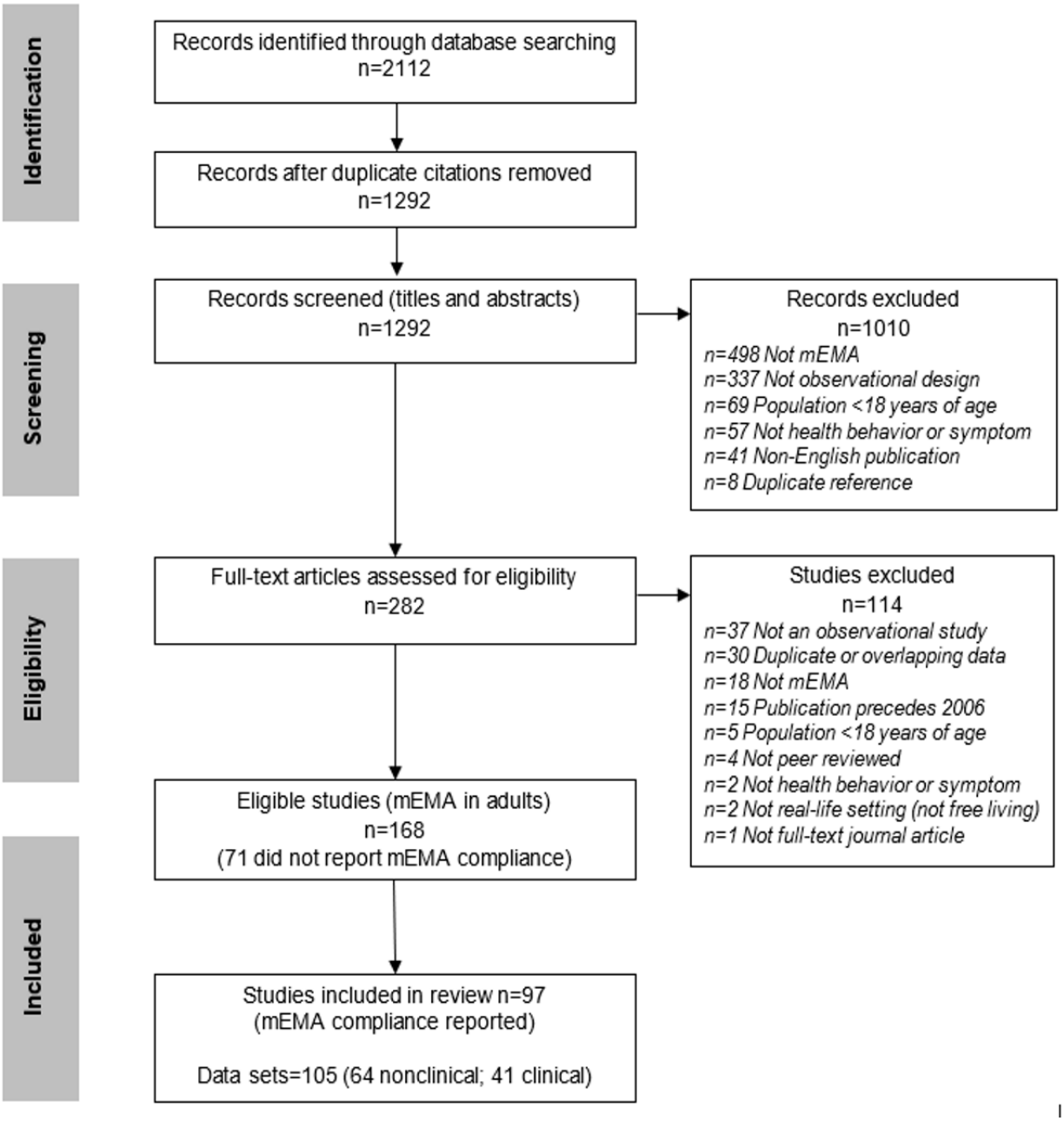
Results

Overview

Figure 1 presents the outcome of the search strategy. Of the 282 studies reviewed as full text, 168/282 (59.6%) included mEMA; however, 42.3% (71/168) were excluded because mEMA compliance was not reported. The majority of the 97 studies retained for this review comprised studies that recruited or reported a single nonclinical group (61/97, 63%) or a clinical (31/97, 32%) group. Two studies included 2 [19] or 3 clinical groups [20]. In addition, 3 studies included clinical and nonclinical comparator groups (4 groups [21], 2 groups [22,23]). Overall, 105 data sets were included in this review (nonclinical: 64/105, 61%; clinical: 41/105, 39%). A description of all included data sets is presented in Multimedia Appendix 1 [19-114].

A total of 44,796 participants were included in the analyses (nonclinical: 42,338/44,796, 94.51%; clinical: 2431/44,796, 5.43%) with a median sample size of 62 (nonclinical: n=89; clinical: n=40; Multimedia Appendix 2). Two data sets (nonclinical) were outliers because of the sample size (n=21,947; n=11,572) [24,25]. The main sources of recruitment for nonclinical data sets were educational institutions (30/64, 47%) and community (26/64, 41%), whereas clinical data sets were predominantly recruited from medical/health services (21/41, 51%) and community (17/41, 41%). For clinical data sets, the most common health conditions were psychiatric or mental health (12/41, 29%), chronic pain and fibromyalgia (6/41, 15%), and eating disorders (5/41, 12%). Multimedia Appendix 2 presents a summary of the study characteristics grouped by primary mEMA target.

Figure 1. Search strategy process and final outcomes (hand searching of reference list-eligible studies and review papers did not identify additional studies to those returned by database searches). mEMA: mobile ecological momentary assessment.



Objective 1: Health-Related Behaviors and Psychological Constructs Assessed With mEMA

Using the primary mEMA target, data sets were grouped into 2 broad domains: *Behavior* or *Psychological construct*. Within the *Behavior* domain, the *Other* category reflects single studies (7), where the primary mEMA target did not align with more common behavior targets (social interactions/activities [26,27], sexual [28], leisure [29], nonsuicidal self-injurious [30], HIV prevention [31], and oral behaviors) [32].

The most frequent primary mEMA target across all domains for nonclinical and clinical data sets was affect (31/105, 29.5% of data sets; nonclinical n=15/64, 14%, clinical n=16/41,15%). The most common primary mEMA target in nonclinical data sets (n=64) reflected the Behavior domain (total 38/64, 59%), whereas clinical data sets (n=41) reflected the Psychological domain (total 32/41, 78%).

With the exception of 1 clinical study (fatigue) [33], the remaining data sets included mEMA items/questions beyond the primary mEMA target. The most frequent secondary targets assessed were affect (50/105, 47.6%), social environment (33/105, 31.4%), physical activity (25/105, 23.8%), cognition (24/105, 22.8%), and physical environment (20/105, 19%). Multimedia Appendix 2 presents a summary of secondary mEMA targets and participant characteristics grouped by the primary mEMA target.

Objective 2: mEMA Protocol and Compliance Reporting

Multimedia Appendix 3 presents a summary of mEMA protocols grouped by primary mEMA target. Among the included studies,

mEMA data were most commonly collected using handheld computer/PDAs (61/105, 58.1%) with mobile phones accounting for approximately one-third (37/105, 35.2%). Participant training in mEMA was reported by most studies (nonclinical: 49/64, 77%; clinical: 35/41, 85%). The provision of incentive (financial or other) was more frequent in nonclinical protocols (nonclinical: 46/64, 72%; clinical: 20/41, 49%).

Across all data sets (n=105), the median monitoring duration for mEMA protocols was 7 days (range: 1-182 days), with durations differing between nonclinical (median 7 days, range 1-49 days) and clinical protocols (median 12 days, range 1-182 days). Most studies included a single prompt type (overall data sets: 69/105, 65.7%; nonclinical: 40/64, 63%; clinical: 29/41, 71%), with random signals being the most common in nonclinical protocols (49/64, 77%) and interval in clinical protocols (25/41, 61%). Of the remaining study protocols, 23% (24/105) of studies included 2 prompt types and 11% (12/105) protocols included all 3 prompt types (random signal, interval, and event-based). The frequency of time-based prompts (signal or interval) ranged from 1 to 42 per day (median: nonclinical=5, range 1-36; clinical=4, range=1-42). The number of specific questions/items within a standard prompt varied markedly across study protocols; it ranged between 1 and 73 (median: nonclinical=10; clinical=8).

Table 1 presents a summary of reporting for compliance metrics for mEMA time-based prompts (ie, signal and fixed prompts). Participant attrition (dropout) rates were reported or could be calculated for half of the 105 data sets (nonclinical: 31/64, 48%; clinical: 22/41, 54%). Less than half of the data sets reported the number of prompts delivered (overall: 22/105, 21%; nonclinical: 14/64, 22%; clinical: 8/41, 20%) or answered (overall: 43/105, 41%; nonclinical: 29/64, 45%; clinical: 14/41, 34%). Approximately one-third of the data sets reported a criterion for valid mEMA data or reasons for data exclusions (overall: 37/105, 35%; nonclinical: 25/64, 39%; clinical: 12/41, 29%). Criteria for valid EMA data fell into 2 main groups, with the most common based on assessment completion (ie, specified threshold for number of prompts completed per day or percentage of overall compliance), followed by response latency period threshold (eg, prompt required to be answered within 30 min). Of the data sets reporting a criterion for response time (overall: 38/105, 36%; nonclinical: 16/64, 25%; clinical: 22/41, 54%), this ranged from 1.5 to 60 min (median 15 min; Multimedia Appendix 3). Other reasons for data exclusion were based on specific time of day prompts (excluding the first or last of the day), technical malfunctions, or unspecified (eg, general statements on participants' limited or poor compliance).

Of the 105 data sets, 82/105 (78.1%) reported compliance using a single metric (cohort, average per person or other), with compliance at the cohort level most common (overall: 62/105, 59%; nonclinical: 34/64, 53%; clinical: 28/41, 68%). Compliance was less frequently reported using the single metric of average per person (overall: 20/105, 19%; nonclinical: 14/64, 22%; clinical: 6/41, 15%) or compliance for both cohort and average per person (overall: 18/105, 17%; nonclinical: 12/64, 19%; clinical: 6/41, 15%). The remaining data sets (n=5; nonclinical: n=4, clinical: n=1) reported compliance after combining event/time-based signals [34] or separate tasks [35], number of completed protocol days [36], total number of prompts (data) available [37], or proportion of completed questions/items per prompt [38]. Cohort compliance reported before data exclusions ranged from 38% to 98% (median 82%) and after data exclusions from 50% to 97% (median 81%; Table 1).

Table 1. Summary of mobile ecological momentary assessment (mEMA) compliance reporting.

Primary mEMA[a] target	NC[b] or C[c] (n)	Reported N=data sets (%)							Cohort compliance (%)	
		Attrition rate	Total prompts delivered	Total prompts answered	Criteria for valid data	Compliance predata exclusions	Compliance postdata exclusions	Average per-person compliance	Predata exclusion, median (range)	Postdata exclusion, median (range)
Smoking	NC (12)	8 (66)	4 (33)	5 (42)	4 (33)	7 (58)	1 (8)	5 (42)	83 (69-93)	83 (74-91)
	C (1)	1 (100)	1 (100)	1 (100)	0 (0)	1 (100)	0 (0)	1 (100)	68 (NA[d])	N/A[e]
Alcohol	NC (8)	3 (37)	1 (12)	3 (37)	3 (37)	4 (50)	4 (50)	2 (25)	90 (86-97)	79 (69-80)
	C (0)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Eating behaviors	NC (10)	6 (60)	2 (20)	5 (50)	5 (50)	4 (40)	3 (3)	5 (50)	90 (40-96)	67 (50-71)
	C (3)	2 (66)	1 (33)	2 (66)	1 (33)	2 (66)	1 (33)	0 (0)	N/A	78 (68-87)
Physical activity	NC (5)	1 (20)	1 (20)	4 (80)	0 (0)	3 (60)	0 (0)	3 (60)	82 (75-95)	N/A
	C (1)	0 (0)	0 (0)	0 (0)	0(0)	1 (100)	0 (0)	0 (0)	N/A	97 (NA)
Other	NC (3)	0 (0)	2 (66)	3 (100)	0 (0)	2 (66)	0 (0)	2 (66)	61 (38-84)	N/A
	C (4)	4 (100)	0 (0)	3 (75)	1 (25)	2 (50)	1 (25)	2 (50)	74 (72-74)	N/A
Personality traits	NC (7)	4 (57)	1 (14)	3 (42)	3 (42)	3 (42)	1 (14)	3 (42)	75 (55-90)	N/A
	C (0)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Affect	NC (15)	7 (46)	2 (13)	5 (33)	9 (60)	6 (40)	6 (40)	7 (46)	78 (63-90)	77 (73-81)
	C (16)	9 (56)	2 (12)	5 (31)	7 (44)	5 (31)	6 (37)	9 (56)	80 (69-96)	83 (79-87)
Cognitions	NC (2)	1 (50)	0 (0)	1 (50)	0 (0)	2 (100)	0 (0)	0 (0)	83 (77-89)	N/A
	C (0)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Symptoms	NC (2)	1 (50)	1 (50)	0 (0)	1 (50)	0 (0)	1 (50)	1 (50)	N/A	N/A
	C (16)	6 (37)	4 (25)	3 (19)	3 (19)	11 (69)	4 (25)	4 (25)	90 (68-98)	86 (86-93)
Total	NC (64)	31 (48)	14 (22)	29 (45)	25 (39)	31 (48)	16 (25)	28 (44)	82 (38-97)	74 (50-91)
	C (41)	22 (54)	8 (20)	14 (34)	12 (29)	22 (54)	12 (29)	16 (39)	80 (68-98)	87 (68-97)
	t (105)	53	22	43	37	53	28	44	82 (38-98)	81 (50-97)
	%	50.4	20.9	40.9	35.2	50.4	26.6	41.9	N/A	N/A

[a]mEMA: mobile ecological momentary assessment.

[b]NC: nonclinical.

[c]C: clinical.

[d]NA: not available as domain includes a single study.

[e]N/A: not applicable.

Question 3: Associations Between Key Features of mEMA Protocols and mEMA Compliance

Of the 105 data sets included in this review, 65% reported sufficient data for inclusion in the meta-analysis (n=68 data sets: 41/105 [39%] ES nonclinical and 27/105 [26%] ES clinical); Multimedia Appendix 1) [20,21,23,26,27,29-31,33,36,39-90]. The remaining data sets did not report cohort compliance but reported average per-person compliance [19,24,25,91-106,28,32] or other [34,35,37,38,107-110], or where cohort compliance was reported, a variable required for the meta-analysis was not [111-114].

The overall compliance rate across all 68 ESs was 81.9% (95% CI 79.1-84.4). There was sizable heterogeneity across the compliance rates (I^2=98). Sensitivity analysis exploring the impact of pre and postdata exclusion compliance rates showed no significant difference (*P*=.67*;* before exclusion: n=50, 81.6%; after exclusion: n=18, 82.8%). There was no significant difference (*P*=.16) between the pooled compliance of nonclinical studies (80.4%; 95% CI 76.1-83.9; I^2=98.6) and clinical studies (84.2%; 95% CI 80.1-87.4; I^2=95.7). Three studies included more than 1 data set and reported compliance ESs for each (data sets n=2 [23], n=3 [20], and n=4 [21]). Sensitivity analysis was undertaken to explore the impact of double counting of mEMA protocol factors within the meta-analysis, where multiple ESs were reported within single studies. When a single ES was retained for each of these studies (lowest ES of the 2 [23], median of 3 [20], ES closest to the average for 4 [21]), the

pooled 62 ESs (81.3%, 95% CI 78.2-84.2) and reported variance (I^2=98) were essentially the same as the full data set (68 ESs: 81.9%; 95% CI 79.1-84.4; I^2=98). To ensure that subgroup analysis was not affected, all analyses were conducted without duplicate ESs, and all relationships were consistent with those of the full data set.

For nonclinical studies, 2 factors (prompt frequency and items/prompt) were significantly related to mEMA compliance. For prompt frequency, the overall model was nonsignificant (*P*=.07), but the coefficient was significant (*P*<.001). Prompting 1 to 3 times per day was associated with higher compliance (87%; 95% CI 82.5-90.4) compared with studies with more than 3 prompts per day (76.9%) and 6 or more prompts per day (79.4%). The number of items per prompt was significant for both the overall model (*P*=.04) and the coefficient (*P*<.001). Factor analysis showed that prompts with more than 26 items had significantly lower compliance (63%; 95% CI 42.3-79.7) compared with prompts with ≤26 items (categories: ≤5; >5 to ≤9; >9.5 to ≤26; compliance range: 84%-78.6%).

For clinical data sets (n=27), no factors were significantly related to compliance. The number of items per prompt approached significance (*P*=.05). Compliance appeared to be lower in studies with 9.5-26 items per prompt (71.1%; 95% CI 62.5-78.6). Significant heterogeneity was reported for all significant findings (nonclinical and clinical), with I^2 values in excess of 90%, suggesting that although some variance can be explained by the significant factors, a large amount of variance remained unexplained. The burden score was not significantly related to compliance. The meta-analysis factor analysis compliance proportions are presented in Table 2.

Table 2. Meta-analysis results for clinical and nonclinical data sets.

Characteristics	Clinical data sets, n=27		Nonclinical data sets, n=41	
Protocol factors	n (%)	Pooled compliance (95% CI)	n (%)	Pooled compliance (95% CI)
Monitoring period, day				
<7	12 (44)	81.6 (74.1-87.3)	24 (58)	77.4 (71.3-85.5)
>7 to ≤14	4 (15)	84.4 (74.3-91.1)	9 (22)	82.1 (71.30-89.5)
>14	11 (41)	86.7 (81.2-91.0)	8 (19)	85.3 (80.5-89.1)
Device[a]				
Mobile	5 (19)	88.6 (71.5-96.1)	17 (41)	78.6 (71.9-84.0)
PDA	18 (66)	81.9 (77.4-85.8)	22 (54)	80.2 (74.2-84.9)
Other	4 (15)	88.8 (82.4-93.1)	2 (5)	92.2 (86.3-95.7)
Training				
Yes	23 (85)	84.4 (79.7-88.4)	36 (88)	80.4 (76.0-84.3)
No	0 (0)	N/A[b]	0 (0)	N/A
NR[c]	4 (15)	82.8 (78.4-86.4)	6 (15)	77.7 (73.1-82.0)
Incentives				
Yes	13 (48)	83.6 (77.7-88.3)	35 (85)	80.4 (79.0-84.3)
No	0 (0)	N/A	6 (15)	77.9 (73.1-82.0)
NR	18 (66)	85.7 (81.3-89.3)	0 (0)	N/A
Prompt frequency, per day				
1-3	8 (30)	85.3 (77.6-90.7)	8 (19)	87.0 (82.5-90.4)
4-5	12 (44)	81.5 (75.8-85.9)	16 (39)	76.9 (70.1-82.5)
≥6	6 (22)	86.3 (74.1-92.4)	17 (41)	79.4 (71.1-85.5)
UTD[d]	1 (4)	90.6 (N/A)	0 (0)	N/A
Burden score				
0-283.5	4 (15)	86.2 (76.9-92.4)	11 (27)	80.5 (75.7-84.6)
284-810	7 (26)	86.4 (75.4-93.0)	10 (24)	79.6 (73.7-84.7)
811-1806	3 (11)	88.8 (64.8-97.1)	13 (31)	82.8 (73.7-89.1)
≥1807	7 (26)	85.3 (80.5-89.0)	4 (10)	79.1 (51.5-93.1)
Items per prompt				
<5	8 (30)	87.2 (80.7-91.9)	10 (24)	82.8 (77.2-87.2)
5 to ≤9.5	7 (26)	88.4 (76.9-94.6)	8 (19)	78.6 (67.5-86.8)
9.5 to ≤26	2 (7)	71.1 (62.5-78.6)	16 (39)	84.0 (79.0-88.0)
>26	6 (22)	87.2 (82.9-90.7)	4 (10)	63.0 (42.3-79.7)
NR	5 (19)	72.7 (68.4-76.9)	3 (7)	70.3 (40.4-89.2)
Number of prompt types				
1	18 (66)	82.6 (78.1-86.5)	25 (61)	79.6 (75.0-83.5)
2	6 (22)	86.4 (71.3-94.2)	11 (27)	83.3 (71.7-90.9)
3	3 (11)	87.2 (85.5-88.8)	5 (12)	77.7 (65.7-86.5)

[a]Device type included with categories: Mobile phone (total n=22; smartphone: clinical n=1; nonclinical n=14; mobile: clinical n=4, nonclinical n=3); PDA (total n=45; clinical n=22, nonclinical n=23); Other (total n=6; electronic diary: clinical n=2, nonclinical n=1; iPod: clinical n=1, nonclinical n=1; watch device: clinical n=1).

[b]N/A: not applicable.

[c]NR: not reported.

[d]UTD: unable to be determined.

Discussion

Principal Findings

This systematic review of observational studies aimed to describe protocols and compliance with mEMA for self-reported health-related behaviors and psychological constructs in adults. Across 105 unique data sets, the key findings of this review were as follows: (1) a variety of health-related behaviors and psychological constructs were assessed, with affect being the most common mEMA target; (2) mEMA protocols varied widely across studies; (3) compliance was inconsistently reported across studies; (4) meta-analysis estimated an overall compliance rate of 81.9% (95% CI 79.1-84.4), with no significant difference between nonclinical and clinical data sets or estimates before or after data exclusions; (5) compliance was associated with prompts per day and items per prompt (nonclinical); and 6) no compelling relationship was identified between key features of mEMA protocols representing *burden* and mEMA compliance.

mEMA Use in Adults for Health-Related Behaviors and Psychological Constructs

The mEMA targets identified in this review reflect those reported in previous systematic reviews: affect/mood [7,12,14,15], cognitions [13], symptoms [15], eating or dietary behaviors [10,11], physical activity [10], and smoking or alcohol consumption [5,6]. Likewise, clinical populations identified in this review (psychiatric or mental health conditions, chronic pain and fibromyalgia, eating disorders, and substance use) were generally consistent with those reported previously [5,7,11,12,14-16]. However, there were chronic conditions unique to this review: oral or dental health, cancer, stroke and traumatic brain injury (for each n=3, 9/41, 22%), HIV, and upper abdominal surgery (for each n=1, 2/41, 5%). The small number of studies identified for these clinical groups may suggest that the potential for mEMA has not yet been realized in these populations.

Reporting of mEMA Protocols and Compliance

Most studies included in this review provided information around the EMA protocol used (device, monitoring duration, frequency and type of prompts, provision of training, and use of incentives). Consistent with previous systematic reviews of both youth and adults, there was considerable heterogeneity across studies for EMA protocols (Multimedia Appendix 3). Heterogeneity may be expected given the various potential applications of this survey approach. The mEMA protocol required to obtain sufficient or appropriate self-reported data on daily habitual behaviors in the general population is not likely to be the same as that for obtaining self-reported data on psychological responses to events or stimuli in clinical contexts. For example, the average EMA monitoring duration for studies of nonclinical adults in this review was 7 days (range: 1-49 days) compared with 12 days (range: 1-182 days) for clinical populations and 30 days (range: 3-730 days) in a review of EMA in substance users [5]. Likewise, prompt type, frequency, and complexity are expected to differ depending on the EMA target and population. Reviews of studies of EMA for diet and physical activity (common behaviors) report a daily average prompt frequency of 20 [10] compared with less than 4 prompts per day in substance use [5]. For these reasons, in systematic reviews of EMA use—including this one—reporting of summary metrics (mean, SD, median, range) for protocol components could be interpreted as a reflection of diversity in EMA application rather than a lack of protocol standardization.

The same rationale cannot be applied to the inconsistencies identified in reporting of EMA protocol compliance. Compliance is problematic to determine for event-based prompts (eg, those completed with smoking or consumption of alcohol). Compliance for time-based notifications, especially when the EMA is conducted using mobile devices, is relatively simple (number of prompts answered out of the total number of prompts delivered). However, participants may respond to a notification but may not complete all survey items or may not respond in a timely manner, affecting the momentary aspect of the EMA. In both of these cases, the act of responding might appropriately contribute to compliance rates, but the data are unlikely to be valid. These concepts were evident in the earliest recommendations for reporting compliance in EMA studies [17], which predate the sampling frame of this systematic review (2006-2016 inclusive). Considering that 71 studies were excluded from this review because of the absence of reporting mEMA compliance, less than half of the studies included in this review complied with recommendations put forward by Stone and Shiffman [17], such as reporting the proportion of delivered prompts answered (43/105, 41%) or defining a criterion for valid EMA data (37/105, 35%). Similarly, less than half of the data sets included in this review reported an average number of prompts answered per person (44/105, 42%), as recommended by more recently published guidelines for reporting EMA [8,10].

With the growth of systematic review methodologies (meta-synthesis, meta-regression, etc), one aspect of reporting for EMA warrants further consideration. EMA allows collection of self-report data across multiple survey items reflecting a range of behavioral, psychological, and contextual factors. It is not uncommon for data collected in the original, primary study to be reported in several publications. The foci of these *offspring* publications may include the total original sample of participants recruited (eg, unpublished data for specific mEMA items or other variables) or explore a subset of the original study participants (eg, patterns associated with participant characteristics). Although this is a reasonable and defensible use of the original study's resources, identification of duplicate or overlapping data in studies can be problematic. Where ambiguity exists, contacting the study authors is one way to clarify which publication should be considered the primary report (and which report overlapping or duplicate data). However, this option becomes less practical as time and people move on. The alternative is for authors to include an explicit statement concerning the existence of publications that include overlapping or duplicate data. There were a number of exemplars of this aspect of reporting in studies included [67,68,96] and excluded from this review [115-118].

Associations Between Key Components of mEMA Protocols and Compliance: Meta-analysis

In our meta-analysis (68 data sets), which replicates and was guided by the authors of 2 previous meta-analyses on this topic [5,9], the overall compliance rate was 81.9% (95% CI 79.1-84.4). This was slightly higher than that reported by Wen et al [9] (78.26%; 95% CI 75.49-80.78) and Jones et al [5] (75.06%; 95% CI 72.37-77.65). Although concerns have been expressed about the relationship between EMA burden and compliance, it remains unclear whether, or which, EMA protocol factors affect participant compliance. In our meta-analysis, for nonclinical data sets, prompt frequency per day and the number of items per prompt were significantly related to compliance (noting that it is not unusual for coefficients derived within a model to be significant even when the overall model is not). However, the findings are likely affected by the number of data sets in some categories. For nonclinical data sets, frequencies of 1-3 prompts per day were associated with small but significantly higher mean cohort compliance. Higher compliance with lower number of prompts perhaps seems intuitive, yet the evidence is inconsistent. Wen et al [9] reported opposite patterns of significance when nonclinical and clinical population data were investigated, and Jones et al [5] and Ono et al [119] reported no relationship with prompt frequency and compliance among substance users and those affected by chronic pain, respectively.

The relationship between the number of items included within each prompt and compliance has not been explored in previous systematic reviews or meta-analyses of mEMA. In this review, the number of items respondents were required to complete in a standard prompt ranged from 1 to 73 (median 10), with a greater number of items more common in the mEMA of psychological constructs (Multimedia Appendix 3). Our analysis showed an intuitive relationship with compliance among nonclinical data (ie, ≥26 items per prompt had the lowest mean cohort compliance of 63%; 95% CI 42.3-79.7), but not with clinical data.

When aiming to identify protocol factors affecting compliance, the inconsistencies in reporting of EMA compliance and the likely publication bias (studies with lower compliance rates may not be submitted or accepted for publication) must also be considered [5]. These factors limit the inclusion of potentially eligible studies in meta-analyses (68/105, 64.8% data sets in this review; 36/42, 86% studies in a previous review [9]). In addition, studies included in meta-analyses privilege *best compliers* through exclusion of participants not meeting criteria for valid EMA data or compliance thresholds (determined a priori or posteriori). Jones et al [5] attempted to address this latter point by exploring protocol factors associated with participant data exclusions (monitoring duration and prompt frequency). Finally, aggregate level compliance may not be sensitive enough or provide sufficient resolution to identify factors associated with higher or lower compliance. While accepting these caveats, there are 2 ways to consider the results of the 3 meta-analyses undertaken by Wen et al [9], Jones et al [5], and this study:

1. There is insufficient resolution to identify associations—if they exist—at the aggregate data level.
2. Although confidence limits might be reduced by adding further studies, the meta-analyses are essentially correct, and the notion of protocol burden imposed on participants has little to no impact on compliance [4,5].

In studies using EMA, the issue of what constitutes an acceptable rate of compliance or missing data is debatable. Although several studies included in this review cite a criterion or commonly used threshold of 80%, we, similar to Jones et al [5], could not identify the derivation of this criterion. For authors currently planning, conducting, or writing papers or protocols on EMA to monitor health-related behaviors of psychological constructs, adequate recording and reporting of compliance data following recommendations by Liao et al [10] and Heron et al [8] should enable future meta-analyses to explore protocol factors affecting participant compliance rates.

This systematic review prospectively aimed to sample a decade of mEMA use (protocol registered in November 2016; sampling frame of 2006 to 2016) in observational studies including adults from clinical and nonclinical populations. As one of the first EMA reporting documents was published in 2002 [17], this sampling frame assumed that researchers planning or reporting studies including mEMA would be aware of these reporting recommendations. The time frame required for the uptake of EMA reporting recommendations is unknown, although estimates of the time required for uptake of translational research ranges between 2 and 17 years [120]. Our sampling frame and review, however, does not capture studies published from 2017 to date. It is possible that more recent publications differ from those included in our review (greater mobile phone use, better reporting of mEMA schedules, and compliance).

There are no universally accepted recommendations concerning the updating of systematic searches or incorporation of the newer studies into the review results. Systematic reviews—depending on the specific question and volume of studies eligible for inclusion—are time- and labor-intensive. For larger reviews, it is not uncommon for these to take >2 years [121], with updates of Cochrane Collaboration systematic reviews taking up to 3.3 years [122]. The current Cochrane Collaboration policy infers that the decision to update a systematic review should consider the importance of the review question and the volume of new information (studies) [122]. Early in the review process (postsearch completion), 2 papers were identified, published in 2016 [10] and 2017 [8], providing updated recommendations for EMA reporting. Although the volume of mEMA studies published from 2017 is substantial and growing, we opted not to undertake an updated search/meta-analysis to *quarantine* mEMA studies published before the availability of the more recent EMA reporting recommendations.

Strengths and Limitations

This review was strengthened by the broad eligibility criteria used, including studies across nonclinical and clinical contexts in adults. The meta-analysis method was replicated from previous studies [5,9], enabling direct comparison of findings. To the best of the authors' knowledge, this review is the first to propose and explore *burden* as a compound effect of the

various EMA factors (monitoring duration, prompt frequency and prompt type, item per prompt) on participant compliance. We have proposed this novel metric as a starting point for conversations, critique, and further development. In its current form, the burden metric does not include all factors likely to contribute to burden (unfamiliarity with technology, adjunctive use of wearable technologies such as accelerometers), the proposed weighting is rudimentary, and the accuracy of study design features was not confirmed by the study authors.

Limitations of this review include a search strategy focused on the use of mEMA and excluding interventions delivered using EMA (EMI). Consequently, the findings of this review should not be extrapolated or assumed to be similar in studies using EMI. Most studies included in this review provided a clear statement of the primary outcome of interest within each observational study, and we are confident that our categorization of primary mEMA targets is defensible. However, when observational studies did not clearly identify or infer a primary outcome of interest and given mEMA survey items can include multiple items for both self-reported behavioral and psychological constructs, for a small number of studies, misclassification may exist with respect to categorization of mEMA targets as primary or secondary. In the absence of explicit statements by the authors on the number of items within each standard notification, we adopted a conservative approach by estimating the minimum compulsory number of items based on either the information provided by authors within publications or reviewing the instruments reported by authors for inclusion within surveys. The impact of including only studies published in English is unknown.

Conclusions

This review suggests that there is substantial interest in the use of mEMA in adults to collect self-reported health-related behavior and psychological construct data in nonclinical and clinical contexts. Across mEMA studies, there was considerable heterogeneity in protocol design, which may reflect a concerted effort by researchers to tailor mEMA protocols for the intended target and/or population to facilitate compliance. However, the number of studies reporting participant compliance with EMA is concerning. As a result of no or underreporting of compliance, pooled compliance rates may be skewed in favor of overall higher EMA compliance rates. This may dampen associations between compliance rates and EMA protocol factors or burden, making it difficult to ascertain which, if any, protocol factors (such as prompt frequency and number of items within prompts, as identified in this analysis) improve compliance and data collection.

Acknowledgments

The authors sincerely thank Dr Cheng Wen and Dr Andrew Jones for their guidance regarding the meta-analysis process. This study was supported by the University of South Australian High Achiever Vacation scholarship scheme (authors AG and JI). This study was not sponsored.

Authors' Contributions

All authors contributed to this systematic review through the initiation and development of the original protocol (MW, HL, and FF), search and screening (AG and JI), data extraction (AG, JI, MW, HL, FF, and KF), synthesis and meta-analysis (KF, HL, FF, and MW), manuscript development, and final review (MW, HL, KF, FF, AG, and JI).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Ecological momentary assessment (EMA) population and compliance characteristics for studies included within review.
[DOCX File , 81 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Summary of mobile ecological momentary assessment (mEMA) targets and participant characteristics in nonclinical and clinical mEMA studies.
[DOCX File , 67 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Summary of mobile ecological momentary assessment protocols.
[DOCX File , 63 KB-Multimedia Appendix 3]

References

1. Burke LE, Shiffman S, Music E, Styn MA, Kriska A, Smailagic A, et al. Ecological momentary assessment in behavioral research: addressing technological and human participant challenges. J Med Internet Res 2017 Mar 15;19(3):e77 [FREE Full text] [doi: 10.2196/jmir.7138] [Medline: 28298264]

2. Dunton GF. Ecological momentary assessment in physical activity research. Exerc Sport Sci Rev 2017 Dec;45(1):48-54 [FREE Full text] [doi: 10.1249/JES.0000000000000092] [Medline: 27741022]
3. Rintala A, Wampers M, Myin-Germeys I, Viechtbauer W. Response compliance and predictors thereof in studies using the experience sampling method. Psychol Assess 2019 Feb;31(2):226-235. [doi: 10.1037/pas0000662] [Medline: 30394762]
4. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
5. Jones A, Remmerswaal D, Verveer I, Robinson E, Franken IH, Wen CK, et al. Compliance with ecological momentary assessment protocols in substance users: a meta-analysis. Addiction 2019 Apr;114(4):609-619. [doi: 10.1111/add.14503] [Medline: 30461120]
6. Freedman MJ, Lester KM, McNamara C, Milby JB, Schumacher JE. Cell phones for ecological momentary assessment with cocaine-addicted homeless patients in treatment. J Subst Abuse Treat 2006 Mar;30(2):105-111. [doi: 10.1016/j.jsat.2005.10.005] [Medline: 16490673]
7. Baltasar-Tello I, Miguélez-Fernández C, Peñuelas-Calvo I, Carballo JJ. Ecological momentary assessment and mood disorders in children and adolescents: a systematic review. Curr Psychiatry Rep 2018 Aug 01;20(8):66. [doi: 10.1007/s11920-018-0913-z] [Medline: 30069650]
8. Heron KE, Everhart RS, McHale SM, Smyth JM. Using mobile-technology-based ecological momentary assessment (EMA) methods with youth: a systematic review and recommendations. J Pediatr Psychol 2017 Nov 01;42(10):1087-1107. [doi: 10.1093/jpepsy/jsx078] [Medline: 28475765]
9. Wen CKF, Schneider S, Stone AA, Spruijt-Metz D. Compliance with mobile ecological momentary assessment protocols in children and adolescents: a systematic review and meta-analysis. J Med Internet Res 2017 Apr 26;19(4):e132 [FREE Full text] [doi: 10.2196/jmir.6641] [Medline: 28446418]
10. Liao Y, Skelton K, Dunton G, Bruening M. A systematic review of methods and procedures used in ecological momentary assessments of diet and physical activity research in youth: an adapted STROBE checklist for reporting EMA studies (CREMAS). J Med Internet Res 2016 Jun 21;18(6):e151 [FREE Full text] [doi: 10.2196/jmir.4954] [Medline: 27328833]
11. Schembre SM, Liao Y, O'Connor SG, Hingle MD, Shen S, Hamoy KG, et al. Mobile ecological momentary diet assessment methods for behavioral research: systematic review. JMIR Mhealth Uhealth 2018 Nov 20;6(11):e11170 [FREE Full text] [doi: 10.2196/11170] [Medline: 30459148]
12. Miguelez-Fernandez C, de Leon SJ, Baltasar-Tello I, Peñuelas-Calvo I, Barrigon ML, Capdevila AS, et al. Evaluating attention-deficit/hyperactivity disorder using ecological momentary assessment: a systematic review. Atten Defic Hyperact Disord 2018 Dec;10(4):247-265. [doi: 10.1007/s12402-018-0261-1] [Medline: 30132248]
13. Moore RC, Swendsen J, Depp CA. Applications for self-administered mobile cognitive assessments in clinical research: a systematic review. Int J Methods Psychiatr Res 2017 Dec;26(4):e1562 [FREE Full text] [doi: 10.1002/mpr.1562] [Medline: 28370881]
14. Bos FM, Schoevers RA, aan het Rot M. Experience sampling and ecological momentary assessment studies in psychopharmacology: a systematic review. Eur Neuropsychopharmacol 2015 Nov;25(11):1853-1864. [doi: 10.1016/j.euroneuro.2015.08.008] [Medline: 26336868]
15. May M, Junghaenel DU, Ono M, Stone AA, Schneider S. Ecological momentary assessment methodology in chronic pain research: a systematic review. J Pain 2018 Jul;19(7):699-716. [doi: 10.1016/j.jpain.2018.01.006] [Medline: 29371113]
16. Bell IH, Lim MH, Rossell SL, Thomas N. Ecological momentary assessment and intervention in the treatment of psychotic disorders: a systematic review. Psychiatr Serv 2017 Nov 01;68(11):1172-1181. [doi: 10.1176/appi.ps.201600523] [Medline: 28669284]
17. Stone AA, Shiffman S. Capturing momentary, self-report data: a proposal for reporting guidelines. Ann Behav Med 2002;24(3):236-243. [Medline: 12173681]
18. Sampson M, McGowan J, Cogo E, Grimshaw J, Moher D, Lefebvre C. An evidence-based practice guideline for the peer review of electronic search strategies. J Clin Epidemiol 2009 Sep;62(9):944-952. [doi: 10.1016/j.jclinepi.2008.10.012] [Medline: 19230612]
19. Solhan MB, Trull TJ, Jahng S, Wood PK. Clinical assessment of affective instability: comparing EMA indices, questionnaire reports, and retrospective recall. Psychol Assess 2009 Sep;21(3):425-436 [FREE Full text] [doi: 10.1037/a0016869] [Medline: 19719353]
20. Sorbi MJ, Peters ML, Kruise DA, Maas CJ, Kerssens JJ, Verhaak PF, et al. Electronic momentary assessment in chronic pain I: psychological pain responses as predictors of pain intensity. Clin J Pain 2006 Jan;22(1):55-66. [doi: 10.1097/01.ajp.0000148624.46756.fa] [Medline: 16340594]
21. Johnson EI, Grondin O, Barrault M, Faytout M, Helbig S, Husky M, et al. Computerized ambulatory monitoring in psychiatry: a multi-site collaborative study of acceptability, compliance, and reactivity. Int J Methods Psychiatr Res 2009;18(1):48-57. [doi: 10.1002/mpr.276] [Medline: 19195050]
22. Kimhy D, Vakhrusheva J, Khan S, Chang RW, Hansen MC, Ballon JS, et al. Emotional granularity and social functioning in individuals with schizophrenia: an experience sampling study. J Psychiatr Res 2014 Jun;53:141-148 [FREE Full text] [doi: 10.1016/j.jpsychires.2014.01.020] [Medline: 24561000]

23. Ritz T, Rosenfield D, Steptoe A. Physical activity, lung function, and shortness of breath in the daily life of individuals with asthma. Chest 2010 Oct;138(4):913-918. [doi: 10.1378/chest.08-3073] [Medline: 20472861]
24. MacKerron G, Mourato S. Happiness is greater in natural environments. Glob Environ Change 2013 Oct;23(5):992-1000. [doi: 10.1016/j.gloenvcha.2013.03.010]
25. Trampe D, Quoidbach J, Taquet M. Emotions in everyday life. PLoS One 2015;10(12):e0145450 [FREE Full text] [doi: 10.1371/journal.pone.0145450] [Medline: 26698124]
26. Jean FA, Swendsen JD, Sibon I, Fehér K, Husky M. Daily life behaviors and depression risk following stroke: a preliminary study using ecological momentary assessment. J Geriatr Psychiatry Neurol 2013 Sep;26(3):138-143. [doi: 10.1177/0891988713484193] [Medline: 23584854]
27. Granholm E, Ben-Zeev D, Fulford D, Swendsen J. Ecological momentary assessment of social functioning in schizophrenia: impact of performance appraisals and affect on social interactions. Schizophr Res 2013 Apr;145(1-3):120-124 [FREE Full text] [doi: 10.1016/j.schres.2013.01.005] [Medline: 23402693]
28. Wray TB, Kahler CW, Monti PM. Using Ecological Momentary Assessment (EMA) to study sex events among very high-risk men who have sex with men (MSM). AIDS Behav 2016 Jan 8:2231-2242. [doi: 10.1007/s10461-015-1272-y] [Medline: 26746212]
29. Zawadzki MJ, Smyth JM, Costigan HJ. Real-time associations between engaging in leisure and daily health and well-being. Ann Behav Med 2015 Aug;49(4):605-615. [doi: 10.1007/s12160-015-9694-3] [Medline: 25724635]
30. Armey MF, Crowther JH, Miller IW. Changes in ecological momentary assessment reported affect associated with episodes of nonsuicidal self-injury. Behav Ther 2011 Dec;42(4):579-588. [doi: 10.1016/j.beth.2011.01.002] [Medline: 22035987]
31. Cook PF, McElwain CJ, Bradley-Springer LA. Feasibility of a daily electronic survey to study prevention behavior with HIV-infected individuals. Res Nurs Health 2010 Jun;33(3):221-234. [doi: 10.1002/nur.20381] [Medline: 20499392]
32. Kaplan SE, Ohrbach R. Self-report of waking-state oral parafunctional behaviors in the natural environment. J Oral Facial Pain Headache 2016;30(2):107-119 [FREE Full text] [doi: 10.11607/ofph.1592] [Medline: 27128474]
33. Hacker ED, Ferrans CE. Ecological momentary assessment of fatigue in patients receiving intensive cancer therapy. J Pain Symptom Manage 2007 Mar;33(3):267-275. [doi: 10.1016/j.jpainsymman.2006.08.007] [Medline: 17349496]
34. Chandra S, Scharf D, Shiffman S. Within-day temporal patterns of smoking, withdrawal symptoms, and craving. Drug Alcohol Depend 2011 Sep 01;117(2-3):118-125 [FREE Full text] [doi: 10.1016/j.drugalcdep.2010.12.027] [Medline: 21324611]
35. Yoshiuchi K, Cook DB, Ohashi K, Kumano H, Kuboki T, Yamamoto Y, et al. A real-time assessment of the effect of exercise in chronic fatigue syndrome. Physiol Behav 2007 Dec 05;92(5):963-968 [FREE Full text] [doi: 10.1016/j.physbeh.2007.07.001] [Medline: 17655887]
36. Thomas JG, Doshi S, Crosby RD, Lowe MR. Ecological momentary assessment of obesogenic eating behavior: combining person-specific and environmental predictors. Obesity (Silver Spring) 2011 Aug;19(8):1574-1579 [FREE Full text] [doi: 10.1038/oby.2010.335] [Medline: 21273995]
37. Ottaviani C, Medea B, Lonigro A, Tarvainen M, Couyoumdjian A. Cognitive rigidity is mirrored by autonomic inflexibility in daily life perseverative cognition. Biol Psychol 2015 Apr;107:24-30. [doi: 10.1016/j.biopsycho.2015.02.011] [Medline: 25749107]
38. Kuntsche E, Labhart F. ICAT: development of an internet-based data collection method for ecological momentary assessment using personal cell phones. Eur J Psychol Assess 2013;29(2):140-148 [FREE Full text] [doi: 10.1027/1015-5759/a000137] [Medline: 24285917]
39. Ambwani S, Roche MJ, Minnick AM, Pincus AL. Negative affect, interpersonal perception, and binge eating behavior: an experience sampling study. Int J Eat Disord 2015 Sep;48(6):715-726. [doi: 10.1002/eat.22410] [Medline: 25946681]
40. Andersson C, Söderpalm Gordh AH, Berglund M. Use of real-time interactive voice response in a study of stress and alcohol consumption. Alcohol Clin Exp Res 2007 Nov;31(11):1908-1912. [doi: 10.1111/j.1530-0277.2007.00520.x] [Medline: 17949395]
41. Berg KC, Peterson CB, Crosby RD, Cao L, Crow SJ, Engel SG, et al. Relationship between daily affect and overeating-only, loss of control eating-only, and binge eating episodes in obese adults. Psychiatry Res 2014 Jan 30;215(1):185-191 [FREE Full text] [doi: 10.1016/j.psychres.2013.08.023] [Medline: 24200217]
42. Buckner JD, Crosby RD, Wonderlich SA, Schmidt NB. Social anxiety and cannabis use: an analysis from ecological momentary assessment. J Anxiety Disord 2012 Mar;26(2):297-304 [FREE Full text] [doi: 10.1016/j.janxdis.2011.12.006] [Medline: 22246109]
43. Burt SA, Donnellan MB. Evidence that the subtypes of antisocial behavior questionnaire (STAB) predicts momentary reports of acting-out behaviors. Pers Individ Dif 2010 Jun;48(8):917-920. [doi: 10.1016/j.paid.2010.02.021]
44. Businelle MS, Ma P, Kendzor DE, Reitzel LR, Chen M, Lam CY, et al. Predicting quit attempts among homeless smokers seeking cessation treatment: an ecological momentary assessment study. Nicotine Tob Res 2014 Oct;16(10):1371-1378 [FREE Full text] [doi: 10.1093/ntr/ntu088] [Medline: 24893602]
45. Clasen PC, Fisher AJ, Beevers CG. Mood-reactive self-esteem and depression vulnerability: person-specific symptom dynamics via smart phone assessment. PLoS One 2015;10(7):e0129774 [FREE Full text] [doi: 10.1371/journal.pone.0129774] [Medline: 26131724]

46. Courvoisier DS, Eid M, Lischetzke T, Schreiber WH. Psychometric properties of a computerized mobile phone method for assessing mood in daily life. Emotion 2010 Feb;10(1):115-124. [doi: 10.1037/a0017813] [Medline: 20141308]
47. Doherty ST, Lemieux CJ, Canally C. Tracking human activity and well-being in natural environments using wearable sensors and experience sampling. Soc Sci Med 2014 Apr;106:83-92. [doi: 10.1016/j.socscimed.2014.01.048] [Medline: 24549253]
48. Fitzsimmons-Craft EE, Bardone-Cone AM, Crosby RD, Engel SG, Wonderlich SA, Bulik CM. Mediators of the relationship between thin-ideal internalization and body dissatisfaction in the natural environment. Body Image 2016 Sep;18:113-122 [FREE Full text] [doi: 10.1016/j.bodyim.2016.06.006] [Medline: 27391791]
49. Fonareva I, Amen AM, Ellingson RM, Oken BS. Differences in stress-related ratings between research center and home environments in dementia caregivers using ecological momentary assessment. Int Psychogeriatr 2012 Jan;24(1):90-98 [FREE Full text] [doi: 10.1017/S1041610211001414] [Medline: 21777503]
50. Heron KE, Scott SB, Sliwinski MJ, Smyth JM. Eating behaviors and negative affect in college women's everyday lives. Int J Eat Disord 2014 Dec;47(8):853-859 [FREE Full text] [doi: 10.1002/eat.22292] [Medline: 24797029]
51. Heron KE, Smyth JM. Body image discrepancy and negative affect in women's everyday lives: an ecological momentary assessment evaluation of self-discrepancy theory. J Soc Clin Psychol 2013 Mar;32(3):276-295. [doi: 10.1521/jscp.2013.32.3.276]
52. Hofmann W, Wisneski DC, Brandt MJ, Skitka LJ. Morality in everyday life. Science 2014 Sep 12;345(6202):1340-1343 [FREE Full text] [doi: 10.1126/science.1251560] [Medline: 25214626]
53. Juth V, Dickerson SS, Zoccola PM, Lam S. Understanding the utility of emotional approach coping: evidence from a laboratory stressor and daily life. Anxiety Stress Coping 2015;28(1):50-70 [FREE Full text] [doi: 10.1080/10615806.2014.921912] [Medline: 24804564]
54. Kashdan TB, Collins RL. Social anxiety and the experience of positive emotion and anger in everyday life: an ecological momentary assessment approach. Anxiety Stress Coping 2010 May;23(3):259-272. [doi: 10.1080/10615800802641950] [Medline: 19326272]
55. Kirchner TR, Cantrell J, Anesetti-Rothermel A, Ganz O, Vallone DM, Abrams DB. Geospatial exposure to point-of-sale tobacco: real-time craving and smoking-cessation outcomes. Am J Prev Med 2013 Oct;45(4):379-385 [FREE Full text] [doi: 10.1016/j.amepre.2013.05.016] [Medline: 24050412]
56. Komulainen E, Meskanen K, Lipsanen J, Lahti JM, Jylhä P, Melartin T, et al. The effect of personality on daily life emotional processes. PLoS One 2014;9(10):e110907 [FREE Full text] [doi: 10.1371/journal.pone.0110907] [Medline: 25343494]
57. Lange S, Süß H. Measuring slips and lapses when they occur - ambulatory assessment in application to cognitive failures. Conscious Cogn 2014 Feb;24:1-11. [doi: 10.1016/j.concog.2013.12.008] [Medline: 24384496]
58. Liao Y, Intille SS, Dunton GF. Using ecological momentary assessment to understand where and with whom adults' physical and sedentary activity occur. Int J Behav Med 2015 Feb;22(1):51-61. [doi: 10.1007/s12529-014-9400-z] [Medline: 24639067]
59. Ramirez J, Miranda R. Alcohol craving in adolescents: bridging the laboratory and natural environment. Psychopharmacology (Berl) 2014 Apr;231(8):1841-1851 [FREE Full text] [doi: 10.1007/s00213-013-3372-6] [Medline: 24363093]
60. Riediger M, Wrzus C, Schmiedek F, Wagner GG, Lindenberger U. Is seeking bad mood cognitively demanding? Contra-hedonic orientation and working-memory capacity in everyday life. Emotion 2011 Jun;11(3):656-665. [doi: 10.1037/a0022756] [Medline: 21534659]
61. Robertson BM, Piasecki TM, Slutske WS, Wood PK, Sher KJ, Shiffman S, et al. Validity of the hangover symptoms scale: evidence from an electronic diary study. Alcohol Clin Exp Res 2012;36(1):171-177. [doi: 10.1111/j.1530-0277.2011.01592.x] [Medline: 21762183]
62. Rowan PJ, Cofta-Woerpel L, Mazas CA, Vidrine JI, Reitzel LR, Cinciripini PM, et al. Evaluating reactivity to ecological momentary assessment during smoking cessation. Exp Clin Psychopharmacol 2007 Aug;15(4):382-389. [doi: 10.1037/1064-1297.15.4.382] [Medline: 17696685]
63. Rutledge T, Stucky E, Dollarhide A, Shively M, Jain S, Wolfson T, et al. A real-time assessment of work stress in physicians and nurses. Health Psychol 2009 Mar;28(2):194-200. [doi: 10.1037/a0013145] [Medline: 19290711]
64. Schuster RM, Mermelstein RJ, Hedeker D. Ecological momentary assessment of working memory under conditions of simultaneous marijuana and tobacco use. Addiction 2016 Aug;111(8):1466-1476. [doi: 10.1111/add.13342] [Medline: 26857917]
65. Seto E, Hua J, Wu L, Shia V, Eom S, Wang M, et al. Models of individual dietary behavior based on smartphone data: the influence of routine, physical activity, emotion, and food environment. PLoS One 2016;11(4):e0153085 [FREE Full text] [doi: 10.1371/journal.pone.0153085] [Medline: 27049852]
66. Setodji CM, Martino SC, Scharf DM, Shadel WG. Quantifying the persistence of pro-smoking media effects on college students' smoking risk. J Adolesc Health 2014 Apr;54(4):474-480 [FREE Full text] [doi: 10.1016/j.jadohealth.2013.09.011] [Medline: 24268361]
67. Shiffman S, Balabanis MH, Gwaltney CJ, Paty JA, Gnys M, Kassel JD, et al. Prediction of lapse from associations between smoking and situational antecedents assessed by ecological momentary assessment. Drug Alcohol Depend 2007 Dec 1;91(2-3):159-168 [FREE Full text] [doi: 10.1016/j.drugalcdep.2007.05.017] [Medline: 17628353]

68. Simons JS, Emery NN, Simons RM, Wills TA, Webb MK. Effects of alcohol, rumination, and gender on the time course of negative affect. Cogn Emot 2017 Nov;31(7):1405-1418 [FREE Full text] [doi: 10.1080/02699931.2016.1226162] [Medline: 27609298]
69. Spook JE, Paulussen T, Kok G, Van EP. Monitoring dietary intake and physical activity electronically: feasibility, usability, and ecological validity of a mobile-based Ecological Momentary Assessment tool. J Med Internet Res 2013;15(9):e214 [FREE Full text] [doi: 10.2196/jmir.2617] [Medline: 24067298]
70. Thielsch C, Andor T, Ehring T. Do metacognitions and intolerance of uncertainty predict worry in everyday life? An ecological momentary assessment study. Behav Ther 2015 Jul;46(4):532-543. [doi: 10.1016/j.beth.2015.05.001] [Medline: 26163716]
71. Tiplady B, Oshinowo B, Thomson J, Drummond GB. Alcohol and cognitive function: assessment in everyday life and laboratory settings using mobile phones. Alcohol Clin Exp Res 2009 Dec;33(12):2094-2102. [doi: 10.1111/j.1530-0277.2009.01049.x] [Medline: 19740132]
72. Waters AJ, Szeto EH, Wetter DW, Cinciripini PM, Robinson JD, Li Y. Cognition and craving during smoking cessation: an ecological momentary assessment study. Nicotine Tob Res 2014 May;16 Suppl 2:S111-S118 [FREE Full text] [doi: 10.1093/ntr/ntt108] [Medline: 23901053]
73. Witkiewitz K, Desai SA, Steckler G, Jackson KM, Bowen S, Leigh BC, et al. Concurrent drinking and smoking among college students: an event-level analysis. Psychol Addict Behav 2012 Sep;26(3):649-654 [FREE Full text] [doi: 10.1037/a0025363] [Medline: 21895348]
74. Zenk SN, Horoi I, McDonald A, Corte C, Riley B, Odoms-Young AM. Ecological momentary assessment of environmental and personal factors and snack food intake in African American women. Appetite 2014 Dec;83:333-341. [doi: 10.1016/j.appet.2014.09.008] [Medline: 25239402]
75. Aaron LA, Turner JA, Mancl LA, Sawchuk CN, Huggins KH, Truelove EL. Daily pain coping among patients with chronic temporomandibular disorder pain: an electronic diary study. J Orofac Pain 2006;20(2):125-137. [Medline: 16708830]
76. Dewey D, McDonald MK, Brown WJ, Boyd SJ, Bunnell BE, Schuldberg D. The impact of ecological momentary assessment on posttraumatic stress symptom trajectory. Psychiatry Res 2015 Dec 15;230(2):300-303. [doi: 10.1016/j.psychres.2015.09.009] [Medline: 26381184]
77. Dhingra LK, Homel P, Grossman B, Chen J, Scharaga E, Calamita S, et al. Ecological momentary assessment of smoking behavior in persistent pain patients. Clin J Pain 2014 Mar;30(3):205-213. [doi: 10.1097/AJP.0b013e31829821c7] [Medline: 23689351]
78. Epstein DH, Tyburski M, Craig IM, Phillips KA, Jobes ML, Vahabzadeh M, et al. Real-time tracking of neighborhood surroundings and mood in urban drug misusers: application of a new method to study behavior in its geographical context. Drug Alcohol Depend 2014 Jan 1;134:22-29 [FREE Full text] [doi: 10.1016/j.drugalcdep.2013.09.007] [Medline: 24332365]
79. Fitzsimmons-Craft EE, Accurso EC, Ciao AC, Crosby RD, Cao L, Pisetsky EM, et al. Restrictive eating in anorexia nervosa: examining maintenance and consequences in the natural environment. Int J Eat Disord 2015 Nov;48(7):923-931 [FREE Full text] [doi: 10.1002/eat.22439] [Medline: 26310991]
80. Hachizuka M, Yoshiuchi K, Yamamoto Y, Iwase S, Nakagawa K, Kawagoe K, et al. Development of a personal digital assistant (PDA) system to collect symptom information from home hospice patients. J Palliat Med 2010 Jun;13(6):647-651. [doi: 10.1089/jpm.2009.0350] [Medline: 20509795]
81. Juengst SB, Graham KM, Pulantara IW, McCue M, Whyte EM, Dicianno BE, et al. Pilot feasibility of an mHealth system for conducting ecological momentary assessment of mood-related symptoms following traumatic brain injury. Brain Inj 2015 Aug;29(11):1351-1361. [doi: 10.3109/02699052.2015.1045031] [Medline: 26287756]
82. Kulich K, Keininger DL, Tiplady B, Banerji D. Symptoms and impact of COPD assessed by an electronic diary in patients with moderate-to-severe COPD: psychometric results from the SHINE study. Int J Chron Obstruct Pulmon Dis 2015;10:79-94 [FREE Full text] [doi: 10.2147/COPD.S73092] [Medline: 25609942]
83. Kuroi R, Minakuchi H, Hara ES, Kawakami A, Maekawa K, Okada H, et al. A risk factor analysis of accumulated postoperative pain and swelling sensation after dental implant surgery using a cellular phone-based real-time assessment. J Prosthodont Res 2015 Jul;59(3):194-198. [doi: 10.1016/j.jpor.2015.05.003] [Medline: 26077378]
84. Lavender JM, De Young KP, Wonderlich SA, Crosby RD, Engel SG, Mitchell JE, et al. Daily patterns of anxiety in anorexia nervosa: associations with eating disorder behaviors in the natural environment. J Abnorm Psychol 2013 Aug;122(3):672-683 [FREE Full text] [doi: 10.1037/a0031823] [Medline: 23647124]
85. Merwin RM, Dmitrieva NO, Honeycutt LK, Moskovich AA, Lane JD, Zucker NL, et al. Momentary predictors of insulin restriction among adults with type 1 diabetes and eating disorder symptomatology. Diabetes Care 2015 Nov;38(11):2025-2032 [FREE Full text] [doi: 10.2337/dc15-0753] [Medline: 26384389]
86. Munsch S, Meyer AH, Milenkovic N, Schlup B, Margraf J, Wilhelm FH. Ecological momentary assessment to evaluate cognitive-behavioral treatment for binge eating disorder. Int J Eat Disord 2009 Nov;42(7):648-657. [doi: 10.1002/eat.20657] [Medline: 19197978]
87. Okifuji A, Bradshaw DH, Donaldson GW, Turk DC. Sequential analyses of daily symptoms in women with fibromyalgia syndrome. J Pain 2011 Jan;12(1):84-93 [FREE Full text] [doi: 10.1016/j.jpain.2010.05.003] [Medline: 20591745]

88. Sohl SJ, Friedberg F. Memory for fatigue in chronic fatigue syndrome: relationships to fatigue variability, catastrophizing, and negative affect. Behav Med 2008;34(1):29-38 [FREE Full text] [doi: 10.3200/BMED.34.1.29-38] [Medline: 18400687]
89. Thielsch C, Ehring T, Nestler S, Wolters J, Kopei I, Rist F, et al. Metacognitions, worry and sleep in everyday life: studying bidirectional pathways using Ecological Momentary Assessment in GAD patients. J Anxiety Disord 2015 Jun;33:53-61. [doi: 10.1016/j.janxdis.2015.04.007] [Medline: 26005837]
90. Weaver A, Young AM, Rowntree J, Townsend N, Pearson S, Smith J, et al. Application of mobile phone technology for managing chemotherapy-associated side-effects. Ann Oncol 2007 Nov;18(11):1887-1892 [FREE Full text] [doi: 10.1093/annonc/mdm354] [Medline: 17921245]
91. Asselbergs J, Ruwaard J, Ejdys M, Schrader N, Sijbrandij M, Riper H. Mobile phone-based unobtrusive ecological momentary assessment of day-to-day mood: an explorative study. J Med Internet Res 2016 Mar 29;18(3):e72 [FREE Full text] [doi: 10.2196/jmir.5505] [Medline: 27025287]
92. Dunbar MS, Scharf D, Kirchner T, Shiffman S. Do smokers crave cigarettes in some smoking situations more than others? Situational correlates of craving when smoking. Nicotine Tob Res 2010 Mar;12(3):226-234 [FREE Full text] [doi: 10.1093/ntr/ntp198] [Medline: 20133379]
93. Hofmann W, Adriaanse M, Vohs KD, Baumeister RF. Dieting and the self-control of eating in everyday environments: an experience sampling study. Br J Health Psychol 2014 Sep;19(3):523-539 [FREE Full text] [doi: 10.1111/bjhp.12053] [Medline: 23751109]
94. Hughes CD, Gunthert K, Wenze S, German R. The subscale specificity of the Affective Control Scale: Ecological validity and predictive validity of feared emotions. Motiv Emot 2015 Jun 5;39(6):984-992. [doi: 10.1007/s11031-015-9497-7]
95. Huguet A, McGrath PJ, Wheaton M, Mackinnon SP, Rozario S, Tougas ME, et al. Testing the feasibility and psychometric properties of a mobile diary (myWHI) in adolescents and young adults with headaches. JMIR Mhealth Uhealth 2015 May 08;3(2):e39 [FREE Full text] [doi: 10.2196/mhealth.3879] [Medline: 25956377]
96. Kanning M, Hansen S. Need satisfaction moderates the association between physical activity and affective states in adults aged 50+: an activity-triggered ambulatory assessment. Ann Behav Med 2017 Feb;51(1):18-29 [FREE Full text] [doi: 10.1007/s12160-016-9824-6] [Medline: 27539030]
97. Kashdan TB, Farmer AS. Differentiating emotions across contexts: comparing adults with and without social anxiety disorder using random, social interaction, and daily experience sampling. Emotion 2014 Jun;14(3):629-638 [FREE Full text] [doi: 10.1037/a0035796] [Medline: 24512246]
98. Kwapil TR, Barrantes-Vidal N, Armistead MS, Hope GA, Brown LH, Silvia PJ, et al. The expression of bipolar spectrum psychopathology in daily life. J Affect Disord 2011 Apr;130(1-2):166-170 [FREE Full text] [doi: 10.1016/j.jad.2010.10.025] [Medline: 21056476]
99. Schüz B, Bower J, Ferguson SG. Stimulus control and affect in dietary behaviours. An intensive longitudinal study. Appetite 2015 Apr;87:310-317. [doi: 10.1016/j.appet.2015.01.002] [Medline: 25579222]
100. Schüz N, Walters JAE, Frandsen M, Bower J, Ferguson SG. Compliance with an EMA monitoring protocol and its relationship with participant and smoking characteristics. Nicotine Tob Res 2014 May;16(Suppl 2):S88-S92. [doi: 10.1093/ntr/ntt142] [Medline: 24052500]
101. Schwerdtfeger A, Eberhardt R, Chmitorz A, Schaller E. Momentary affect predicts bodily movement in daily life: an ambulatory monitoring study. J Sport Exerc Psychol 2010 Oct;32(5):674-693. [Medline: 20980710]
102. Warthen MW, Tiffany ST. Evaluation of cue reactivity in the natural environment of smokers using ecological momentary assessment. Exp Clin Psychopharmacol 2009 Apr;17(2):70-77 [FREE Full text] [doi: 10.1037/a0015617] [Medline: 19331483]
103. Vasconcelos E Sa D, Wearden A, Hartley S, Emsley R, Barrowclough C. Expressed Emotion and behaviourally controlling interactions in the daily life of dyads experiencing psychosis. Psychiatry Res 2016 Nov 30;245:406-413. [doi: 10.1016/j.psychres.2016.08.060] [Medline: 27611070]
104. Ebner-Priemer UW, Kuo J, Schlotz W, Kleindienst N, Rosenthal MZ, Detterer L, et al. Distress and affective dysregulation in patients with borderline personality disorder: a psychophysiological ambulatory monitoring study. J Nerv Ment Dis 2008 Apr;196(4):314-320. [doi: 10.1097/NMD.0b013e31816a493f] [Medline: 18414126]
105. Green KT, Dennis PA, Neal LC, Hobkirk AL, Hicks TA, Watkins LL, et al. Exploring the relationship between posttraumatic stress disorder symptoms and momentary heart rate variability. J Psychosom Res 2016 Mar;82:31-34 [FREE Full text] [doi: 10.1016/j.jpsychores.2016.01.003] [Medline: 26944396]
106. Mazure CM, Weinberger AH, Pittman B, Sibon I, Swendsen J. Gender and stress in predicting depressive symptoms following stroke. Cerebrovasc Dis 2014;38(4):240-246 [FREE Full text] [doi: 10.1159/000365838] [Medline: 25401293]
107. Kothari DJ, Davis MC, Yeung EW, Tennen HA. Positive affect and pain: mediators of the within-day relation linking sleep quality to activity interference in fibromyalgia. Pain 2015 Mar;156(3):540-546 [FREE Full text] [doi: 10.1097/01.j.pain.0000460324.18138.0a] [Medline: 25679472]
108. Burns JW, Gerhart JI, Bruehl S, Peterson KM, Smith DA, Porter LS, et al. Anger arousal and behavioral anger regulation in everyday life among patients with chronic low back pain: Relationships to patient pain and function. Health Psychol 2015 May;34(5):547-555 [FREE Full text] [doi: 10.1037/hea0000091] [Medline: 25110843]

109. Smyth JM, Wonderlich SA, Heron KE, Sliwinski MJ, Crosby RD, Mitchell JE, et al. Daily and momentary mood and stress are associated with binge eating and vomiting in bulimia nervosa patients in the natural environment. J Consult Clin Psychol 2007 Aug;75(4):629-638. [doi: 10.1037/0022-006X.75.4.629] [Medline: 17663616]
110. Ainsworth J, Palmier-Claus JE, Machin M, Barrowclough C, Dunn G, Rogers A, et al. A comparison of two delivery modalities of a mobile phone-based assessment for serious mental illness: native smartphone application vs text-messaging only implementations. J Med Internet Res 2013;15(4):e60 [FREE Full text] [doi: 10.2196/jmir.2328] [Medline: 23563184]
111. Tomiyama AJ, Mann T, Comer L. Triggers of eating in everyday life. Appetite 2009 Feb;52(1):72-82 [FREE Full text] [doi: 10.1016/j.appet.2008.08.002] [Medline: 18773931]
112. Luczak SE, Rosen IG, Wall TL. Development of a real-time repeated-measures assessment protocol to capture change over the course of a drinking episode. Alcohol Alcohol 2015 Mar;50(2):180-187 [FREE Full text] [doi: 10.1093/alcalc/agu100] [Medline: 25568142]
113. Monk RL, Heim D, Qureshi A, Price A. 'I have no clue what I drunk last night' using Smartphone technology to compare in-vivo and retrospective self-reports of alcohol consumption. PLoS One 2015;10(5):e0126209 [FREE Full text] [doi: 10.1371/journal.pone.0126209] [Medline: 25992573]
114. Santangelo P, Reinhard I, Mussgay L, Steil R, Sawitzki G, Klein C, et al. Specificity of affective instability in patients with borderline personality disorder compared to posttraumatic stress disorder, bulimia nervosa, and healthy controls. J Abnorm Psychol 2014 Feb;123(1):258-272 [FREE Full text] [doi: 10.1037/a0035619] [Medline: 24661176]
115. Piasecki TM, Cooper ML, Wood PK, Sher KJ, Shiffman S, Heath AC. Dispositional drinking motives: associations with appraised alcohol effects and alcohol consumption in an ecological momentary assessment investigation. Psychol Assess 2014 Jun;26(2):363-369 [FREE Full text] [doi: 10.1037/a0035153] [Medline: 24274049]
116. Goldschmidt AB, Wonderlich SA, Crosby RD, Engel SG, Lavender JM, Peterson CB, et al. Ecological momentary assessment of stressful events and negative affect in bulimia nervosa. J Consult Clin Psychol 2014 Feb;82(1):30-39 [FREE Full text] [doi: 10.1037/a0034974] [Medline: 24219182]
117. Trela CJ, Piasecki TM, Bartholow BD, Heath AC, Sher KJ. The natural expression of individual differences in self-reported level of response to alcohol during ecologically assessed drinking episodes. Psychopharmacology (Berl) 2016 Jun;233(11):2185-2195 [FREE Full text] [doi: 10.1007/s00213-016-4270-5] [Medline: 27037938]
118. Tomko RL, Solhan MB, Carpenter RW, Brown WC, Jahng S, Wood PK, et al. Measuring impulsivity in daily life: the momentary impulsivity scale. Psychol Assess 2014 Jun;26(2):339-349 [FREE Full text] [doi: 10.1037/a0035083] [Medline: 24274047]
119. Ono M, Schneider S, Junghaenel DU, Stone AA. What affects the completion of ecological momentary assessments in chronic pain research? An individual patient data meta-analysis. J Med Internet Res 2019 Feb 05;21(2):e11398 [FREE Full text] [doi: 10.2196/11398] [Medline: 30720437]
120. Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. J R Soc Med 2011 Dec;104(12):510-520 [FREE Full text] [doi: 10.1258/jrsm.2011.110180] [Medline: 22179294]
121. Khangura S, Konnyu K, Cushman R, Grimshaw J, Moher D. Evidence summaries: the evolution of a rapid review approach. Syst Rev 2012;1:10 [FREE Full text] [doi: 10.1186/2046-4053-1-10] [Medline: 22587960]
122. Elkins MR. Updating systematic reviews. J Physiother 2018 Jan;64(1):1-3 [FREE Full text] [doi: 10.1016/j.jphys.2017.11.009] [Medline: 29289593]

Abbreviations

EMA: ecological momentary assessment
ES: effect size
ESS: effective sample size
mEMA: mobile ecological momentary assessment
mEMI: mobile ecological momentary intervention

Edited by G Eysenbach; submitted 12.11.19; peer-reviewed by M May, G Dunton, K Heron; comments to author 23.12.19; revised version received 01.03.20; accepted 31.10.20; published 03.03.21

Please cite as:
Williams MT, Lewthwaite H, Fraysse F, Gajewska A, Ignatavicius J, Ferrar K
Compliance With Mobile Ecological Momentary Assessment of Self-Reported Health-Related Behaviors and Psychological Constructs in Adults: Systematic Review and Meta-analysis
J Med Internet Res 2021;23(3):e17023
URL: https://www.jmir.org/2021/3/e17023
doi: 10.2196/17023
PMID: 33656451

Original Paper

Artificial Intelligence Techniques That May Be Applied to Primary Care Data to Facilitate Earlier Diagnosis of Cancer: Systematic Review

Owain T Jones[1], MPhil; Natalia Calanzani[1], PhD; Smiji Saji[1], MBBCHIR; Stephen W Duffy[2], MSc; Jon Emery[3], DPhil; Willie Hamilton[4], MD; Hardeep Singh[5], MD, MPH; Niek J de Wit[6], MD; Fiona M Walter[1], MD

[1]Primary Care Unit, Department of Public Health & Primary Care, University of Cambridge, Cambridge, United Kingdom

[2]Wolfson Institute for Preventive Medicine, Queen Mary University of London, London, United Kingdom

[3]Centre for Cancer Research and Department of General Practice, University of Melbourne, Victoria, Australia

[4]College of Medicine and Health, University of Exeter, Exeter, United Kingdom

[5]Center for Innovations in Quality, Effectiveness and Safety, Michael E DeBakey Veterans Affairs Medical Center and Baylor College of Medicine, Houston, TX, United States

[6]Julius Center for Health Sciences and Primary Care, UMC Utrecht, Utrecht, Netherlands

Corresponding Author:
Owain T Jones, MPhil
Primary Care Unit
Department of Public Health & Primary Care
University of Cambridge
2 Wort's Causeway
Cambridge, CB1 8RN
United Kingdom
Phone: 44 1223762554
Email: otj24@medschl.cam.ac.uk

Abstract

Background: More than 17 million people worldwide, including 360,000 people in the United Kingdom, were diagnosed with cancer in 2018. Cancer prognosis and disease burden are highly dependent on the disease stage at diagnosis. Most people diagnosed with cancer first present in primary care settings, where improved assessment of the (often vague) presenting symptoms of cancer could lead to earlier detection and improved outcomes for patients. There is accumulating evidence that artificial intelligence (AI) can assist clinicians in making better clinical decisions in some areas of health care.

Objective: This study aimed to systematically review AI techniques that may facilitate earlier diagnosis of cancer and could be applied to primary care electronic health record (EHR) data. The quality of the evidence, the phase of development the AI techniques have reached, the gaps that exist in the evidence, and the potential for use in primary care were evaluated.

Methods: We searched MEDLINE, Embase, SCOPUS, and Web of Science databases from January 01, 2000, to June 11, 2019, and included all studies providing evidence for the accuracy or effectiveness of applying AI techniques for the early detection of cancer, which may be applicable to primary care EHRs. We included all study designs in all settings and languages. These searches were extended through a scoping review of AI-based commercial technologies. The main outcomes assessed were measures of diagnostic accuracy for cancer.

Results: We identified 10,456 studies; 16 studies met the inclusion criteria, representing the data of 3,862,910 patients. A total of 13 studies described the initial development and testing of AI algorithms, and 3 studies described the validation of an AI algorithm in independent data sets. One study was based on prospectively collected data; only 3 studies were based on primary care data. We found no data on implementation barriers or cost-effectiveness. Risk of bias assessment highlighted a wide range of study quality. The additional scoping review of commercial AI technologies identified 21 technologies, only 1 meeting our inclusion criteria. Meta-analysis was not undertaken because of the heterogeneity of AI modalities, data set characteristics, and outcome measures.

Conclusions: AI techniques have been applied to EHR-type data to facilitate early diagnosis of cancer, but their use in primary care settings is still at an early stage of maturity. Further evidence is needed on their performance using primary care data,

implementation barriers, and cost-effectiveness before widespread adoption into routine primary care clinical practice can be recommended.

(J Med Internet Res 2021;23(3):e23483) doi: 10.2196/23483

KEYWORDS

artificial intelligence; machine learning; electronic health records; primary health care; early detection of cancer

Introduction

Background

Cancer control is a global health priority, with 17 million new cases diagnosed worldwide in 2018. In high-income countries such as the United Kingdom, approximately half the population over the age of 50 years will be diagnosed with cancer in their lifetime [1]. Although the National Health Service (NHS) currently spends approximately £1 billion (US $1.37 billion) on cancer diagnostics per year [2], the United Kingdom lags behind comparable European nations with their cancer survival rates [3].

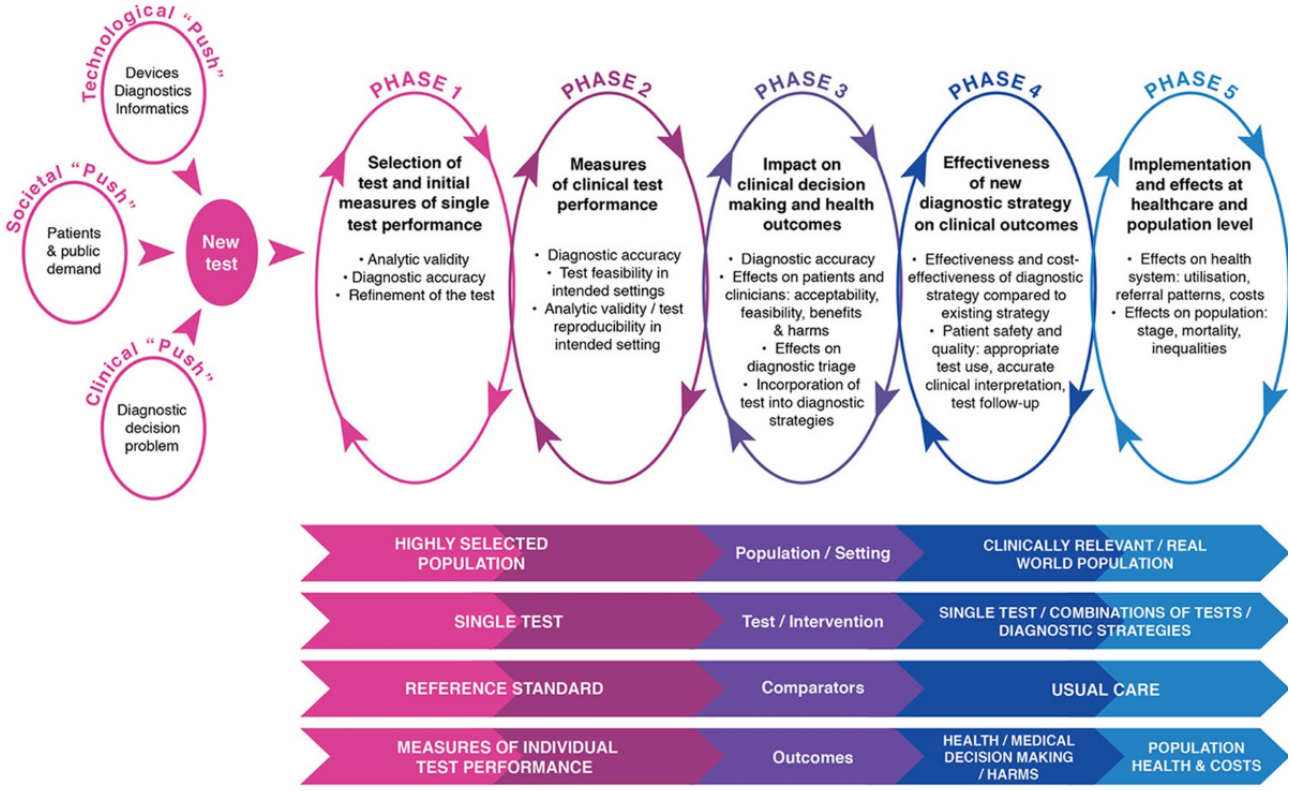
In gatekeeper health care systems such as the United Kingdom, most people diagnosed with cancer first present in primary care [4], where general practitioners evaluate (often vague) presenting symptoms and decide on an appropriate management strategy, including investigations, specialist referral, or reassurance. More accurate assessment of these symptoms, especially for patients with multiple consultations, could lead to earlier diagnosis of cancer and improved outcomes for patients, including improved survival rates [5,6].

There is accumulating evidence that artificial intelligence (AI) can assist clinicians in making better clinical decisions or even replace human judgment, in certain areas of health care. This is due to the increasing availability of health care data and the rapid development of big data analytic methods. There has been increasing interest in the application of AI in medical diagnosis, including machine learning and automated analysis approaches. Recent studies have applied AI to patient symptoms to improve diagnosis [7,8], to retinal images for the diagnosis of diabetic retinopathy [9], to mammography images for breast cancer diagnosis [10,11], to computed tomography (CT) scans for the diagnosis of intracranial hemorrhages [12], and to images of blood films for the diagnosis of acute lymphoblastic leukemia [13].

Few AI techniques are currently implemented in routine clinical care. This may be due to uncertainty over the suitability of current regulations to assess the safety and efficacy of AI systems [14-16], a lack of evidence about the cost-effectiveness and acceptability of AI systems [14], challenges to implementation into existing electronic health records (EHRs) and routine clinical care, and uncertainty over the ethics of using AI systems. A recent review of AI and primary care reported that research on AI for primary care is at an early stage of maturity [17], although research on AI-driven tools such as symptom checkers for patient and clinical users are more mature [18-21].

The CanTest framework [22] (Figure 1) establishes the developmental phases required to ensure that new diagnostic tests or technologies are fit for purpose when introduced into clinical practice. It provides a roadmap for developers and policy makers to bridge the gap from the development of a diagnostic test or technology to its successful implementation. We used this framework to guide the assessment of the studies identified in this review.

Figure 1. The CanTest Framework [22].



Objectives

Few studies of AI-based techniques for the early detection of cancer have been undertaken in primary care settings [17]. Therefore, the aim of this systematic review is to identify AI techniques that facilitate the early detection of cancer and could be applied to primary care EHR data. We also aim to summarize the diagnostic accuracy measures used to evaluate existing studies and evaluate the quality of the evidence, the phase of development the AI technologies have reached, the gaps that exist in the evidence, and the potential for use in primary care. As many commercial technological developments are not documented in academic publications, we also performed a parallel scoping review of commercially available AI-based technologies for the early detection of cancer that may be suitable for implementation in primary care settings.

Methods

Search Strategy and Selection Criteria

This study was conducted in accordance with PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analysis) guidelines [23], and the protocol was registered with PROSPERO (an international prospective register of systematic reviews) before conducting the review (CRD42020176674) [24]. All aspects of the protocol were reviewed by the senior research team.

We included all primary research articles published in peer-reviewed journals, without language restrictions, from January 01, 2000, to June 11, 2019. Studies were included if they provided evidence around the accuracy, utility, acceptability, or cost-effectiveness of applying AI techniques to facilitate the early detection of cancer and could be applied to primary care EHRs (ie, to the types of data found in primary care EHRs) [22]. We included AI techniques based on any type of data that were relevant to primary care settings, including coded data and free text. We included all types of study design, as we anticipated that there would be few relevant randomized controlled trials. We kept our search terms broad to not miss relevant studies and carefully considered evidence from any health care system to assess whether the evidence could be applied to primary care settings.

As our aim is to identify AI techniques that would be applicable in primary care clinical settings, we excluded studies that incorporated data not typically available in primary care EHRs in the early diagnostic stages (eg, histopathology images, magnetic resonance imaging, or CT scan images). We also excluded studies that only described the development of an AI technique without any testing or evaluation data, studies that did not incorporate an element of machine learning (ie, with training and testing or validation steps), studies that used AI techniques for biomarker discovery alone, and studies that were based on sample sizes of less than 50 cases or controls. Machine learning techniques and neural networks have been described since the 1960s [25,26]; however, they were initially limited by computing power and data availability. We chose to start our search in 2000, as this was when the earliest research describing the new wave of machine learning techniques emerged [27].

We searched MEDLINE, Embase, SCOPUS, and Web of Science bibliographic databases, using keywords related to AI, cancer, and early detection. We extended these systematic

searches through manual searching of the reference lists of the included studies. We contacted study authors, where required. Where studies were not published in English, we identified suitably qualified native speakers to help assess these studies. We performed a parallel scoping review to look for commercially developed AI technologies that were not identified through systematic searches, thus unpublished and not scientifically evaluated. This included manually searching commercial research archives and networks (eg, arXiv [28], Google [29], Microsoft [30], and IBM [31]), reviewing the computer-based technologies identified in 3 recent reviews [19-21], and manually searching for further technologies mentioned in the text or references of the studies and websites included in these reviews.

Following duplicate removal, 1 author (OJ) screened titles and abstracts to identify studies that fit the inclusion criteria. Of the titles and abstracts, 17.42% (1838/10,456) were checked by 2 other authors (SS and NC); interrater reliability was excellent at 96.24% (1769/1838). Any disagreements were discussed by the core research team (OJ, SS, NC, and FW), and a consensus was reached. Three reviewers (OJ, SS, and NC) independently assessed the full-text articles for inclusion in the review. Any disagreements were resolved by a consensus-based decision.

Data Analysis

Data extraction was undertaken independently by at least two reviewers (OJ, SS, and NC) into a predesigned data extraction spreadsheet. The research team met regularly to reach consensus by discussing and resolving any differences in data extraction. One author (OJ) amalgamated the data extraction spreadsheets, summarizing the data where possible.

The main summary measures collected included sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), area under the receiver operating characteristic (AUROC) curve, and any other diagnostic accuracy measures of the AI techniques. Secondary outcomes include the types of AI used, the type of data used to train and test the algorithms, and how these algorithms were evaluated. We also collected data, where identified, on cost-effectiveness and patient or clinician acceptability.

Risk of bias assessment was undertaken for all full-text papers by 2 independent researchers (OJ and NC) using the quality assessment of diagnostic accuracy studies-2 (QUADAS-2) critical appraisal tool [32]. OJ assessed all studies, and 50% (40/79) of them were cross-checked by NC. Any disagreements in the assessment were resolved by consensus discussion.

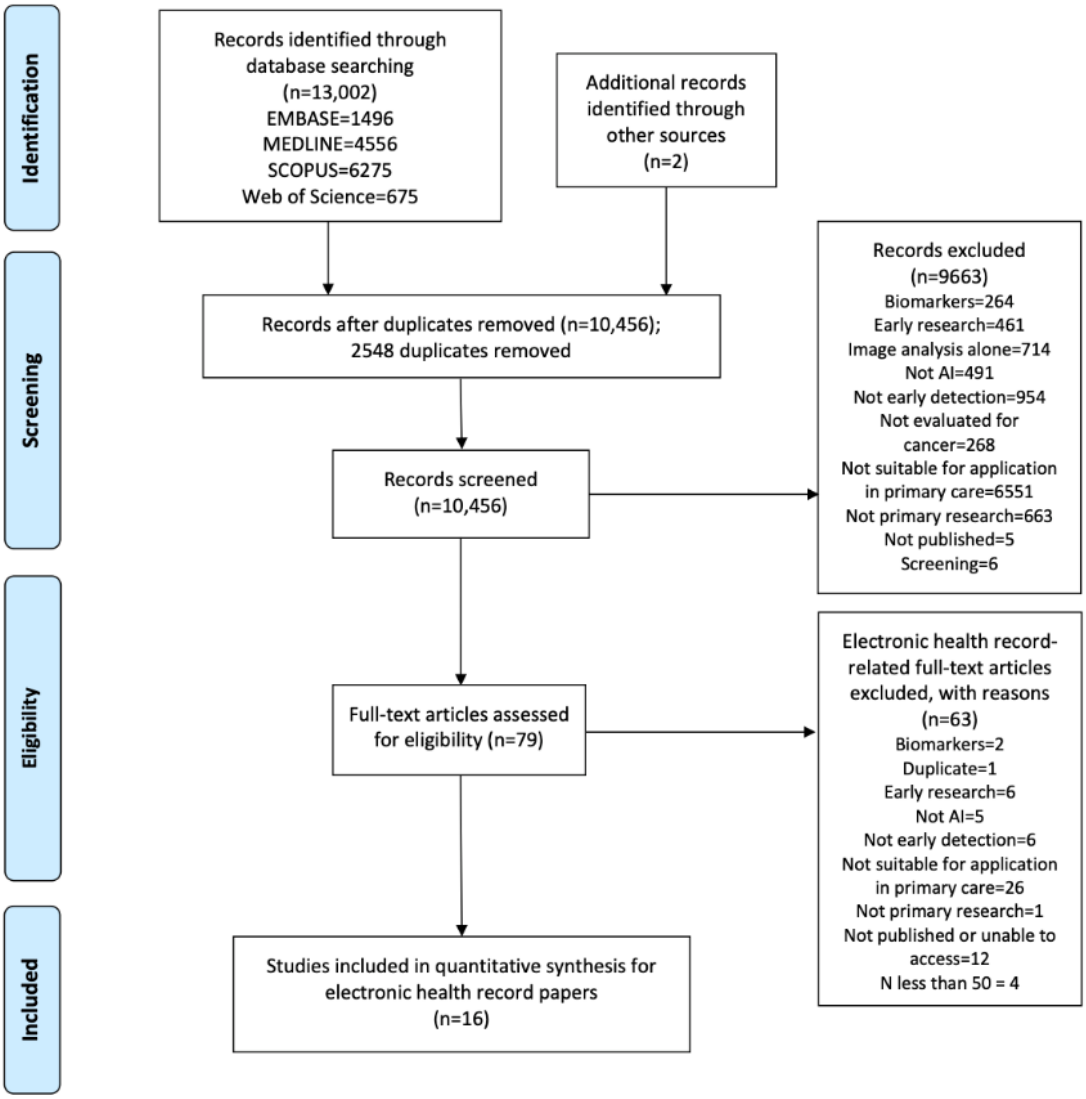
The studies identified were heterogeneous, employing various AI techniques and using different outcome measures for evaluation. Hence, a meta-analysis of the data was not possible, and we chose to use a narrative synthesis approach, following established guidance on its methodology [33]. We aimed to summarize the findings of the identified studies using primarily a textual approach, while also providing an overview of the quantitative outcome measures used in the studies. Once data extraction was completed, we explored the relationships that emerged within the data.

Full details of our review question, search strategy, inclusion or exclusion criteria, and data extraction methodology are described in Multimedia Appendices 1 [1-5,7-9,11-13,34-38] and 2, and the full list of excluded studies is provided in Multimedia Appendix 3 [34,39-114].

Results

A total of 13,004 articles were identified in database searches (including 2548 duplicates), and 793 articles underwent full-text review. Of the 79 articles that were related to EHRs, 16 met the inclusion criteria and were included in this analysis (Figure 2), representing the data of 3,862,910 patients. No articles identified through other sources or reference lists met the inclusion criteria.

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analysis) flow diagram for studies included in the review. AI: artificial intelligence.



Tables 1 and 2 show the main study characteristics for the 16 included studies, including the modality of AI used. Supplementary information on the variables included in the AI techniques is available in Multimedia Appendix 4 [34,39-53]. We categorized the variables included into the following categories: demographics, symptoms, comorbidities, lifestyle history, examination findings, blood results, and other. Most studies (n=13) described the initial development and testing of an AI technique [39-51]. Three studies validated the AI technique developed by Kinar et al [48] in independent data sets from 3 different countries (Israel, United States, and United Kingdom) [34,52,53].

Table 1. Study details including modality of artificial intelligence and adopted comparison or control.

Study	Authors' origin	Cancer	Modality of artificial intelligence	Comparison or control			
				Histopathology	Specialist	Not stated	Other
Development studies							
Alzubi et al, 2019 [39]	Jordan and India	Lung cancer	WONN-MLB[a]	X[b]	—[c]	—	1[d]
Chang et al, 2009 [40]	Taiwan	Pancreatic Cancer	BPNN[e]; LR[f]	—	—	X	2[g]; 3[h]
Cooper et al, 2018 [41]	United Kingdom	Colorectal Cancer	ANN[i]; CVT[j]; LR	X	X	—	4[k]
Cowley et al, 2013 [42]	United Kingdom	Colorectal Cancer	BPANN[l]	—	X	—	2; 5[m]
Daqqa et al, 2017 [43]	Gaza, Palestine	Leukemia	SVM[n]; DT[o]; K-NN[p]	X	—	—	2
Goryński et al, 2014 [44]	Poland	Lung cancer	MLP-ANN[q]	X	X	—	—
Hart et al, 2018 [45]	United States	Lung cancer	BPANN	—	—	X	2; 6[r]
Kalra et al, 2003 [46]	United States	Prostate cancer	BPNN	X	—	—	2; 3
Kang et al, 2017 [47]	China	Any cancer	BPNN; CVT; SVM; DT	X	X	—	2
Kinar et al, 2016 [48]	Israel and United States	Colorectal Cancer	DT/RF[s]; GBM[t]; CVT	X	X	—	3; 6
Kop et al, 2016 [49]	The Netherlands	Colorectal Cancer	CART[u]; RF; LR; CVT	X	X	—	—
Miotto et al, 2016 [50]	United States	Multiple diseases and cancers	DNN[v]; RF	—	X	—	2; 3
Payandeh et al, 2009 [51]	Iran	CML[w] and lymphoproliferative disorders	MLP-ANN	X	X	—	3
Validation studies							
Birks et al, 2017 [52]	United Kingdom	Colorectal Cancer	DT/RF; GBM; CVT	X	X	—	—
Hornbrook et al, 2017 [34]	United States	Colorectal Cancer	DT/RF; GBM; CVT	X	X	—	—
Kinar et al, 2017 [53]	Israel	Colorectal Cancer	DT/RF; GBM; CVT	X	X	—	—

[a]WONN-MLB: weight optimized neural network with maximum likelihood boosting.

[b]X: corresponding control used in this study.

[c]Not used in this study.

[d]1: previously developed artificial intelligence methods.

[e]BPNN: back propagation neural network.

[f]LR: logistic regression.

[g]2: other artificial intelligence methods developed by this author.

[h]3: other statistical (ie, non-artificial intelligence) techniques.

[i]ANN: artificial neural network.

[j]CVT: cross-validation techniques.

[k]4: colonoscopy.

[l]BPANN: back propagation artificial neural network.

[m]5: primary care clinicians.

[n]SVM: support vector machine.

[o]DT: decision tree.

[p]K-NN: K-nearest neighbor.

[q]MLP-ANN: multilayer perceptron artificial neural network.

[r]6: screening tests (eg, low-dose computed tomography scan and fecal occult blood test).

[s]RF: random forest.

[t]GBM: gradient boosting model.

[u]CART: classification and regression trees.

[v]DNN: deep neural network.

[w]CML: chronic myeloid leukemia.

The study authors originated from a variety of countries, including the United States (n=5), countries in the Middle East (n=5), Europe (n=5), and Asia (n=3), with some studies involving multiple countries. The AI techniques were most commonly developed to identify colorectal cancer (n=7) [34,41,42,48,49,52,53], although they also addressed lung cancer (n=3) [39,44,45], hematological cancers (n=2) [43,51], pancreatic cancer (n=1) [40], prostate cancer (n=1) [46], and multiple cancers (n=2) [47,50].

Neural networks were the dominant technique employed (n=10) [39-42,44-47,50,51], with many neural network subtypes mentioned. The study by Miotto et al [50] was the only study to include a processed form of the free text notes in the data used by the AI technique, although the work described by Kop et al [49] was developed in a subsequent study to include clinical free text data [115].

The majority of studies (n=9) used a combination of histopathological diagnoses and expert opinion as the control for their study [34,41,44,47-49,51-53]. The clinical control group was unclear in 2 studies [40,45]. Many studies used multiple AI techniques and then compared them with each other (n=8) [40,42,43,45-47,49,50]. Some studies used non-AI techniques, such as logistic regression and screening tests, as comparators for the performance of the AI technique that was being developed [40,41,45,46,48-51].

Table 2. Study details: patient variables.

Study	Patient variables									
	Age	Sex	Demographics	Symptoms	Comorbidities	Lifestyle	Examination	FBC[a]	Other blood tests	Other[b]
Development studies										
Alzubi et al, 2019 [39]	X[c]	—[d]	—	X	X	X	—	—	—	X
Chang et al, 2009 [40]	X	X	—	X	X	X	—	X	X	—
Cooper et al, 2018 [41]	X	X	X	—	—	—	—	—	—	X
Cowley et al, 2013 [42]	—	—	—	X	X	X	—	—	—	X
Daqqa et al, 2017 [43]	—	—	—	—	—	—	—	X	—	—
Goryński et al, 2014 [44]	X	X	X	X	X	X	X	X	X	X
Hart et al, 2018 [45]	X	X	X	—	X	X	X	—	—	—
Kalra et al, 2003 [46]	X	—	X	X	X	—	X	—	X	—
Kang et al, 2017 [47]	X	X	—	—	—	—	X	X	X	X
Kinar et al, 2016 [48]	X	X	—	—	—	—	—	X	—	—
Kop et al, 2016 [49]	X	X	—	X	X	X	X	X	X	X
Miotto et al, 2016 [50]	—	—	X	X	X	X	X	—	X	X
Payandeh et al, 2009 [51]	—	—	—	—	—	—	—	X	—	—
Validation studies										
Birks et al, 2017 [52]	X	X	—	—	—	—	—	X	—	—
Hornbrook et al, 2017 [34]	X	X	—	—	—	—	—	X	—	—
Kinar et al, 2017 [53]	X	X	—	—	—	—	—	X	—	—

[a]FBC: full blood count.

[b]More detail on other variables included is available in Multimedia Appendix 4.

[c]X: corresponding variable used in this study.

[d]Not used in this study.

Most of the studies (n=12) included blood test results, all suitable for use in primary care settings. Age was also commonly included (n=12). Other variables used were sex (n=10), demographics (n=5), symptoms (n=7), comorbidities (n=8), lifestyle history (n=7), examination findings (n=6), medication or prescription history (n=3), spirometry results (n=2), urine dipstick results (n=1), fecal immunochemical test results (n=1), x-ray text reports (n=1), and referrals (n=1).

Table 3 shows the study designs and populations. Most studies used data sets originating from specialist care settings (n=7) [39,40,42-44,46,51], with only 3 studies using solely primary care patient data [41,49,52]. Kinar et al [48] included a follow-up validation study based on the health improvement network (THIN) database, also using primary care data. Several studies used a mixture of primary and secondary care patient data (n=5) [34,47,48,50,53].

Table 3. Study population and study design.

Study details	Population from health care setting	Database used	Disease positive population (patients)	Disease negative population (patients)	Training set (patients)	Testing set (patients)
Development studies						
Alzubi et al, 2019 [39]	Specialist care	Wroclaw Thoracic Surgery Centre	1200 in total; numbers of disease positive and negative unclear	1200 in total; numbers of disease positive and negative unclear	N/S[a]	1000
Chang et al, 2009 [40]	Specialist care (unclear)	"a certain medical center"	194	157[b]	234	117
Cooper et al, 2018 [41]	Primary care	NHS[c] Bowel Cancer Screening Programme comparative study [116]	549	1261	N/S	N/S
Cowley et al, 2013 [42]	Specialist care	2-week wait colorectal referrals to Castle Hill Hospital	74	703	777	100
Daqqa et al, 2017 [43]	Specialist care	Complete Blood Count test repository, European Gaza Hospital	2000	2000	N/S	N/S
Goryński et al, 2014 [44]	Specialist care	Patients treated at Kuyavia and Pomerania Centre of pulmonology	103	90	97	48
Hart et al, 2018 [45]	Other (survey)	National Health Interview Survey	649	488,418	342,347	146,719
Kalra et al, 2003 [46]	Specialist care	Men whose samples were tested at 6 sites in the United States[d]	348	N/S	218	144
Kang et al, 2017 [47]	Mixed	Database of Ci Ming Health Checkup Center	650	1650	N/S	N/S
Kinar et al, 2016 [48][e]	Mixed	Maccabi Health Services EMRs[f] linked to the Israel Cancer Registry	2437	463,670	466,107	139,205
Kop et al, 2016 [49]	Primary care	6 anonymized data sets from 3 urban regions, each covering a GP[g] recording system	1292	263,879	N/S	N/S
Miotto et al, 2016 [50]	Mixed	Mount Sinai Data Warehouse	276,214 patients with 78 diseases	276,214 patients with 78 diseases	200,000	76,214
Payandeh et al, 2009 [51]	Specialist care	Blood test results from patients at the Taleghani Hospital	450	N/S	360	132
Validation studies						
Birks J et al, 2017 [52]	Primary care	Clinical Practice Research Datalink	5141	2,220,108	N/A[h]	N/A
Hornbrook et al, 2017 [34]	Mixed	Kaiser Permanente North West EHR[i] system, Kaiser Permanente Tumor Registry	900	16,195	N/A	N/A
Kinar et al, 2017 [53]	Mixed	Maccabi Health Services EMRs, linked to the Israel Cancer Registry	133	112,451	N/A	N/A

[a]N/S: not stated.

[b]Cases of acute pancreatitis.

[c]NHS: National Health Service.

[d]Hospitals included: Northwest Prostate Institute Seattle, the University of Washington Seattle, the Johns Hopkins Hospital Baltimore, Memorial Sloan-Kettering Cancer Institute New York, Brigham and Women's Hospital Boston, and The University of Texas MD Anderson Cancer Center

[e]NB: this study also included a small validation study in the Health Improvement Network database in the United Kingdom (n=25,613)

[f]EMR: electronic medical record.

[g]GP: general practitioner.

[h]N/A: not applicable

[i]EHR: electronic health record.

Almost all the studies used different data sets, with the exception of the Maccabi Health Services EHR, which was used in 2 studies [48,53]. The data set sizes ranged from 193 to 2,225,249 patients, with a mean of 241,585 (SD 555,953), median of 3,150, and IQR of 267,237 patients. The wide range is primarily due to the large data set used by Birks et al [52]. Of the 13 development studies, 3 provided no information on the control population used [39,46,51]. Five of the development studies did not provide full information on how they partitioned their data set for the training and testing of the algorithm [39,41,43,47,49]. Five studies appeared to have independent training and testing data sets, with most split in ratios ranging from 60:40 to 70:30 [40,44-46,50].

Three studies [34,52,53] validated a previously developed AI technique [48] in independent data sets. Kinar et al [48] reported both the initial development of an AI technique and a subsequent validation study in an independent data set. The study by Cooper et al [41] was the only study that developed an AI technique based on prospectively collected clinical data, with the data originating from a pilot study of fecal immunochemical testing by the NHS Bowel Cancer Screening Programme [116].

Table 4 summarizes the main reported outcome measures. Specificity (n=11), AUROC (n=11), and sensitivity (n=10) were the most frequently reported; others included PPV (n=6), NPV (n=5), diagnostic accuracy (n=4), and odds ratios (n=3). Specificity results range from 80.6% [45] to 100% [51], sensitivity results from 0% [51] to 96.7% [40], and AUROC results from 0.55 [45] to 0.9896 [44].

Table 4. Outcome measures.

Study	Cancer type	Outcome measures for each modality of AI[a]
Development studies		
Alzubi et al, 2019 [39]	Lung cancer	• Specificity: 92%, Accuracy: 93% • False positive rate: 9%, F-1 score: 92%
Chang et al, 2009 [40]	Pancreatic cancer	• Sensitivity: BPNN[b] 88.3%, genetic algorithm LR[c] 96.7%, stepwise LR 96.7% • Specificity: BPNN 84.2%, genetic algorithm LR 82.5%, stepwise LR 73.7% • AUROC[d]: BPNN 0.895, genetic algorithm LR 0.921, stepwise LR 0.882
Cooper et al, 2018 [41]	Colorectal cancer	• Sensitivity: 35.15% (at FIT[e] threshold 160 μg g^{-1}) • Specificity: 85.57% • PPV[f]: 51.47%, NPV[g]: 75.19%, AUROC: 0.69, cancer detection rate: 10.66%
Cowley et al, 2013 [42]	Colorectal cancer	• Sensitivity: 90% • Specificity: 96% • PPV: 62%, NPV: 99%
Daqqa et al, 2017 [43]	Leukemia	• Sensitivity: SVM[h] 69.7%, K-NN[i] 60.0%, decision tree 62.4% • Specificity: SVM 81.5%, K-NN 82.8%, decision tree 87.1% • PPV: SVM 71.3%, K-NN 68.1%, decision tree 76.1% • NPV: SVM 80.4%, K-NN 74.1%, decision tree 87.1% • Accuracy: SVM 76.82%, K-NN 72.15%, decision tree 77.3% • F-measure: SVM 70%, K-NN 60%, decision tree 67%
Goryński et al, 2014 [44]	Lung cancer	• AUROC: 0.9896
Hart et al, 2018 [45]	Lung cancer	• Sensitivity: ANN[j] 75.30% • Specificity: ANN 80.60% • AUROC: ANN 0.86, RF[k] 0.81, SVM 0.55
Kalra et al, 2003 [46]	Prostate cancer	• Specificity: 92% • AUROC: 0.825
Kang et al, 2017 [47]	Any cancer	• Sensitivity: DNN[l] 64.07%, SVM 54.46%, decision tree 60.00% • Specificity: DNN 94.77%, SVM 95.27%, decision tree 91.50% • AUROC: DNN 0.882, SVM 0.928, decision tree 0.824 • Accuracy: DNN 86.00%, SVM 83.83%, decision tree 83.60% • Using fuzzy interval of threshold with DNN achieves sensitivity 90.20%, specificity 94.22%, accuracy 93.22%
Kinar et al, 2016 [48]	Colorectal cancer	• Specificity: Testing set 88% overall (at a sensitivity of 50%). Higher for proximal colon tumors. Validation set 94% (at a sensitivity of 50%) • AUROC: Testing set 0.82, validation set 0.81 • OR[m] 26 at false +ve rate of 0.5% (testing set), OR 40 at false +ve rate of 0.5% (validation set). Algorithm identified 48% more CRC[n] cases than gFOBT[o]
Kop et al, 2016 [49]	Colorectal cancer	• Sensitivity: CART[p] 53.9%, RF 63.7%, LR 64.2% • PPV: CART 2.6%, RF 3%, LR 3% • AUROC: CART 0.885, RF 0.889, LR 0.891 • F1-score: CART 0.049, RF 0.057, LR 0.058. • Drugs for constipation most important predictor of CRC, followed by iron deficiency anemia

Study	Cancer type	Outcome measures for each modality of AI[a]
Miotto et al, 2016 [50]	Multiple diseases and cancers	• Specificity: 92% • AUROC: 0.773 for classification of all diseases (cancer and other diagnoses). Rectal or anal cancer 0.887, liver or intrahepatic bile duct cancer 0.886, prostate cancer 0.859, multiple myeloma 0.849, ovarian cancer 0.824, bladder cancer 0.818, testicular cancer 0.811, pancreatic cancer 0.795, leukemia 0.774, uterine cancer 0.771, non-Hodgkin lymphoma 0.771, bronchial or lung cancer 0.770, colon cancer 0.767, breast cancer 0.762, kidney or renal pelvis cancer 0.753, brain or nervous system cancer 0.742, Hodgkin disease 0.731, cervical cancer 0.675 • Accuracy index: 0.929 overall for classification of all diseases • F-score: 0.181 for classification of all diseases • Deep patient obtained approximately 55% correct predictions when suggesting 3 or more diseases per patient, regardless of time interval
Payandeh et al, 2009 [51]	CML[q] and lymphopro-liferative disorders	• Sensitivity: CML 0%, lymphoproliferative disorder 0% • Specificity: CML 100%, lymphoproliferative disorder 99.2% • PPV: CML 0%, lymphoproliferative disorder 0% • NPV: CML 99.2%, lymphoproliferative disorder 100% • Error % for convoluted neural network 0.33, error % for LR 0.78
Validation studies		
Birks et al, 2017 [52]	Colorectal cancer	• AUROC: analyzed at various time intervals before diagnosis, 3-6 months 0.844, 18-24 months 0.776
Hornbrook et al, 2017 [34]	Colorectal cancer	• Sensitivity: 0-180 days (test to diagnosis): 50-75 years: 34.5%, 40-89 years: 39.9%; 181-360 days: 50-75 years: 18.8%, 40-89 years: 27.4% • AUROC: 0.80, OR: 34.7 at 99% specificity, 19.7 at 97%, 14.6 at 95%, 10.0 at 90%
Kinar et al, 2017 [53]	Colorectal cancer	• Sensitivity: 17.0% at 1% +ve rate, 24.4% at 3% +ve rate • PPV: 2.1% at 1% +ve rate, 1.0% at 3% +ve rate • NPV: 99.9% at 1% +ve rate, 99.9% at 3% +ve rate • OR: 21.8% at 1% +ve rate, 10.9% at 3% +ve rate

[a]AI: artificial intelligence.
[b]BPNN: back propagation neural network.
[c]LR: logistic regression.
[d]AUROC: area under the receiver operating characteristic.
[e]FIT: fecal immunochemical test.
[f]PPV: positive predictive value.
[g]NPV: negative predictive value.
[h]SVM: support vector machine.
[i]K-NN: K-nearest neighbor.
[j]ANN: artificial neural network.
[k]RF: random forest.
[l]DNN: deep neural network.
[m]OR: odds ratio.
[n]CRC: colorectal cancer.
[o]gFOBT: guaiac fecal occult blood test.
[p]CART: classification and regression trees.
[q]CML: chronic myeloid leukemia.

We looked for other secondary outcomes, including implementation barriers to AI techniques in primary care settings, but did not find any evidence related to patient or clinician acceptability or cost-effectiveness.

Table 5 shows the outcomes of the risk of bias assessment using the QUADAS-2 tool. The studies demonstrated a wide range in quality; however, no studies were excluded based on their risk of bias assessment. The identified limitations were acknowledged in the relative contribution of the studies to the conclusions of the review.

Table 5. Critical appraisal results using the Quality Assessment of Diagnostic Accuracy Studies-2 tool.

Study	Risk of bias				Applicability concerns		
	Patient selection	Index test	Reference standard	Flow and timing	Patient selection	Index test	Reference standard
Alzubi et al, 2019 [39]	a	b		c			
Birks et al, 2017 [52]							
Chang et al, 2009 [40]							
Cooper et al, 2018 [41]							
Cowley et al, 2013 [42]							
Daqqa et al, 2017 [43]							
Goryński et al, 2014 [44]							
Hart et al, 2018 [45]							
Hornbrook et al, 2017 [34]							
Kalra et al, 2003 [46]							
Kang et al, 2017 [47]							
Kinar et al, 2016 [48]							
Kinar et al, 2017 [53]							
Kop et al, 2016 [49]							
Miotto et al, 2016 [50]							
Payandeh et al, 2009 [51]							

[a]High risk.

[b]Low risk.

[c]Unclear risk.

Table 6 summarizes the computer-based technologies identified in our parallel scoping review of commercial AI technologies. We identified 21 commercial computer-based technologies. Of these, 11 were clinician-facing differential diagnosis technologies that did not appear to be integrated into the EHR [117-127]. Ten of the technologies were linked to, or integrated into, the EHR in some way [8,128-136]. Nine of the technologies did not use AI algorithms incorporating an element of machine learning, as was required in our inclusion criteria [118,120-127]. It was also not clear from the websites and studies of 3 further technologies whether they met our AI inclusion criteria [117,130,134]. There were 8 technologies that met our inclusion criteria for AI (Abtrace [128], Babylon [8], Cthesigns [129], Isabel [131], Medial EarlySign [132], symcat [119], symptomate [135], and the unnamed technology evaluated by Liang et al [136]). Only the Medial EarlySign tool was evaluated for its performance in the diagnosis or triage of potential cancer [132]; 4 of the studies developing and validating this technology were included in this systematic review [34,48,52,53]. Cthesigns is specifically designed to aid the early diagnosis of cancer but has not been the subject of any studies we could identify [129].

Table 6. Summarizing scoping review of commercial artificial intelligence technologies.

Technology identified (origin) websites and associated academic studies	Not AI[a]	Not cancer	Not primary care based	Not early detection or diagnosis	Early research	Not published	Not primary research	<50 cases or controls
Abtrace (United Kingdom)								
Abtrace website [128]	—[b]	—	—	—	—	X[c]	—	—
Babylon (United Kingdom)								
Babylon health website [8]	—	—	—	—	—	—	—	—
Zhelezniak et al [137]	—	X	X	X	X	—	—	—
Douglas et al [138]	—	X	X	X	X	—	—	—
Smith et al [139]	—	X	X	X	X			
National Health Service 111 powered by Babylon - Outcomes Evaluation [140]	—	X	—	X	—	—	—	—
Middleton et al [141]	—	X	—	X	—	—	—	—
Cthesigns (United Kingdom)								
Cthesigns website [129]	—	—	—	—	—	X	—	—
Diagnosis Pro (United States)								
No website identified	—	—	—	—	—	—	—	—
Bond et al [117]	N/C[d]	X	—	—	—	—	—	—
DocResponse (United States)								
Docresponse website [130]	N/C	—	—	—	—	X	—	—
DxPlain (United States)								
Dxplain website [118]	N/C	—	—	—	—	—	—	—
Barnett et al [142]	X	—	—	—	X	—	X	—
Barnett et al [143]	X	X	—	—	—	—	—	—
Bauer et al [144]	X	X	—	—	—	—	—	—
Berner et al [145]	X	X	X	—	—	—	—	—
Bond et al [117]	X	X	—	—	—	—	—	X
Elhanan et al [146]	X	—	—	—	X	—	—	—
Elkin et al [147]	X	X	X	—	—	—	—	—
Feldman et al [148]	X	X	X	—	—	—	—	X
Hammersley et al [149]	X	X	X	—	—	—	—	—
Hoffer et al [150]	X	—	—	X	—	—	—	—
London et al [151]	X	—	—	—	X	—	—	—
Iliad (United States)								
No website identified	—	—	—	—	—	—	—	—
Berner et al [145]	X	X	X	—	—	—	—	—
Elstein et al [152]	X	X	X	—	—	—	—	X
Friedman et al [153]	X	—	X	—	—	—	—	X
Gozum et al [154]	X	—	X	—	—	—	—	X
Graber et al [155]	X	—	X	—	—	—	—	X
Heckerling et al [120]	X	—	X	—	—	—	—	X
Lange et al [156]	X	—	—	—	—	—	—	X
Lau et al [157]	—	—	—	—	—	—	X	—

Technology identified (origin) websites and associated academic studies	Not AI[a]	Not cancer	Not primary care based	Not early detection or diagnosis	Early research	Not published	Not primary research	<50 cases or controls
Li et al [158]	X	X	X	—	—	—	—	X
Lincoln et al [159]	X	X	X	—	—	—	—	X
Murphy et al [160]	X	X	X	—	—	—	—	X
Wolf et al [161]	X	X	X	—	—	—	—	X
Internist-1 (United States)								
No website identified	—	—	—	—	—	—	—	—
Miller et al [121]	X	X	X	—	—	—	—	X
Miller et al [122]	X	—	X	—	—	—	—	X
Isabel (United Kingdom)								
Isabel healthcare website – Isabel pro [131]	—	—	—	—	—	—	—	—
Bond et al [117]	—	X	—	—	—	—	—	—
Ramnarayan et al [162]	—	X	—	—	—	—	—	—
Ramnarayan et al [163]	—	X	—	—	—	—	—	—
Carlson et al [164]	—	X	—	—	—	—	—	—
Graber et al [165]	—	—	—	—	—	—	X	—
Graber et al [166]	—	X	—	—	—	—	—	—
Ramnarayan et al [167]	—	X	—	—	—	—	—	—
Bavdekar et al [168]	—	X	—	—	—	—	—	—
Ramnarayan et al [169]	—	X	—	—	—	—	—	—
Semigran et al [20]	—	X	—	—	—	—	—	—
Meyer et al [170]	—	X	—	—	—	—	—	—
Meditel (United States)								
No website identified	—	—	—	—	—	—	—	—
Berner et al [145]	X	X	X	—	—	—	—	—
Hammersley et al [149]	X	X	X	—	—	—	—	—
Waxman et al [171]	X	X	X	—	—	—	—	—
Wexler et al [123]	X	X	X	—	—	—	—	X
Medial Early sign (United States/Israel)								
Earlysign website [132]	—	—	—	—	—	—	—	—
Kinar et al [53][e]	—	—	—	—	—	—	—	—
Birks et al [52][e]	—	—	—	—	—	—	—	—
Hornbrook et al [34][e]	—	—	—	—	—	—	—	—
Goshen et al [172]	—		X	—	—	—	—	—
Zack et al [173]	—	X	—	—	—	—	—	—
Cahn et al [174]	—	X	—	—	—	—	—	—
Multilevel Diagnosis Decision Support System (Spain)								
No website identified	—	—	—	—	—	—	—	—
Rodriguez-Gonzalez et al [124]	X	X	—	—	—	—	—	X
Online webGP (United Kingdom; later became eConsult)								
Emis health online-triage website [175][f]	—	—	—	—	—	—	—	—

Technology identified (origin) websites and associated academic studies	Not AI[a]	Not cancer	Not primary care based	Not early detection or diagnosis	Early research	Not published	Not primary research	<50 cases or controls
Hurleygroup website [176][g]	—	—	—	—	—	—	—	—
Edwards et al [133]	X	X	—	X	—	—	—	—
Carter et al [177]	X	X	—	X	—	—	—	—
Cowie et al [178]	X	X	—	X	—	—	—	—
Pepid (United States)								
Pepid website [125][h]	N/C	—	—	—	—	—	—	—
Bond et al [117]	X	X	X	—	—	—	—	—
Problem Knowledge Couplers (PKC; United States)								
No website identified	—	—	—	—	—	—	—	—
Apkon et al [126]	X	—	—	X	—	—	—	—
Quick Medical Reference (QMR) (United States; developed from Internist-1)								
No website identified	—	—	—	—	—	—	—	—
Arene et al [179]	X	—	X	—	—	—	—	X
Bacchus et al [180]	X	—	X	—	—	—	—	X
Bankowitz et al [181]	X	—	X	—	—	—	—	X
Berner et al [145]	X	X	X	—	—	—	—	—
Berner et al [182]	X	—	—	—	—	—	—	X
Friedman et al [153]	X	—	X	—	—	—	—	X
Gozum et al [154]	X	—	X	—	—	—	—	X
Graber et al [155]	X	—	X	—	—	—	—	X
Miller et al [122]	X	—	X	—	—	—	—	X
Lemaire et al [183]	X	—	X	—	—	—	—	—
Reconsider (United States)								
No website identified	—	—	—	—	—	—	—	—
Nelson et al [127]	X	X	X	—	—	—	—	—
Symcat (United States)								
Symcat website [119]	—	—	—	—	—	X	—	—
Symptify (United States)								
Symptify website [134]	N/C	—	—	—	—	X	—	—
Symptomate (Poland)								
Symptomate website [135]	—	—	—	—	—	X	—	—
Unnamed								
No website identified	—	—	—	—	—	—	—	—
Liang H et al [136]	—	X	X	—	—	—	—	—

[a]AI: artificial intelligence.

[b]Not applicable or no data.

[c]Study excluded for the reason specified in the column label.

[d]N/C: not clear.

[e]These studies met the inclusion criteria of the systematic review and were therefore included.

[f]Edwards et al [133] suggests that this Egton Medical Information Systems (EMIS) application is powered by the eConsult system.

[g]Carter et al [177] suggests that this is the group who developed webGP.

[h]Several published studies are linked in the research section of the website, none involved use of the differential diagnosis or decision support tools. Some case studies audited the use of these tools.

Discussion

Principal Findings

We identified 16 studies reporting AI techniques that could facilitate the early detection of cancer and could be applied to the types of data found in primary care EHRs. However, heterogeneity of AI modalities, data set characteristics, outcome measures, conduct of these studies, and quality assessment meant that we were unable to draw strong conclusions about the utility of these techniques in primary care settings. There was a notable paucity of evidence on performance using primary care data. Coupled with the lack of evidence on implementation barriers or cost-effectiveness, this may help explain why AI techniques have not been adopted widely into primary care clinical practice to date. The study by Kinar et al [48] and its subsequent validation in independent data sets [34,52,53], including primary care data sets, is a valuable example of a staged evaluation of an AI technique from early development, via validation data sets, to evaluation in the population for intended use [22]. The work by Kop and collaborators [49,115,184] also represents a good example of the staged development of an AI technique, with sequential peer-reviewed, published evaluations at each stage.

We also identified 21 commercial AI technologies, many of which have not been evaluated and reported in peer-reviewed, published studies. Many other technologies that were patient-facing and designed for the triage of symptoms were identified but had not been applied to EHRs. Eight of these technologies appeared to be based on newer machine learning AI techniques, with the majority appearing to be driven by knowledge-based decision tree algorithms. Only one of the identified technologies has been evaluated specifically for cancer, although it may be more efficacious for these technologies to be very general in scope and to be widely used, rather than to have a narrow focus on cancer alone. With wider adoption, these technologies have a greater potential for raising patient and clinician awareness of cancer. However, it remains important to fully understand their diagnostic accuracy and safety, including for the triage of potential cancer symptoms. AI technologies applied to EHRs are potentially useful for primary care clinicians; however, they need to be designed in a way that is appropriate for the type and origin of the data found in primary care EHRs and to have been thoroughly and transparently evaluated in the population the technology is intended for.

Strengths and Limitations

The strengths of this systematic review include the following: a broad and inclusive search strategy to avoid missing studies; guidance of an international expert panel in the development of the protocol and search strategy; independent screening, quality assessment, and data extraction processes; followed PRISMA guidance; and a parallel scoping review for commercial AI technologies. As only a few heterogeneous studies were identified, it was not possible to synthesize the data and evaluate the utility of these AI techniques. Furthermore, only one commercially available AI technology was identified via the systematic review. Many of the technologies identified in the parallel scoping review lacked sufficient academic detailing and evidence for their accuracy or safety. This is a rapidly evolving research area, which will require further review over time.

Conclusions

Worldwide, there is a great deal of interest in AI techniques and their potential in medicine, not least in the United Kingdom where politicians and NHS leaders have publicly prioritized the incorporation of AI into clinical settings. Our findings support those of Kueper et al [17], namely, that although some AI techniques have good initial validation reports, they have not yet been through the steps for full application in clinical practice. Validation using independent data is preferable to splitting a single data set [185] and could be the next step in the development of many AI techniques identified in this review. Much of the research is at an early stage, with variable reporting and conduct, and requires further validation in prospective clinical settings and assessment of cost-effectiveness after clinical implementation before it can be incorporated into daily practice safely and effectively [186].

Consensus is required on how AI techniques designed for clinical use should be developed and validated to ensure their safety for patients and clinicians in their intended settings. Good internal and external validity is required in these experiments to avoid bias, most notably spectrum bias [187] and distributional shift [16], and to ensure that the appropriate data are used to develop the AI technique in keeping with its anticipated clinical setting and diagnostic function. The CanTest framework provides an outline for further studies aiming to develop this evidence base for AI techniques in clinical settings; to prove their safety and efficacy to commissioners, clinicians, and patients; and to enable them to be implemented in clinical practice [22]. Prospective evaluation in the clinical setting for which the AI technique is intended is essential: AI aimed at primary care clinics must be evaluated in primary care settings, where cancer prevalence is low compared with specialist settings, to accurately evaluate their future performance [187,188]. Further research around the acceptability of AI techniques for patients and clinicians and their cost-effectiveness will also be important to facilitate rapid implementation. Once these AI techniques are ready for implementation, they will require careful design to ensure effective integration into health information systems [189]. Data governance and protection must also be addressed, as they may present significant barriers to the implementation of these technologies [190,191].

In conclusion, AI techniques have the potential to aid the interpretation of patient-reported symptoms and clinical signs and to support clinical management, doctor-patient communication, and informed decision making. Ultimately, in the context of early cancer detection, these techniques may help reduce missed diagnostic opportunities and improve safety netting. However, although there are a few good examples of staged validation of these AI techniques, most of the research is at an early stage. We found numerous examples of the implementation of AI technologies without any or sufficient evidence for their accuracy or safety. Further research is required to build up the evidence base for AI techniques applied to EHRs

and to reassure commissioners, clinicians, and patients that they are safe and effective enough to be incorporated into routine clinical practice.

Acknowledgments

This research was funded by the National Institute for Health Research (NIHR) Policy Research Programme, conducted through the Policy Research Unit in Cancer Awareness, Screening, and Early Diagnosis, PR-PRU-1217-21601. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. This work was also supported by the CanTest Collaborative (funded by Cancer Research UK C8640/A23385), of which FW and WH are directors and JE, HS, and NdW are associate directors. HS is additionally supported by the Houston Veterans Administration Health Services Research and Development Center for Innovations in Quality, Effectiveness, and Safety (CIN13-413) and the Agency for Healthcare Research and Quality (R01HS27363). The funding sources had no role in the study design, data collection, data analysis, data interpretation, writing of the report, or the decision to submit for publication. The authors would like to thank Isla Kuhn, Reader Services Librarian, University of Cambridge Medical Library, for her help in developing the search strategy.

Authors' Contributions

OJ developed the protocol, completed the search, screened the articles for inclusion, extracted the data, synthesized the findings, interpreted the results, and drafted the manuscript. NC screened the articles for inclusion, extracted the data, and critically revised the manuscript. SS screened the articles for inclusion, extracted the data, and critically revised the manuscript. WH developed the protocol, interpreted the results, and critically revised the manuscript. SD, JE, HS, and NdW critically revised the manuscript. FW developed the protocol, synthesized the findings, interpreted the results, and critically revised the manuscript. All authors approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Protocol for the study.
[DOCX File , 34 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Search strategies.
[DOCX File , 16 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Results of the full-text article review.
[DOCX File , 38 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Supplementary information to table 1.
[DOCX File , 36 KB-Multimedia Appendix 4]

References

1. Cancer statistics for the UK. Cancer Research UK. URL: https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk [accessed 2020-11-30]
2. Hamilton W. Diagnosing symptomatic cancer in the NHS. Br Med J 2015 Oct 13;351:5311. [doi: 10.1136/bmj.h5311] [Medline: 26466605]
3. Coleman M, Forman D, Bryant H, Butler J, Rachet B, Maringe C, et al. Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995–2007 (the International Cancer Benchmarking Partnership): an analysis of population-based cancer registry data. The Lancet 2011 Jan;377(9760):127-138. [doi: 10.1016/s0140-6736(10)62231-3]
4. Hiom SC. Diagnosing cancer earlier: reviewing the evidence for improving cancer survival. Br J Cancer 2015 Mar 31;112 Suppl 1(S1):S1-S5 [FREE Full text] [doi: 10.1038/bjc.2015.23] [Medline: 25734391]
5. Garbe C, Peris K, Hauschild A, Saiag P, Middleton M, Bastholt L, European Dermatology Forum (EDF), European Association of Dermato-Oncology (EADO), European Organisation for ResearchTreatment of Cancer (EORTC). Diagnosis and treatment of melanoma. European consensus-based interdisciplinary guideline - update 2016. Eur J Cancer 2016 Aug;63:201-217. [doi: 10.1016/j.ejca.2016.05.005] [Medline: 27367293]

6. Lyratzopoulos G, Wardle J, Rubin G. Rethinking diagnostic delay in cancer: how difficult is the diagnosis? Br Med J 2014 Dec 09;349(dec09 3):7400-7400 [FREE Full text] [doi: 10.1136/bmj.g7400] [Medline: 25491791]
7. Isabel Differential Diagnosis Generator. Isabel Healthcare. 2018. URL: https://www.isabelhealthcare.com [accessed 2020-11-30]
8. Artificial intelligence. Babylon Health. URL: https://www.babylonhealth.com/ai [accessed 2020-11-30]
9. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. J Am Med Assoc 2016 Dec 13;316(22):2402-2410. [doi: 10.1001/jama.2016.17216] [Medline: 27898976]
10. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature 2020 Jan 1;577(7788):89-94. [doi: 10.1038/s41586-019-1799-6] [Medline: 31894144]
11. Li Z, Yu L, Wang X, Yu H, Gao Y, Ren Y, et al. Diagnostic performance of mammographic texture analysis in the differential diagnosis of benign and malignant breast tumors. Clin Breast Cancer 2018 Aug;18(4):621-627. [doi: 10.1016/j.clbc.2017.11.004] [Medline: 29199085]
12. Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal VK, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. The Lancet 2018 Dec;392(10162):2388-2396. [doi: 10.1016/s0140-6736(18)31645-3]
13. Shafique S, Tehsin S. Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks. Technol Cancer Res Treat 2018 Jan 01;17:1533033818802789 [FREE Full text] [doi: 10.1177/1533033818802789] [Medline: 30261827]
14. Preparing the healthcare workforce to deliver the digital future. The Topol Review.: NHS Health Education England; 2019. URL: https://topol.hee.nhs.uk/ [accessed 2020-11-30]
15. Artificial intelligence and primary care. Royal College of General Practitioners. URL: https://www.rcgp.org.uk/-/media/Files/CIRC/CIRC-AI-REPORT.ashx?la=en [accessed 2020-11-30]
16. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf 2019 Mar 12;28(3):231-237 [FREE Full text] [doi: 10.1136/bmjqs-2018-008370] [Medline: 30636200]
17. Kueper JK, Terry AL, Zwarenstein M, Lizotte DJ. Artificial intelligence and primary care research: a scoping review. Ann Fam Med 2020 May 01;18(3):250-258 [FREE Full text] [doi: 10.1370/afm.2518] [Medline: 32393561]
18. Millenson M, Baldwin J, Zipperer L, Singh H. Beyond Dr. Google: the evidence on consumer-facing digital tools for diagnosis. Diagnosis (Berl) 2018 Sep 25;5(3):95-105 [FREE Full text] [doi: 10.1515/dx-2018-0009] [Medline: 30032130]
19. Riches N, Panagioti M, Alam R, Cheraghi-Sohi S, Campbell S, Esmail A, et al. The Effectiveness of Electronic Differential Diagnoses (DDX) Generators: a systematic review and meta-analysis. PLoS One 2016 Mar 8;11(3):0148991 [FREE Full text] [doi: 10.1371/journal.pone.0148991] [Medline: 26954234]
20. Semigran H, Linder J, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. Br Med J 2015 Jul 08;351:3480 [FREE Full text] [doi: 10.1136/bmj.h3480] [Medline: 26157077]
21. Chambers D, Cantrell AJ, Johnson M, Preston L, Baxter SK, Booth A, et al. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. BMJ Open 2019 Aug 01;9(8):027743 [FREE Full text] [doi: 10.1136/bmjopen-2018-027743] [Medline: 31375610]
22. Walter FM, Thompson MJ, Wellwood I, Abel GA, Hamilton W, Johnson M, et al. Evaluating diagnostic strategies for early detection of cancer: the CanTest framework. BMC Cancer 2019 Jun 14;19(1):586 [FREE Full text] [doi: 10.1186/s12885-019-5746-6] [Medline: 31200676]
23. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, PRISMA-P Group. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev 2015 Jan 01;4(1):1 [FREE Full text] [doi: 10.1186/2046-4053-4-1] [Medline: 25554246]
24. Jones O, Ranmuthu C, Prathivadi K, Saji S, Calanzani N, Emery J, et al. Establishing which modalities of artificial intelligence (AI) for the early detection and diagnosis of cancer are ready for implementation in primary care: a systematic review. Prospero: International prospective register of systematic reviews 2020 [FREE Full text] [doi: 10.15124/CRD42020176674]
25. McCarthy J, Minsky M, Rochester N, Shannon C. A proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. AI Magazine,27(4), 12. 2006. URL: https://doi.org/10.1609/aimag.v27i4.1904 [accessed 2021-01-25]
26. Muehlhauser L. What should we learn from past AI forecasts? Open Philanthropy Project. 2016. URL: https://www.openphilanthropy.org/focus/global-catastrophic-risks/potential-risks-advanced-artificial-intelligence/what-should-we-learn-past-ai-forecasts [accessed 2021-01-25]
27. AI in the UK?: Ready, Willing and Able. HOUSE OF LORDS: Select Committee on Artificial Intelligence. 2018. URL: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf [accessed 2021-01-25]
28. arXiv.org e-Print archive. Cornell University. URL: https://arxiv.org/ [accessed 2020-11-30]
29. Research. Google AI - research. URL: https://ai.google/research/ [accessed 2020-11-30]
30. Emerging technology, computer, and software research. Microsoft Research. URL: https://www.microsoft.com/en-us/research/ [accessed 2020-11-30]
31. Artificial intelligence. IBM Research. URL: https://www.research.ibm.com/artificial-intelligence/ [accessed 2020-11-30]

32. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536 [FREE Full text] [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
33. Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M, et al. Guidance on the conduct of narrative synthesis in systematic reviews. ESRC Methods Program Swindon.: University of Lancaster; 2006. URL: https://tinyurl.com/19ok1m1j [accessed 2021-01-25]
34. Hornbrook MC, Goshen R, Choman E, O'Keeffe-Rosetti M, Kinar Y, Liles EG, et al. Early colorectal cancer detected by machine learning model using gender, age, and complete blood count data. Dig Dis Sci 2017 Oct 23;62(10):2719-2727. [doi: 10.1007/s10620-017-4722-8] [Medline: 28836087]
35. Survival three times higher when cancer is diagnosed early. Cancer Research UK. URL: https://www.cancerresearchuk.org/about-us/cancer-news/press-release/2015-08-10-survival-three-times-higher-when-cancer-is-diagnosed-early [accessed 2019-12-17]
36. NHS Long Term Plan. URL: https://www.longtermplan.nhs.uk/ [accessed 2021-02-08]
37. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017 Feb 02;542(7639):115-118. [doi: 10.1038/nature21056] [Medline: 28117445]
38. Marchetti MA, Codella NC, Dusza SW, Gutman DA, Helba B, Kalloo A, International Skin Imaging Collaboration. Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. J Am Acad Dermatol 2018 Feb;78(2):270-277.e1 [FREE Full text] [doi: 10.1016/j.jaad.2017.08.016] [Medline: 28969863]
39. ALzubi JA, Bharathikannan B, Tanwar S, Manikandan R, Khanna A, Thaventhiran C. Boosted neural network ensemble classification for lung cancer disease diagnosis. Applied Soft Computing 2019 Jul;80:579-591. [doi: 10.1016/j.asoc.2019.04.031]
40. Chang C, Hsu M. The study that applies artificial intelligence and logistic regression for assistance in differential diagnostic of pancreatic cancer. Expert Systems with Applications 2009 Sep;36(7):10663-10672. [doi: 10.1016/j.eswa.2009.02.046]
41. Cooper JA, Parsons N, Stinton C, Mathews C, Smith S, Halloran SP, et al. Risk-adjusted colorectal cancer screening using the FIT and routine screening data: development of a risk prediction model. Br J Cancer 2018 Jan 2;118(2):285-293 [FREE Full text] [doi: 10.1038/bjc.2017.375] [Medline: 29096402]
42. Cowley J. The use of knowledge discovery databases in the identification of patients with colorectal cancer. University of Hull [Dissertation]. 2012 Jul 01. URL: https://core.ac.uk/download/pdf/18526844.pdf [accessed 2020-11-30]
43. Daqqa K, Maghari A. Prediction and diagnosis of leukemia using classification algorithms. In: Proceedings of 8th International Conference on Information Technology. 2017 Presented at: 8th International Conference on Information Technology (ICIT); May 17-18, 2017; Amman, Jordan p. 638-643. [doi: 10.1109/icitech.2017.8079919]
44. Gorynski K, Safian I, Gradzki W. Artificial neural networks approach to early lung cancer detection. Cent Eur J Med 2014;9(5):632-641. [doi: 10.2478/s11536-013-0327-6]
45. Hart GR, Roffman DA, Decker R, Deng J. A multi-parameterized artificial neural network for lung cancer risk prediction. PLoS One 2018 Oct 24;13(10):0205264 [FREE Full text] [doi: 10.1371/journal.pone.0205264] [Medline: 30356283]
46. Kalra P, Togami J, Bansal BSG, Partin AW, Brawer MK, Babaian RJ, et al. A neurocomputational model for prostate carcinoma detection. Cancer 2003 Nov 01;98(9):1849-1854 [FREE Full text] [doi: 10.1002/cncr.11748] [Medline: 14584066]
47. Kang G, Ni Z. Research on early risk predictive model and discriminative feature selection of cancer based on real-world routine physical examination data. In: Proceedings of IEEE Int Conf Bioinforma Biomed BIBM. 2016 Presented at: IEEE Int Conf Bioinforma Biomed BIBM; 2016; Shenzhen, China p. 1512-1519. [doi: 10.1109/bibm.2016.7822746]
48. Kinar Y, Kalkstein N, Akiva P, Levin B, Half EE, Goldshtein I, et al. Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study. J Am Med Inform Assoc 2016 Sep 15;23(5):879-890 [FREE Full text] [doi: 10.1093/jamia/ocv195] [Medline: 26911814]
49. Kop R, Hoogendoorn M, Teije AT, Büchner FL, Slottje P, Moons LM, et al. Predictive modeling of colorectal cancer using a dedicated pre-processing pipeline on routine electronic medical records. Comput Biol Med 2016 Sep 01;76:30-38. [doi: 10.1016/j.compbiomed.2016.06.019] [Medline: 27392227]
50. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep 2016 May 17;6(1):26094 [FREE Full text] [doi: 10.1038/srep26094] [Medline: 27185194]
51. Payandeh M, Aeinfar M, Aeinfar V, Hayati M. A new method for diagnosis and predicting blood disorder and cancer using artificial intelligence (Artificial Neural Networks). Int J Hematol Stem Cell Res 2009;3(4):25-33. [doi: 10.1109/isie.2009.5213591]
52. Birks J, Bankhead C, Holt TA, Fuller A, Patnick J. Evaluation of a prediction model for colorectal cancer: retrospective analysis of 2.5 million patient records. Cancer Med 2017 Oct 21;6(10):2453-2460 [FREE Full text] [doi: 10.1002/cam4.1183] [Medline: 28941187]
53. Kinar Y, Akiva P, Choman E, Kariv R, Shalev V, Levin B, et al. Performance analysis of a machine learning flagging system used to identify a group of individuals at a high risk for colorectal cancer. PLoS One 2017 Feb 9;12(2):0171759 [FREE Full text] [doi: 10.1371/journal.pone.0171759] [Medline: 28182647]

54. Malik N, Idris W, Gunawan TS, Olanrewaju RF, Ibrahim SN. Classification of normal and crackles respiratory sounds into healthy and lung cancer groups. Int J Electr Comput Eng 2018 Jun 01;8(3):1530. [doi: 10.11591/ijece.v8i3.pp1530-1538]
55. Adams K, Sideris M, Papagrigoriadis S. Lunchtime Posters-Can we make “Straight to Test” decisions in Two Week Wait (2WW) patients with the help of an Artificial Neural Network (ANN)? Colorectal Dis 2014 Aug 22;16:41-68. [doi: 10.1111/codi.12643]
56. Ahmed A, Shah M, Wahid A, ul Islam S, Abbasi MK, Asghar MN. Big data analytics using neural networks for earlier cancer detection. J Med Imaging Hlth Inform 2017 Oct 01;7(6):1469-1474. [doi: 10.1166/jmihi.2017.2189]
57. Ahmed K, Emran AA, Jesmin T, Mukti RF, Rahman MZ, Ahmed F. Early detection of lung cancer risk using data mining. Asian Pac J Cancer Prev 2013;14(1):595-598 [FREE Full text] [doi: 10.7314/apjcp.2013.14.1.595] [Medline: 23534801]
58. Ahmen U, Rasool G, Zafar S, Maqbool HF. Fuzzy Rule Based Diagnostic System to Detect the Lung Cancer. 2018 Presented at: 2018 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube); November 12-13, 2018; Quetta, Pakistan URL: https://ieeexplore.ieee.org/document/8610976 [doi: 10.1109/ICECUBE.2018.8610976]
59. Alaa A, Moon K, Hsu W, van der Schaar M. ConfidentCare: a clinical decision support system for personalized breast cancer screening. IEEE Trans Multimedia 2016 Oct;18(10):1942-1955. [doi: 10.1109/TMM.2016.2589160]
60. Alharbi A, Tchier F, Rashidi M. Using a GeneticFuzzy algorithm as a computer aided breast cancer diagnostic tool. Asian Pac J Cancer Prev 2016;17(7):3651-3658 [FREE Full text] [Medline: 27510026]
61. Ayeldeen H, Elfattah MA, Shaker O, Hassanien AE, Kim TH. Case-Based Retrieval Approach of Clinical Breast Cancer Patients. 2015 Presented at: 2015 3rd International Conference on Computer, Information and Application; May 21-23, 2015; Yeosu, South Korea. [doi: 10.1109/CIA.2015.17]
62. Balachandran K. An efficient optimization based lung cancer pre-diagnosis system with aid of Feed Forward Back Propagation Neural Network (FFBNN). J Theor Appl Inf Technol 2013 Oct;56(2):263-271 [FREE Full text]
63. Bhar JA, George V, Malik B. Cloud Computing with Machine Learning Could Help Us in the Early Diagnosis of Breast Cancer. 2015 Presented at: Second International Conference on Advances in Computing and Communication Engineering; May 1-2, 2015; Dehradun, India. [doi: 10.1109/ICACCE.2015.62]
64. CHauhan R, Kaur H, Sharma S. A Feature Based Approach for Medical Databases. In: Proceedings of the International Conference on Advances in Information Communication Technology & Computing. 2016 Presented at: AICTC '16; Aug 2016; Bikaner, India. [doi: 10.1145/2979779.2979873]
65. Chen Y, Joo EM. Biomedical diagnosis and prediction using parsimonious fuzzy neural networks. 2012 Presented at: 38th Annual Conference on IEEE Industrial Electronics Society; December 24, 2012; Montreal, QC, Canada. [doi: 10.1109/IECON.2012.6388524]
66. Choudhury T, Kumar V, Nigam D, Vashisht V. Intelligent Classification of Lung & Oral Cancer through Diverse Data Mining Algorithms. Presented at: 2016 International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE); 2016; Ghaziabad p. 133-138. [doi: 10.1109/ICMETE.2016.24]
67. Çınar M, Engin M, Engin EZ, Ziya Ateşçi Y. Early prostate cancer diagnosis by using artificial neural networks and support vector machines. Expert Systems with Applications 2009 Apr;36(3):6357-6361. [doi: 10.1016/j.eswa.2008.08.010]
68. Del Grossi AA, De Mattos Senefonte HC, Quaglio VG. Prostate cancer biopsy recommendation through use of machine learning classification techniques. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Switzerland: Springer; 2014:710-721.
69. Durga S, Kasturi K. Lung disease prediction system using data mining techniques. J Adv Res in Dynamical and Contr Sys 2017;9(5):62-66 [FREE Full text]
70. Elhoseny M, Bian G, Lakshmanaprabu S, Shankar K, Singh AK, Wu W. Effective features to classify ovarian cancer data in internet of medical things. Computer Networks 2019 Aug;159:147-156. [doi: 10.1016/j.comnet.2019.04.016]
71. Elshazly HI, Elkorany AM, Hassanien AE. Ensemble-based classifiers for prostate cancer diagnosis. In: Proocedings of the 9th International Computer Engineering Conference: Today Information Society What’s Next?, ICENCO 2013.: IEEE Computer Society; 2013 Presented at: 9th International Computer Engineering Conference: Today Information Society What’s Next?, ICENCO 2013; 2013; 9th International Computer Engineering Conference: Today Information Society What’s Next?, ICENCO 2013 p. 49-54. [doi: 10.1109/ICENCO.2013.6736475]
72. Fan Y, Chaovalitwongse WA. Optimizing feature selection to improve medical diagnosis. Ann Oper Res 2009 Jan 6;174(1):169-183. [doi: 10.1007/s10479-008-0506-z]
73. Gaebel J, Cypko MA, Lemke HU. Accessing patient information for probabilistic patient models using existing standards. Stud Health Technol Inform 2016;223:107-112. [Medline: 27139392]
74. Gao Z, Gong J, Qin Q, Lin J. [Application of support vector machine in the detection of early cancer]. Sheng Wu Yi Xue Gong Cheng Xue Za Zhi 2005 Oct;22(5):1045-1048. [Medline: 16294750]
75. Gelnarová E, Šafařík L. Comparison of three statistical classifiers on a prostate cancer data. Neural Network World 2005;15(4):311-318 [FREE Full text]
76. Ghaderzadeh M. Clinical decision support system for early detection of prostate cancer from benign hyperplasia of prostate. In: Proceedings of the 14th World Congress on Medical and Health Informatics, Pts 1 and 2. 2013 Presented at: Proceedings of the 14th World Congress on Medical and Health Informatics, Pts 1 and 2; 2013; Netherlands p. 928. [doi: 10.3233/978-1-61499-289-9-928]

77. Ghany KKA, Ayeldeen H, Zawbaa HM, Shaker O, IEEE. A rough set-based reasoner for medical diagnosis. In: Proceedings of the International Conference on Green Computing and Internet of Things. 2015 Presented at: International Conference on Green Computing and Internet of Things; 2015; Beni-Suef University, Egypt p. 429-434. [doi: 10.1109/ICGCIoT.2015.7380502]
78. Ghany KKA, Ayeldeen H, Zawbaa HM, Shaker O, Ayedeen G, IEEE. Diagnosis of breast cancer using secured classifiers. In: Proceedings of the International Conference on Electrical and Computing Technologies and Applications. 2017 Presented at: International Conference on Electrical and Computing Technologies and Applications; 2017; Beni-Suef University, Egypt p. 680-684. [doi: 10.1109/ICECTA.2017.8251947]
79. Goraneseu F, Gorunescu M, El-Darzi E, Ene M, Gorunescu S. Statistical comparison of a probabilistic neural network approach in hepatic cancer diagnosis. In: EUROCON 2005 - The International Conference on Computer as a Tool. 2005 Presented at: EUROCON 2005 - The International Conference on Computer as a Tool; 2005; Belgrade; Yugoslavia p. 237-240. [doi: 10.1109/EURCON.2005.1629904]
80. Gorunescu F, Belciug S. Boosting backpropagation algorithm by stimulus-sampling: application in computer-aided medical diagnosis. J Biomed Inform 2016 Oct;63:74-81 [FREE Full text] [doi: 10.1016/j.jbi.2016.08.004] [Medline: 27498068]
81. Gorunescu M, Gorunescu F, Revett K. A neural computing-based approach for the early detection of hepatocellular carcinoma. Proceedings of World Academy of Science, Engineering and Technology 2006;17:65 [FREE Full text]
82. Govinda K, Singla K, Jain K. Fuzzy based uncertainty modeling of Cancer Diagnosis System. In: Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS). 2017 Presented at: International Conference on Intelligent Sustainable Systems (ICISS); Dec 7-8, 2017; Palladam, India p. 740-743. [doi: 10.1109/ISS1.2017.8389272]
83. Halpern Y, Horng SK, Choi Y, Sontag D. Electronic medical record phenotyping using the anchor and learn framework. J Am Med Inform Assoc 2016 Jul;23(4):731-740 [FREE Full text] [doi: 10.1093/jamia/ocw011] [Medline: 27107443]
84. Hart GR, Roffman DA, Decker R. Scientific abstracts and sessions. Med Phys 2018 Jun 11;45(6):e120-e706. [doi: 10.1002/mp.12938]
85. Hornbrook MC, Goshen R, Choman E, O'Keeffe-Rosetti M, Kinar Y, Liles EG, et al. Correction to: early colorectal cancer detected by machine learning model using gender, age, and complete blood count data. Dig Dis Sci 2018 Jan;63(1):270. [doi: 10.1007/s10620-017-4859-5] [Medline: 29181742]
86. Hsu JL, Hung PC, Lin HY, Hsieh CH. Applying under-sampling techniques and cost-sensitive learning methods on risk assessment of breast cancer. J Med Syst 2015 Apr;39(4):210. [doi: 10.1007/s10916-015-0210-x] [Medline: 25712814]
87. Ilhan HO, Celik E. The mesothelioma disease diagnosis with artificial intelligence methods. In: Proceedings of the 10th International Conference on Application of Information and Communication Technologies (AICT). 2016 Presented at: 2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT); Oct 12-14, 2016; Baku, Azerbaijan. [doi: 10.1109/ICAICT.2016.7991825]
88. Ji Z, Wang B. Identifying potential clinical syndromes of hepatocellular carcinoma using PSO-based hierarchical feature selection algorithm. Biomed Res Int 2014;2014:127572 [FREE Full text] [doi: 10.1155/2014/127572] [Medline: 24745007]
89. Kong Q, Wang D, Wang Y, Jin Y, Jiang B. Multi-objective neural network-based diagnostic model of prostatic cancer. System Engineering Theory and Practice 2018;38(2):532-544. [doi: 10.12011/1000-6788(2018)02-0532-13]
90. Kou L, Yuan Y, Sun J, Lin Y. Prediction of cancer based on mobile cloud computing and SVM. In: Proceedings of the International Conference on Dependable Systems and Their Applications (DSA). 2017 Presented at: 2017 International Conference on Dependable Systems and Their Applications (DSA); 2017; Beijing, China. [doi: 10.1109/DSA.2017.20]
91. Kshivets O. P2.11-13 Precise early detection of lung cancer and blood cell circuit. J Thoracic Oncol 2018 Oct;13(10):S783. [doi: 10.1016/j.jtho.2018.08.1360]
92. Liu S, Gaudiot J, Cristini V. Prototyping virtual cancer therapist (VCT): a software engineering approach. Conf Proc IEEE Eng Med Biol Soc 2006;2006:5424-5427. [doi: 10.1109/IEMBS.2006.259230] [Medline: 17945900]
93. Liu Y, Pan Q, Zhou Z. Improved feature selection algorithm for prognosis prediction of primary liver cancer. In: Intelligence Science II. Switzerland: Springer; 2018:422-430.
94. Meng J, Zhang R, Chen D. Utilizing narrative text from electronic health records for early warning model of chronic disease. In: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC). 2018 Presented at: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC); Oct 7-10, 2018; Miyazaki, Japan. [doi: 10.1109/SMC.2018.00713]
95. Mesrabadi HA, Faez K. Improving early prostate cancer diagnosis by using Artificial Neural Networks and Deep Learning. In: Proceedings of the 4th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS). 2018 Presented at: 2018 4th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS); Dec 25-27, 2018; Tehran, Iran. [doi: 10.1109/ICSPIS.2018.8700542]
96. Morgado P, Vicente H, Abelha A, Machado J, Neves J, Neves J. A case-based approach to colorectal cancer detection. In: Information Science and Applications 2017. Singapore: Springer; 2017:433-442.
97. Nalluri MR, Roy DS. Hybrid disease diagnosis using multiobjective optimization with evolutionary parameter optimization. J Healthc Eng 2017;2017:5907264 [FREE Full text] [doi: 10.1155/2017/5907264] [Medline: 29065626]
98. Nikitaev VG, Pronichev AN, Nagornov OV, Zaytsev SM, Polyakov EV, Romanov NA, et al. Decision support system in urologic cancer diagnosis. J Phys : Conf Ser 2019 Apr 16;1189:012032. [doi: 10.1088/1742-6596/1189/1/012032]

99. Polat K, Senturk U. A Novel ML Approach to Prediction of Breast Cancer: Combining of mad normalization, KMC based feature weighting and AdaBoostM1 classifierA novel ML approach to prediction of breast cancer: combining of mad normalization, KMC based feature weighting and AdaBoostM1 classifier. In: 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). 2018 Presented at: 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT); Oct 19-21, 2018; Ankara, Turkey. [doi: 10.1109/ISMSIT.2018.8567245]
100. Rahman A, Muniyandi RC. Feature selection from colon cancer dataset for cancer classification using Artificial Neural Network. Int J Adv Sci Engi and Info Tech. 2018. URL: https://www.researchgate.net/publication/328924307_Feature_selection_from_colon_cancer_dataset_for_cancer_classification_using_Artificial_Neural_Network [accessed 2021-02-08]
101. Ramya Devi M, Gomathy B. An intelligent system for the detection of breast cancer using feature selection and PCA methods. Int J Appl Engi Res. 2015. URL: https://www.researchgate.net/publication/283232820_An_intelligent_system_for_the_detection_of_breast_cancer_using_feature_selection_and_PCA_methods [accessed 2021-02-08]
102. Richter AN, Khoshgoftaar TM. Melanoma risk modeling from limited positive samples. Netw Model Anal Health Inform Bioinforma 2019 Apr 4;8(1):-. [doi: 10.1007/s13721-019-0186-4]
103. Safdari R, Arpanahi H, Langarizadeh M, Ghazisaiedi M, Dargahi H, Zendehdel K. Design a fuzzy rule-based expert system to aid earlier diagnosis of gastric cancer. Acta Inform Med 2018;26(1):19. [doi: 10.5455/aim.2018.26.19-23]
104. Shalev V, Kinar Y, Kalkstein N, Akiva P, Half E, Goldshtein I, et al. Computational analysis of blood counts significantly increases detection rate of gastric and colorectal cancers: PR0195 Esophageal, Gastric and Duodenal Disorders. J Gastroenterol and Hepatol 2013:761-762.
105. Sobar, Machmud R, Wijaya A. Behavior determinant based cervical cancer early detection with machine learning algorithm. Advanced Science Letters 2016;22(10):3120-3123. [doi: 10.1166/asl.2016.7980]
106. Soliman THA, Mohamed R, Sewissy AA. A hybrid analytical hierarchical process and deep neural networks approach for classifying breast cancer. In: Proceedings of the 11th International Conference on Computer Engineering & Systems (ICCES). 2016 Presented at: 2016 11th International Conference on Computer Engineering & Systems (ICCES); Dec 20-21, 2016; Cairo, Egypt. [doi: 10.1109/ICCES.2016.7822002]
107. Sushma Rani N, Srinivasa Rao P, Parimala P. An efficient statistical computation technique for health care big data using R. IOP Conf Ser : Mater Sci Eng 2017 Sep 07;225:012159. [doi: 10.1088/1757-899x/225/1/012159]
108. Wang D, Quek C, See Ng G. Ovarian cancer diagnosis using a hybrid intelligent system with simple yet convincing rules. Applied Soft Computing 2014 Jul;20:25-39. [doi: 10.1016/j.asoc.2013.12.018]
109. Wang G, Teoh JYC, Choi KS. Diagnosis of prostate cancer in a Chinese population by using machine learning methods. Annu Int Conf IEEE Eng Med Biol Soc 2018 Jul;2018:1-4. [doi: 10.1109/EMBC.2018.8513365] [Medline: 30440319]
110. Xu W, Zhang R, Qimin E, Liu J, Laing C. [The clinical application of data mining in laryngeal cancer]. Lin Chung Er Bi Yan Hou Tou Jing Wai Ke Za Zhi 2015 Jul;29(14):1272-1275. [Medline: 26672241]
111. Yasodha P, Ananthanarayanan NR. Analysing big data to build knowledge based system for early detection of ovarian cancer. Indian J Sci and Tech 2015;8(14). [doi: 10.17485/ijst/2015/v8i14/65745]
112. Zangooei MH, Habibi J, Alizadehsani R. Disease Diagnosis with a hybrid method SVR using NSGA-II. Neurocomputing 2014 Jul;136:14-29. [doi: 10.1016/j.neucom.2014.01.042]
113. Zhang L, Wang H, Liang J, Wang J. Decision support in cancer base on fuzzy adaptive PSO for feedforward neural network training. In: Proceedings of the International Symposium on Computer Science and Computational Technology. 2008 Presented at: 2008 International Symposium on Computer Science and Computational Technology; Dec 20-22, 2008; Shanghai, China. [doi: 10.1109/iscsct.2008.73]
114. Zhang Z, Zhang H, Bast Jr RC. An application of artificial neural networks in ovarian cancer early detection. In: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. 2000 Presented at: IEEE-INNS-ENNS International Joint Conference on Neural Networks; July 27, 2000; Como, Italy URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-
115. Hoogendoorn M, Szolovits P, Moons LM, Numans ME. Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer. Artif Intell Med 2016 May;69:53-61 [FREE Full text] [doi: 10.1016/j.artmed.2016.03.003] [Medline: 27085847]
116. Moss S, Mathews C, Day TJ, Smith S, Seaman HE, Snowball J, et al. Increased uptake and improved outcomes of bowel cancer screening with a faecal immunochemical test: results from a pilot study within the national screening programme in England. Gut 2017 Sep 07;66(9):1631-1644. [doi: 10.1136/gutjnl-2015-310691] [Medline: 27267903]
117. Bond WF, Schwartz LM, Weaver KR, Levick D, Giuliano M, Graber ML. Differential diagnosis generators: an evaluation of currently available computer programs. J Gen Intern Med 2012 Feb 26;27(2):213-219 [FREE Full text] [doi: 10.1007/s11606-011-1804-8] [Medline: 21789717]
118. DXplain. DXplain. URL: http://www.mghlcs.org/projects/dxplain [accessed 2020-11-30]
119. Symcat Symptom Checker. Symcat Symptom Checker. URL: http://www.symcat.com/ [accessed 2020-11-30]

120. Heckerling PS, Elstein AS, Terzian CG, Kushner MS. The effect of incomplete knowledge on the diagnoses of a computer consultant system. Med Inform (Lond) 1991 Jul 12;16(4):363-370. [doi: 10.3109/14639239109067658] [Medline: 1762472]
121. Miller RA, Pople HE, Myers JD. Internist-I, an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med 1982 Aug 19;307(8):468-476. [doi: 10.1056/nejm198208193070803]
122. Miller RA, McNeil MA, Challinor SM, Masarie FE, Myers JD. The Internist-1/quick medical reference project--status report. West J Med 1986 Dec;145(6):816-822 [FREE Full text] [Medline: 3544509]
123. Wexler JR, Swender PT, Tunnessen WW, Oski FA. Impact of a system of computer-assisted diagnosis. Initial evaluation of the hospitalized patient. Am J Dis Child 1975 Feb 01;129(2):203-205. [doi: 10.1001/archpedi.1975.02120390037008] [Medline: 1091140]
124. Rodríguez-González A, Torres-Niño J, Mayer MA, Alor-Hernandez G, Wilkinson MD. Analysis of a multilevel diagnosis decision support system and its implications: a case study. Comput Math Methods Med 2012 Sep;2012(9):367345-367333 [FREE Full text] [doi: 10.1155/2012/367345] [Medline: 23320043]
125. Clinical decision support. PEPID. URL: https://www.pepid.com/ [accessed 2020-11-30]
126. Apkon M, Mattera JA, Lin Z, Herrin J, Bradley EH, Carbone M, et al. A randomized outpatient trial of a decision-support information technology tool. Arch Intern Med 2005 Nov 14;165(20):2388-2394. [doi: 10.1001/archinte.165.20.2388] [Medline: 16287768]
127. Nelson SJ, Blois MS, Tuttle MS, Erlbaum M, Harrison P, Kim H, et al. Evaluating RECONSIDER. J Med Syst 1985 Dec;9(5-6):379-388. [doi: 10.1007/bf00992575]
128. Our Solutions!. Abtrace. URL: https://www.abtrace.co/solution/ [accessed 2020-11-30]
129. C the signs. C the Signs. URL: https://cthesigns.co.uk/ [accessed 2020-11-30]
130. DocResponse. DocResponse. URL: https://www.docresponse.com/ [accessed 2020-11-30]
131. Isabel Pro - the DDx Generator. Isabel Healthcare. URL: https://uk.isabelhealthcare.com/products/isabel-pro-ddx-generator [accessed 2020-11-30]
132. Medial EarlySign. Medial EarlySign. URL: https://earlysign.com/ [accessed 2020-11-30]
133. Edwards HB, Marques E, Hollingworth W, Horwood J, Farr M, Bernard E, et al. Use of a primary care online consultation system, by whom, when and why: evaluation of a pilot observational study in 36 general practices in South West England. BMJ Open 2017 Nov 22;7(11):e016901 [FREE Full text] [doi: 10.1136/bmjopen-2017-016901] [Medline: 29167106]
134. Simptify. Simptify.com. URL: https://symptify.com/ [accessed 2020-11-30]
135. Check your symptoms online. Symptomate. URL: https://symptomate.com/ [accessed 2020-11-30]
136. Liang H, Tsui BY, Ni H, Valentim CCS, Baxter SL, Liu G, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat Med 2019 Mar;25(3):433-438. [doi: 10.1038/s41591-018-0335-9] [Medline: 30742121]
137. Zhelezniak V, Busbridge D, Shen A, Smith S, Hammerla N. Decoding decoders: finding optimal representation spaces for unsupervised similarity tasks. ICLR 2018 Work Track. arXiv.org. Preprint posted online September 5, 2018. URL: https://arxiv.org/abs/1805.03435 [accessed 2020-11-30]
138. Douglas L, Zarov I, Gourgoulias K, Lucas C, Hart C, Baker A, et al. A universal marginalizer for amortized inference in generative models. NIPS 2017 Work Adv Approx Bayesian Inference. Preprint posted online November 2, 2017. URL: https://arxiv.org/abs/1711.00695 [accessed 2020-11-30]
139. Smith S, Turban D, Hamblin S, Hammerla N. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. arXiv.org. Preprint posted online February 13, 2017. URL: https://arxiv.org/abs/1702.03859 [accessed 2020-11-30]
140. NHS 111 Powered by Babylon - outcomes evaluation. Babylon Health. 2017. URL: https://assets.babylonhealth.com/nhs/NHS-111-Evaluation-of-outcomes.pdf [accessed 2020-11-30]
141. Middleton K, Butt M, Hammerla N, Hamblin S, Mehta K, Parsa A. Sorting out symptoms: design and evaluation of the 'babylon check' automated triage system. arXiv.org. Preprint posted online June 7, 2016. URL: https://arxiv.org/abs/1606.02041 [accessed 2020-11-30]
142. Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain: an evolving diagnostic decision-support system. J Am Med Assoc 1987 Jul 03;258(1):67. [doi: 10.1001/jama.1987.03400010071030]
143. Barnett GO, Famiglietti KT, Kim RJ, Hoffer EP, Feldman MJ. DXplain on the internet. Proc AMIA Symp 1998:607-611 [FREE Full text] [Medline: 9929291]
144. Bauer BA, Lee M, Bergstrom L, Wahner-Roedler DL, Bundrick J, Litin S, et al. Internal medicine resident satisfaction with a diagnostic decision support system (DXplain) introduced on a teaching hospital service. Proc AMIA Symp 2002:31-35 [FREE Full text] [Medline: 12463781]
145. Berner ES, Webster GD, Shugerman AA, Jackson JR, Algina J, Baker AL, et al. Performance of four computer-based diagnostic systems. N Engl J Med 1994 Jun 23;330(25):1792-1796. [doi: 10.1056/nejm199406233302506]
146. Elhanan G, Socratous SA, Cimino JJ. Integrating DXplain into a clinical information system using the World Wide Web. Proc AMIA Annu Fall Symp 1996:348-352 [FREE Full text] [Medline: 8947686]
147. Elkin PL, Liebow M, Bauer BA, Chaliki S, Wahner-Roedler D, Bundrick J, et al. The introduction of a diagnostic decision support system (DXplain™) into the workflow of a teaching hospital service can decrease the cost of service for diagnostically challenging Diagnostic Related Groups (DRGs). Int J Med Inform 2010 Nov;79(11):772-777 [FREE Full text] [doi: 10.1016/j.ijmedinf.2010.09.004] [Medline: 20951080]

148. Feldman MJ, Octo Barnett G. An approach to evaluating the accuracy of DXplain. Computer Methods and Programs in Biomedicine 1991 Aug;35(4):261-266. [doi: 10.1016/0169-2607(91)90004-d]
149. Hammersley J, Cooney K. Evaluating the utility of available differential diagnosis systems. Proc Annu Symp Comput Appl Med Care 1988 Nov 9:229-231 [FREE Full text]
150. Hoffer EP, Feldman MJ, Kim RJ, Famiglietti KT, Barnett GO. DXplain: patterns of use of a mature expert system. AMIA Annu Symp Proc 2005:321-325 [FREE Full text] [Medline: 16779054]
151. London S. DXplain: a web-based diagnostic decision support system for medical students. Medical Reference Services Quarterly 1998 May 07;17(2):17-28. [doi: 10.1300/j115v17n02_02]
152. Elstein A, Friedman C, Wolf F, Murphy G, Miller J, Fine P, et al. Effects of a decision support system on the diagnostic accuracy of users: a preliminary report. J Am Med Inform Assoc 1996;3(6):422-428 [FREE Full text] [doi: 10.1136/jamia.1996.97084515] [Medline: 8930858]
153. Friedman CP, Elstein AS, Wolf FM, Murphy GC, Franz TM, Heckerling PS, et al. Enhancement of clinicians' diagnostic reasoning by computer-based consultation: a multisite study of 2 systems. J Am Med Assoc 1999 Nov 17;282(19):1851-1856. [doi: 10.1001/jama.282.19.1851] [Medline: 10573277]
154. Gozum ME. Emulating cognitive diagnostic skills without clinical experience: a report of medical students using Quick Medical Reference and Iliad in the diagnosis of difficult clinical cases. Proc Annu Symp Comput Appl Med Care 1994:991 [FREE Full text] [Medline: 7950096]
155. Graber MA, VanScoy D. How well does decision support software perform in the emergency department? Emerg Med J 2003 Sep 01;20(5):426-428 [FREE Full text] [doi: 10.1136/emj.20.5.426] [Medline: 12954680]
156. Lange L, Haak S, Lincoln M. Use of Iliad to improve diagnostic performance of nurse practitioner students. J Nurs Educ 1997;36(1):36-45. [doi: 10.3928/0148-4834-19970101-09]
157. Lau L, Warner H, Poulsen A. Research review: a computer-based diagnostic model for individual case review. Top Health Inf Manage 1995 Feb;15(3):67-79. [Medline: 10140306]
158. Li YC, Haug PJ, Lincoln MJ, Turner CW, Pryor TA, Warner HH. Assessing the behavioral impact of a diagnostic decision support system. Proc Annu Symp Comput Appl Med Care 1995:805-809 [FREE Full text] [Medline: 8563402]
159. Lincoln MJ, Turner CW, Haug PJ, Warner HR, Williamson JW, Bouhaddou O, et al. Iliad training enhances medical students' diagnostic skills. J Med Syst 1991 Feb;15(1):93-110. [doi: 10.1007/bf00993883]
160. Murphy GC, Friedman CP, Elstein AS, Wolf FM, Miller T, Miller JG. The influence of a decision support system on the differential diagnosis of medical practitioners at three levels of training. Proc AMIA Annu Fall Symp 1996:219-223 [FREE Full text] [Medline: 8947660]
161. Wolf FM, Friedman CP, Elstein AS, Miller JG, Murphy GC, Heckerling P, et al. Changes in diagnostic decision-making after a computerized decision support consultation based on perceptions of need and helpfulness: a preliminary report. Proc AMIA Annu Fall Symp 1997:263-267 [FREE Full text] [Medline: 9357629]
162. Ramnarayan P, Roberts GC, Coren M, Nanduri V, Tomlinson A, Taylor PM, et al. Assessment of the potential impact of a reminder system on the reduction of diagnostic errors: a quasi-experimental study. BMC Med Inform Decis Mak 2006 Apr 28;6(1):22 [FREE Full text] [doi: 10.1186/1472-6947-6-22] [Medline: 16646956]
163. Ramnarayan P, Winrow A, Coren M, Nanduri V, Buchdahl R, Jacobs B, et al. Diagnostic omission errors in acute paediatric practice: impact of a reminder system on decision-making. BMC Med Inform Decis Mak 2006 Nov 06;6:37 [FREE Full text] [doi: 10.1186/1472-6947-6-37] [Medline: 17087835]
164. Carlson J, Abel M, Bridges D, Tomkowiak J. The impact of a diagnostic reminder system on student clinical reasoning during simulated case studies. Simulation in Healthcare: J Society Simul Healthcare 2011;6(1):11-17. [doi: 10.1097/sih.0b013e3181f24acd]
165. Graber ML. Taking steps towards a safer future: measures to promote timely and accurate medical diagnosis. Am J Med 2008 May;121(5 Suppl):S43-S46. [doi: 10.1016/j.amjmed.2008.02.006] [Medline: 18440355]
166. Graber ML, Tompkins D, Holland JJ. Resources medical students use to derive a differential diagnosis. Med Teach 2009 Jun 27;31(6):522-527. [doi: 10.1080/01421590802167436] [Medline: 19811168]
167. Ramnarayan P, Cronje N, Brown R, Negus R, Coode B, Moss P, et al. Validation of a diagnostic reminder system in emergency medicine: a multi-centre study. Emerg Med J 2007 Sep 01;24(9):619-624 [FREE Full text] [doi: 10.1136/emj.2006.044107] [Medline: 17711936]
168. Bavdekar S, Pawar M. Evaluation of an Internet-Delivered Pediatric Diagnosis Support System (ISABEL®) in a Tertiary Care Center in India. Indian Pediatr 2005;42(11):91. [Medline: 16340049]
169. Ramnarayan P, Britto J. Paediatric clinical decision support systems. Arch Dis Child 2002 Nov;87(5):361-362 [FREE Full text] [doi: 10.1136/adc.87.5.361] [Medline: 12390900]
170. Meyer AND, Giardina TD, Spitzmueller C, Shahid U, Scott TMT, Singh H. Patient perspectives on the usefulness of an artificial intelligence-assisted symptom checker: cross-sectional survey study. J Med Internet Res 2020 Jan 30;22(1):14679 [FREE Full text] [doi: 10.2196/14679] [Medline: 32012052]
171. Waxman HS, Worley WE. Computer-assisted adult medical diagnosis: subject review and evaluation of a new microcomputer-based system. Medicine (Baltimore) 1990 May;69(3):125-136. [Medline: 2189054]

172. Goshen R, Choman E, Ran A, Muller E, Kariv R, Chodick G, et al. Computer-assisted flagging of individuals at high risk of colorectal cancer in a large health maintenance organization using the colonflag test. JCO Clinical Cancer Informatics 2018 Dec(2):1-8. [doi: 10.1200/cci.17.00130]
173. Zack CJ, Senecal C, Kinar Y, Metzger Y, Bar-Sinai Y, Widmer RJ, et al. Leveraging machine learning techniques to forecast patient prognosis after percutaneous coronary intervention. JACC Cardiovasc Interv 2019 Jul 22;12(14):1304-1311 [FREE Full text] [doi: 10.1016/j.jcin.2019.02.035] [Medline: 31255564]
174. Cahn A, Shoshan A, Sagiv T, Yesharim R, Goshen R, Shalev V, et al. Prediction of progression from pre-diabetes to diabetes: development and validation of a machine learning model. Diabetes Metab Res Rev 2020 Feb 14;36(2):e3252. [doi: 10.1002/dmrr.3252] [Medline: 31943669]
175. EMIS Health - online triage. EMIS Health. URL: https://www.emishealth.com/products/partner-products/online-triage/ [accessed 2020-11-30]
176. Hurley Group. Hurley Group. URL: http://hurleygroup.co.uk/ [accessed 2020-11-30]
177. Carter M, Fletcher E, Sansom A, Warren FC, Campbell JL. Feasibility, acceptability and effectiveness of an online alternative to face-to-face consultation in general practice: a mixed-methods study of webGP in six Devon practices. BMJ Open 2018 Feb 15;8(2):018688 [FREE Full text] [doi: 10.1136/bmjopen-2017-018688] [Medline: 29449293]
178. Cowie J, Calveley E, Bowers G, Bowers J. Evaluation of a digital consultation and self-care advice tool in primary care: a multi-methods study. Int J Environ Res Public Health 2018 May 02;15(5):896 [FREE Full text] [doi: 10.3390/ijerph15050896] [Medline: 29724040]
179. Arene I, Ahmed W, Fox M, Barr CE, Fisher K. Evaluation of quick medical reference (QMR) as a teaching tool. MD Comput 1998;15(5):323-326. [Medline: 9753979]
180. Bacchus CM, Quinton C, O'Rourke K, Detsky AS. A randomized crossover trial of quick medical reference (QMR) as a teaching tool for medical interns. J Gen Intern Med 1994 Nov;9(11):616-621. [doi: 10.1007/bf02600304]
181. Bankowitz R, McNeil M, Challinor S, Parker R, Kapoor W, Miller R. A computer-assisted medical diagnostic consultation service. Implementation and prospective evaluation of a prototype. Ann Intern Med 1989 May 15;110(10):824-832 [FREE Full text] [doi: 10.7326/0003-4819-110-10-824] [Medline: 2653156]
182. Berner ES, Maisiak RS, Cobbs CG, Taunton OD. Effects of a decision support system on physicians' diagnostic performance. J Am Med Inform Assoc 1999 Sep 01;6(5):420-427 [FREE Full text] [doi: 10.1136/jamia.1999.0060420] [Medline: 10495101]
183. Lemaire JB, Schaefer JP, Martin LA, Faris P, Ainslie MD, Hull RD. Effectiveness of the Quick Medical Reference as a diagnostic tool. Can Med Asso J 1999 Sep 21;161(6):725-728 [FREE Full text] [Medline: 10513280]
184. Kop R, Hoogendoorn M, Moons L, Numans M, ten Teije A. On the advantage of using dedicated data mining techniques to predict colorectal cancer. In: Holmes J, Bellazzi R, Sacchi L, Peek N, editors. Artificial Intelligence in Medicine. AIME 2015. Lecture Notes in Computer Science, vol 9105. Switzerland: Springer International Publishing; 2015:133-142.
185. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. Br Med J 2009 May 28;338(may28 1):605-605. [doi: 10.1136/bmj.b605] [Medline: 19477892]
186. Singh H, Sittig DF. A sociotechnical framework for Safety-Related Electronic Health Record Research Reporting: The SAFER Reporting framework. Annals of Internal Medicine 2020 Jun 02;172(11_Supplement):92-100. [doi: 10.7326/m19-0879]
187. Usher-Smith JA, Sharp SJ, Griffin SJ. The spectrum effect in tests for risk prediction, screening, and diagnosis. Br Med J 2016 Jun 22;353:3139 [FREE Full text] [doi: 10.1136/bmj.i3139] [Medline: 27334281]
188. Kanagasingam Y, Xiao D, Vignarajan J, Preetham A, Tay-Kearney M, Mehrotra A. Evaluation of artificial intelligence-based grading of diabetic retinopathy in primary care. JAMA Netw Open 2018 Sep 07;1(5):182665 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.2665] [Medline: 30646178]
189. 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]
190. Forcier M, Gallois H, Mullan S, Joly Y. Integrating artificial intelligence into health care through data access: can the GDPR act as a beacon for policymakers? J Law Biosci 2019 Oct;6(1):317-335 [FREE Full text] [doi: 10.1093/jlb/lsz013] [Medline: 31666972]
191. European Parlimentary Research Service (EPRS). The impact of the General Data Protection Regulation (GDPR) on artificial intelligence. STUDY: Panel for the Future of Science and Technology. 2020. URL: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf [accessed 2020-11-30]

Abbreviations

AI: artificial intelligence
AUROC: area under the receiver operating characteristic
CT: computed tomography
EHR: electronic health record
NHS: National Health Service

NIHR: National Institute for Health Research
NPV: negative predictive value
PPV: positive predictive value
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-analysis
QUADAS-2: quality assessment of diagnostic accuracy studies-2

Edited by G Eysenbach; submitted 13.08.20; peer-reviewed by Y Liang, Y Chu, R Verheij; comments to author 01.10.20; revised version received 05.11.20; accepted 30.11.20; published 03.03.21

Please cite as:
Jones OT, Calanzani N, Saji S, Duffy SW, Emery J, Hamilton W, Singh H, de Wit NJ, Walter FM
Artificial Intelligence Techniques That May Be Applied to Primary Care Data to Facilitate Earlier Diagnosis of Cancer: Systematic Review
J Med Internet Res 2021;23(3):e23483
URL: https://www.jmir.org/2021/3/e23483
doi: 10.2196/23483
PMID: 33656443

Original Paper

Assessment of Diagnostic Competences With Standardized Patients Versus Virtual Patients: Experimental Study in the Context of History Taking

Maximilian C Fink[1], MSc; Victoria Reitmeier[1], Dr med; Matthias Stadler[2,3], Dr phil; Matthias Siebeck[1,3], Prof Dr, MME (D); Frank Fischer[2,3], Prof Dr; Martin R Fischer[1], Prof Dr, MME (Bern)

[1]Institute for Medical Education, University Hospital, LMU Munich, Munich, Germany

[2]Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany

[3]Munich Center of the Learning Sciences, Ludwig-Maximilians-Universität München, Munich, Germany

Corresponding Author:
Maximilian C Fink, MSc
Institute for Medical Education
University Hospital, LMU Munich
Pettenkoferstraße 8a
Munich, 80336
Germany
Phone: 49 089 4400 57428
Email: maximilian.fink@yahoo.com

Abstract

Background: Standardized patients (SPs) have been one of the popular assessment methods in clinical teaching for decades, although they are resource intensive. Nowadays, simulated virtual patients (VPs) are increasingly used because they are permanently available and fully scalable to a large audience. However, empirical studies comparing the differential effects of these assessment methods are lacking. Similarly, the relationships between key variables associated with diagnostic competences (ie, diagnostic accuracy and evidence generation) in these assessment methods still require further research.

Objective: The aim of this study is to compare perceived authenticity, cognitive load, and diagnostic competences in performance-based assessment using SPs and VPs. This study also aims to examine the relationships of perceived authenticity, cognitive load, and quality of evidence generation with diagnostic accuracy.

Methods: We conducted an experimental study with 86 medical students (mean 26.03 years, SD 4.71) focusing on history taking in dyspnea cases. Participants solved three cases with SPs and three cases with VPs in this repeated measures study. After each case, students provided a diagnosis and rated perceived authenticity and cognitive load. The provided diagnosis was scored in terms of diagnostic accuracy; the questions asked by the medical students were rated with respect to their quality of evidence generation. In addition to regular null hypothesis testing, this study used equivalence testing to investigate the absence of meaningful effects.

Results: Perceived authenticity (1-tailed t_{81}=11.12; P<.001) was higher for SPs than for VPs. The correlation between diagnostic accuracy and perceived authenticity was very small (r=0.05) and neither equivalent (P=.09) nor statistically significant (P=.32). Cognitive load was equivalent in both assessment methods (t_{82}=2.81; P=.003). Intrinsic cognitive load (1-tailed r=−0.30; P=.003) and extraneous load (1-tailed r=−0.29; P=.003) correlated negatively with the combined score for diagnostic accuracy. The quality of evidence generation was positively related to diagnostic accuracy for VPs (1-tailed r=0.38; P<.001); this finding did not hold for SPs (1-tailed r=0.05; P=.32). Comparing both assessment methods with each other, diagnostic accuracy was higher for SPs than for VPs (2-tailed t_{85}=2.49; P=.01).

Conclusions: The results on perceived authenticity demonstrate that learners experience SPs as more authentic than VPs. As higher amounts of intrinsic and extraneous cognitive loads are detrimental to performance, both types of cognitive load must be monitored and manipulated systematically in the assessment. Diagnostic accuracy was higher for SPs than for VPs, which could potentially negatively affect students' grades with VPs. We identify and discuss possible reasons for this performance difference between both assessment methods.

(J Med Internet Res 2021;23(3):e21196) doi: 10.2196/21196

KEYWORDS

clinical reasoning; medical education; performance-based assessment; simulation; standardized patient; virtual patient

Introduction

Performance-Based Assessment With Standardized Patients and Virtual Patients

Since the turn of the millennium, performance-based assessment has become a mandatory part of medical licensure examinations in various countries [1], complementing traditional assessment formats, such as text vignettes, with methods including standardized patients (SPs) and simulated virtual patients (VPs). SPs have been used for performance-based assessment in health care since the 1960s [2]. However, VPs have only recently become more widely employed in this domain [3].

The term SPs refers to (trained) actors or real former patients who act as if they display symptoms of a disease [4]. Usually, students encounter several SPs in assessment settings to reliably measure clinical variety [5]. Performance is then scored by a trained faculty member or the SPs themselves using a rating scheme. Although we will elaborate on the specific features used for this assessment method later, it should be noted here that organizing an assessment with SPs is relatively resource intensive [6].

VPs are a type of computer simulation and typically include an authentic model of a real-world situation that can be manipulated by the participant [7]. VPs can use avatars or realistic videos with SPs as stimuli and offer varying degrees of interaction [8]. Moreover, assessment through VPs can take place automatically, and a recent study showed that such an automatic assessment corresponds well to ratings from clinician-educators [9]. The production of authentic VPs can frequently produce considerable costs above $10,000 [10]. Although the initial production of VPs is often more resource intensive than organizing SPs, this assessment method is then permanently available and fully scalable to a large audience.

Next, we summarize a conceptual framework. This framework provides, on the one hand, a precise operationalization of diagnostic competences. On the other hand, the framework includes a research agenda that summarizes essential moderators of performance that should be examined systematically in research on simulation-based assessment.

A Framework for the Assessment of Diagnostic Competences With Simulations

The framework developed by Heitzmann et al [10] to facilitate diagnostic competences with simulations operationalizes diagnostic competences in assessment settings as a disposition. This disposition encompasses the components of diagnostic knowledge, diagnostic quality, and diagnostic activities. Diagnostic knowledge includes conceptual and strategic knowledge [11]. Conceptual knowledge encompasses concepts and their relationships. Strategic knowledge comprises possible avenues and heuristics in diagnosing. Diagnostic quality consists of components' diagnostic accuracy and efficiency that can serve as major outcome measures in empirical studies. Diagnostic activities entail the actions of persons assessed during the diagnostic process, such as evidence generation by asking questions in history taking. The framework proposes that context is an important moderator in assessment. Therefore, more research on the effects of the assessment methods SPs and VPs seems to be warranted. A meta-analysis on simulation-based learning of complex skills [12] added to this framework that authenticity should also be explored as an important moderator in assessment and learning. Similarly, a meta-analysis on instructional design features in simulation-based learning indicated that certain types of cognitive load could be detrimental to performance [13]. Therefore, it could be fruitful to explore the relationship between cognitive load and diagnostic competences within SP and VP assessments.

Perceived Authenticity and Diagnostic Competences With SPs and VPs

There is a multitude of conceptualizations of authenticity. In our study, we focus on *perceived authenticity* [14] because this concept can be assessed entirely internally by learners' judgment. Other related concepts such as *thick authenticity* [15] and *fidelity* [16] can, at least to some extent, also be determined externally.

According to a factor analysis by Schubert et al [14], perceived authenticity—sometimes also called presence—comprises the facets of realness, involvement, and spatial presence. Realness describes the degree to which a person believes that a situation and its characteristics resemble a real-life context [14]. Involvement is defined as a feeling of cognitive immersion and judgment that a situation has personal relevancy [17]. Spatial presence denotes the feeling of physical immersion in a situation [14]. SPs are considered highly authentic because they are carefully trained to realistically portray symptoms and allow for natural interactions [18]. Empirical studies support this claim, reporting high values of perceived authenticity for SPs [19,20]. VPs also received rather high perceived authenticity scores in empirical studies [21] but lacked some of the features that may make SPs particularly authentic, such as high interactivity in oral conversations. Thus, VPs could potentially evoke lower perceived authenticity than SPs. Findings on the effect of authenticity on diagnostic competences are mixed. On the one hand, it has been argued that higher authenticity is associated with higher engagement and better performance [22]. On the other hand, literature reviews [23,24] that compared the relationship between perceived authenticity and clinical performance in simulation-based learning only reported minimal effects of authenticity. In addition, an empirical study [25] showed that above a certain threshold, further increases in perceived authenticity do not improve diagnostic accuracy.

Cognitive Load and Diagnostic Competences With SPs and VPs

Cognitive load theory posits that performance can be inhibited through high situational demands that stress working memory and attention [26]. The cognitive load consists of the following 3 different facets [27]: *Intrinsic* load results from the interplay between certain topics and materials and the assessed person's expertise. *Extraneous load* is created exclusively by characteristics of the assessment environment that strain memory and attention without being necessary for performance. *Germane load* refers to the cognitive load created through the assessed person's cognitive processes, including schema construction and abstraction. Intrinsic and extraneous cognitive loads are considered additive and can inhibit performance in complex tasks [27]. Germane load, however, is theorized to bolster performance [27]. A few primary studies from medical education have already contrasted the cognitive load of different assessment methods and reported their relationship with diagnostic competences. Dankbaar et al [28] demonstrated that intrinsic and germane cognitive loads were higher for a group learning emergency skills with a simulation game than for a group learning with a text-based simulation. Extraneous load did not differ between these groups, and none of the groups differed in performance. Haji et al [29] compared surgical skills training with less complex and more complex simulation tasks. The total cognitive load was higher in the more complex simulation than in the less complex simulation, and cognitive load was negatively associated with performance. As a result of these findings, we can conclude that SPs and VPs generally do not differ in different facets of cognitive load if the assessment methods are of equal complexity, and the main characteristics related to the facets are similar. The literature summarized earlier also shows that intrinsic and extraneous cognitive loads are negatively associated with diagnostic competences.

Assessment Method and Diagnostic Competences

Before we discuss diagnostic accuracy and evidence generation—2 important aspects of diagnostic competences—it should be noted that diagnostic competences are only a part of the broader concept of clinical reasoning. Clinical reasoning emphasizes the process of diagnosing and encompasses the full process of making clinical decisions, including the selection, planning, and reevaluation of a selected intervention [30]. In line with the conceptual framework by Heitzmann et al [10] for facilitating diagnostic competences, *diagnostic accuracy* denotes the correspondence between the learner's diagnoses and the solutions determined by experts for the same cases. According to this framework, *evidence generation* (ie, actions related to the gathering of data in a goal-oriented way) is also an important quality criterion for the diagnostic process and a crucial aspect of diagnostic competences.

Diagnostic Accuracy

Currently, there are only a few studies in the health care domain that contrast assessments using VPs and SPs directly in one experiment. Edelstein et al [1] investigated assessments with SPs and computer-based case simulations in advanced medical students using a repeated measures design. A moderate positive correlation was found between diagnostic accuracy in the two assessment formats that used different cases. Guagnano et al [31] examined SPs and computer-based case simulations in a medical licensing exam. Participants first completed the computer-based case simulations and then completed the SPs. The two assessment methods correlated positively with each other. Hawkins et al [32] compared the assessment of patient management skills and clinical skills with SPs and computer-based case simulations in a randomized controlled trial. Participating physicians completed both assessment methods, and a positive correlation of diagnostic accuracy with both assessment methods was reported. Outside the health care domain, a meta-analysis of studies from different domains reported a robust modality effect for students in problem-solving tasks. Students who solved problems presented in the form of illustrations accompanied by text were more successful than students who solved problems presented merely in text form [33]. Similarly, it seems reasonable to assume that one assessment method could lead to higher diagnostic accuracy than the other assessment method because of its different characteristics. The described findings from the health care domain tentatively indicate that SPs and VPs could result in relatively equivalent diagnostic accuracy. Such a finding would contradict the modality effect reported in other domains.

Evidence Generation

Comparable empirical studies on evidence generation for SPs and VPs are lacking. Nevertheless, we can assume that the quantity of evidence generation should be higher for SPs than for VPs. The main reason for this is that students can ask questions of SPs more quickly orally than by selecting questions from a menu of options with VPs. Apart from this difference in evidence generation between the 2 assessment methods, the relationships between evidence generation and diagnostic accuracy are interesting. The relationship between the quantity of evidence generation and diagnostic accuracy is relatively complex. The ideal amount of evidence generation may depend strongly on the case difficulty, the diagnostic cues contained in the evidence, and learner characteristics. For these reasons, the framework by Heitzmann et al [10] for facilitating diagnostic competences argues that the sheer quantity of evidence generation is not a dependable quality criterion for the diagnostic process. However, the quality of evidence generation is hypothesized by Heitzmann et al [10] to be a rather dependable quality criterion for the diagnostic process. This agrees with the literature, as we know from studies on SPs using observational checklists that the quality of evidence generation is positively associated with diagnostic accuracy [34]. Moreover, one study with specialists in internal medicine and real patients demonstrated that asking specific questions in history taking correlated positively with clinical problem solving [35].

Study Aim, Research Questions, and Hypotheses

We aim to compare the perceived authenticity, cognitive load, and diagnostic competences in SPs and VPs. We also aim to examine the relationships of perceived authenticity, cognitive load, and quality of evidence generation with diagnostic accuracy. Thus, we address the following 3 research questions: To what extent does perceived authenticity differ across the 2

assessment methods, and how is it associated with diagnostic accuracy (RQ1)? We hypothesize that SPs induce higher perceived authenticity than VPs (H1.1). Moreover, we expect to be able to demonstrate with equivalence tests for correlations (given in the *Statistical Analyses* section) that perceived authenticity is not associated meaningfully with diagnostic accuracy (H1.2). Next, is cognitive load equivalent for SPs and VPs, and how is it related to diagnostic accuracy (RQ2)? We assume to find equivalent cognitive load for SPs and VPs (H2.1). Moreover, we expect that intrinsic and extraneous loads are negatively related to diagnostic accuracy (H2.2-H2.3). To what extent are the diagnostic competences components diagnostic accuracy, quantity of evidence generation, and quality of evidence generation equivalent or differ for SPs and VPs, and how are they related to each other (RQ3)? We hypothesize that SPs and VPs evoke equivalent diagnostic accuracy (H3.1). In addition, we assume that the quantity of evidence generation is higher for SPs than for VPs (H3.2). We also expect that the quality of evidence generation is positively related to diagnostic accuracy (H3.3).

Methods

Participant Characteristics and Sampling Procedures

A sample of 86 German medical students (with a mean age of 26.03 years, SD 4.71) made up the final data set. This sample consisted of 63% (54/86) females and 37% (32/86) males. Medical students in years 3-6 of a 6-year program with a good command of German were eligible. Medical students in years 3-5 (44/86, 51%) were considered novices, as they were still completing the clinical part of the medical school. Medical students in year 6 (42/86, 49%) were regarded as intermediates as they had passed their second national examination and worked full time as interns in a medical clinic or practice. We provide a detailed overview of participant characteristics across all conditions and a CONSORT (Consolidated Standards of Reporting Trials)–style diagram of participant flow in Multimedia Appendix 1.

We collected data from October 20, 2018, to February 20, 2019, in the medical simulation center of the University Hospital, LMU Munich. We recruited participants via on-campus and web-based advertising. Participants were randomly assigned to conditions by the first author by drawing a pin code to log in to an electronic learning environment without knowing the condition assigned to the pin. In the final data collection sessions, the conditions were filled by the first author with random participants from specific expertise groups (novices vs intermediates). This procedure was applied to achieve a comparable level of expertise in all conditions. As expected, the proportion of participants from different expertise groups did not differ across conditions (χ^2_3=0.2; P=.99).

Research Design

The study used a repeated measures design with assessment method (SPs vs VPs) as the key factor. In addition, we varied the between-subjects factor case group (CG) order and assessment method order. In total, students encountered 6 different cases. We provide an overview of the experiment in Table 1. Details of the succession through cases and medical content in the experimental conditions are provided in Table 2. We attempted to ensure similar topics and difficulty for both CGs by conducting an expert workshop and adapting cases based on the experts' feedback as part of creating the experimental materials.

Table 1. General overview of the experiment.

Part of the experiment	Activity or test	Duration (min)
Pretest	Briefing	10
	Conceptual knowledge test	40
	Strategic knowledge test	40
Break	—[a]	10
Assessment phase I (cases 1-3)	VPs[b] or SPs[c]	70
Break and change of modality	—	5
Assessment phase II (cases 4-6)	VPs or SPs	70
Posttest and debriefing	Working memory test	15
	End-debriefing	5

[a]No activity or test takes place.

[b]VP: virtual patient.

[c]SP: standardized patient.

Table 2. Succession through cases and medical content in the experimental conditions[a,b].

Cases	Condition 1A	Condition 1B	Condition 2A	Condition 2B
1-3	CG[c] A (SPs[d])	CG B (VPs[e])	CG B (SPs)	CG A (VPs)
4-6	CG B (VPs)	CG A (SPs)	CG A (VPs)	CG B (SPs)

[a]Case group A: (1) pulmonary embolism with lymphoma, (2) congestive heart failure with atrial fibrillation, and (3) hyperventilation tetany caused by a panic attack.

[b]Case group B: (1) pulmonary embolism with coagulation disorder, (2) community-acquired pneumonia, and (3) hypertrophic obstructive cardiomyopathy.

[c]CG: case group.

[d]SP: standardized patient.

[e]VP: virtual patient.

Procedure and Materials

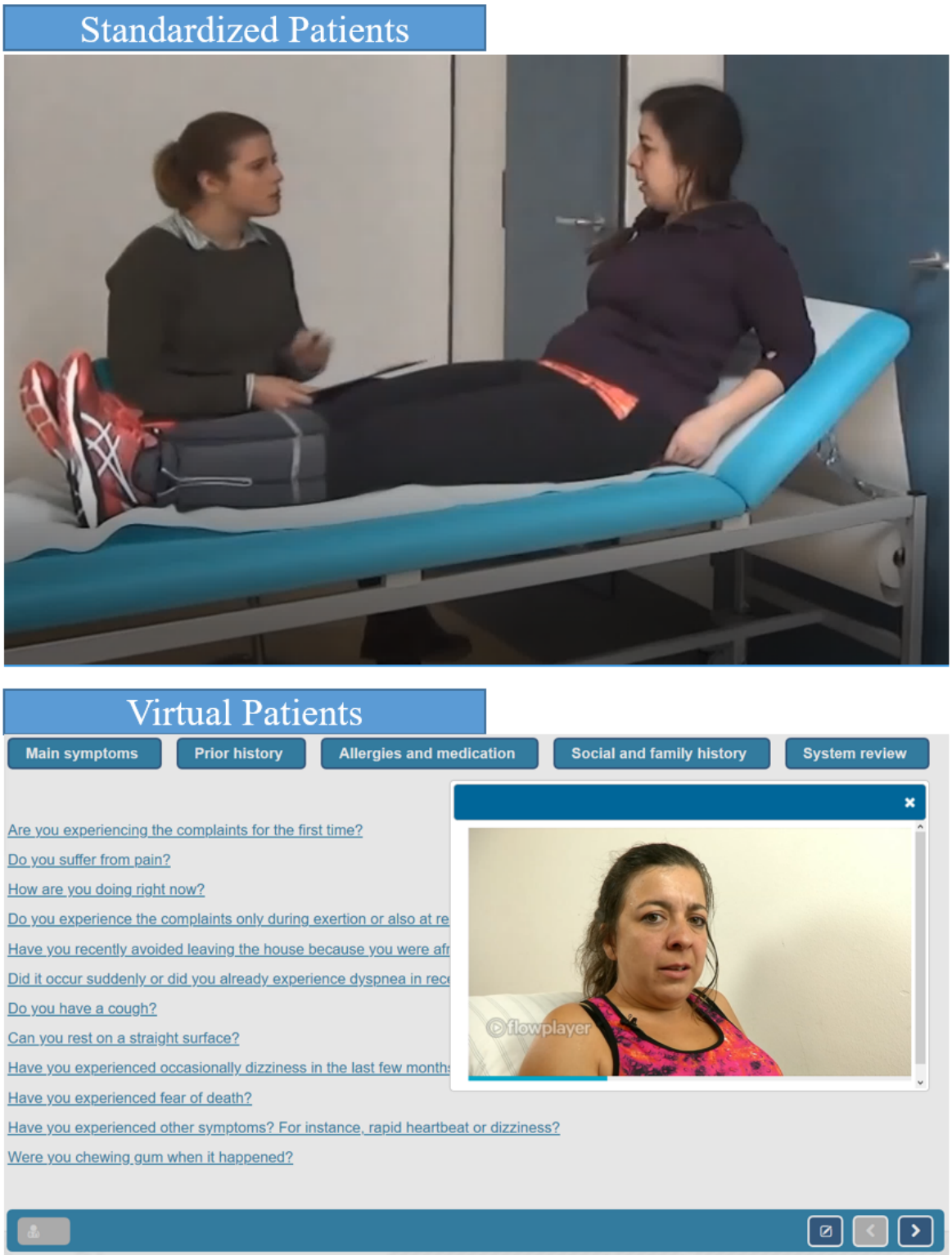
Participants completed a pretest of conceptual knowledge and strategic knowledge at the beginning of the experiment. Afterward, participants took part in the assessment phase, solving the first 3 cases with SPs and the next 3 cases with VPs or vice versa. All cases were drafted by a specialist in general practice and evaluated positively by an expert panel. The cases were not adapted from real clinical cases but based on cases from textbooks and symptoms reported in guidelines. A short familiarization phase preceded each assessment phase and included a motivational scale. For all cases in both assessment methods, assessment time was held constant at 8 minutes and 30 seconds for history taking and 5 minutes for writing up a diagnosis for the case in an electronic patient file. At the end of the experiment, participants were debriefed. A more detailed overview of the procedure can be found in Multimedia Appendix 2.

Assessment with SPs was conducted in a simulated emergency room. All SPs were (semi-) professional actors who were financially compensated; most had previous experience working in an SP program. All SPs were extensively trained by an acting coach and a physician, memorized their symptoms and scripts, and were not aware of their patient's diagnosis. Participants first received prior information (eg, electrocardiogram and lab results) and presentation of the chief complaint for each case. Next, participants formulated and asked questions independently, and the SPs responded. The interaction was recorded on a video. After each case, the participants completed a patient file, including measures of diagnostic accuracy and other scales. A screenshot of this assessment method is provided in Figure 1.

The assessment with the VPs was carried out in a simulated assessment environment in a computer room. First, participants received prior information and a video with a chief complaint for each case. The participants then selected questions independently from a menu with up to 69 history-taking questions. The VP's answer was streamed as a video, including a recorded response by an actor. After each case, the participants completed a patient file, including a measure of diagnostic accuracy and other scales. A screenshot of this assessment method is provided in Figure 1.

The VPs, patient file, and other measures were implemented in the electronic assessment environment CASUS [36]. The questions provided for the VPs were based on a structural and topical analysis of history-taking forms by Bornemann [37] and are displayed in Multimedia Appendix 3. According to this analysis, physician questions in history taking can fall under the 5 categories of main symptoms, prior history, allergies and medication, social and family history, and system review. Participants with SPs received empty history-taking forms for all cases and time to formulate possible history-taking questions during the familiarization phase, at which point participants in the VPs only read all questions from the menu. Without this additional structuring support in the SP condition, the participants in the VP condition would have received additional support in the form of a list of questions in the menu.

Figure 1. History-taking with standardized patients and virtual patients.



Measures and Covariates

Perceived Authenticity

Perceived authenticity was operationalized as a construct with the 3 dimensions of realness, involvement, and spatial presence [14]. All 3 authenticity scales used a 5-point scale ranging from (1) *disagree* to (5) *agree* and were taken from multiple validated questionnaires [14,38-40]. The items were slightly adapted to simulation-based assessment and are included in Multimedia Appendix 4. A combined score for all 3 dimensions was built by calculating the mean. This scale achieved a reliability of Cronbach α=.88.

Cognitive Load

The cognitive load scale by Opfermann [41] used in this study assessed the extraneous cognitive load with 3 items and germane and intrinsic cognitive loads with 1 item each. A 5-point scale from (1) *very easy*, (2) *rather easy*, (3) *neutral*, (4) *rather hard*,

to (5) *very hard* was used. The scale is included in Multimedia Appendix 4. A combined score for all 3 facets was built by calculating the mean. This scale achieved a reliability of Cronbach α=.88.

Motivation, Diagnostic Knowledge, and Other Control Variables

We assessed motivation as a control variable because it could differ between assessment methods and potentially affect performance. The expectancy component of motivation was assessed with a 4-item, 7-point scale adapted from Rheinberg et al [42]. The motivation expectancy scale ranged from (1) strongly disagree to (7) strongly agree. The value component of motivation was measured with a 4-item, 5-point scale based on a questionnaire by Wigfield [43]. The motivation value scale ranged from (1) strongly disagree to (5) strongly agree. The full scales are provided in Multimedia Appendix 4. Diagnostic knowledge was also measured in this study but later not taken into account in the analyses because it was similar in VPs and SPs because of the repeated measures design. We measured diagnostic knowledge using a conceptual and strategic knowledge test. Both types of knowledge have been identified as predictors of clinical reasoning [44]. The maximum testing time was set to 40 minutes per test. More details on both diagnostic knowledge tests are reported in Multimedia Appendix 4. Apart from this, demographic data were collected, including participants' sex, age, and expertise (year of medical school).

Diagnostic Competences

Diagnostic Accuracy

Diagnostic accuracy was assessed based on the answer to the prompt "Please choose your final diagnosis after history taking" from a long menu containing 239 alternative diagnoses. Two physicians created a coding scheme for scoring diagnostic accuracy in all cases (Multimedia Appendix 4). To do that, the physicians rated all 239 alternative diagnoses for all cases and resolved the disagreements until they reached full agreement. One of the physicians was a specialist in general practice who also drafted the cases. The other physician was a board-certified doctor familiar with medical assessment through her dissertation. The latter physician, who is also the second author of this paper, then scored diagnostic accuracy based on the coding scheme: 1 point was allocated for the designated correct answer, 0.5 point for a partially correct answer, and 0 point for an incorrect answer. Due to having only 1 rater to score the diagnostic accuracy with the comprehensive coding scheme, a reliability estimate cannot be reported. However, this is also not necessary because the exact diagnostic accuracy score for all selectable diagnoses included in the electronic assessment environment was determined upfront in the coding scheme.

Evidence Generation

The second author classified the quality of evidence generation by determining the essential questions relevant for the correct diagnosis for each VP case (the coding scheme is given in Multimedia Appendix 4). This process took part before looking at the experimental data. All solutions were discussed with a specialist in general practice, and all disagreements were resolved. Student assistants transcribed all utterances recorded in the videos of the SP encounters, and the electronic assessment environment stored all selected questions during the VP encounters. The *R* scripts automatically classified the log data from the VPs using the coding scheme. Student assistants had no medical background and were trained by the second author to code the transcripts from the SP encounters. This task mainly implied recognizing the intent of history-taking questions and linking them, if possible, to the most similar question in the coding scheme. After training the raters, 20% of this complex and extensive SP data were coded by 2 raters to check interrater agreement. This data set encompassed SP data from 18 of the 86 participants of our study with all three SP cases in which the participants took part. Fleiss κ=0.74 demonstrated that agreement was substantial, and the rest of the data were coded by the same raters individually. The score for quantity of evidence generation corresponded to the total number of questions posed for each case. To calculate the score for quality of evidence generation for each case, we counted the number of relevant questions posed and divided this score by the number of relevant questions that could potentially be posed.

Scale Construction

Diagnostic accuracy and evidence generation scales for each assessment method and combining the 2 methods were built by calculating the mean of the included cases. Case 1 in CS A was excluded from all analyses because of high difficulty (mean diagnostic accuracy 0.05, SD 0.18).

Statistical Analyses

This study answers the proposed research questions using traditional null hypothesis significance testing (NHST) and equivalence testing. In contrast to NHST, equivalence testing can be used to investigate "whether an observed effect is surprisingly small, assuming that a meaningful effect exists in the population" [45]. For this type of test, first, the smallest effect size of interest, that is, the threshold for a meaningful effect, is specified based on the literature. The null hypothesis that the effect is more extreme than the smallest effect size of interest is then investigated. To do this, 2 separate 1-sided tests (TOST; eg, *t* tests) are conducted [46]. These tests examine whether the observed effect is more extreme than the specified smallest effect size of interest. If both 1-sided tests are significant, the null hypothesis that there is a meaningful effect that is more extreme than the smallest effect size of interest is rejected. Thus, equivalence is supported. For more convenient reporting, only the *t* test with a higher P value is reported. In cases in which equivalence cannot be supported, NHST is performed for follow-up analyses.

All statistical analyses were performed using *R* version 3.6.1 [47]. The TOST procedure and the corresponding package TOSTER [45] were used to conduct the equivalence tests. In all statistical analyses, the alpha level was set to 5%; 1-tailed tests were used where applicable. The Bonferroni-Holm method [48] was used to correct P values for multiple comparisons in post hoc and explorative tests.

For all equivalence tests, the smallest effect size of interest was determined based on the discussed literature. For H1.2 and related post hoc tests, the smallest effect size of interest was set

to be more extreme than $r=\pm0.20$, which corresponds to the effect size of small but meaningful correlations typically encountered in the social sciences [49]. For H2.1 and related post hoc tests, a meaningful effect was determined as an effect of Cohen $d=0.35$. This effect size lies between a small effect (Cohen $d=0.20$) and a medium effect (Cohen $d=0.50$) [49] and occurs frequently in the social sciences. For H3.1, we determined that a meaningful effect exists in the case of a difference of ±0.125 points in diagnostic accuracy. This was based on supposing a pass cutoff of 0.50 for diagnostic accuracy (ranging from 0 to 1) and setting 4 equal intervals for the hypothetical passing grades A-D.

Power Analysis

We conducted a priori power analysis for dependent samples *t* tests (H1.1 and H3.2). This power analysis was based on a small to medium effect of Cohen $d=0.30$, 2-tailed testing, an error probability of 5%, and 80% power, resulting in a targeted sample of 90 participants. Moreover, we carried out a priori power analyses for 1-tailed correlations with $r=\pm0.25$, an error probability of 5%, and 80% power (H2.2-H2.3 and H3.3). This power analysis resulted in a planned sample size of 95 participants. A post hoc power analysis for the main equivalence test (H3.1) with 86 participants, the observed effect of Cohen $d=0.26$, and an error probability of 5% resulted in a power of 78%. All power analyses were conducted using G*Power software [50].

Results

Descriptive Statistics and Analysis of Control Variables

Descriptive statistics are provided in Table 3. The perceived authenticity variables were rated as very high for SPs and relatively high for VPs. Cognitive load variables were reported to be moderate in both assessment methods. The average diagnostic accuracy was medium. The quantity of evidence generation was higher for SPs than for VPs. The quality of evidence generation was medium for both assessment methods. Motivational variables were rated rather highly for both SPs and VPs. A post hoc comparison showed that the value aspect of motivation was higher for SPs than for VPs (2-tailed $t_{83}=2.89$; $P=.01$; Cohen $d=0.31$), whereas the expectancy aspect did not differ between assessment methods (2-tailed $t_{83}=0.44$; $P=.66$; Cohen $d=0.05$). Participants demonstrated slightly above medium performance on the conceptual and strategic knowledge tests. Multimedia Appendix 5 provides an additional visualization of the results using boxplots and bee swarm plots.

Table 3. Descriptive statistics.

Variable	Both methods, mean (SD)	SPs[a], mean (SD)	VPs[b], mean (SD)
Perceived authenticity[c]	3.62 (0.67)	4.02 (0.67)	3.23 (0.84)
Realness[c]	3.71 (0.79)	4.13 (0.74)	3.28 (1.07)
Involvement[c]	3.82 (0.66)	4.03 (0.73)	3.61 (0.83)
Spatial presence[c]	3.35 (0.80)	3.89 (0.83)	2.80 (1.05)
Cognitive load[c]	2.88 (0.61)	2.88 (0.74)	2.90 (0.69)
Intrinsic load[c]	3.18 (0.68)	3.20 (0.78)	3.14 (0.80)
Extraneous load[c]	2.84 (0.65)	2.82 (0.79)	2.87 (0.76)
Germane load[c]	2.74 (0.76)	2.73 (0.88)	2.76 (0.84)
Diagnostic competences			
Diagnostic accuracy[d]	0.46 (0.18)	0.51 (0.28)	0.41 (0.24)
Quantity of evidence generation	22.26 (4.88)	29.01 (8.03)	17.34 (4.21)
Quality of evidence generation[d]	0.40 (0.11)	0.37 (0.18)	0.43 (0.13)
Control variables			
Motivation expectancy aspect[e]	5.07 (0.91)	5.10 (0.88)	5.05 (1.08)
Motivation value aspect[c]	4.44 (0.51)	4.54 (0.54)	4.34 (0.67)
Conceptual knowledge[d]	0.65 (0.14)	—[f]	—
Strategic knowledge[d]	0.66 (0.15)	—	—

[a]SP: standardized patient.

[b]VP: virtual patient.

[c]Scale range: 1-5.

[d]Scale range: 0-1.

[e]Scale range: 1-7.

[f]Knowledge was assessed before taking part in SPs and VPs.

Perceived Authenticity and Diagnostic Accuracy (RQ1)

A paired sample *t* test demonstrated that in line with hypothesis H1.1, perceived authenticity was considered higher for SPs than VPs in terms of the combined score (1-tailed t_{81}=11.12; *P*<.001; Cohen *d*=1.23). Post hoc tests showed that this was also the case for realness (t_{80}=8.83; *P*<.001; Cohen *d*=0.98), involvement (t_{81}=4.60; *P*<.001; Cohen *d*=0.51), and spatial presence (t_{79}=10.65; *P*<.001; Cohen *d*=1.19). Our expectation in H1.2 was that perceived authenticity would not be meaningfully associated with diagnostic accuracy. The TOST procedure for correlations showed that the relationship between diagnostic accuracy and the combined perceived authenticity score (*r*=0.05; *P*=.09) was outside the equivalence bounds of a meaningful effect of *r*=±0.20. Post hoc equivalence tests demonstrated that this also holds for the relationship of diagnostic accuracy with realness (*r*=0.03; *P*=.06), involvement (*r*=0.07; *P*=.11), and spatial presence (*r*=0.05; *P*=.08). Reanalyzing these correlations with regular 1-tailed NHST tests also yielded nonsignificant results for the combined score (*P*=.32), realness (*P*=.39), involvement (*P*=.28), and spatial presence (*P*=.33). These results mean that there is neither evidence for the absence of meaningful correlations nor evidence for significant correlations. These inconclusive findings may stem from the lack of statistical power because of the relatively small sample size [45].

Cognitive Load and Diagnostic Accuracy (RQ2)

We hypothesized in H2.1 that we would find equivalent cognitive load scores for SPs and VPs. Equivalence testing with the TOST procedure for paired samples indicated that for both assessment methods, the scores for combined cognitive load (t_{82}=2.81; *P*=.003) were significantly within the equivalence bounds of an effect of Cohen *d*=0.35. Adjusted post hoc equivalence tests showed that this is also the case for intrinsic load (t_{82}=−2.47; *P*=.008), extraneous load (t_{82}=2.55; *P*=.01), and germane load (t_{82}=2.64; *P*=.01). We expected in H2.2-H2.3 to uncover negative correlations between diagnostic accuracy and intrinsic cognitive load and extraneous load. As assumed, intrinsic cognitive load (1-tailed *r*=−0.30; *P*=.003) and extraneous load (1-tailed *r*=−0.29; *P*=.003) correlated negatively with the combined score for diagnostic accuracy. Adjusted explorative follow-up analyses showed that germane load (*r*=−0.25; *P*=.010) and the total score for cognitive load

($r=-0.31$; $P=.004$) also correlated negatively with the combined score for diagnostic accuracy.

Assessment Method and Diagnostic Competences (RQ3)

Diagnostic Accuracy

In H3.1, we hypothesized finding equivalent diagnostic accuracy scores for SPs and VPs. H3.1 was first examined by applying a paired samples TOST procedure. According to our data, we cannot reject hypothesis H3.1 that a difference in diagnostic accuracy of at least ±0.125 points (1 grade) exists between the 2 assessment methods ($t_{85}=-0.60$; $P=.28$). A follow-up 3-way mixed design analysis of variance demonstrated that neither the CG order nor the assessment method order ($F_{3,82}=2.49$; $P=.12$; $\eta^2=0.03$, respectively, $F_{3,82}=0.02$; $P=.88$; $\eta^2=0.01$) had a significant effect on diagnostic accuracy. The assessment method itself, however, had a significant main effect ($F_{3,82}=6.30$; $P=.01$; $\eta^2=0.07$), indicating that diagnostic accuracy was higher for SPs than for VPs. The finding that diagnostic accuracy was higher for SPs than for VPs also corresponds to the result of a paired sample *t* test (2-tailed $t_{85}=2.49$; $P=.01$; Cohen $d=0.27$).

Evidence Generation

H3.2 that students display an increased quantity of evidence generation with SPs than with VPs was supported (1-tailed $t_{69}=12.26$; $P<.001$; Cohen $d=1.47$). However, in an explorative follow-up analysis, we found no evidence that the *quantity* of evidence generation was related to diagnostic accuracy (1-tailed $r=0.11$; $P=.15$). This finding holds equally for SPs ($r=-0.09$; $P=.76$) and VPs ($r=-0.10$; $P=.82$). Moreover, H3.3 that the *quality* of evidence generation is positively related to diagnostic accuracy in both assessment methods was not supported (1-tailed $r=0.18$; $P=.05$). Corrected post hoc analyses showed, however, that the quality of evidence generation was positively related to diagnostic accuracy for VPs ($r=0.38$; $P<.001$); this finding did not hold for SPs ($r=0.05$; $P=.32$). Additional post hoc exploratory analyses revealed that the quality of evidence generation was higher for VPs than for SPs (2-tailed $t_{74}=-2.47$; $P=.02$; Cohen $d=0.29$).

Discussion

Principal Findings

With regard to perceived authenticity, our results showed that SPs and VPs achieved high scores on all 3 dimensions of realness, involvement, and spatial presence. Despite this high level of perceived authenticity in both assessment methods, perceived authenticity was higher for SPs than for VPs on all 3 dimensions. This finding is in line with the literature, which has long claimed that SPs achieve a very high level of perceived authenticity [18-20]. Other studies on perceived authenticity have so far focused on comparing formats such as SPs, video presentations, and text vignettes and different levels of authenticity within VPs [21]. Our study extends this literature by directly comparing SPs and VPs with respect to 3 frequently used perceived authenticity variables. This comparison seems particularly relevant, as both assessment formats are becoming increasingly popular. Our findings on the relationship between perceived authenticity and diagnostic accuracy are mixed. The equivalence test on correlations was not significant; therefore, we could not confirm the hypothesis that perceived authenticity is not meaningfully associated with diagnostic accuracy. However, a regular correlation between perceived authenticity and diagnostic accuracy that was calculated afterward was close to 0. Taken together, these findings of nonequivalence and nonsignificance indicate that we did not have sufficient power to draw a conclusion [45]. Nevertheless, we have found some indication that the correlation between perceived authenticity and diagnostic competences is rather small. This finding is in accordance with literature reviews [23,24], which reported small correlations between perceived authenticity and performance.

With regard to cognitive load, we found that the combined score is equivalent for SPs and VPs that use the same clinical cases. This finding substantiates the literature suggesting that cognitive load depends mainly on task complexity [29]. Moreover, the fact that the extraneous load was equivalent for SPs and VPs indicates that user interaction through a software menu does not substantially increase cognitive load. This finding is important because decreasing the cognitive load by allowing for user input using natural language processing [21] is still highly expensive. Our study also adds to the literature that the level of cognitive load is similar in SPs and VPs as assessment methods if the different types of cognitive load are systematically controlled for during the design process. In addition, we demonstrated that intrinsic and extraneous cognitive loads correlate negatively with diagnostic accuracy. The finding on intrinsic cognitive load corroborates that the interplay between materials and the assessed person's expertise is associated with performance. The finding on extraneous cognitive load shows that unnecessary characteristics of the assessment environment can strain memory and attention and be detrimental to performance in assessment settings. Together, these findings fit well with the literature, which has repeatedly reported negative effects of intrinsic and extraneous cognitive loads on complex problem solving in medical education [27] and other domains [51]. Our study unveils that a negative relationship between intrinsic and extraneous cognitive loads and performance in a simulation-based measure of diagnostic competences already shows when overall cognitive load is medium on average.

Our study found no evidence that diagnostic accuracy was equivalent for SPs and VPs. In contrast, higher diagnostic accuracy was achieved for SPs than for VPs. The small number of studies comparing both assessment methods so far [1,31,32] have reported medium correlations, not taking into account different case content or testing time. Using the TOST procedure as a novel methodological approach, our study contributes to the literature by finding that grading was not equivalent, as participants received a better hypothetical grade when the simulation-based assessment was administered with SPs than with VPs. On the one hand, we cannot rule out that this finding may be explained by additional support from the actors in the SP assessment. To avoid and mitigate such an effect, actors were trained by an acting coach and a physician, memorized their symptoms and scripts, and did not know the diagnosis of

their case. Moreover, student assistants screened all SP assessments, and no additional systematic support by actors was discovered. On the other hand, this finding can be explained by the lower appraisal of motivational value and the lower quantity of evidence generation reported for VPs. Participants solving VP cases may thus have been less engaged and may have collected a smaller number of important diagnostic cues that supported their diagnostic process.

Contrary to our expectations, the quality of evidence generation was not positively correlated with the *combined* diagnostic accuracy score. Closer inspection of the data revealed that the quality of evidence generation was positively correlated with diagnostic accuracy in *VPs*. This confirmed relationship is in line with the theoretical assumptions of Heitzmann et al [10]. In *SPs*, however, the quality of evidence was not correlated with diagnostic accuracy. This finding contradicts the theoretical assumptions of Heitzmann et al [10] and empirical results from studies using observational checklists with SPs [34] and real patients [36]. There are 2 explanations for these conflicting findings. First, the quality of evidence generation was, as an exploratory follow-up *t* test indicated, higher in VPs than in SPs. This higher quality of evidence generation could have been caused by a slightly different process of history taking in both assessment methods. Participants working with VPs selected questions from a menu. In contrast, participants working with SPs formulated questions during history taking freely. Second, SPs could have offered additional support to assessed persons who displayed a low quality of evidence generation, whereas VPs reacted in a completely standardized way to all assessed persons.

Limitations

One methodological limitation of our study might be the low statistical power for the analysis of hypothesis H1.2 and related post hoc analyses that addressed the relationship between the perceived authenticity variables and diagnostic accuracy. This lack of statistical power can primarily be attributed to our investigation of whether a correlation of $r=\pm0.20$ or more extreme exists. As recommended by Lakens [46], the smallest effect size of interest was selected based on findings from the literature. Specifying the smallest effect size of interest to be larger would have increased power but not have contributed findings from a valuable equivalence test to the literature. This is the case because the literature already assumes a small effect size [23,24].

One theoretical limitation of the study is that the results on perceived authenticity may not generalize without restrictions to other related concepts of authenticity. Shaffer et al [15] argue that thick authenticity consists of four different aspects. An authentic task, situation, or material should (1) exist in real life, (2) be meaningful, (3) allow the learner to engage in professional activities of the discipline, and (4) be conducted rather similar in instruction and assessment. The authors assume that thick authenticity can only be achieved when all aspects of authenticity are adequate and that VPs could potentially achieve similar authenticity to SPs. Hamstra et al [16] proposed distinguishing fidelity using the terms physical resemblance and functional task alignment. The authors report weak evidence for the relationship between physical resemblance and performance, and strong evidence for the relationship between functional task alignment and performance. In our study, the concepts of thick authenticity and fidelity were not measured for two reasons. First, these concepts can, to some extent, only be judged externally by experts. Second, the repeated measures design of the study forced us to keep aspects such as thick authenticity, physical resemblance, and functional task alignment as similar as possible in SPs and VPs. Nevertheless, we believe that the relationship between different authenticity concepts and diagnostic competences still requires further research. Future studies should attempt to untangle the relationship between different authenticity concepts and diagnostic competences by measuring these systematically.

Conclusions

Our findings on the relationship between perceived authenticity and diagnostic accuracy contribute to the debate on the costs and benefits of perceived authenticity in performance-based assessments. These results relativize the importance of perceived authenticity in assessment. Increasing the perceived authenticity of assessment methods above a certain necessary threshold and thus raising their costs [23] does not seem to be of much benefit. Such spending could potentially squander a large share of the medical education budget [52] that could be put to more valuable use. Our results on cognitive load highlight its importance as a process variable in assessment settings. Performance-based assessment should thus attempt to reduce extraneous load and control for intrinsic load to measure performance in a standardized way that is still close to clinical practice [53].

Finally, the findings on diagnostic competences have some practical implications if VPs are used as an alternative to SPs in assessment. In particular, we found that VPs could lead to lower diagnostic accuracy scores than SPs, which could, in turn, negatively affect students' grades. There are 2 different mechanisms that could explain this finding: assessment with SPs could overestimate true performance or assessment with VPs could underestimate true performance. In accordance with SPs overestimating performance, we could not rule out additional support from the actors. In fact, the low, nonsignificant correlation between the quality of evidence generation and diagnostic accuracy in SPs, together with the higher diagnostic accuracy in SPs, could indicate that actors provided some additional support (eg, to participants who displayed low quality of evidence generation). Careful training [54] and screening thus seem to be of great importance to avoid additional support from actors during SP assessment to match the high level of standardization that VPs provide. The mechanism of possible underestimation of performance with VPs could be substantiated by the lower motivational value and quantity of evidence generation discovered for VPs. We suggest taking the following measures: students could be motivated additionally in VP assessment by more interactive environments (eg, using natural language processing) or providing automated elaborated feedback directly after the assessment. Moreover, the assessment time can be extended when menu-based VPs are used in practice. This way, the quantity of evidence generation could be raised to a level similar to that in the SP assessment.

Acknowledgments

The authors would like to thank Hannah Gerstenkorn, who developed the case vignettes. In addition, the authors would like to thank Ana Maria Semm, Renke Biallas, Jessica Feichtmayr, and Johannes Kissel, who assisted in conducting the study and analyzing the data, and Keri Hartman for proofreading. Finally, the first author (M Fink) would like to thank Larissa Kaltefleiter for her advice. This work was funded by the German Research Association (Deutsche Forschungsgemeinschaft; project number FOR2385).

Authors' Contributions

M Fink wrote the first draft of the manuscript, took part in conducting the study, and conducted data analysis and visualization. VR took part in conducting the study and provided feedback and editing. M Stadler conducted data analysis and visualization and provided feedback and assisted with editing. M Siebeck conceptualized and designed the study, provided feedback and editing, and acquired funding. FF conceptualized and designed the study, provided feedback and editing, and acquired funding. M Fischer conceptualized and designed the study, provided feedback and editing, and acquired funding. All authors approved the final manuscript for submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Participant characteristics across all conditions and CONSORT (Consolidated Standards of Reporting Trials)–style diagram of participant flow.
[DOCX File , 55 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Overview of the experimental procedure and simulation phases.
[DOCX File , 22 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Table containing the questions provided with all virtual patients. These questions were allocated to the five history-taking categories of main symptoms, prior history, allergies and medication, social and family history, and system review.
[DOCX File , 27 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Authenticity scales, cognitive load scales, coding scheme for diagnostic accuracy, coding scheme for the quality of evidence generation, motivation scales, and details of the diagnostic knowledge tests.
[DOCX File , 33 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Boxplots and bee swarm plots for authenticity, cognitive load, and clinical reasoning variables for standardized patients and virtual patients.
[DOCX File , 73 KB-Multimedia Appendix 5]

References

1. Edelstein RA, Reid HM, Usatine R, Wilkes MS. A comparative study of measures to evaluate medical students' performance. Acad Med 2000 Aug;75(8):825-833. [doi: 10.1097/00001888-200008000-00016] [Medline: 10965862]
2. Barrows HS, Abrahamson S. The programmed patient: a technique for appraising student performance in clinical neurology. J Med Educ 1964 Aug;39:802-805. [Medline: 14180699]
3. Botezatu M, Hult H, Tessma MK, Fors UGH. Virtual patient simulation for learning and assessment: superior results in comparison with regular course exams. Med Teach 2010;32(10):845-850. [doi: 10.3109/01421591003695287] [Medline: 20854161]
4. Vu NV, Barrows HS. Use of standardized patients in clinical assessments: recent developments and measurement findings. Educational Researcher 2016 Jul;23(3):23-30. [doi: 10.3102/0013189x023003023]
5. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J 1975 Mar 22;1(5955):447-451 [FREE Full text] [doi: 10.1136/bmj.1.5955.447] [Medline: 1115966]

6. Ziv A. Simulators and simulation-based medical education. In: A practical guide for medical teachers. Vol. 2. 3rd ed. Amsterdam: Elsevier; 2009.
7. de JT. Instruction based on computer simulations. In: Handbook of research on learning and instruction. New York: Routledge; 2011:446-466.
8. Villaume WA, Berger BA, Barker BN. Learning motivational interviewing: scripting a virtual patient. Am J Pharm Educ 2006 Apr 15;70(2):33 [FREE Full text] [doi: 10.5688/aj700233] [Medline: 17149413]
9. Setrakian J, Gauthier G, Bergeron L, Chamberland M, St-Onge C. Comparison of assessment by a virtual patient and by clinician-educators of medical students' history-taking skills: exploratory descriptive study. JMIR Med Educ 2020 Mar 12;6(1):14428 [FREE Full text] [doi: 10.2196/14428] [Medline: 32163036]
10. Heitzmann N, Seidel T, Hetmanek A, Wecker C, Fischer MR, Ufer S, et al. Facilitating diagnostic competences in simulations in higher education a framework and a research agenda. Frontline Learning Research 2019 Dec 3:1-24. [doi: 10.14786/flr.v7i4.384]
11. Kopp V, Stark R, Fischer MR. Fostering diagnostic knowledge through computer-supported, case-based worked examples: effects of erroneous examples and feedback. Med Educ 2008 Aug;42(8):823-829. [doi: 10.1111/j.1365-2923.2008.03122.x] [Medline: 18564096]
12. Chernikova O, Heitzmann N, Stadler M, Holzberger D, Seidel T, Fischer F. Simulation-based learning in higher education: a meta-analysis. Review of Educational Research 2020 Jun 15;90(4):499-541. [doi: 10.3102/0034654320933544]
13. Cook DA, Brydges R, Hamstra SJ, Zendejas B, Szostek JH, Wang AT, et al. Comparative effectiveness of technology-enhanced simulation versus other instructional methods: a systematic review and meta-analysis. Simul Healthc 2012 Oct;7(5):308-320. [doi: 10.1097/SIH.0b013e3182614f95] [Medline: 23032751]
14. Schubert T, Friedmann F, Regenbrecht H. The experience of presence: factor analytic insights. Presence: Teleoperators & Virtual Environments 2001 Jun;10(3):266-281. [doi: 10.1162/105474601300343603]
15. Shaffer DW, Resnick M. Thick authenticity: new media and authentic learning. J Interact Learn Res 1999;10(2):195-216 [FREE Full text]
16. Hamstra SJ, Brydges R, Hatala R, Zendejas B, Cook DA. Reconsidering fidelity in simulation-based training. Acad Med 2014 Mar;89(3):387-392 [FREE Full text] [doi: 10.1097/ACM.0000000000000130] [Medline: 24448038]
17. Hofer M. Presence und involvement. 1st ed. Baden-Baden: Nomos; 2016:978-973.
18. Barrows HS. An overview of the uses of standardized patients for teaching and evaluating clinical skills. AAMC. Acad Med 1993 Jun;68(6):443-451. [doi: 10.1097/00001888-199306000-00002] [Medline: 8507309]
19. Luctkar-Flude M, Wilson-Keates B, Larocque M. Evaluating high-fidelity human simulators and standardized patients in an undergraduate nursing health assessment course. Nurse Educ Today 2012 May;32(4):448-452. [doi: 10.1016/j.nedt.2011.04.011] [Medline: 21565436]
20. Rethans JJ, Sturmans F, Drop R, van der Vleuten C. Assessment of the performance of general practitioners by the use of standardized (simulated) patients. Br J Gen Pract 1991 Mar;41(344):97-99 [FREE Full text] [Medline: 2031767]
21. Friedman CP, France CL, Drossman DD. A randomized comparison of alternative formats for clinical simulations. Med Decis Making 1991;11(4):265-272. [doi: 10.1177/0272989X9101100404] [Medline: 1766329]
22. Padgett J, Cristancho S, Lingard L, Cherry R, Haji F. Engagement: what is it good for? The role of learner engagement in healthcare simulation contexts. Adv Health Sci Educ Theory Pract 2019 Oct;24(4):811-825. [doi: 10.1007/s10459-018-9865-7] [Medline: 30456474]
23. Norman G, Dore K, Grierson L. The minimal relationship between simulation fidelity and transfer of learning. Med Educ 2012 Jul;46(7):636-647. [doi: 10.1111/j.1365-2923.2012.04243.x] [Medline: 22616789]
24. Schoenherr JR, Hamstra SJ. Beyond fidelity: deconstructing the seductive simplicity of fidelity in simulator-based education in the health care professions. Simul Healthc 2017 Apr;12(2):117-123. [doi: 10.1097/SIH.0000000000000226] [Medline: 28704289]
25. La Rochelle JS, Durning SJ, Pangaro LN, Artino AR, van der Vleuten CPM, Schuwirth L. Authenticity of instruction and student performance: a prospective randomised trial. Med Educ 2011 Aug;45(8):807-817. [doi: 10.1111/j.1365-2923.2011.03994.x] [Medline: 21752077]
26. Sweller J, van Merrienboer JJG, Paas FGWC. Cognitive architecture and instructional design. Educational Psychology Review 1998;10(3):251-296. [doi: 10.1023/a:1022193728205]
27. Young JQ, Van Merrienboer J, Durning S, Ten Cate O. Cognitive Load Theory: implications for medical education: AMEE Guide No. 86. Med Teach 2014 May;36(5):371-384. [doi: 10.3109/0142159X.2014.889290] [Medline: 24593808]
28. Dankbaar MEW, Alsma J, Jansen EEH, van Merrienboer JJG, van Saase JLCM, Schuit SCE. An experimental study on the effects of a simulation game on students' clinical cognitive skills and motivation. Adv Health Sci Educ Theory Pract 2016 Aug;21(3):505-521 [FREE Full text] [doi: 10.1007/s10459-015-9641-x] [Medline: 26433730]
29. Haji FA, Cheung JJH, Woods N, Regehr G, de Ribaupierre S, Dubrowski A. Thrive or overload? The effect of task complexity on novices' simulation-based learning. Med Educ 2016 Sep;50(9):955-968. [doi: 10.1111/medu.13086] [Medline: 27562895]
30. Young M, Thomas A, Lubarsky S, Ballard T, Gordon D, Gruppen LD, et al. Drawing boundaries: the difficulty in defining clinical reasoning. Acad Med 2018 Jul;93(7):990-995. [doi: 10.1097/ACM.0000000000002142] [Medline: 29369086]

31. Guagnano MT, Merlitti D, Manigrasso MR, Pace-Palitti V, Sensi S. New medical licensing examination using computer-based case simulations and standardized patients. Acad Med 2002 Jan;77(1):87-90. [doi: 10.1097/00001888-200201000-00020] [Medline: 11788331]
32. Hawkins R, MacKrell Gaglione M, LaDuca T, Leung C, Sample L, Gliva-McConvey G, et al. Assessment of patient management skills and clinical skills of practising doctors using computer-based case simulations and standardised patients. Med Educ 2004 Sep;38(9):958-968. [doi: 10.1111/j.1365-2929.2004.01907.x] [Medline: 15327677]
33. Hu L, Chen G, Li P, Huang J. Retracted article: multimedia effect in problem solving: a meta-analysis. Educ Psychol Rev 2019 Jul 11;32(3):901. [doi: 10.1007/s10648-019-09490-4]
34. Stillman PL, Swanson DB, Smee S, Stillman AE, Ebert TH, Emmel VS, et al. Assessing clinical skills of residents with standardized patients. Ann Intern Med 1986 Nov;105(5):762-771. [doi: 10.7326/0003-4819-105-5-762] [Medline: 3767153]
35. Woolliscroft JO, Calhoun JG, Billiu GA, Stross JK, MacDonald M, Templeton B. House officer interviewing techniques: impact on data elicitation and patient perceptions. J Gen Intern Med 1989;4(2):108-114. [doi: 10.1007/BF02602349] [Medline: 2709168]
36. Casus computer software. 2018. URL: https://www.instruct.eu/en/ [accessed 2021-01-30]
37. Bornemann BM. Documentation forms of internal medicine and surgery for history taking and the physical examination for the medical training of students in Germany: An analysis of content and structure. Diss. München: Institut für Didaktik und Ausbildungsforschung in der Medizin der Ludwig-Maximilians-Universität München; 2016. URL: https://edoc.ub.uni-muenchen.de/19166/1/Bornemann_Barbara.pdf [accessed 2021-02-16]
38. Seidel T, Stürmer K, Blomberg G, Kobarg M, Schwindt K. Teacher learning from analysis of videotaped classroom situations: does it make a difference whether teachers observe their own teaching or that of others? Teaching and Teacher Education 2011 Feb;27(2):259-267. [doi: 10.1016/j.tate.2010.08.009]
39. Vorderer P, Wirth W, Gouveia F, Biocca F, Saari T, Jäncke F, et al. MEC spatial presence questionnaire (MEC-SPQ): short documentation and instructions for application. Report to the European Community, Project Presence: MEC (IST--37661). 2001. URL: http://www.ijk.hmt-hannover.de/presence [accessed 2021-01-30]
40. Frank B. Validation. Measuring Presence in Laboratory-Based Research with Microworlds 2014:51-61. [doi: 10.1007/978-3-658-08148-5_6]
41. Opfermann M. There's more to it than instructional design: the role of individual learner characteristics for hypermedia learning. Berlin: Logos Verlag; 2008:1-295.
42. Rheinberg F, Vollmeyer R, Burns BD. FAM: Ein fragebogen zur erfassung aktueller motivation in lern- und leistungssituationen. Diagnostica 2001 Apr;47(2):57-66. [doi: 10.1026//0012-1924.47.2.57]
43. Wigfield A. Expectancy-value theory of achievement motivation: a developmental perspective. Educ Psychol Rev 1994 Mar;6(1):49-78. [doi: 10.1007/bf02209024]
44. Schmidmaier R, Eiber S, Ebersbach R, Schiller M, Hege I, Holzer M, et al. Learning the facts in medical school is not enough: which factors predict successful application of procedural knowledge in a laboratory setting? BMC Med Educ 2013 Mar 22;13(1):28 [FREE Full text] [doi: 10.1186/1472-6920-13-28] [Medline: 23433202]
45. Lakens D, Scheel AM, Isager PM. Equivalence testing for psychological research: a tutorial. Advances in Methods and Practices in Psychological Science 2018 Jun 01;1(2):259-269. [doi: 10.1177/2515245918770963]
46. Lakens D. Equivalence tests: a practical primer for tests, correlations, and meta-analyses. Soc Psychol Personal Sci 2017 May;8(4):355-362 [FREE Full text] [doi: 10.1177/1948550617697177] [Medline: 28736600]
47. R Foundation for statistical computing. R [Computer software]. Vienna, Austria: R Foundation for Statistical Computing; 2019. URL: https://www.r-project.org/ [accessed 2021-02-16]
48. Holm S. A simple sequentially rejective multiple test procedure. Scandinavian J Stat. 1979. URL: https://www.jstor.org/stable/4615733?seq=1 [accessed 2021-02-16]
49. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale: Lawrence Erlbaum; 1988.
50. G*Power computer software. 2014. URL: https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html [accessed 2021-01-30]
51. Sweller J, van Merriënboer JJG, Paas F. Cognitive architecture and instructional design: 20 years later. Educ Psychol Rev 2019 Jan 22;31(2):261-292. [doi: 10.1007/s10648-019-09465-5]
52. Lapkin S, Levett-Jones T. A cost-utility analysis of medium vs. high-fidelity human patient simulation manikins in nursing education. J Clin Nurs 2011 Dec;20(23-24):3543-3552. [doi: 10.1111/j.1365-2702.2011.03843.x] [Medline: 21917033]
53. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990 Sep;65(9 Suppl):S63-S67. [doi: 10.1097/00001888-199009000-00045] [Medline: 2400509]
54. Lewis KL, Bohnert CA, Gammon WL, Hölzer H, Lyman L, Smith C, et al. The Association of Standardized Patient Educators (ASPE) Standards of Best Practice (SOBP). Adv Simul (Lond) 2017;2:10 [FREE Full text] [doi: 10.1186/s41077-017-0043-4] [Medline: 29450011]

Abbreviations

CG: case group

NHST: null hypothesis significance testing
SP: standardized patient
TOST: 2 separate 1-sided test
VP: virtual patient

Edited by G Eysenbach, R Kukafka; submitted 25.06.20; peer-reviewed by J Cheung, P Bergl, A Kononowicz, S Edelbring; comments to author 08.08.20; revised version received 01.10.20; accepted 27.12.20; published 04.03.21

Please cite as:
Fink MC, Reitmeier V, Stadler M, Siebeck M, Fischer F, Fischer MR
Assessment of Diagnostic Competences With Standardized Patients Versus Virtual Patients: Experimental Study in the Context of History Taking
J Med Internet Res 2021;23(3):e21196
URL: https://www.jmir.org/2021/3/e21196
doi: 10.2196/21196
PMID: 33661122