PATIENT-CENTERED OUTCOMES
RESEARCH INSTITUTE
WWW.PCORI.ORG | INFO@PCORI.ORG | FOLLOW US @PCORI

EMERGING TECHNOLOGIES AND THERAPEUTICS REPORT

NARRATIVE REVIEW AND EVIDENCE MAPPING

Artificial Intelligence in Clinical Care

Narrative Review and Evidence Mapping

Artificial Intelligence in Clinical Care

Federico Girosi, Sean Mann, Vishnupriya Kareddy

RAND Corporation

February 2021

Suggested Citation: Girosi F, Mann S, Kareddy V. Narrative Review and Evidence Mapping: Artificial Intelligence in Clinical Care. Patient-Centered Outcomes Research Institute; February 2021. Prepared by RAND under Contract No. IDIQ-TO#22-RAND-ENG-AOSEPP-04-01-2020.

All statements, findings, and conclusions in this publication are solely those of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI) or its Board of Governors. This publication was developed through a contract to support PCORI's work. Questions or comments may be sent to PCORI at info@pcori.org or by mail to 1828 L Street NW, Suite 900, Washington, DC 20036.

Table of Contents

Figures

Tables

Abbreviations

AED	automated external defibrillator
AF	atrial fibrillation
AI	artificial intelligence
AUC	area under the curve
CART	classification and regression tree
CDRH	Center for Devices and Radiological Health
COPD	chronic obstructive pulmonary disease
CRC	Cooperative Research Center
DL	deep learning
ECG	electrocardiogram
ED	emergency department
EEG	electroencephalogram
EHR	electronic health records
FDA	US Food and Drug Administration
ICU	intensive care unit
IEEE	Institute of Electrical and Electronics Engineers
ML	machine learning
NLP	natural language processing
PCORI	Patient-Centered Outcomes Research Institute
RCT	randomized control trial
VA	Department of Veterans Affairs

Acknowledgments

We are grateful to Natalie Benda, Laura Faherty, Paul Koegel, and Jeannie Ringel for their helpful reviews of this report. We also thank Andrea Richardson for input on evidence maps, Bob Rudin for input on the applications of artificial intelligence to health, Jody Larkin for assistance with the literature search, Pedro Nascimento de Lima for providing all the visualizations associated with this report, and Massimo Piccardi for input on machine learning methodology. We thank PCORI for its support, and in particular Bill Lawrence, Gowri Raman, Rachel Andricosky, and Santosh Rao.

Executive Summary

- Aside from areas related to medical imaging, we are in still in the early stages of implementing AI in health care. The number of AI applications actually in use is small compared with the amount of available research and interest in the topic.
- We found 109 non-imaging-based AI applications in current or potential near-future use. These applications most commonly address cardiovascular conditions, diabetes, or general patient care.
- While several evaluations found benefits associated with use of specific AI applications—and virtually none found any harms—the quality of evidence varied significantly. Potential users need more high-quality evaluations of AI applications' direct impact on patient care.
- Stakeholders recognize that there are potential risks associated with the widespread adoption of AI in health, including the introduction of bias, an unequal distribution of benefits and harms, violation of individual privacy, and the threat of malicious attacks on data and devices.
- Despite the early stages and concerns about potential risks, there is no evidence suggesting that the future of AI applications in health is less than bright if an effort to produce high-quality evidence is made.

From self-driving cars to smart supply chains, artificial intelligence (AI) has already changed many aspects of human life. Some areas of health care, such as medical imaging, are benefiting from AI, but questions remain on the role of AI in the broader health sector. Recognizing the importance of this emerging technology and the potential of AI to improve health for all, the Patient-Centered Outcomes Research Institute commissioned this report, which focuses on applications of AI in health that are currently in use or may be in use in the near future.

The report contains 3 major components: a broad overview of all applications of AI in health care; a narrative review describing a specific subset of these applications; and finally a systematic description—or mapping—of the evidence surrounding these specific applications' use. The scope of the narrative review and evidence mapping portions of this report, which constitute the core of this study, includes applications that aim to improve care and outcomes for individual patients and that address a defined range of health conditions, while excluding medical imaging applications, as these have already been adopted and studied to a wider extent.

The overall purpose of this report is to characterize the applications of AI in health care that are in the scope of the study and describe the evidence, or lack thereof, on the associated benefits, harms, and risks. To accomplish this task, we reviewed not only the academic literature but also US Food and

Drug Administration (FDA) approval documents, clinical trials documentation, and web pages of commercial products. We interviewed many stakeholder representatives to validate the scope of the study and gather views and concerns regarding the adoption of AI technologies in health care.

Our search found 109 AI applications—that is, specific AI-based products or tools—that are either in use already or that could be adopted for use in the near future because they are being tested in a real-world setting. Most of the AI applications were either specific to cardiovascular disease or diabetes or geared toward the general population, often targeting users of services such as hospitals or emergency departments.

Most of the AI applications we found (61%) aimed to provide information concerning an individual patient's health status, for example by formulating a diagnosis, by assigning to an individual the risk of an adverse event, or by assessing the symptoms displayed. Several AI applications (30%) provide health recommendations—that is, a list of actions that can be taken to address an underlying health issue. Prominent in this group were consumer-empowering applications, which help individuals manage or prevent certain health conditions, as well as applications recommending specific treatment options, such as personalized medication dosage. The remaining 9% of AI applications were truly part of a treatment delivery process, such as intelligent insulin pumps or smartphone applications that deliver cognitive behavioral therapy. Out of all the applications, only 20% were cleared for use by the FDA; the remaining applications either were in use without FDA clearance or were still in development for potential near-future use.

For most of the AI applications that we reviewed (86%), we were able to find some evidence relative to the performance of the applications. We collected 173 evaluation studies. The evidence varied greatly in quality, and only half of the evaluation studies were published in peer-reviewed journals; the rest was a combination of non-peer-reviewed papers, FDA or clinical trial documents, and web pages of commercial products. The evaluation studies also differed greatly in terms of their design and the outcomes measured: only a quarter of the evaluations were rigorous randomized control trials in which the effect of an AI application on a health-related outcome was reported. Approximately half of the evaluations were studies that did not report health outcomes at all, but rather reported the accuracy of the AI method used in the application or the usability of the application itself. Overall, despite dissimilarity in the quality of the evidence, we found that many patients, especially those with cardiovascular disease or diabetes, are already benefiting, or are likely to benefit in the near future, from AI applications to the health domain.

This study highlights the fact that this initial evidence base is not very broad and needs to expand along more than one dimension. What is needed is not just more studies, but more high-quality studies, such as randomized control trials or prospective studies, that could provide a solid link between the use of an AI application and improvement in well-defined health-related outcomes.

The small size of the evidence base contrasts with the common perception that AI is already highly prevalent in health—undoubtedly supported by the thousands of research papers published on this topic every year. Our literature search has confirmed that the vast majority of studies on AI and health

describe the *potential* benefits of certain solutions. Very few of the published AI solutions are ready to enter clinical care, suggesting that, from an implementation point of view, the field is still in its early days, barriers need to be overcome, and concerns must be addressed. In addition to privacy and safety, the most common concern among our stakeholder representatives was the possibility of introducing or perpetuating existing bias into AI solutions, thus creating or worsening an unequal distribution of outcomes across the population. The literature shows that researchers are well aware of this important ethical issue, although there is no clear way to address it yet. That said, we reiterate the fact that applications of AI in health care, other than those related to imaging, are still in their early days, and it is normal for new technologies to face ethical issues at this stage of adoption. Therefore, there is no evidence to suggest that the path for AI applications in health, while complex and long, should be anything less than bright.

1. Introduction

The large number of publications on the topic of Artificial Intelligence (AI) and health care, and the multitude of health-related commercial applications that contain some elements of AI, suggest that AI holds great potential in the field of health care.[1-3] However, the evidence base necessary to justify applications of AI to the health domain is limited. The vast majority of publications on AI and health do not describe AI applications (ie, AI-based products or tools) that are actually in use,[4] and many commercial applications that are currently in use have not been evaluated in peer-reviewed studies.

To help understand the use of this rapidly developing technology in health care, this report seeks to fulfil 3 primary aims:

> Aim 1. Provide a broad overview of the full range of AI applications with potential for use in health care.
>
> Aim 2. Conduct a narrative review that characterizes a specific subset (see study scope below) of AI applications in current or potential near-future use in clinical care.
>
> Aim 3. Map the evidence that exists for evaluating these specific AI applications and their potential benefits.

In keeping with the broad scope of aim 1, this report begins by giving an overview of the various types of AI applications with potential for use in health care. The remainder of the report then narrows in scope to address aims 2 and 3, the narrative review and evidence map, which assess in greater detail a particular set of these applications.

To address aim 3, we follow an evidence mapping approach,[5] in which we identify, categorize, and provide a visual depiction of published studies evaluating the AI applications identified in aim 2.

This report is intended for a broad audience of health care stakeholders—whether patients or clinicians, hospital administrators, insurers, or researchers—interested in the state of AI in clinical care.

Study Scope

Because the full range of potential AI applications in health care is so broad, we begin with an overview (aim 1) and then establish clear boundaries to determine which applications are included in the following narrative review and evidence map (aims 2 and 3). We organize these boundaries according to several dimensions, including an application's current development and regulatory and adoption status, as well as function, AI type, data input, and health conditions addressed. These dimensions are used throughout this report to characterize different types of AI applications and will be discussed further in the presentation of our conceptual framework in chapter 2.

We present in Table 1.1 the exact dimensions of the scope of the narrative review and evidence map as compared with the scope of our broad overview of AI in health care.

Table 1.1. Scope of the Broad Overview vs Scope of the Narrative Review and Evidence Map

Dimension	Scope of the broad overview of AI in health care (aim 1)	Scope of the narrative review (aim 2) and evidence map (aim 3)
Current status	Applications in current use in any setting or in any phase of development, including early research	Applications in current use or that may be adopted for potential near-future use in the next 5 years
Function	Applications used to perform any health care function, including health system administration, resource management, research, epidemiology, and clinical care	Applications used to perform clinical care functions of providing patient evaluation, health recommendations, or treatment delivery
AI type	Applications that use any type of artificial intelligence	Machine learning–based applications only
Health conditions	Applications used to address general patient health or any health condition	Applications that address general patient health or any of 9 specific types of health conditions: cancer, cerebrovascular, cardiovascular, dementia, diabetes, kidney disease, mental health, respiratory, and substance abuse
Data input	Applications that use any type of data input, including imaging data	Applications that use non-imaging-based data inputs
User	Applications designed for use by clinicians, patients, or any other user type	Applications designed for use by clinicians, patients, or any other user type
Setting	Applications used in health facilities, at home, or in any other setting	Applications used in health facilities, at home, or in any other setting
Platform	Applications embedded in any device or computing platform	Applications embedded in any device or computing platform

The Patient-Centered Outcomes Research Institute (PCORI) and the RAND study team collaboratively developed the scope of the narrative review and evidence map to focus on a diverse set of applications that are directly relevant to patient care. We discussed this scope with a wide range of health care stakeholders to ensure that it reflected their priorities prior to beginning the review. Multiple stakeholders suggested that we include the following, which we incorporated into the narrative review scope described above:

- Applications focused on general patient health (in addition to those focused on the 9 specific health conditions)
- Applications used by patients for self-monitoring and management (in addition to applications used by clinicians or hospitals)
- Applications focused on prevention (in addition to applications focused on diagnosis or treatment)

Stakeholders agreed that the scope should exclude imaging-based applications, such as those related to radiology or dermatology, as they represent a more mature set of technologies that have already been extensively studied and are at a more advanced stage of adoption.

The scope of the narrative review and evidence map focuses on applications in current or potential near-future use to distinguish these most immediately relevant applications from the much larger number of applications that are in earlier stages of development. We further narrow our review to applications that are used in clinical care, rather than to perform other health care system functions, as this group most directly affects patient health. We focus on machine learning–based applications, as this type has become the dominant kind of AI in both use and development today.

Although we were unable to cover all potential health conditions, we do examine applications that address general patient health as well as a set of 9 specific health conditions. These conditions represent a diverse set of health challenges that affect large numbers of patients and can have potentially serious effects. We consider applications designed for any type of user, setting, or platform in our review.

Organization of the Report

The remainder of this introductory chapter provides background on AI and machine learning, followed by a description of US government regulations concerning the use of AI in health care.

Chapter 2 then presents the methods used in this study. This chapter also describes the 2 conceptual frameworks we used in this study: a framework to classify different types of AI applications in health care and a framework to group different types of evaluation studies that provide data on these AI applications.

Chapter 3 addresses aim 1 by giving a broad overview of the use of AI in health care, focusing on the applications not discussed in the narrative review of aim 2.

Chapter 4 addresses aim 2 by presenting the findings of our narrative review of AI applications in current or potential near-future use in clinical care. These findings are organized according to the conceptual framework provided in chapter 2, with applications characterized by each of the framework dimensions. Chapter 4 concludes with a visual depiction of these AI applications according to their current status, function, health conditions, and users.

Chapter 5 addresses aim 3 by examining the evidence base surrounding the use of the AI applications that fall within the scope of our narrative review. Evaluations are characterized by publication type, study design, sample size, and outcomes measured. This chapter ends with 2 evidence maps—that is, visualizations that aim to identify presence or absence of knowledge and to depict evaluation studies according to multiple dimensions.[5]

Chapter 6 summarizes views and concerns expressed by stakeholders regarding the application of AI in health.

Chapter 7 turns to the implications of our narrative review and evidence map findings. It concludes by considering the potential research activities that could address the gaps identified in our map of the evidence base surrounding the use of these applications.

Background on Artificial Intelligence and Machine Learning

As noted by a 2016 National Science and Technology Council report, "there is no single definition of AI that is universally accepted by practitioners."[6] In this report, we define *AI* as a system with the ability to mimic or simulate the completion of tasks we usually attribute to humans, such as reasoning, learning from examples, communicating, displaying or understanding emotions, and planning and making decisions. This definition is quite similar to that used by the National Institute of Standards and Technology as well as in a recent US Food and Drug Administration (FDA) discussion paper.[7,8] Depending on which human task is being simulated, one obtains a different type of AI. Some examples of AI types and their corresponding human tasks are shown in Table 1.2.

Table 1.2. Correspondence Between Human Tasks and AI Types

Human task	Type of AI
Reasoning	Expert or rule-based systems
Communicating by voice and text	Natural language processing and conversational AI
Displaying and understanding emotions	Affective computing
Planning and making decisions	Reinforcement learning
Learning to perform a task from examples	Machine learning (supervised)
Clustering and pattern discovery	Machine learning (unsupervised)

It is important to recognize that these are very broad categories that are not mutually exclusive and share many methodologies and technologies. For the purpose of this report, the type of human, intelligent task that we expect an artificial system to perform is the ability to learn to solve a problem from a set of examples. The resulting type of AI falls under the label of *supervised machine learning*, and it is the AI type that is, currently, most commonly used for applications in health and health care. There is also a related type—unsupervised machine learning—that we originally included in the scope of this study; however, we did not identify any applications of this kind in our narrative review. Therefore, the only type of machine learning considered here is supervised machine learning (ML). We provide a brief description of ML in the next section.

Machine Learning

Humans use a variety of strategies to solve a problem. One common approach is to apply reasoning and rules to the set of observations defining the problem to obtain a solution. A different problem-

solving strategy, however, may involve looking at data that show how problems of the same type already have been solved, finding a problem that is similar to the one at hand, and using the same solution or a comparable one.

As an example, consider the task of pronouncing words in a foreign language. Here, an instance of the problem to solve is a word and the solution is the correct pronunciation. One approach to this task consists of acquiring a textbook containing the phonetic rules of pronunciation and applying them when needed. This tactic is effective because it allows someone to solve any instance of a problem correctly—as long as the pronunciation rules are complete and cover every possible case.

An alternative approach, however, consists of drawing a sample of words from the vocabulary and asking native speakers to pronounce them. After going through this procedure many times, an individual (the learner) would eventually learn how to pronounce these words correctly and would be able to generalize—that is, to correctly pronounce words unknown to him or her. This "generalizing" approach could be as effective as the first one—if the learner has been exposed to a sufficiently large number of examples and as long as similar words have like pronunciations. If similar words were associated with radically different pronunciations, with no discernible patterns, then learning would be impossible, and the only effective strategy would be to memorize the pronunciation of each word in the lexicon.

The 2 approaches are radically different: in the first case, a preexisting and complete set of pronunciation rules, codified in one book, exist; in the second, learners have developed in their heads an association that maps any word to a pronunciation. This association map plays the role of the phonetic rules, but unlike those rules it was derived empirically from a set of examples and is never explicitly codified (although it is stored in the learner's brain).

Algorithms that mimic the process described in the second approach, in which a class of problems is solved by learning to associate an instance of a problem (eg, a specific word) to its solution (eg, the pronunciation of that word), based on a set of examples, fall in the category of supervised machine learning. Supervised ML algorithms have the following characteristics:

- **The task to be performed is to associate an instance of a problem with its solution.** For example, a physician receives the clinical records of a patient and aims to confirm a diagnosis of systemic lupus erythematosus (lupus). We view the clinical observations as an instance of the problem—that is, the observational data that someone can use to answer the question of whether the patient has lupus. The diagnosis of lupus (yes/no) is the solution of that problem, and the task is to associate the records of any patient to the corresponding diagnosis. Note that, for simplicity of language, we often refer to *an instance of a problem* as *a problem* unless the distinction is important.
- **Machine learning relies on a data set of examples of problems and the corresponding solutions**. For example, someone has access to electronic health record (EHR) data of patients who were evaluated for lupus. Again, the problem is the set of clinical data that must be associated with the presence or absence of lupus. For each patient, we also know whether the

lupus diagnosis was confirmed (the solution). The data set is a collection of pairs (problem, solution), one for each patient.

- **Similar problems are assumed to have similar solutions (under some meaningful notions of similarity).** This is what is usually called the "smoothness" assumption, and it is the key assumption that allows generalization—that is, to solve problems that have never been encountered. For example, generalization takes place when someone correctly diagnoses lupus in a new patient: this task is possible only if patients "who look like this" tend to share the same diagnosis. If, instead of diagnosing lupus, someone were trying to determine the patients' favorite colors, this assumption would not be satisfied, since patients with similar clinical characteristics do not share the same favorite colors. In such cases, generalization is not possible.
- **A machine learning model is an association between problems and solutions**—that is, a set of empirical rules that associate any instance of a problem to a solution. The defining feature of ML is the fact that these rules are derived entirely from the data. In the lupus example above, each patient is represented by a long list of numbers corresponding to test results, and the association is represented by a complex formula that takes those numbers, combines them using numerical coefficients, and produces a binary output: 1 for lupus and 0 otherwise. To design an ML model, someone starts with a generic formula, whose coefficients are not specified, and estimates those coefficients based on the observed data. In common parlance, the ML model has "learned" the association between problems and solutions, or it has been "trained" on a set of examples.

A summary of the machine learning approach to problem solving is shown in Figure 1.1 below.

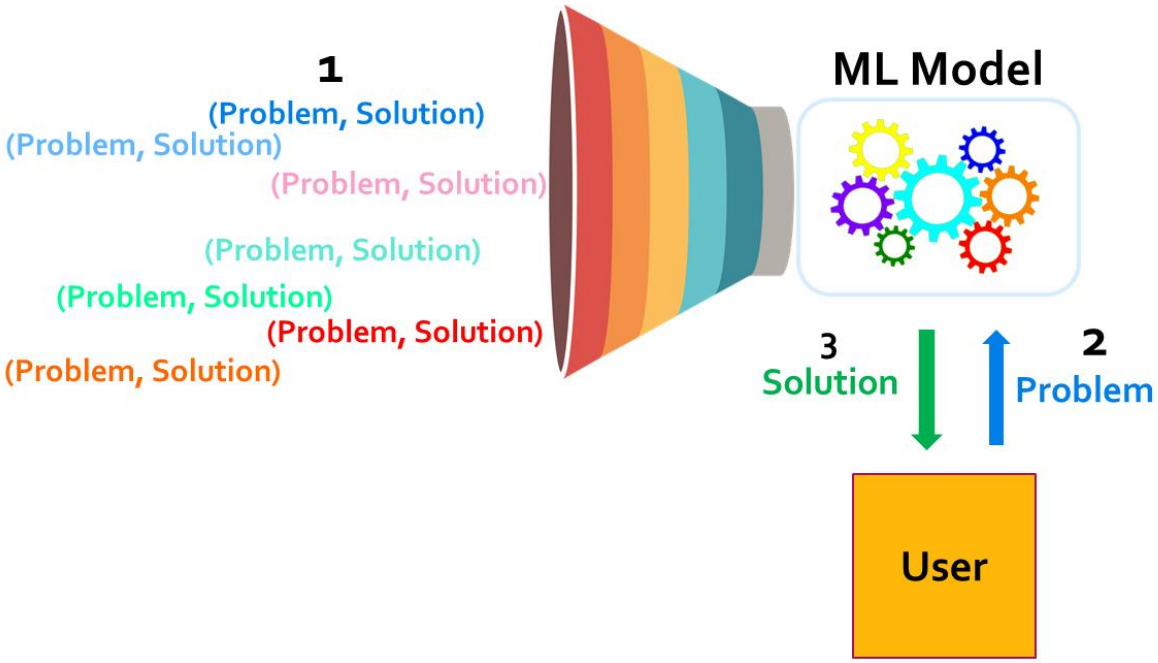
As mentioned above, in this report we use the term *machine learning* to denote *supervised machine learning*; for completeness we also provide a short definition of *unsupervised machine learning*. This latter type of ML mimics the human ability to observe a set of data (such as the EHRs of a group of patients, or a set of images, or documents) and understand that there is an underlying structure, usually represented by several groups. A clinician may look at a group of patients and realize that they can be divided in clusters, with the characteristics that patients in each cluster are similar, but patients in different clusters are instead dissimilar. Similarly, a researcher analyzing documents for a literature review may discover that they can be grouped according to certain topics and subtopics—and therefore that researcher can define a taxonomy that assigns each document to a specific class. In some cases, the clusters detected contain only one point that is different from all the others. This is the case, for example, of an auditor who finds unusually high level of opioid prescriptions associated with a specific provider, making that provider an outlier and worth investigating further. Unsupervised ML models can perform tasks similar to the ones described above and detect both clusters and outliers in data sets.

More generally, unsupervised ML models can reconstruct the distribution underlying a set of observations, or at least to capture certain features of it, and therefore are helpful in understanding the

structure of the data. They share with supervised ML models the fact that the only input to the algorithm is a set of observations. With supervised ML the observations are labeled, in which the label refers to the solution of the problem at hand (usually the prediction of an event, such a diagnosis), while with unsupervised ML models the data do not have any labels attached to them.

Being able to predict an event is highly useful in clinical care, since it can clearly inform a decision. Understanding the structure underlying the data may be useful, but it is less directly applicable, which explains why supervised ML is much more common in AI applications to clinical care. Unsupervised ML, however, plays a very important role in other applications of AI to health care and in particular in fraud and error detection,[9-11] where the ability to identify outliers and unusual events is critical.[12-14]

Figure 1.1. The Supervised Machine Learning Approach to Problem Solving



Note: An ML model can be thought of as an engine that has been fed a "training" data set with many instances of problems and corresponding solutions (1) and has "learned" the correspondence between a problem and its solution. A user queries the ML engine with a specific instance of a new problem (2) and the system returns a solution (3) that is consistent with the problem in the data set.

Advantages of Machine Learning

What makes the machine learning approach particularly useful is that it is applicable in situations in which there is no preexisting set of rules—because they are unknown, are not available, or are too complex to be codified in a usable way. ML allows researchers to extract the rules that map an instance of a problem onto its solution from a set of past observations. ML excels at tasks that are difficult for

humans: analyzing very large data sets, finding correlations across hundreds of variables, and exploiting these correlations to make accurate predictions. Therefore, ML algorithms often find patterns that humans cannot identify either because they involve too many variables or because it would be too time consuming to perform.

Conversely, ML systems are usually unable to generalize from limited data, and although they are able to establish a connection among events, they are typically not able to assess whether that connection is causal. Humans, instead, excel at this type of task: clinicians do not need to see thousands of cases to make a correct diagnosis and are much better positioned to establish causality because of their deep understanding of human anatomy and physiology and their ability to apply sophisticated reasoning.

The complementarity between what humans can do and what ML can do makes ML a powerful tool in any setting where people need support in tasks they already perform, to increase accuracy, or for tasks that people are not currently performing because those tasks are too complex or require a large amount of resources. Combining this complementarity with the increased availability of large clinical and administrative data sets, and the improved computing power available to analyze them, explains the explosion of interest in applications of ML to health care, which is discussed in the next section.

AI and Health Care

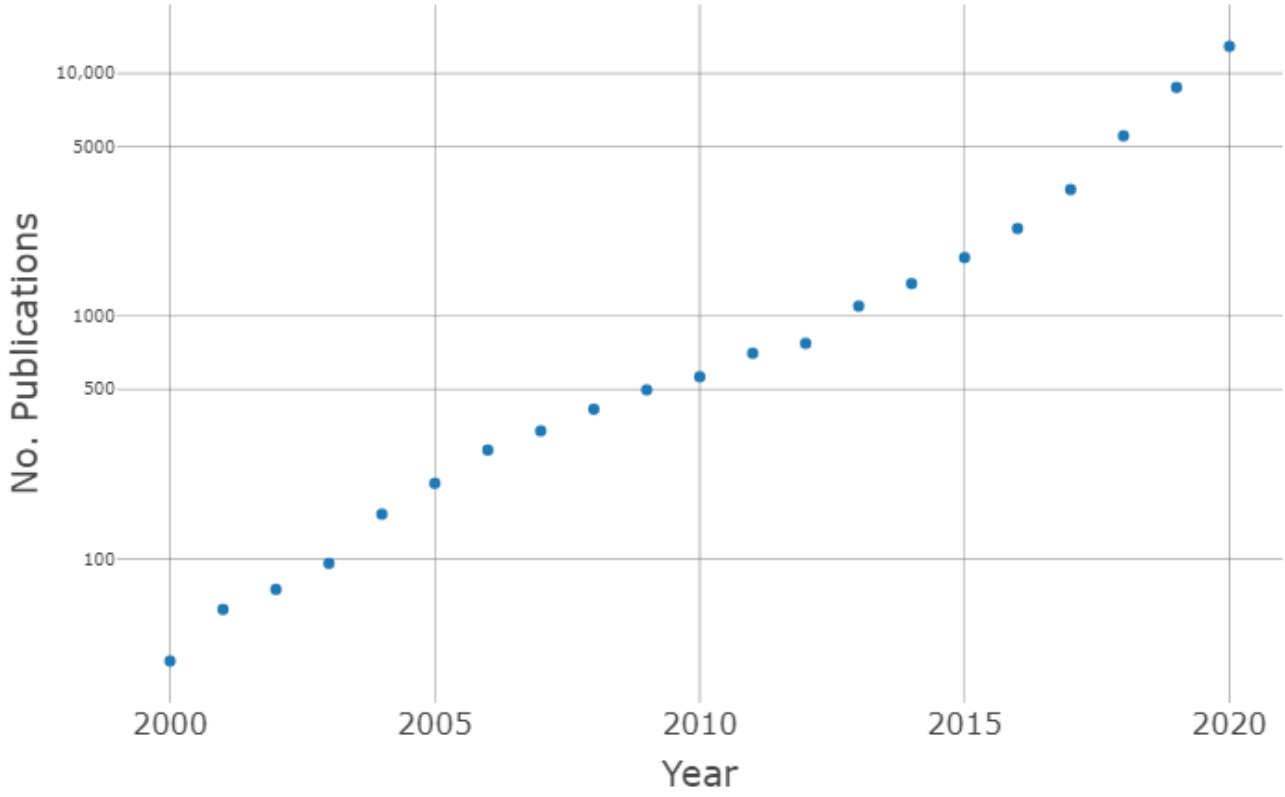
Studies on the topic of AI and health care have grown exponentially over the past 20 years. A simple PubMed search for the terms *health* and (*artificial intelligence* OR *machine learning*) since the year 2000 returns more than 30 000 documents. To convey how quickly the number of publications has grown over time, we show in Figure 1.2 the cumulative number of documents returned by the PubMed search as a function of time. On the horizontal axis we list years, from 2000 to 2020, and on the vertical axis we give, on a logarithmic scale, the number of documents published until a given a year. Because the vertical axis is on a logarithmic scale, the fact that the plot is almost linear means that the growth has been exponential. In fact, a more detailed analysis shows that the number of publications has been doubling approximately every 2.7 years.

It is crucial to point out that the vast majority of these documents do not describe a real-world application of AI and ML methods to health care, in which an algorithm is actually implemented in a clinical setting. Rather, they generally discuss the *potential* benefits of such applications by applying algorithms to clinical or administrative data and reporting their performances in a more exploratory fashion.

As the rest of the document will discuss, the number of applications whose use in clinical care is documented to some extent is much smaller,[4] and the number of cases in which peer-reviewed evidence about the effectiveness of those algorithms is even smaller. The exception to this general statement is the applications of AI and ML to medical imaging, for which some of those potential

benefits are already realized and are out of the scope of this review. Examples of such applications are presented in chapter 3.

Figure 1.2. Publications in PubMed on AI and Health Since 2000



Regulation of AI in Health

Some AI applications, together with other algorithms and software programs, are regulated as medical devices by the FDA Center for Devices and Radiological Health (CDRH). This is the case for AI applications that are embedded in specific medical devices (such as an insulin pump or electronic stethoscope) as well as applications that exist as algorithms for use on web-based, mobile, or other computing platforms.[15] AI applications are not regulated differently than are other software programs; although the FDA has proposed potential modifications to the regulatory process to allow for approval of more frequent or even continuous updates to already-cleared AI algorithms, these modifications have not yet been adopted.[7]

The FDA regulates only those AI applications that perform clinical functions. Under the FDA's statutory authority established by the Food, Drug, and Cosmetics Act, regulated medical devices include any "apparatus, implement, machine . . . or other similar or related article, including a component part, or accessory, which is . . . intended for use in the diagnosis of disease or other conditions, or in the cure, mitigation, treatment, or prevention of disease."[15] The 21st Century Cures

Act enacted in 2016 clarified that software that focuses on health care administration, resource management, epidemiology, research, and other nonclinical functions is not subject to FDA regulation.[16] Software focused on "maintaining or encouraging a healthy lifestyle and . . . unrelated to the diagnosis, cure, mitigation, prevention, or treatment of a disease or condition" is also exempt.[16]

The FDA classifies medical devices based on the risk they pose to patients as Class I (lowest risk), Class II, or Class III (greatest risk).[17] Most Class I and some Class II devices are exempt from FDA regulatory clearance processes.[18]

The FDA maintains 3 primary pathways for clearing devices for use in the United States. Which pathway a device goes through when seeking regulatory approval depends on 2 factors: the risk the device might pose to patients and whether the device is novel, meaning not "substantially equivalent" to any already cleared devices.[19]

The De Novo classification request pathway covers novel medical devices that pose lower or more moderate risks to patients.[19] The 510(k) premarket notification pathway exists for devices that are deemed to be "substantially equivalent" to an already approved device.[20] The premarket approval pathway is the "most stringent" and is required for devices that pose the greatest potential risk to patients.[21] By issuing an emergency use authorization, the FDA can also clear devices for use in addressing a specific public health crisis.[22]

2. Methods and Conceptual Frameworks

This chapter describes the methods and conceptual frameworks used in our study. It begins by describing the stakeholder interviews we conducted at the outset of the study effort to inform the study aims and scope. This discussion is followed by a description of the methods we used to identify and examine a wide range of documents to address the study aims. We then present the conceptual framework we used to classify AI applications in health care for our narrative review, followed by the conceptual framework we employed to classify evaluation studies for our mapping of the evidence base surrounding these applications.

We adopt an evidence mapping approach in this study, meaning "a systematic search of a broad field to identify gaps in knowledge and/or future research needs that presents results in a user-friendly format, often a visual figure or graph, or a searchable database."[5] Our systematic document search focused on identifying all in-scope AI applications and all published evaluations of these applications. We then categorize applications and evaluation studies according to formal conceptual frameworks and present the results visually at the end of chapter 4 (for applications) and chapter 5 (for evaluation studies). We provide full information on each individual application and evaluation study in the tables in appendices A and B, respectively.

Stakeholder Interviews

We began our study by conducting a series of interviews to elicit feedback on our proposed review design as well as to gather initial information on the impact of AI in clinical care. These interviews focused on understanding a broad range of stakeholder perspectives on the use of AI in health care, including the issues and concerns that stakeholders view as most relevant for further study. These perspectives helped ensure that we considered stakeholder priorities when determining the review scope and the guiding questions. Stakeholder representatives also served as key informants to help us understand the impact of AI use in clinical care from a variety of viewpoints.

We began the interviewee recruitment process by identifying a broad set of stakeholder groups we sought to include, drawing on PCORI's list of stakeholders.[23] These groups were patients/patient advocates, clinicians, hospitals/health systems, payers/insurers, purchasers/employers, public policymakers, industry, and researchers. We then identified potential stakeholder representatives to contact for each of these groups.

We reached out to a total of 13 individuals via email to request interviews. Nine individuals responded and agreed to participate: 1 patient advocate, 1 clinician, 1 hospital/health system manager, 2 health payers/insurers, 1 purchaser/employer, 1 public policymaker, 1 industry analyst, and 1 researcher.

Our application to RAND's Human Subjects Protection Committee was approved as exempt from further review on April 28, 2020. Interviews were approximately an hour long, were conducted by

phone or videoconference, and were not recorded. We sent interviewees via email a study description and an informed consent protocol prior to the interview. These documents can be found in appendix C.

We developed an internal interview guide, with a list of interview questions, that can also be found in appendix C. Although we had planned to use these questions to guide our interview discussions, we did not treat them as strict scripts for interviews; rather, we tailored our exact interview approach to each individual interviewee, taking into account the varying perspectives and backgrounds that different stakeholder groups and individuals bring to this topic. Some questions were not applicable to all interviewees, in which case we moved on to another question or topic. Throughout the interview we encouraged interviewees to expand on their answers or raise additional topics they felt important for discussion.

Two researchers were present in each interview: one as lead interviewer and one as lead note taker. These roles were occasionally exchanged for a portion of the interview so that the note taker could participate in the interview.

For each interview, to obtain one final summary document for each interview, the lead interviewer reviewed and revised as necessary the unstructured written notes from the lead note taker. Because the goal of the interviews was to extract themes of interest to stakeholders, we took an inductive, rather than deductive, approach, and each of the 3 authors performed a simple thematic analysis on the set of all interviews. We then reconciled and finalized emerging themes during live discussions among the researchers.

Document Sources and Screening

We collected for use in our study academic, government, and gray literature documents from 4 different sources:

- Systematic reviews from academic research databases
- FDA records
- Clinical trial records
- Additional academic and gray literature from targeted searches

As Table 2.1 shows, the systematic reviews were used to inform our broad overview of AI in health care (aim 1) and identify clinical AI applications for our narrative review (aim 2). Documents from the other 3 sources were also used to identify applications and in addition provided evidence evaluating these applications' use (aim 3).

Table 2.1. Document Sources and Primary Use

Documents and sources	Broad overview of AI in health (aim 1)	Narrative review of AI applications (aim 2)	Map of evidence on AI applications (aim 3)
Systematic reviews from academic research databases	X	X	
FDA records		X	X
Clinical trial records		X	X
Additional documents from targeted searches		X	X

We conducted a systematic search of several document databases as part of our review. The full list of terms used in each of these searches can be found in appendix D. We selected the search terms by leveraging the researchers' experience in the area of both health care and AI, RAND experience on similar projects, and search terms used in similar systematic reviews. Figure 2.1 depicts the number of documents found in each search, together with the results of our document screening.

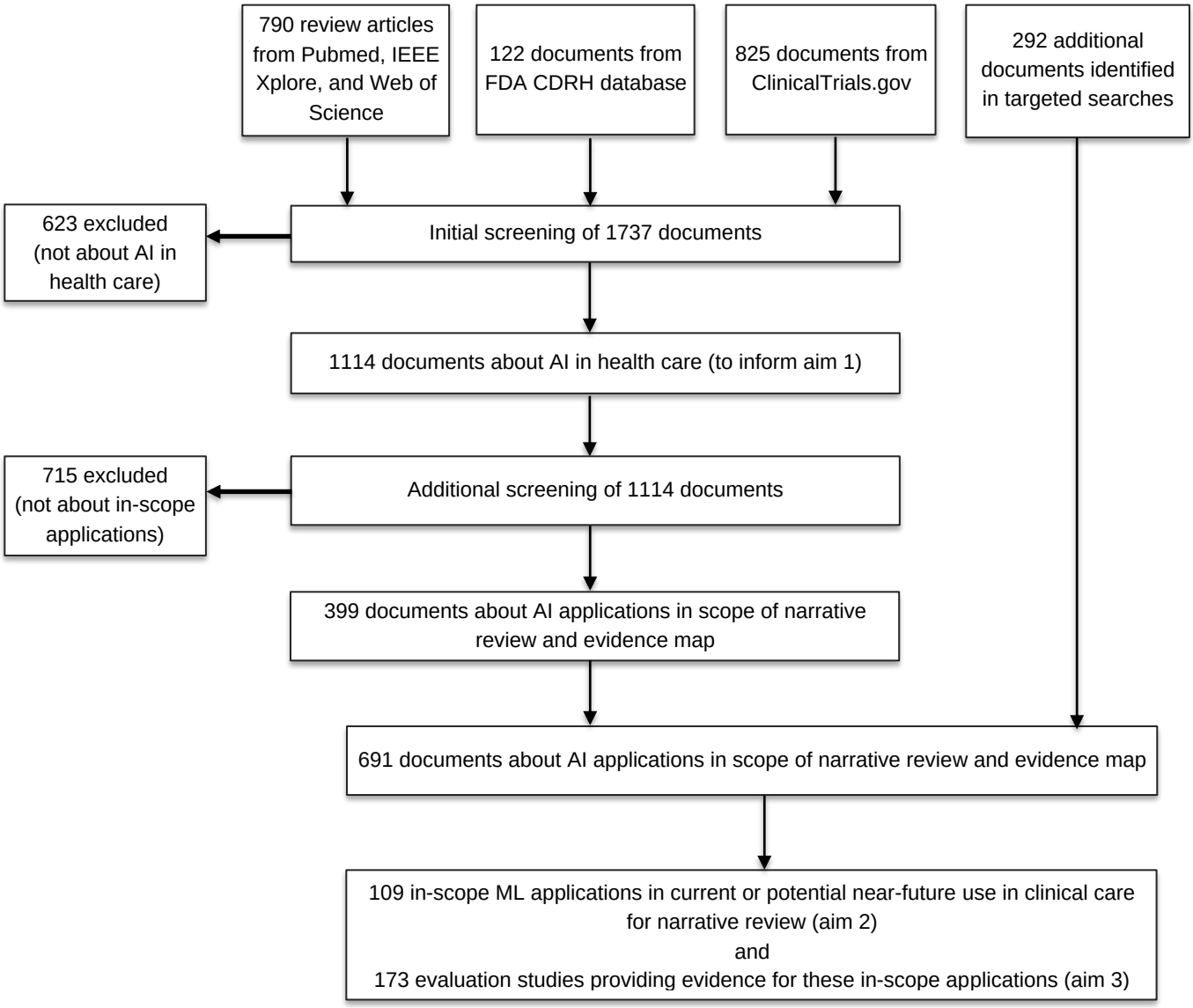
To obtain systematic reviews on AI in health care, we searched 3 research databases: PubMed, Web of Science, and the Institute of Electrical and Electronics Engineers (IEEE) Xplore Digital Library. We identified in these databases all documents whose title or abstract included (1) at least one search term focused on AI and ML, (2) at least one search term focused on health care, and (3) the word *review*. Given the rapidly developing nature of the field, we restricted our search to just the most recent reviews of AI in health care—those that date from 2019 and 2020.

We searched the FDA CDRH document library for all documents, from any year, that mentioned the terms *artificial intelligence*, *machine learning*, *neural network*, or *deep learning* (DL) in the text. The CDRH document library includes all FDA records approving the use of medical devices, including software that relies on AI.[24]

We also searched the ClinicalTrials.gov database for study records that mentioned any AI or ML search term. This US National Library of Medicine database serves as a registry of "privately and publicly funded clinical studies conducted around the world."[25] We restricted this search to records from 2012 to 2020, to capture both ongoing and recently completed clinical trials of AI applications that might potentially be adopted for use in near future.

We then completed targeted searches of the academic and gray literature, seeking to identify additional documents relevant to in-scope ML applications in clinical care. This effort included 2 Google searches on July 30, 2020, using the following search strings: *FDA-approved artificial intelligence* and *artificial intelligence clinical adoption* as well as targeted Google, Google Scholar, and PubMed searches (July-September 2020) using specific application names.

Figure 2.1. Document Sources and Screening



Screening Process

The documents returned by the searches detailed above went through 2 screening steps. In the first, broader, screening step, we sought to identify all documents that discussed the use of AI in health care. Documents that passed this step were then subject to a second screening step in which we sought to pinpoint just those documents that were fully in scope for our narrative review and evidence map—namely, those that specifically discussed the use of non-imaging-based ML in clinical care to address general patient health or a set of 9 common health conditions.

In the first screening step, we examined the titles and abstracts of the systematic reviews and the full text of the FDA approval documents and clinical trial records to determine whether each of these documents discussed topics relevant to AI in health care broadly. To pass this screening step, the document had to meet 2 inclusion criteria:

1. Discuss the use of AI.
2. Discuss any topics related to health or health care.

The 1114 documents that met these 2 criteria were used to inform our broad overview of AI in health care (aim 1) presented in chapter 3 and also provided context for our concluding discussion of regulatory, ethical, and practical implications in chapter 7.

In the second screening step, we examined the full text of the systematic reviews, clinical trial records, and FDA documents that passed initial screening, to determine whether they discussed AI applications in the scope of our narrative review and evidence map. To pass this screening step, the document had to meet all 4 of the following inclusion criteria:

1. Discuss AI applications that relied on data inputs other than imaging.
2. Discuss AI applications that specifically used machine learning rather than some other technique referred to as AI.
3. Discuss AI applications intended for use in clinical care to provide patient evaluation, health recommendations, or treatment delivery.
4. Discuss AI applications that addressed either general patient health or at least 1 of the 9 specific patient health conditions within the scope of our narrative review.

All 3 members of the study team participated in both screening steps. To assess screening consistency across team members, 2 team members screened independently a random sample of 526 documents. Independent decisions in the first screening step agreed 84% of the time, with an interrater reliability Cohen's kappa of 0.62 (kappa values can range from –1.0, lowest reliability, to 1.0, highest reliability). Independent screening decisions in step 2 agreed 85% of the time, with an interrater reliability kappa of 0.58. In cases of disagreement, final screening decisions were arrived at by consensus of the full team following discussion.

A total of 399 documents passed the second screening step. We combined these documents with the 292 additional documents found in our targeted searches, for a total of 691 documents used to inform our narrative review and evidence map.

To conduct our narrative review of AI applications in health care, we examined these 691 documents to identify in-scope applications that were currently in use or that could be adopted for use within the next 5 years. To be included in the latter category, an application needed to have been either (1) actually implemented in patient care as part of a field evaluation study or (2) described as submitting for FDA approval within the next year.

We identified a total of 109 applications in this way. The final stage of our review assessed the evidence base surrounding these applications. This effort involved further examination of these 691 documents to identify evaluations of these 109 applications. Although a few applications had no published evaluations associated with them, most applications were the subject of one or more evaluations. In the end we identified a total of 173 evaluation studies from both the academic and gray literature in this manner.

Conceptual Frameworks

This report aims to provide a picture of the type of AI applications in clinical care that are in current or near-future use, subject to some scope restrictions, (aim 2) as well as the evidence surrounding them (aim 3). We consider applications to be in near-future use if they have been either (1) actually implemented in patient care as part of a field evaluation study or (2) described as submitting for FDA approval within the next year. The presence of multiple aims implies that 2 distinct narratives need to be presented, each with its own unit of analysis and reporting framework:

- **Aim 2**: The unit of analysis is the AI application itself. The reporting framework is intended to capture structural characteristics, such as the function of the application, the targeted health condition, or the regulatory status. The characteristics reported are independent of any evidence about the performance of the application and about the benefit and risks associated with it.
- **Aim 3**: The unit of analysis is an evaluation study containing some form of evidence about the effectiveness of an AI application. The reporting framework mirrors standard frameworks used in systematic reviews of health interventions. It is meant to capture measurable outcomes associated with the use of the application as well as characteristics of the evidence that relate to its quality.

In both cases, information about each unit of analysis (applications or documents) must be extracted and categorized so that it can then be tabulated and analyzed. The frameworks used to perform these operations are described in the sections below.

Framework for Narrative Review of AI Applications in Clinical Care (Aim 2)

There was no obvious guidance on which dimensions should be used to characterize AI applications, determine which ones fall within the scope of this review, and report informative statistics about them. Therefore, guided by PCORI and by our reading of the literature, we developed a conceptual framework to characterize applications according to 8 key dimensions that we summarize in Table 2.2 and describe in more detail in chapter 4: AI type, function, current adoption status, health condition, data input, user, setting, and platform. These dimensions were chosen to reflect the various ways that these applications are characterized in the literature as well as to capture the information identified by stakeholders as of particular importance.

Our framework includes some of the same dimensions found in general frameworks proposed for health information technology more broadly.[26,27] A challenge in adopting any existing framework is that the information available on current and near-future applications is often patchy and may not necessarily fit in that framework. Therefore, the design of the framework we use in this report partly reflects PCORI's interest in understanding what composes the broader field of AI in health care, partly aligns with existing work, and partly is driven by the data, as described in the rest of this section.

For each of the dimensions of the framework, we had to define which categories to which an application could be assigned. For some dimensions, the choice was clear, as in the case of health conditions, for which the categories correspond to the priority conditions described in chapter 1. For other dimensions, we arrived at the choice of the corresponding categories through the following iterative process:

1. We started from a tentative set of categories that we arrived at using a combination of basic principles and reasoning, conversations with PCORI and stakeholders, and an initial reading of the literature.
2. Each researcher independently reviewed about 20 applications and either used the current set of tentative categories or added new ones.
3. We then met as a team, compared the categories we each had defined, and merged them or modified them until we reached a consensus about a new set of tentative categories. Key criteria for the definition of the categories were (a) that each category should include a sufficiently large number of applications, (b) that categories should be limited in number, and (c) that categories should be easily interpretable.
4. We repeated this process, starting from item 2 above, until no new categories were created.

Several iterations of the procedure described above were needed to come to definitions of categories that represented the data well and were also easy to interpret. One of the dimensions, the function of the AI application, turned out to be quite complex and therefore we broke it down into both categories and subcategories so that we could report the results at a finer level of detail.

It is important to underscore that, although the categories corresponding to each dimension are as distinct as possible, an application can belong to multiple categories. For example, the dimension

"User" contains the categories "Health care professionals" and "Patients," which are clearly distinct, but there are applications that are used jointly by health care professionals and patients, and therefore we categorize these applications as belonging to the "composite" category "Health care professionals/patients."

From an operational point of view, the methodology described above allows us to attach to each AI application 8 labels, one for each dimension of the framework, where each label describes one or more categories. The resulting data set forms the basis for the analysis of chapter 4, where we dedicate a section to each of the dimensions of the framework. Because the definition of the framework was influenced by the set of AI applications that are included in this review, the framework itself is a result of this study, and therefore more details about the specific definition of dimensions and categories are presented in the narrative review of chapter 4 together with the analysis of the data. Here we report a summary of the framework in Table 2.2 and Table 2.3 to make the methodological discussion above more concrete and to allow for the reader to be able to refer to it in the next chapter.

Table 2.2. Dimensions and Categories Used to Characterize AI Applications in Health Care Within the Scope of Our Narrative Review (Aim 2)

Current status	
FDA cleared/approved	AI applications in health care that can be granted FDA clearance or approval via any of the regulatory pathways covering medical devices, as discussed in chapter 1
In use	AI applications that are currently in use as part of patient care. Many, but not all, of these applications are FDA cleared/approved.
Fielded for evaluation	AI applications that have been used as part of patient care as part of a field evaluation study
Health care function	
Patient evaluation	Applications that are used to evaluate an individual patient's health status, including for patient diagnosis, monitoring, prognosis evaluation, risk stratification, or assessment of therapeutic harms and benefits
Health recommendations	Applications that provide recommendations for treatment or health-related behaviors, usually based on patient evaluation
Treatment delivery	Applications that directly provide treatment to a patient, without human involvement, usually based on patient evaluation

Patient health conditions	
Cerebrovascular	Includes stroke as well as poststroke rehabilitation
Cardiovascular	Includes chronic conditions such as hypertension as well as acute events such as cardiac arrest, atrial fibrillation and heart failure
Kidney disease	Includes chronic kidney disease and issues surrounding kidney transplant
Diabetes	Includes diabetes prevention and self-management, and diabetes-related complications such as diabetic retinopathy
Mental health	All mental health–related conditions, including autism, attention deficit disorder, phobia, posttraumatic stress disorder, or suicidal ideation
Substance abuse	Includes all addiction and substance abuse, including alcohol and tobacco
Dementia	Includes Alzheimer's disease and other dementias, including loss of memory or cognitive ability due to brain disorders such as Parkinson's disease
Cancer	Includes all types of cancer
Respiratory	Includes all respiratory diseases and breathing disorders
General health	Includes general health of patients regardless of their specific disease or condition. Examples include general symptom checkers as well as applications that seek to reduce adverse events across the entire hospital inpatient population.
AI type	
ML: parametric	Artificial neural networks, deep learning networks, support vector machines, and Bayesian networks
ML: decision trees	Classification and regression tree and random forest
ML: regression	Linear, logit, multinomial logit, Cox proportional hazards
AI: conversational AI	The combination of methods and technologies used in chatbots (such as natural language processing, text mining, speech recognition, and different forms of ML)

Data inputs	
Nonimagery sensor	Includes accelerometers, electrocardiograms, electroencephalograms, pulse oximeters, photoplesmography, mass spectrometry, continuous glucose monitoring devices, and other sensors that can provide biomarker information
Audio	Includes all forms of audio data, including human voice and heart sounds
Patient-entered information	Structured information entered by a patient, for example, by answering an online questionnaire or filling in a patient-reported outcome measure
EHR structured variables	The structured clinical and administrative information found in an EHR system, not including free text or images
EHR free text	Natural language text found in the clinical notes of EHRs
Free text	Natural language text from sources other than EHRs, such as text from medical literature
Smartphone communication	The combination of natural language text, structured information (like checking boxes or menus), and emojis found on smartphones
Genetic	Includes all forms of genetic data obtained from biological samples
User	
Health care professional	Applications designed to be used by health care professionals. This includes applications that, while they might rely on patient-collected data, deliver analytical outputs solely to health care professionals.
Patient	Applications designed to be used by patients
Nonprofessional caregiver	Applications designed to be used by a nonprofessional engaged in patient care, such as a family member or friend
Other user groups	Applications that perform functions outside the scope of this review are often designed for use by other stakeholder groups, such as health system administrators, insurers, government regulators, or researchers.
Setting	
Hospital inpatient	Applications used as part of care for patients admitted to a hospital
Outpatient	Applications used as part of outpatient care during patient visits to a health facility
Home	Applications used in the patient's home or other nonmedical settings
Platform	
Smartphone	Mobile smartphone applications
Wearable	Applications embedded in devices worn by the patient
Other	Applications that can be designed for use on a wide variety of platforms, including computers, tablets, or medical equipment

Table 2.3. Dimensions and Categories Used to Characterize AI Applications in Health Care Not Included in the Scope of Our Narrative Review

Current status	
Preimplementation	AI applications in development as well as developed applications that have never been fielded for actual use in patient care
Health care function	
Epidemiology and public health	Applications that evaluate health at the population level, such as disease surveillance
Research and development	Applications that aid in biomedical research, including support to systematic reviews, disease-related biomarker identification, and drug discovery
Administrative tasks	Applications that aid in the performance of administrative tasks, such as chart documentation, clinical handovers, or assignment of billing codes
Resource management	Applications that aid in the management of financial, personnel, or other health system resources
Error, fraud, and neglect	Applications that detect error, fraud, or neglect in the delivery of health care
Education and training	Applications that aid in the training of health care professionals, for example, by generating simulated patient cases or by offering instructional feedback
Patient health conditions	
Out-of-scope conditions	A wide range of specific health conditions that are not covered by the 9 in-scope categories listed above in Table 2.2, including gastrointestinal disease, musculoskeletal conditions, epilepsy, and many others. Though important, they are outside the scope of this review.
Data inputs	
Imagery	Imagery and video from medical devices including ultrasound, CT, PET, MRI, X-ray, colonoscopy, etc; less commonly includes imagery or video from nonmedical cameras such as smartphones

Description of AI Type Categories

Within ML are a large number of methods, and the exact details of how they work are mainly irrelevant for the purposes of this report, except for the fact that some methods are more interpretable (less of a "black box") than others. We describe below the broad subcategories of ML that we have used to characterize the applications in this report, outlining their interpretability—a topic that will be discussed further at the end of this report.

Parametric models (such as artificial neural networks,[28] deep learning networks,[29] support vector machines[30]): What these methods have in common is that the variable to be predicted, or output (say, whether someone is at risk for developing diabetes), is related to the input variables (say, the observed characteristics of the individual) by a complex formula that contains many unknown coefficients. Past data, where both the input and output are observed, are used to estimate the value of these coefficients

and maximize the probability of giving a correct answer. The methods differ on the specific type of formula being used. For example, deep learning networks correspond to many formulas nested within one another, and the word *deep* refers to a high degree of nesting. They are commonly used in health partly because of their great success when applied to image data. Because the formulas are very complicated and often contain thousands of parameters, these methods are not easily interpretable.[31]

Decision trees (such as classification and regression tree [CART] models, and random forests[32,33]): These methods arrive at a solution by applying a sequence of "splitting" rules to the variables used to make the prediction. For example, "If age is larger than 15 and gender is male and glucose level is normal, predict this outcome." The simpler versions of these methods (such as CART) are easy to interpret because the rules can usually be visualized as a "tree" of decisions. More complex versions such as random forests tend to lose interpretability, although methods can be applied to make them more interpretable.[34]

Regression models (such as logistic regression, linear regression, proportional hazard regression models, and lasso[35]): Although regression models are very often used for the purpose of analyzing data and establishing associations in statistics, they can also be used as predictive models in ML. For example, technically speaking, logits and linear regression models are the simplest type of artificial neural networks. They fall in the set of parametric methods listed above, but we single them out here because they tend to be highly interpretable.

We focus mostly on ML methods, but the world of AI applications is complex and often a multitude of AI methods can be used in the same application. This is the case for chatbot[36] applications, which typically use a smartphone and text and/or speech input to interact with a user in a conversational manner. Chatbots rely on a wealth of methods, often a combination of ML and natural language processing. This ensemble of methods goes under the name of *conversational AI*, and we will use this term to denote the type of AI that is used by chatbots.

Defining and Measuring Relevance of AI Applications to Stakeholders

While in the long term all stakeholders may benefit from a successful application of AI in health, not all AI applications are equally relevant to stakeholders, as some applications may have a more direct impact on them than others. Whether an application is relevant to a stakeholder is a complex concept that is more difficult to define than the application characteristics presented in the framework of Table 2.2. Yet, it seems important to be able to make statements about which proportion of AI applications are relevant to specific stakeholders. Therefore, in this section we provide a simple definition of relevance that can be operationalized and reported on together with other dimensions of the framework of Table 2.2. We did not include this measure in Table 2.2 because we acknowledge that it has some limitations.

We include in the notion of relevance of an application both the potential to have an impact and the ability of the stakeholder to take an action, such as purchasing or using the application. For patients and individual providers, the notion of *relevant* overlaps with the notion of *user*, while for employers

and health plans it relates to the possibility of deploying the application to employees or plan members. Consider, for example, a smartphone application that helps a patient manage anxiety. This application is directly relevant to the patient, who benefits from it and has control over whether to purchase it, and it is this notion of relevance that we wish to capture.

A key stakeholder we have not mentioned so far is the government. We have not attempted to define what *relevant* means for the government because the government has myriad objectives and roles. For example, as a payer its objectives are aligned with those of health plans, but as a promoter of public health its objectives are aligned with those of patients. We also have assumed that the objectives of patient advocates are exactly aligned with those of patients. From an operational point of view, to assign the relevance of an application to a stakeholder, we proceed as follows:

Employers and health plans: If a commercial application currently on the market can be deployed by an employer or a health plan, this information is apparent from the main website. For vendors of different products, marketed to different stakeholders, it is common for the website to have separate sections with labels such as "For Health Plans" or "For Employers," or both.[37-39] Therefore, we label an application *relevant* to employers and health plans if there is such a clear indication on the website. This is a strict definition, and its key limitation is that it does not capture applications that are not commercial products, such as those that are currently being tested in a clinical trial. We opted for this stricter definition because its operationalization is unambiguous and it captures a well-defined concept.

Patients, providers, and caregivers: For patients (or caregivers), an application is *relevant* if they are users and they are directly affected by it (examples include applications that help patients manage their conditions or that monitor health status). For providers, broadly defined to include individual providers as well as health systems, *relevant* means that the application is a tool that can be used in their practice (such as diagnostic or risk assessment tools). We underscore that, as with employers and health plans, an application can be of interest to multiple stakeholders.

Framework for Mapping the Evidence on AI Applications in Clinical Care (Aim 3)

To map the evidence base surrounding the use of in-scope AI applications, we identified and characterized evaluation studies according to their publication type, study design, study population and sample size, and outcomes measured. We developed this framework via the same iterative process to identify dimensions and categories of interest (presented above).

Publication Type

Given the broad range of applications, and the fact that much of the pertinent information comes from the gray literature, we considered evidence from 5 types of publications, which are described below.

- **Peer-reviewed articles**: This includes evidence from peer-reviewed journal articles and conference proceedings.
- **Other academic documents**: This includes conference posters, abstracts, and oral presentations as well as academic preprints from online repositories such as *arXiv*, *medRχiv*, and *bioRχiv*.
- **FDA summary documents**: Every application with an FDA clearance has one or more FDA clearance documents, which contain varying degrees of evidence.
- **Clinical trial records**: Many evaluations of clinical AI applications are described in study records accessible on the ClinicalTrials.gov registry. These include descriptions of ongoing studies whose findings have not yet been published.
- **Other gray literature**: This include all other types of documents, including corporate white papers and websites as well as industry publications.

In this context, the term *evaluation study* refers to any publicly available information evaluating the effectiveness and safety of the application. Evaluations published in peer-reviewed academic journals are subject to a more consistently rigorous quality assurance process than are the other publication types—and may be more likely to report findings that cast an AI application in a negative light than would a document posted on a company website. Because a full evaluation of the quality of the evidence was not in the scope of this study, we use publication type as a proxy for evidence quality in the rest of the report.

Study Design

The evaluation of the effectiveness of AI applications has some complexities that are not usually encountered in other clinical evaluation studies, where the effectiveness of a health intervention is estimated. Many AI applications have a predictive component that provides information that is then acted on, and therefore both the accuracy of the prediction and the effects of acting on that prediction must be evaluated.

Performance tests are evaluation studies that examine the accuracy of an AI application, using real or simulated patient data without fielding these applications as an actual intervention that affects patient care. In contrast, field evaluation studies examine the effects of an AI application on patient care. These are like traditional field intervention studies, in which one observes the effect of the intervention on a group of subjects.

We distinguish between 3 types of field evaluation studies: *Randomized controlled trials* (RCTs) divide subjects randomly into multiple groups, including at least one group whose care includes the use

of an AI application and one group that receives some other form of care. *Pre-post evaluations* do not compare outcomes between randomly assigned groups of patients and instead assess outcomes by comparing how patient health differs after receiving care involving an AI application compared with health before receiving such care. *Implementation studies* evaluate the effects of an intervention on nonhealth outcomes such as user satisfaction or cost of care.

Study Population and Sample Size

For each evaluation study, we recorded the characteristics of the patient population and the study sample size. We distinguished between patient populations defined by specific settings, ages, or health conditions. We also captured the total number of patients included in the study sample and the number of health records used as inputs to the AI application, when provided. Occasionally a publication reports the number of records but does not report the number of distinct patients corresponding to those records. When possible, we have estimated the number of unique patients, but in some cases this variable is just missing. In addition, there are some studies in which the unit of observation is neither a patient nor a record but rather, for example, a vignette, and therefore the concept of number of patients does not apply. In these cases, we have labeled the observations as *not applicable*.

Outcomes Measured

Given the broad scope of this review, the range of outcomes measured in the studies is vast. To keep the analysis manageable, we grouped outcome measures into 6 categories, listed below ordered by their prevalence in our data:

- **Accuracy:** Many studies fall in the category of *performance test* and therefore measure some notion of accuracy, such as the area under the curve (AUC), sensitivity, specificity, and positive/negative predictive values; however, accuracy is also reported in other types of studies.
- **Health indicator**: This is a large category that includes self-reported measures of health status (such as patient-reported outcome measures), any type of biomarker (such as HbA_{1c}), and the diagnosed presence of a health condition (such as atrial fibrillation).
- **Resource utilization:** This category includes any type of cost, independent of payer, and standard utilization measures such as number of visits, hospital admission, and hospital readmissions.
- **Appropriate treatment:** Many applications aim to ensure that patients receive the right treatment and care, and therefore study outcomes often are measured in these terms. Examples include applications that provide correct drug dosage,[40] that lead to an adequate level of palliative care,[41] and that improve adherence[42] and screening for depression.[43]

- **User satisfaction:** In some studies, one of the outcomes measured is the user's level of satisfaction with the application itself.[44-46] This dimension is important because it is a determinant of adoption. In addition, for applications that are patient oriented, usability is one of the factors that determines how long the patient will make use of it, which in turn determines whether health benefits will be realized and sustained.
- **Time to treatment:** In some cases, studies report time to treatment because one of the goals of an evaluated application is to provide more timely treatment. Examples include the provision of emergency services for patients with cardiac arrest[47] and the provision of palliative care.[48].

Later in the document we will define the term *health outcomes* as the 2 outcome measures that most directly relate to health: health indicator and appropriate treatment.

3. Broad Overview of Artificial Intelligence in Health Care

This chapter addresses aim 1 of our study: to provide a broad overview of the full range of AI applications with potential for use in health care. Though the remainder of this report (chapters 4 and 5) focuses on ML applications that assist with clinical functions as part of delivering individual patient care, a number of other functions could be performed by AI applications in health care, and those functions will be the main focus of this chapter.

Administrative Tasks

There is great potential for AI to improve the efficiency of the health care system by automating or aiding tasks performed by health care professionals or administrators and that do not necessarily involve patient care. Some of these tasks are time consuming[49] and critical for the provision of high-quality care and for appropriate billing. Far from replacing humans, AI applications with speech recognition and natural language processing capabilities can make clinical documentation more efficient,[50] facilitate medical chart reviewing[51] or the handover of clinical information,[52] and automate the assignment of billing codes.[53] Systems of this type are expected to allow health care professionals to spend more time in productive activities, enhancing both the productivity of the workplace as well as the quality of care, and at the same time providing an element of consistency and reducing random variation.

Education and Training

AI can affect patient care in indirect ways, one of which is through its applications to medical education and training.[54,55] For example, ML can analyze data acquired during surgery, such as the motion of the hand or the eye, and correctly predict the skills of the surgeon.[56] This capability opens the path to replacing the evaluation of surgeon skills currently performed by more senior surgeons, which is subject to variation, with a more consistent and objective evaluation. Similarly, it has been proposed that ML could "listen" to recordings of patient/clinician communications and automatically assess the quality of the communication skills of the clinician.[57]

Detection of Error, Fraud, and Neglect

ML is often applied to estimate the probability of future harmful events, so that they can be prevented. Harmful events are not limited to the clinical setting but can also include financial harm as well as violence or neglect. Financial harm can be unintentional, caused by error, or intentional, caused by fraud. Because intention is difficult to observe, error and fraud are usually lumped together in the

literature, and often authors simply refer to fraud detection. Detection of fraud and error in the health domain is a very active research area[14,58-64] that uses both administrative claims data and EHRs to identify potential for harm. ML has an advantage over humans in performing these tasks because it can process millions of claims very quickly. However, humans have an advantage over ML systems because they have valuable prior knowledge about the context in which error or fraud take place and about criminal patterns and intents, none of which are represented in the data. Therefore, hybrid systems that attempt to combine the best of what humans and machines have to offer are likely to become increasingly common.[65,66]

Fraud detection is an area for which *unsupervised* ML, as described in chapter 2, is commonly used.[13,67,68] This is because, often, researchers have disposal only a data set without a clear indication of which cases are fraudulent, making it impossible to apply supervised ML and predictive models. Unsupervised ML allows researchers to analyze the data and find outliers or unusual complex patterns that, by virtue of being different from the norm, are *potentially* fraudulent. Applications of this type simply point researchers and auditors to the suspicious records for further investigation.

The methodology used to detect fraud and error can also be used in a different area, one that lies at the intersection of health and social service provision: the identification of individuals likely to exhibit harmful behavior or to be subjected to it, and the identification of unusual patterns that may point to neglect, rather than fraud. Examples include the identification of potentially violent individuals among psychiatric patients,[69,70] the detection of elder abuse and neglect,[71,72] and the prediction of the risk of harm to children monitored by Child Protective Services.[73]

Epidemiology and Public Health

There are many applications of AI that affect people's health outside of the clinical care context and whose function fits in the framework of public health. For example, because one of the tasks that ML performs well is the detection of anomalous events, a natural area of application is the design of health surveillance systems,[74-76] including those that use social media as input.[77] ML has also been applied to public safety. For example, one very active area of research concerns safety for workers, for which there are applications to provide continuous monitoring[78] at construction sites, to develop improved safety indicators,[79] and to predict occupational accidents.[80] In the transportation sector, many applications have been developed to improve road safety, including some that detect driver stress levels[78] or drowsiness[81] and predict crash severity.[82,83] ML has also been applied in the area of environmental health, for which it has been used to study air pollution,[84-87] as well as in toxicology research.[88,89]

Research and Development

AI has also been used to improve the way research is performed. Recent advances in natural language processing (NLP) and text mining allow researchers to use ML to make systematic reviews more efficient.[90,91] For example the Cochrane Transform Project's Evidence Pipeline[92] uses AI to screen thousands of documents and identify which ones are most likely to be relevant for inclusion in Cochrane Reviews. In addition, ML and NLP are used to analyze large bodies of biomedical literature for evidence on issues such as drug-drug interactions[93] or other biomedical relations.[94]

Deep learning artificial neural networks have been particularly successful in aiding the task of drug design,[95-100] making predictions concerning drug binding,[101] and extending the drug design framework from small molecules to antibodies.[102] One recent review concluded that most new approaches for identifying drug targets fall within the category of ML methods.[103-105]

Imaging

Medical imaging is the area in which AI has already made a significant impact—and where many applications are currently in use. One way to quantify this statement is to consult the website on FDA-cleared medical imaging AI algorithms maintained by the Data Science Institute of the American College of Radiology.[106] As of October 2020, the list contains 80 entries spanning a wide range of applications. The list is not necessarily complete, and additional applications can be found, for example, in the Medical Futurist online database of FDA-cleared AI applications in health, that has a broader scope.[107]

In the medical imaging sector, AI applications support professionals at all stages of their workflow. Some applications target the process of image acquisition, by making it possible to maintain high image quality while reducing exposure to X-rays[108,109] and radioactive tracers,[110] or reducing the time spent in the scanner.[111] AI applications can also guide professionals during the image acquisition phase, for example, when performing an ultrasound, to enhance image quality.[112]

Other applications play a role in the process of triaging and risk stratification, interpreting scans as they are acquired so that the workflow of radiologists can be prioritized,[113-115] or alerting health professionals to patients at high risk of cardiovascular, lung, bone, and other diseases.[116] A large number of applications aim to provide real-time support to radiologists by identifying suspicious findings[117] or fractures,[118,119] providing volumetric quantification of segmented structures,[120] and automatically generating reports.[113,121,122]

The adoption of AI applications in the medical imaging sector is expanding rapidly and generating great interest[123-125] as well as some anxieties among health care professionals.[126-128] To put the state of the art in perspective, it is important to note that all these applications implement some form of ML algorithm. Therefore, the human task they mimic is only the ability to classify patterns similar to those that exist in the data. In most cases these algorithms do not mimic any other type of intelligence: they do not reason about other types of information about the patient, nor do they take into account

anatomical or physiological knowledge. As a consequence, these applications can compete with only one of the skills of the imaging professionals—that is, the ability to classify visual patterns such as those encountered in radiology, dermatology, or pathology—and are unable to process all the additional information that goes into formulating a diagnosis or evaluation. In addition, these algorithms tend to solve very specific problems. For example, algorithms for fracture detection do not detect all types of fractures.[129,130] Rather, there may be an algorithm for vertebrae and a completely different one for hips. As a result, AI applications in imaging are still spanning only a fraction of all possible ailments.

Although this observation regarding the limitation of ML is particularly relevant in the imaging field, given the greater adoption rate, it applies to all other applications described later in this report, and it is useful for maintaining a perspective on the possibilities of ML in health.

4. Narrative Review of Machine Learning Applications in Clinical Care

This chapter addresses aim 2 of our study by presenting the findings of our narrative review of AI applications in current or potential near-future use in clinical care.

We identified a total of 109 non-imaging-based ML applications in current or potential near-future use in clinical care to address general patient health or the 9 health conditions within the scope of our review. Full details on each of these applications can be found in appendix A. These 109 in-scope applications represent all that we were able to identify in our search of the academic literature, FDA approval documents, clinical trial records, and gray literature, with one exception. As there are a very large number of AI-based symptom checker chatbots available for use, we included only the 7 that appeared at least twice in the reviewed documents.

In this chapter, we focus on the description of these 109 applications, which are categorized according to the framework presented in Table 2.2. Each of the dimensions defined in Table 2.2 has a dedicated section below, titled with the dimension name. The chapter begins by discussing the current development, regulatory, and adoption status of these applications, followed by an examination of their functions, targeted health conditions, ML methods, users, settings, and platforms. We discuss evaluation studies that provide evidence on these applications' effectiveness, accuracy, and safety in the next chapter.

Current Status

Of the 109 applications we identified in our review, 22 (20%) were cleared for use by the FDA, 44 (40%) were in current use without FDA clearance, and 43 (40%) were in development for potential near-future use.

All 22 of the FDA-cleared applications were designed for use by health care professionals or by a patient as directed by health care professionals. These include most arrythmia detection applications, some patient deterioration monitors, and some diabetes self-management apps.

A small number of applications (2 out of 22) were approved through the De Novo classification pathway. One application, CLEWICU, which identifies intensive care unit (ICU) patients at higher risk of respiratory failure, was cleared under an FDA emergency use authorization for the duration of the coronavirus crisis.[131]

The great majority of applications (19 out of 22) were cleared through the 510(k) premarket notification pathway.[20] Of these, only the Advisor Pro/MD-Logic and the Ahead applications received their 510(k) clearance based on "substantial equivalence" to an earlier ML-based device that went through the De Novo application process.[132,133] The remaining 17 applications trace their 510(k)

"substantial equivalence" lineage back to one or more non-machine-learning-based applications approved earlier by the FDA.

Classification information was unavailable for 2 applications: eMurmur ID, due to incomplete FDA documentation, and CLEWICU, due to its emergency authorization. All 20 other in-scope applications received Class II designations, the most common FDA designation, which covers devices that pose a moderate risk to patients. No applications were classified as Class I devices, though this is not necessarily surprising as nearly all lower-risk devices that would fall in this category are exempt from FDA classification entirely.

No approved applications were classified as falling in the highest risk category of Class III devices, which are subject to the full device premarket approval process and include high-risk devices such as automated external defibrillators and implantable devices such as pacemakers or artificial pancreas systems.[17,134]

FDA-approved applications are briefly summarized in Table 4.1. Full details on these applications can be found in appendix A.

Table 4.1. Applications Approved by the FDA

Application	Function	Health condition	User
Advisor Pro/MD-Logic	Health recommendations	Diabetes	Health care professionals
Ahead	Patient evaluation	General	Health care professionals
AI-ECG Tracker	Patient evaluation	Cardiovascular	Health care professionals
Biovitals Analytics Engine/Biovitals HF	Patient evaluation	General	Health care professionals
BlueStar	Health recommendations	Diabetes	Patient
Cardiologs Platform	Patient evaluation	Cardiovascular	Health care professionals
CLEWICU	Patient evaluation	Respiratory	Health care professionals
Current Platform	Patient evaluation	General	Health care professionals
Eko Analysis Software	Patient evaluation	Cardiovascular	Health care professionals
eMurmur AI	Patient evaluation	Cardiovascular	Health care professionals
EnsoSleep	Patient evaluation	Respiratory	Health care professionals
FibriCheck	Patient evaluation	Cardiovascular	Health care professionals
KardiaAI/Kardia Mobile	Patient evaluation	Cardiovascular	Patient
Loop System	Patient evaluation	Cardiovascular	Health care professionals
MATRx Plus	Treatment delivery	Respiratory	Patient

Application	Function	Health condition	User
Pathwork Tissue of Origin Test	Patient evaluation	Cancer	Health care professionals
PhysIQ Personalized Physiology Engine	Patient evaluation	Cardiovascular	Health care professionals
Rhythm Express RX-1 MDSP Technology	Patient evaluation	Cardiovascular	Health care professionals
RhythmAnalytics	Patient evaluation	Cardiovascular	Health care professionals
VITEK MS	Patient evaluation	General	Health care professionals
WAVE Clinical Platform: Visensia, the Safety Index	Patient evaluation	General	Health care professionals
Zio AT ECG Monitoring System	Patient evaluation	Cardiovascular	Health care professionals

A total of 45 applications are available for use without having received FDA approval. Five of these applications—Cardiomatics, Corti, Karantis360, Nectarine Health, and ResApp—are solely in use outside of the United States, primarily in Europe, though at least one of them, ResApp, is applying for FDA clearance. The remaining 40 applications are in use within the United States without having undergone the FDA approval process. This category includes all symptom checkers, all mental health apps, all online risk calculators, all genetic analysis apps, some diabetes self-management apps, and some patient deterioration monitors.

Applications in use without FDA approval are listed below in Table 4.2. Full details on these applications can be found in appendix A.

Table 4.2. Applications in Use Without FDA Approval

Application	Function	Health conditions	User
Ada	Health recommendations	General	Patient
Apple Watch 4: Fall Detection App	Treatment delivery	General	Patient
Babylon Health	Health recommendations	General	Patient
Buoy Health	Patient evaluation	General	Patient
Cardiomatics	Patient evaluation	Cardiovascular	Health care professionals
Corti	Patient evaluation	Cardiovascular	Health care professionals
DayTwo	Health recommendations	Diabetes	Patient

Application	Function	Health conditions	User
eCART (electronic cardiac arrest triage)	Patient evaluation	Cardiovascular	Health care professionals
EPIC Deterioration Index	Patient evaluation	Cerebrovascular	Health care professionals
Ginger.io	Patient evaluation	Mental health	Health care professionals
HealthTap AI	Health recommendations	General	Patient
Heart Failure Risk Calculator	Patient evaluation	Cardiovascular	Health care professionals
Ibis	Patient evaluation	General	Patient
Intermountain Healthcare Readmission/Mortality Prediction	Patient evaluation	Cardiovascular	Health care professionals
JVION Machine	Patient evaluation	General	Health care professionals
K Health	Patient evaluation	General	Patient
Karantis360	Patient evaluation	Alzheimer/dementia; general	Health care professionals
KDPI-EPTS Survival Benefit Estimator	Patient evaluation	Kidney disease	Health care professionals
Lark Diabetes Care	Health recommendations	Diabetes	Patient
Lark DPP	Health recommendations	Diabetes	Patient
Lark for Hypertension	Health recommendations	Cardiovascular	Patient
Nectarine Health	Patient evaluation	Alzheimer/dementia; general	Health care professionals
Omada Health	Health recommendations	Diabetes	Patient
One Drop	Health recommendations	Diabetes	Patient
Owlytics	Patient evaluation	Alzheimer/dementia; general	Health care professionals
Preventice BeatLogic Platform	Patient evaluation	Cardiovascular	Health care professionals
Qventus	Patient evaluation	General	Health care professionals
rapid Whole Genome Sequencing	Patient evaluation	General	Health care professionals

Application	Function	Health conditions	User
Rare Disease Auxiliary Diagnosis System	Patient evaluation	General	Health care professionals
REACH VET	Patient evaluation	Mental health	Health care professionals
ResApp	Patient evaluation	Respiratory	Health care professionals
Seattle Heart Failure Model	Patient evaluation	Cardiovascular	Health care professionals
Sepsis Watch	Patient evaluation	General	Health care professionals
SOPHiA GENETICS	Patient evaluation	Cancer	Health care professionals
Steth IO Software	Patient evaluation	Cardiovascular	Health care professionals
Sugar.IQ	Patient evaluation	Diabetes	Patient
Symptomate	Patient evaluation	General	Patient
Targeted Real-time Early Warning Score	Patient evaluation	General	Health care professionals
Tempus Oncology Testing	Patient evaluation	Cancer	Health care professionals
Tess	Treatment delivery	Mental health	Patient
Virta	Health recommendations	Diabetes	Patient
Watson for Oncology and Genomics	Health recommendations	Cancer	Health care professionals
Woebot	Treatment delivery	Mental health	Patient
Wysa	Treatment delivery	Mental health	Patient
Your.MD	Patient evaluation	General	Patient

We identified an additional 42 applications as potential candidates for adoption within the next 5 years. To be included in this category, these applications needed to have been either (1) actually implemented in patient care as part of a field evaluation study or (2) described as submitting for FDA approval within the next year. A total of 41 applications met the first criterion, and 1 application, HeartHero, met the second criterion.

We recognize that these criteria will undoubtably miss applications that will in fact be adopted in clinical practice in the next 5 years. Nevertheless, the list below represents a broad range of applications that appear closer to actual adoption—as distinguished from the hundreds of algorithms that have been conceived, developed, and tested on clinical data over the past 10 years that have not ended up in actual clinical use.

Though this list includes many applications similar to those already in use, it also includes 3 automated treatment delivery devices that would likely require FDA premarket approval as higher risk Class III devices: HeartHero Automated External Defibrillator (AED), Beta Bionic's Bionic Pancreas, and Medtronic's Minimed 780G artificial pancreas.

Applications that may be adopted for use within the next 5 years are listed below in Table 4.3. Full details on these applications can be found in appendix A.

Table 4.3. Applications in Potential Near-Future Use

Application	Function	Health condition	User
Advanced Electronic Safety of Prescriptions Model	Patient evaluation	General	Health care professionals
AI-Assisted Insulin Titration System	Health recommendations	Diabetes	Patient
AIM@BP	Health recommendations	Cardiovascular	Patient
Anemia Control Model	Health recommendations	Kidney disease	Health care professionals
Anticoagulation Management Service	Patient evaluation	Cardiovascular	Health care professionals
ASSIST	Treatment delivery	Mental health	Patient
Assisted Rehabilitation Care	Health recommendations	Cerebrovascular	Patient
Atrial Fibrillation Risk Prediction	Patient evaluation	Cardiovascular	Health care professionals
Bionic Pancreas	Treatment delivery	Diabetes	Patient
BQ Device	Treatment delivery	Cerebrovascular	Health care professionals
BrightArm Compact	Health recommendations	Cerebrovascular	Patient
Companion	Patient evaluation	Mental health	Patient
Control Tower	Health recommendations	General	Health care professionals

Application	Function	Health condition	User
COVID-19 Alert System	Health recommendations	Respiratory	Patient
Dashboard for Diabetes Care	Health recommendations	Diabetes	Health care professionals
Diabetes Prevention App	Health recommendations	Diabetes	Patient
Diagnostic AI for Pediatric Diseases	Patient evaluation	General	Health care professionals
ECG AI-Guided Screening	Patient evaluation	Cardiovascular	Health care professionals
FIND FH Algorithm	Patient evaluation	Cardiovascular	Health care professionals
Heart Failure Medication Reminder App	Health recommendations	Cardiovascular	Patient
Heart Failure Treatment Gap Model	Health recommendations	Cardiovascular	Health care professionals
HeartHero AED	Treatment delivery	Cardiovascular	Caregiver
Hypotension Prediction (HYPE)	Patient evaluation	General	Health care professionals
iDEFCO	Patient evaluation	Cancer	Health care professionals
Jumpstart	Patient evaluation	General	Health care professionals
Kelahealth	Patient evaluation	General	Health care professionals
Medical Early Warning Score ++	Patient evaluation	General	Health care professionals
Minimed 780G / MD-Logic Artificial Pancreas	Treatment delivery	Diabetes	Patient
Neuro Motor Index	Patient evaluation	Dementia	Health care professionals
Optima 4 Blood Pressure	Health recommendations	Cardiovascular	Health care professionals
PD_Manager	Patient evaluation	Dementia	Health care professionals
Pediatric Symptom Checker	Health recommendations	General	Patient
Radiation Therapy Risk Algorithm	Patient evaluation	Cancer	Health care professionals
Rose Platform	Health recommendations	Mental health	Patient
Sepsis Prediction Algorithm	Patient evaluation	General	Health care professionals
Short Arm Human Centrifuge Rehab	Health recommendations	Cerebrovascular	Health care professionals
Sinedie	Health recommendations	Diabetes	Patient
Smart Angel	Patient evaluation	General	Health care professionals
Smoking Cessation App	Health recommendations	Substance abuse	Patient

Application	Function	Health condition	User
t2.coach	Health recommendations	Diabetes	Patient
Warfarin Dosage App	Health recommendations	Cerebrovascular	Patient
Wellthy Diabetes	Health recommendations	Diabetes	Patient

Health Care Function

Many ML applications in current clinical use are designed to assist with patient evaluation, with some applications additionally providing health recommendations or, in a few cases, directly providing treatment to patients. We describe them in detail below.

Patient Evaluation

The most common function we found in our analysis was *patient evaluation* (61% of applications). Applications with this function offer information concerning an individual patient's health status and are used either by a health care professional (79% of applications) or by the patient (21% of applications).

Different types of applications fall within this broad category. Many are built for the purpose of supporting health care professionals in the formulation of a *diagnosis* and assess the presence of a health condition (or a pathogen in a biospecimen[135]), often on the basis of a combination of EHRs, biomarkers, and genetic information. Here we take a broad view and include in this category applications that have a clear diagnostic purpose and applications that monitor a patient over a period of time and send results or alerts to a clinician, who may then use this information to diagnose a condition.

About half of the applications in this category deal with the diagnosis of heart conditions using data from electrocardiograms (ECGs),[136,137] heart sounds, or photoplethysmograms.[138] These applications primarily focus on diagnosis of arrythmia,[139-144] atrial fibrillation (AF),[143,145,146] and heart murmurs.[145,147,148] Other applications that take as input a signal or a sound include those that use electroencephalogram (EEG) data to detect brain injury[149] or to produce sleep scores[150] and applications that use cough and breath sounds to diagnose or assess the severity of respiratory diseases.[151] The applications mentioned above are oriented toward detecting fairly common conditions for which good diagnostic tools may already exist; however, there are also ML applications currently in use for the detection of rare diseases.[152,153] The latter is actually an area of applications where researchers expect ML to perform particularly well, as it requires processing very large amounts of data to find meaningful patterns.[154] Other diagnostic applications include the analysis of genetic material for the identification of tumor type,[155,156] the early detection of Alzheimer's disease,[157] the diagnosis of a broad range of pediatric diseases,[158] and the identification of microorganisms to aid in the diagnosis of bacterial and yeast infections.[135]

Sometimes AI applications in the patient evaluation category stop short of providing a diagnosis and limit themselves to providing a *symptom assessment*. Applications in this category, often referred to as "symptom checkers," are mostly used by patients to obtain the likelihood of having a particular health condition given a set of symptoms, to inform their health care–seeking behavior. Well-known applications in this category are Symptomate,[159,160] Your.MD,[161,162] Ada,[163] Buoy,[164] HealthTap,[165] K Health,[166] and Babylon.[167]

Another common function of ML applications is *risk stratification*, in which information about a patient is used to estimate the probability that patient will experience a certain event, so that an appropriate personalized course of action can be devised. In the inpatient setting, common risks are represented by sepsis,[168-171] heart failure,[172,173] hospital readmission within 30 days,[172,174] postoperative complications,[175] hospital-acquired conditions,[174] and health care–associated infections,[174] as well as respiratory failure and hemodynamic instability for COVID-19 patients. Though these applications predict specific types of events, many within this category also more generally predict the risk of patient deterioration or the risk of exacerbation of current conditions. Some are used exclusively in a hospital setting, such as the WAVE platform,[176] the Epic Deterioration Index,[177] eCART,[171] Medical Early Warning Score ++,[178] and CLEWICU.[179] Other applications monitor patients at home through wearable or other sensor devices.[180-182]

Because prediction is a task that ML performs particularly well, it is not surprising that it is found in many applications across different settings and different types of inputs. For example, advances in speech recognition and NLP have made it possible to apply ML algorithms to phone calls to emergency dispatchers and quickly recognize that the callers may be experiencing cardiac arrest. EHRs are an excellent candidate to provide input to predictive models, and in addition to some of the applications presented above, relevant ones include REACH VET,[183] a platform that identifies veterans at risk for suicide, hospitalization, illness, or other adverse outcomes, and the IQ-MATCH platform, which identifies primary care patients with AF who are not on anticoagulation therapy and at high risk of stroke.[184]

It is important to note that, though the word *risk* tends to denote a negative event, in the context of risk stratification it also refers to positive events, such as benefiting from a treatment or a service. Examples of such treatments/services include early palliative care,[48] goals of care discussions,[185] and social worker engagement.[186]

Health Recommendations

Applications whose function falls in the *patient evaluation* category described above provide information about the health status of the patient, such as a diagnosis or a level of risk, but do not provide a recommendation to the user concerning what to do with that information. That said, a fairly large group of applications (30% of all applications) goes one step further and provides the user with a list of actions that can be taken to address an underlying health issue. We group these applications under the label of *health recommendations*. Within this broad category, about half of the applications relate to *patient empowerment*, in that they aid patients in managing or preventing a chronic condition. The most commonly targeted condition is diabetes,[37,39,187-194] but cardiovascular conditions are also targeted,[42,195,196] as well as mental health[197] and smoking.[198] Some applications are also able to deal with a broad range of chronic diseases.[199] The remaining half of the applications includes those that aim to recommend *optimal treatment options* (33%), which are almost evenly divided between those recommending a particular medication dosage[200-203] (such as of insulin medication[204]) and those recommending a whole treatment plan.[205-208] The remaining 17% of the applications are dedicated to providing *personalized physical therapy recommendations.*[209-211]

Treatment Delivery

The group of AI applications discussed above provides the user with recommendations but do not act on them directly. A smaller group of applications (9% of the total) directly delivers some form of clinical treatment to patients, and 4 of the 10 applications in this category relate to mental health. One is a wearable that delivers mindfulness meditation training to caregivers in distress,[212] while the other 3 are chatbots that deliver mental health talk therapy.[38,213,214] The therapy model is slightly different across these 3 apps: Wysa[213] and Tess[38] also provide the option of human communication, while Woebot[214] is a smartphone application focused on cognitive behavioral therapy.

Other applications provide different forms of treatment. The artificial pancreas applications provided by Medtronic[215] and Beta Bionics[216] are insulin pumps that use AI to decide the timing and dosage of insulin injections. This is also the case for 2 applications that automatically call 911 in an emergency. One detects whether the user has fallen and remains immobile,[217] and the other is a "smart" automated external defibrillator that administers a shock only after an ML algorithm detects that the patient is experiencing cardiac arrest.[218] Other applications in this category include an oral sleep appliance that adapts itself while in use[219] and a device that delivers personalized low-frequency and low-intensity electromagnetic fields therapy to patients with recent stroke.[220]

Patient Health Conditions

Most of the applications included are specific to patients with 1 of the 9 conditions or groups of conditions shown in Table 2.2; however, there is also a large number of applications (27%) that target broad categories of patients regardless of their specific health condition, aiming to benefit individuals who use health care services. Typical examples are general symptom checkers as well as inpatient monitoring applications that seek to identify patients at risk of sepsis, ICU transfer, or other adverse events. Many patients affected by one of the priority conditions are likely to display symptoms of some sort and are also likely to use hospital, ICU, and other health care services. Therefore, they would be directly affected by the adoption of technologies targeting health care services users. Hence, we have included these applications in this report, and to distinguish them from those targeting specific conditions, we have assigned them a *general* health condition label.

In Table 4.4 we show the distribution of health conditions targeted by the applications. Given the very small number of applications with dual health conditions, we ignore this issue in the rest of the report and assign a unique health condition to each application, based on our best understanding of its intent.

Table 4.4. Number of Applications Targeting Specific Health Conditions

Health condition	Number of applications	% of applications
Cardiovascular	32	29.4
General	29	26.6
Diabetes	17	15.6
Mental health	9	8.3
Cerebrovascular	7	6.4
Respiratory	7	6.4
Cancer	6	5.5
Alzheimer/dementia	5	4.6
Kidney disease	2	1.8
Substance abuse	1	0.9

Note: A few applications target more than one condition, such as diabetes and cardiovascular disease, and therefore some applications are counted more than once (which explains why the number of applications sums to 115 instead of 109).

AI Type

We show the distribution of applications by the type of AI used in Table 4.5.

Table 4.5. Distribution of AI Type

AI type	Number of applications	% of applications
Unspecified	67	61.5
Parametric	22	20.2
Conversational AI	10	9.2
Decision trees	6	5.5
Regression	4	3.7

Strikingly, in a majority (61.5%) of applications it is unclear which ML method was used. Although knowing the type of ML is not always necessary for the evaluation of the effectiveness or safety of an application, this matter points to a lack of transparency in the documentation of these applications. In fact, in a few cases, many additional documents and website had to be consulted to establish whether an application used any type of ML at all.

For applications using a known type of ML, the distribution of methods are as expected. The largest group, parametric methods, is mostly composed of artificial neural network architectures, often of the DL type. The next group, conversational AI, corresponds to chatbot applications. We do find this type of application more commonly used in the context of mental health and substance abuse, which is expected since they are highly interactive and often provide some form of messaging-based therapy.[38,214]

Data Inputs

For each of the applications that we reviewed, we captured the type of data input used. The list of data types and their definitions is presented in Table 2.2. Some applications use more than one data type as input; to give a full picture, we provide in Table 4.6 the distribution of all the combinations of input found in the data.

Table 4.6. Distribution of Data Input Types

Data input	Number of applications	% of applications
Nonimagery sensor	41	37.6
EHR	19	17.4
Smartphone communication	14	12.8
Patient-entered information	9	8.3
Smartphone communication; nonimagery sensor	6	5.5
Audio	3	2.8
Genetic	3	2.8
Nonimagery sensor; audio	3	2.8
Nonimagery sensor; EHR	3	2.8
Smartphone communication; EHR	2	1.8
Text: EHR; EHR	2	1.8
Patient-entered information; nonimagery sensor	1	0.9
Text: EHR	1	0.9
Text: EHR; genetic	1	0.9
Text; genetic	1	0.9

The most common data input type was data obtained by a passive nonimagery sensor, used both by patients and health care professionals. For health care professionals, applications that use ECG data as input are particularly prevalent given that cardiovascular conditions are targeted by 29.4% of all applications. For patients, common sensors include continuous glucose monitors and sensors that are commonly found on smartphones and wearables, such as those that measure heart rate and other vital signs.

EHR and smartphone communication are a distant second and third in the list of data inputs. We underscore that by EHR we mean the component of the EHR that does not include free text, such as clinical notes. That component is denoted by "Text: EHR" in the table to highlight the fact that it is used infrequently, despite the fact that existing literature shows that it has the potential to greatly enhance the performance of ML algorithms.[221,222] A factor that contributes to explaining the low usage of clinical notes is the well-known difficulty with deidentifying clinical notes.[223-225] They are likely to contain names of patients and family members, references to places or events, and other information that can easily identify a patient in the data. This difficulty implies that researchers and developers are much less likely to have access to this type of data, which currently seems to be mostly untapped.

User

AI applications target both patients and health care professionals as users but not in equal proportions. As Table 4.7 shows, most of the applications (55%) are designed for health care professionals only, while 33% are designed for patients only. A smaller proportion (11%) is designed to be used by both patients and health care professionals, and only 1% is meant for caregivers; however, these proportions change as a function of the health condition considered. For applications that deal with cardiovascular disease, health care professionals represent 71% of the users, but for diabetes-related applications, 76% of the users are patients. This breakdown reflects the fact that many of the diabetes applications are designed to empower consumers and help them better manage their condition, while for cardiovascular disease many applications provide clinical decision support and therefore are oriented toward health care professionals.

Interestingly, the set of applications for which the intended users are both the patient and the health care professional is relatively small (11% of all applications). These applications tend to monitor patients and provide alerts or other assessments of patient health status to both patients and their health care professionals.[37,138,197] Only one application, the HeartHero AED, is designed for use by a nonprofessional caregiver.[218]

Table 4.7. Distribution of User Types

User	Number of applications	% of applications
Health care professionals	60	55
Patients	36	33
Patients; health care professionals	12	11
Caregiver	1	1

Setting

We use the variable *setting* to capture where the application is deployed. We are particularly interested in understanding how many of the applications are deployed in a home setting. We notice that, although setting and user often overlap, and an application deployed at home is likely to be used by the patient, this is not always the case. For example, there are smart monitoring devices for conditions such as heart disease and mental health that gather information from the patient at home and then send it to a clinician for analysis.[138,140,182,226-228]

We find that half of all applications are used in a home setting and tend to be smartphone or wearable based. The next largest group is applications used in an outpatient setting, which captures 34% of the applications, followed by the inpatient setting (14%), which typically consists of applications such as those for hospital readmission,[172] clinical deterioration,[176] or early sepsis detection.[229,230]

One expects to find a correlation between the setting and the health condition being targeted. Indeed, we find that applications targeting diabetes or mental health are almost exclusively home based, while only 40 % of applications for cardiovascular disease belong to this category, with the remaining 60% mostly concentrated in the outpatient setting (eg, automated analysis of ECG). We show the full distribution of setting and health condition in Table 4.8.

Table 4.8. Number of Applications by Health Condition and Setting

Health condition	Home	Inpatient	Other	Outpatient
Cancer	1	0	0	5
Cardiovascular	13	3	0	15
Cerebrovascular	1	1	0	3
Dementia	3	0	1	1
Diabetes	15	0	1	1
General	12	10	0	7
Kidney disease	0	0	0	2
Mental health	7	0	0	1
Respiratory	2	1	0	2
Substance abuse	1	0	0	0
Total number of applications	55	15	2	37
% of applications	50.5	13.8	1.8	33.9

Platform

The applications we consider in this report run on a variety of platforms. About 30% are hosted on smartphones, such as chatbots for the management of mental health[38,213,214] and other chronic conditions,[37,187,195] text messaging applications that act as medication reminders,[42,196] and applications that use the phone as a device to detect heart problems.[138,143,148]

Intelligent wearable devices are becoming increasingly common[231-233] and constitute 15% of all applications described in this report. They include, for example, the artificial pancreas,[215,216] fall-detection devices,[217,234] and devices that assess the risk of patient deterioration.[181,182] Within this group, approximately half of the applications consist of devices with some built-in AI, such as the artificial pancreas[215,216] or an oral appliance to treat sleep apnea.[219] For the remaining half of wearable applications, the role of the device is mainly to collect and transmit the data, which are then analyzed with AI-powered software on another computing platform, often in real time.

The remaining 55 % of applications tend to run on institutional computer systems, with a small group of applications that are web-based tools, such as some symptom checkers[159] or risk

calculators.[173,235] A few of these applications are linked to a specific device, such as a mass spectrometer[135] or an automated electronic defibrillator; however, the vast majority of these software applications can process input data acquired from any device that collects data of a certain type. In many cases, this allows health care professionals to run AI-powered applications using computing platforms and sensor devices they already have.

As expected, we find that most of the applications based on a wearable device or a smartphone are characterized by the home setting. In addition, we find that wearables are predominantly used in applications that provide patient evaluations (because they act as monitoring devices), while smartphones are predominantly used in applications whose function is to provide health recommendations.

Stakeholder Relevance

We did not include in the analysis framework of Table 2.2 the relevance of applications to stakeholders because its definition has some limitations, described in chapter 2, and underestimates the relevance to employers and health plans. However, we do report it here, since it still provides helpful information. Because an application can be relevant to a group of stakeholders, such as patients and providers, we show in Table 4.9 the number of applications that are relevant to all the stakeholder combinations found in the data.

Table 4.9. Relevance of Applications to Stakeholders, by Stakeholder Groups

Stakeholder(s)	Number of applications
Providers	47
Providers, patients	28
Patients	14
Patients, health plans, employers	6
Providers, Patients, employers	5
Patients, employers	4
Caregivers	2
Providers, health plans	1
Providers, patients, health plans	1
Providers, patients, health plans, employers	1

The most salient feature of Table 4.9 is the fact that there is a fairly large number of applications that target both providers and patients. Examples of such applications include some symptom checkers,[166,167] which share the information entered by a patient with a provider, or applications that

monitor the patient in their home environment (for falls,[234] patient deterioration,[182] or unusual behaviors pointing to possible dementia[236]) and communicate such information to clinical personnel. Though the table is useful because it shows the combinations of stakeholders to which applications are relevant, it also makes it difficult to count how many applications are relevant to a specific stakeholder. Therefore, we provide this information in Table 4.10.

Table 4.10. Relevance of Applications to Stakeholders, by Individual Stakeholder

Stakeholder	Number of applications
Providers	83
Patients	59
Employers	16
Health plans	9
Caregivers	2

The large number of applications that are relevant to providers, shown in Table 4.10, is not surprising, as many applications are tools that can be used to improve diagnoses and provide better health recommendations. Patients constitute the next stakeholder in terms of relevance; 59 applications directly involve the patient. As Table 4.9 shows, approximately half of them are also relevant to providers, but the other half is relevant either to patients only or to a combination of patients, employers, and health plans. Examples of applications that are relevant to patients only include the artificial pancreas,[215,216] chatbots that deliver mental health treatment without assistance from any clinician,[214] and applications that help patients manage their glucose levels[194,237] or medication adherence.[189,196]

Employers and health plans do not have many current applications that are directly relevant to them, compared with providers and patients; however, here we are counting only the applications that explicitly target these stakeholders on their website. There are many applications for which there is no commercial product yet that might become highly relevant to employers and health plans in the near future, and therefore we are most likely underestimating the options available to these stakeholders.

Visualization of Application Characteristics

We summarize some of the dimensions that we have studied in this chapter in Figure 4.1. We report health conditions on the vertical axis and function on the horizontal axis, and then use shapes to represent users and colors to represent current development status.

The visualization makes it clear that cardiovascular disease and diabetes are addressed most frequently among our 9 in-scope conditions; very few AI applications address substance abuse or kidney disease.

The visualization also shows that many applications are not targeted to a specific disease group in the population. These applications often pursue patients who use certain services, such as inpatient or emergency department, and members of our priority population may benefit from them because they tend to have high utilization levels of those services. Typical examples of these applications include hospital readmission models and early sepsis detection systems.

Figure 4.1 also shows major gaps in the availability of ML applications for specific groups of patients. In particular, patients with kidney disease or with issues of substance abuse have very few options. The substance abuse application is a smoking cessation chatbot,[198] integrated with other behavioral and drug interventions. One of the 2 applications for kidney disease estimates the survival benefit to kidney transplant recipients,[235] while the other uses AI to assist in the anemia management of patients with chronic kidney disease who are undergoing hemodialysis.[203]

Figure 4.1. Application Characteristics

Note: We report health condition on the vertical axis and function on the horizontal axis. We use shapes to represent users and colors to represent current development status.

5. Mapping the Evidence on Machine Learning Applications in Clinical Care

This chapter addresses aim 3 of our study by developing an evidence map for the 109 applications described in the previous chapter.

We identified 173 evaluation studies, providing information on 94 of these 109 applications. We categorized these studies by publication type, study design, study population and sample size, and outcomes measured according to the framework described at the end of chapter 2. The full set of information we captured for each of these evaluation studies is presented in appendix B.

The chapter concludes with 2 evidence maps that depict the evaluation studies according to publication type, study design, and sample size as well as the health conditions addressed by the evaluated applications.

Publication Type

Table 5.1 gives an overview of the number of applications discussed in evaluation studies, grouped by publication type and current adoption status.

Table 5.1. Applications Examined in Evaluation Studies, by Study Design and Application Status

	Number and percentage of applications that are the subject of . . .			
Application status	No published evaluations	An evaluation study	A peer-reviewed study	A peer-reviewed RCT
FDA approved (out of 22)	4 (18%)	18 (82%)	11 (50%)	4 (18%)
In use without FDA approval (out of 45)	10 (22%)	35 (78%)	27 (60%)	4 (9%)
In potential near-future use (out of 42)	1 (2%)	41 (98%)	14 (33%)	6 (14%)
Total applications (out of 109)	15 (14%)	94 (86%)	52 (48%)	14 (13%)

The proportion of applications for which we found no published evidence at all is relatively low (14%). Applications in this category tend to be commercial products whose website does not contain any information about how the product works. Some of these websites contain pages relating "user stories," but we did not count them as evaluation studies.

For the 86% of applications for which some evidence is available, we have looked in more detail at the publication type, which is summarized in Table 5.2.

Table 5.2. Distribution of Publication Types

Publication type	Number of publications	% of publications
Peer-reviewed articles	85	49
Other academic documents	19	11
FDA summary documents	11	6
Clinical trial records	38	22
Other gray literature	20	12

Note: This table applies to the 86% of applications for which some evidence is available.

The figures reported in the table refer to the number of publications and show that almost half of the publications are peer reviewed. An application could have more than one peer-reviewed publication; we find that a somewhat smaller proportion (40%) of applications have one or more peer-reviewed publications. The application with the most peer-reviewed publications in our data set was the sepsis prediction algorithm developed by Dascena,[170,] which had 4.[238-241]

There is no obvious pattern explaining which types of applications have peer-reviewed evidence, other than the fact that evidence for applications produced by small commercial developers (with 10 or fewer employees) is less likely to be peer reviewed. This finding is consistent with the view that small developers have less time and fewer resources to invest in the production of high-quality evidence. We found no significant pattern relating the presence of peer-reviewed evidence to AI type, current status, and health condition.

The least common form of evidence is the FDA summary. We found only 11 FDA summaries in our evidence database, even though there are 22 applications that are FDA cleared. This disparity is justified by the fact that not all FDA documents associated with FDA clearances contain information that can be counted as evidence.

Study Design

We found both performance tests and field evaluations in our search for studies evaluating the AI applications identified in our narrative review. Approximately 40% of studies were performance tests that evaluated the accuracy of an AI application's analytic outputs without actually implementing the application in patient care. Many of these tests consisted of researchers acquiring retrospective patient data from EHRs or other existing data sets, applying the AI application's predictive model, and then checking the model output against known patient outcomes to estimate accuracy. Some of these tests followed a prospective study design, in which patient data are collected and analyzed during the study period to assess an application's accuracy using new data. An example of such an evaluation is the Mayo Clinic study on the validation of an ML algorithm for the detection of left ventricular systolic dysfunction.[242]

The remaining 60% of the studies were field evaluations, in which the application is actually implemented and used as part of patient care. RCTs accounted for 26% of all studies. The simpler pre-post design was used in 22% of all studies. Pre-post studies often involved an application that runs on a smartphone, a wearable, or a home device. Subjects were given access to the application and several outcomes were measured, sometimes by the application itself and sometimes by a health care professional. Examples of evaluations of this type include evaluations of the Wellthy Diabetes and Virta smartphone apps for diabetes management[243,244] and Lark personalized health coach for hypertension control.[195]

We also found a smaller portion of the studies (11%) to be field evaluations in which the goal is to acquire measures related to the implementation of the application, such as user satisfaction, training time, or cost, rather than the application's effectiveness. Examples of evaluations of this type include those for a Parkinson's disease management application,[245] the Babylon symptom checker,[246] and the BlueStar diabetes management application.[46]

Table 5.3 gives a view of designs at the level of the evaluation study. However, it is also informative to obtain a view at the level of the applications, allowing researchers to answer questions such as "how many applications had an RCT performed?" or "how many applications had their performance measured in a performance test?" We provide such as a view in Table 5.4.

Table 5.3. Distribution of Study Designs (at Evaluation Level)

Study design	Number of publications	% of publications
Performance test	71	41
Field evaluation: RCT	45	26
Field evaluation: pre-post	37	22
Field evaluation: implementation	20	11

Table 5.4. Distribution of Study Designs (at Application Level)

Study design	Number of applications	% of application
Performance test	45	49
Field evaluation: RCT	36	39
Field evaluation: pre-post	27	29
Field evaluation: implementation	15	16

Note: An application may have several evaluation studies, and therefore the categories on the rows are not exclusive.

Table 5.4 shows that, although only 26% of the evaluation studies are RCTs, almost 40% of the applications had an RCT performed. This difference is attributed to the fact that applications may have multiple evaluation studies but are unlikely to have more than one RCT associated with them, so that when we look at evaluations, RCTs become "diluted" among other studies. Similarly, we also find that almost half of the application had a performance test, although performance tests constitute only 41% of the evaluation studies.

Study Population and Sample Size

Most of the evaluation studies examined applications that address 1 of the 9 specific in-scope health conditions rather than general patient health. Therefore, as expected, we find that study populations are often defined by the health condition associated with a specific application; however, in 27% of cases the application does not have a targeted health condition (say, prediction of hospital readmission), and in those cases the study population is usually driven by the context in which the application is deployed—that is, a combination of setting and user.

The study sample size follows a clear pattern as a function of the study design. For RCTs, recruitment and management are complex, and therefore it is not too surprising that the median sample size for an RCT in our data is only 142. Other field evaluations, such as pre-post studies, have a simpler design and the sample size tends to be larger: the median is 238, but the 75th percentile is 1765, implying that in 25% of these studies the study sample is quite large. Finally, because performance tests often use data collected retrospectively, from data sources such as EHRs, studies in this category have the largest number of patients. The distribution of study sample sizes is shown in Table 5.5.

Table 5.5. Distribution of Sample Size (Number of Patients)

Study design	25th percentile	Median	75th percentile	Maximum
Field evaluation: RCT	48	142	632	51 645
Field evaluation: other	75	238	1765	102 456
Performance test	168	530	5587	120 818

The evaluation study with the largest number of patients was a performance test for the prediction of survival after donor kidney transplant,[235] and the largest RCT was the one for the evaluation of a sepsis prediction algorithm.[247]

Often the unit of analysis is not the patient but a health record, and each patient may contribute more than one record to the study. We show the distribution of records in Table 5.6 for completeness. The pattern is exactly the same as that for patient, but the numbers are larger.

Occasionally a publication reports the number of records but does not report the number of distinct patients corresponding to those records. When possible, we have estimated the number of unique patients, but in some cases this variable is missing. In addition, there are some studies in which the unit of observation is neither a patient nor a record, but rather, for example, a vignette, and therefore the concept of number of patients does not apply. In the evidence maps shown in Figures 5.1 and 5.2, both missing data and not applicable cases are reported as "not applicable" for simplicity.

Table 5.6. Distribution of Sample Size (Number of Records)

Study design	25th percentile	Median	75th percentile	Maximum
Field evaluation: RCT	49	146	686	51 645
Field evaluation: other	75	294	2801	158 000
Performance test	260	818	16 679	51 081 348

Outcome Measures

Table 5.7 below shows how the outcome measures described in chapter 2 are distributed in our database of 173 evaluation studies. Each study may report more than one outcome measure, and each row of Table 5.7 shows both the total number and the proportion of studies reporting a specific outcome. We present details on the outcome measures considered in each study, as well as the conclusion of the authors regarding those measures, in appendix B.

Table 5.7. Distribution of Outcome Measures

Outcome measures	Number of studies	Proportion of studies (%)
Accuracy	82	47
Health indicator	66	38
Resource utilization	29	17
Appropriate treatment	20	12
User satisfaction	16	9
Time to treatment	5	3

The large proportion of studies that report accuracy as one of their outcome measures (47%) is partly driven by the fact that many of the studies we have found belong to the category *performance test*, and therefore they all report accuracy and possibly other outcomes.

Table 5.7 includes all evaluation studies, including performance and implementation studies, whose main outcomes are not related to health. Therefore, it is useful to report what type of outcomes are reported in the 82 field evaluation studies, such as RCTs and pre-post studies, whose primary goal is to report a health-related outcome. This is shown in Table 5.8.

Table 5.8. Distribution of Outcome Measures (Excluding Performance and Implementation Studies)

Outcome measures	Number of studies	Proportion of studies (%)
Health indicator	63	77
Resource utilization	22	27
Appropriate treatment	15	18
Accuracy	8	10
Time to treatment	5	6
User satisfaction	3	4

Not surprisingly, the most commonly reported outcome measure in this group is a health indicator (77% of studies in this group), followed by resource utilization as a distant second (27% of all studies). Only 18% of these studies report an outcome related to the delivery of appropriate treatment. Examples of applications with evaluation studies of this type include systems to improve medication adherence[42,196] and to recommend optimal hypertension treatment,[205] as well as symptom checkers that provide triaging recommendations.[164]

Overall, analyzing the author conclusions for each study, reported in appendix B, we find that introduction of the AI application led to an improvement in outcomes in 84% of the cases in which a comparison with the status quo is performed.

Of all the evaluation studies reviewed, only one found direct evidence of harm caused by an AI application. This study concluded that use of an online symptom checker was associated with higher health anxiety and negative emotional affect, compared with a control group that did not use any online search.[248] The other few exceptions where the outcome associated with an application was worse than the comparator were studies in which the outcome measured was accuracy and the researchers compared the performance of the AI application with human performance.[160,162] In these cases, however, no direct harm was observed, as these studies focused exclusively on accuracy and the recommendations of the applications were not actually implemented in patient care.

Included in the health indicators is also health risks and the safety of the applications. Although RCTs routinely report the presence or absence of adverse events, safety is not mentioned very often in the publications and the websites we analyzed—with some exceptions. For example, Your.MD, a

symptom checker application, displays prominently on its website the fact that the company is allegedly the first in its field to have established a clinical advisory board to ensure that the application is safe.[249] Babylon, another symptom checker, also has a "responsibility hub," which is a website where safety and transparency are discussed.[250]

Despite these limitations, we were able to obtain an estimated number of applications for which an assessment of the safety of the applications has been performed (or is being performed), by combining the analysis of the evaluation studies with 3 assumptions:

1. Applications with FDA clearance have been assessed for safety.
2. Applications that are the subject of an RCTs have been, or will be, assessed for safety.
3. Performance studies do not assess for safety (unless explicitly mentioned).

The results of this analysis are shown in Table 5.9.

Table 5.9. Number and Proportion of Applications Assessed for Safety

Safety assessment	Number of applications	Proportion of applications (%)
Application was (or will be) assessed for safety	50	46
Applications was not assessed for safety	45	41
Safety assessment uncertain	14	13

Overall, we were unable to assess the safety for only 13% of all applications. Applications in this category included several disease management platforms and some applications for which the evidence was found mostly on a commercial website.

The remaining 87% of the applications were almost evenly split between presence or absence of a safety assessment. We found that, for 46% of the applications, some assessment of the safety of the application either has been performed or will be performed. The main reason for which an application was labeled as not having a safety assessment performed was because the only evidence for the application was a performance study.

An additional concern about safety is the relatively small sample size of many of the field evaluation studies: if adverse events are rare, they may not be caught until an application is widely in use, and there is no clear mechanism of postmarket surveillance for these applications.

Accuracy

Most of the applications we considered contain a predictive component whose accuracy is often recorded as the outcome measure. When the outcome is binary, the most common form of reporting accuracy is through the AUC, a number between 0 and 1, where 1 corresponds to perfect classification. Sensitivity and specificity are also often reported, followed by positive and negative predictive values.

The values of the different accuracy measure span a wide range, and AUC values vary from 0.7 to 0.95. We do not report their distribution here because comparing accuracy measures across different applications is not meaningful and is possibly misleading: a sensitivity of 0.8 could be excellent in one context because the corresponding human performance is only 0.7, but there will be cases where any sensitivity below 0.9 would be unacceptable. Meaningful comparisons can be made for a specific application when the accuracy of the ML algorithm is compared with the accuracy of humans or of a comparator. When comparisons of this type are made, they tend to turn out in favor of the applications, but this may be due to publication bias and the fact that, for FDA-cleared applications, developers often have to show that the proposed application is at least as accurate as an existing product. Exceptions to this rule are cases in which an external validation has been performed, as in the case of some symptom checkers that were shown to be not as accurate as humans.[160,162]

What is more interesting about accuracy measures is the fact that—despite their prevalence—we have found no studies that actually explain how inaccuracies affect patient outcomes: what is the consequence of a false positive, and how does it compare with the consequences of a false negative? Not knowing the answers to these questions makes it impossible to understand whether the application is beneficial—unless additional health outcomes or the performance of a comparator are presented. We will come back to this issue in chapter 7.

Table 5.7 presents the prevalence of accuracy measures across the evaluation studies; however, it does not show whether accuracy is usually studied as the only outcome or jointly with other outcomes. In addition, it provides a view at the level of the evaluation study, not of the application. Because it is important to understand how often accuracy and other outcomes are reported for a given application, we answer this question in Table 5.10. We define as *health outcomes* the 2 outcome measures that most directly relate to health: health indicator and appropriate treatment. in the rows of the table we report whether an application has at least one evaluation study that reports accuracy. In the columns we report instead whether an application has at least one evaluation study that reports a health outcome. The numbers in the table represent proportions of all applications.

Table 5.10. Proportion of Applications for Which Evidence on Accuracy and/or Health Outcomes Has Been Reported (%)

	Health outcomes not reported	Health outcomes reported	*Total*
Accuracy not reported	5	39	44
Accuracy reported	33	22	56
Total	38	62	100

Note: In this figure, the term *health outcome* refers to health indicator or appropriate treatment.

Table 5.10 shows that, when we look at the application level, we obtain a slightly different picture on the prevalence of accuracy than the one we get from Table 5.7, which refers to evaluation studies.

The table shows that, for 56% of the applications, some notion of accuracy has been measured, while accuracy is reported in 47% of the evaluation studies (Table 5.7), implying that evaluation studies tend to favor health-related outcomes to some extent.

More important, though, is the fact that accuracy and health outcomes have been jointly reported for only 22% of the applications. For the remaining 78% of applications, either accuracy or health outcomes are reported, with a small 5% of cases in which neither is reported and the user satisfaction is reported instead.

Evidence Maps

We have summarized some features of the evidence on AI applications in health care in the 2 evidence maps below, where the unit of observation is the evaluation study.

In Figure 5.1, we map the evidence along the dimensions of health condition, evidence type, and sample size. The most apparent pattern emerging from the map is the great disparity in how the evidence is distributed across health conditions. In fact, most of the available evidence is concentrated on applications to general health, heart disease, and diabetes, and very little evidence is available on the effectiveness of AI applications for substance abuse and dementia.

Figure 5.1 also shows that the evidence is unevenly distributed along publication type. What is particularly striking is the fact that there is virtually no evidence contained in FDA summaries for applications to cardiovascular disease, even though most FDA-cleared applications concern this condition, as shown in Figure 4.1.

In addition, the figure shows that, though general health is the category with the most evaluation studies, cardiovascular disease is actually the condition with the most peer-reviewed studies and therefore the higher quality of evidence.

In Figure 5.2, we map the evidence along the dimensions of health condition, study design, and sample size. The first pattern that emerges from this map is the relative paucity of RCTs for applications related to general health and cardiovascular disease, for which most of the evaluation studies are performance tests. This implies that, while there is a relative abundance of peer-reviewed evidence for these health conditions, as shown in Figure 5.1, most of the evidence is not about health outcomes but instead about the accuracy of the applications, which is not as informative from a patient perspective.

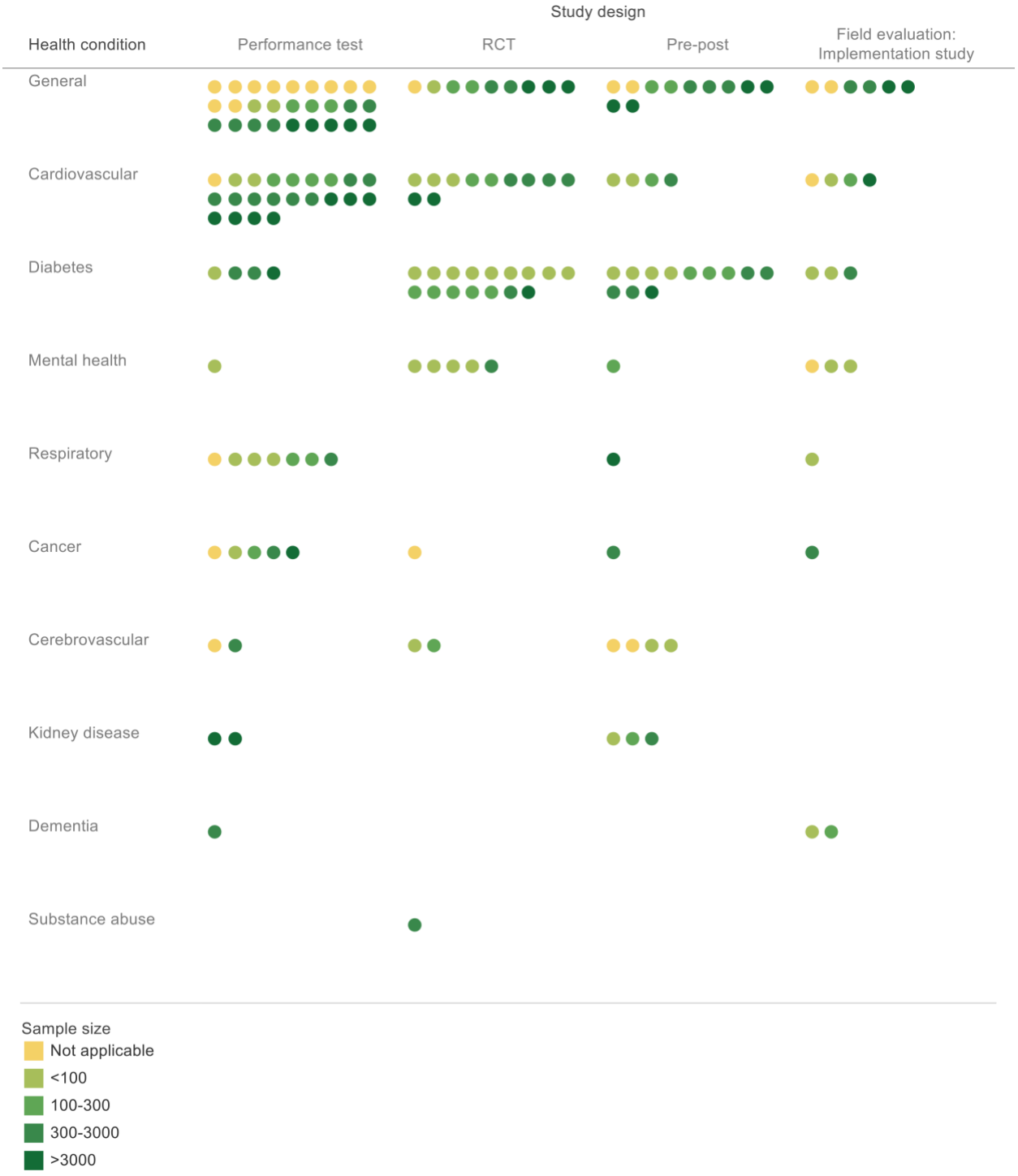
A qualitatively similar pattern, but exacerbated, is observed for cancer and respiratory disease. For these conditions only a single RCT is available, and a large portion of the evidence comes from performance tests.

The map shows that the opposite pattern is true for diabetes: even if the number of evaluation studies for this condition is smaller than what is observed for general health and cardiovascular disease, most of the studies are RCTs or have a pre-post design and therefore report health outcomes.

Figure 5.1. Evidence Map: Health Condition vs Evidence Type

Note: On the vertical axis we show the health conditions, sorted by the number of evaluation studies, and on the horizontal axis we show the publication type. Green shades are used to represent the sample size; yellow is used for studies in which the notion of sample size is not applicable, as described in chapter 2.

Figure 5.2. Evidence Map: Health Condition vs Study Design



Note: On the vertical axis we show the health conditions, sorted by the number of evaluation studies, and on the horizontal axis we show the study design. Green shades are used to represent the sample size; yellow is used for studies in which the notion of sample size is not applicable, as described in chapter 2.

6. Stakeholder Views

In the early stages of this project we performed several interviews with stakeholders to gather their views regarding the application of AI in health and to validate the intended scope of the analysis. While performing a formal analysis of the stakeholder interviews is not one of the aims of this report, not reporting some of the views gathered would be a missed opportunity. Therefore, we use this chapter to accomplish 2 tasks: first, we briefly summarize what we learned from the stakeholder interviews; second, we describe 3 issues that emerged as important concerns to stakeholders. The description of the concerns is not linked to the analysis of the data, and its only purpose is to inform the reader and provide some context to these important topics.

Stakeholder Interviews

We spent part of each stakeholder interview to validate the proposed scope of the project, which is the intention to exclude imaging applications, the list of priority conditions, and the decision to focus on applications related to clinical care.

All interviewees agreed that AI applications to medical imaging are at a more advanced level of adoption and that therefore it was reasonable to exclude them from the analysis and focus on applications in early stages. Few interviewees questioned the usefulness of targeting a prespecified list conditions, but suggested to include applications that are not disease specific. As described in chapter 2, because of this feedback we also included in the scope of the analysis applications that target broad categories of patients and aim to benefit all individuals who use health care services.

A theme that drew common agreement was consumer empowerment.[251-253] All interviewees agreed on the need to include not only applications used by clinicians but also consumer-empowering and consumer-facing applications, such as those designed to improve disease management or to assess symptoms and inform care-seeking behavior. A common view appeared to be that AI applications of this type not only could benefit patients directly but also contribute more generally to make the patient an integral part of the clinical care process.

We used the views expressed above to validate and refine the scope of the analysis. In addition to those views we also collected suggestions on which features of the applications should be reported in the analysis framework, as well as which outcomes would be important to capture. In fact, elements of the framework such as user, setting, and platform were all explicitly mentioned in the interviews, although not necessarily by the same person.

One dimension that was mentioned by all interviewees but did not enter the framework was privacy: all agreed that it would be useful to report how the privacy of patients is protected when their data are used in an application. Though we did consider privacy protection as an element of the framework described in Table 2.2, we concluded, after reviewing the literature, that the information needed to characterize applications along this dimension was simply not available. For the same

reasons we were unable to include a dimension capturing whether social determinants of health were taken in account by the application, even though most interviewees thought it would be highly useful.

Stakeholder Concerns

During the interviews, stakeholders expressed concerns regarding the applications of AI in health, and 3 are described: the first 2 (bias and interpretability) were expressed by most interviewees, and the third (security) was brought up by one stakeholder as an important issue that deserves due attention and should be discussed more often.

Bias and the Value Alignment Problem

A well-known set of concerns regarding the application of AI to health (and other domains such as justice and education) is the so-called value alignment problem,[254-256] which arises because AI algorithms tend to be optimized for global accuracy and are agnostic about where the errors occur. Society, however, values equity[257] and is sensitive to the distribution of errors across the population and may not be willing to accept solutions that favor or disfavor specific population subgroups.[258,259]A typical example in which misalignment arises is when a machine learning predictive model has been developed using training data from a group that is not representative of the general population. The model may perform extremely well on the population used for development and maybe its first implementation, but it may not perform as well on the general population, or it may perform particularly badly on a specific subpopulation. The value on which the misalignment arises in this case is "fairness," since there are social expectations around equal distributions of benefits.

Another common form of misalignment occurs when an ML algorithm has been trained using data that include human decisions affected by bias or has been designed in such a way that incorporates bias. A striking example of how this can happen was recently demonstrated by Obermeyer et al.[260] The authors found that a commercial product widely used to determine access to high-risk health care management programs gives preferential access to white individuals, compared with black individuals of similar health status. The reason for which this bias is encoded in the algorithm is the fact that the algorithm uses health care cost as a proxy for health care risk. Because of existing inequalities in access to care, black individuals experience lower costs than do white individuals, at the same level of health, and therefore appear to the algorithm healthier than they really are, resulting in being assigned a lower risk than white individuals with similar health status.

Issues of value misalignment consistently arose during interviews with stakeholders. The issue of the generalizability of an application developed for a specific population to a broader group of patients is discussed often in the studies we analyzed; however, it is rare that steps are actively taken to understand the potential for harm and unintended consequences.[261] Sepsis Watch,[229] a platform for early sepsis detection, is an example. The managing institution, concerned about the possibility of unknowingly introducing inequality and bias in their algorithm, partnered with 2 research institutions

and designed studies to investigate the sociocultural dimensions of clinical integration of Sepsis Watch.[262]

Overall, given that the health sector is still in the early days of implementation of AI applications, and evidence is available for only a relatively small number of populations, often chosen ad hoc, it would be surprising to see the issue of generalizability being addressed on a regular basis. Entities that are better able to address the issue are health care providers that operate in many sites and that cater to disparate communities.[48]

It is encouraging, however, that the issue of fairness in machine learning is an active research area, and methodologies that allow incorporating principles of distributive justice in the design and implementation of AI applications are being developed.[263-268] During the design stage of the application, the first recommended step to reduce bias is the identification of the group of individuals who might be put at disadvantage and need protection. The next step consists then of modifying the ML algorithm in such a way that its performance along certain dimensions is equalized between the protected and nonprotected groups.[263,264] For example, one could attempt to achieve similar health benefits in the 2 groups. This could be difficult to achieve in practice, and a more practical alternative might be the equalization of the accuracy of the application. In cases in which the ML application is used to allocate resources, one could also design the application in such a way that the distribution of resources is fairly allocated.[263] Choosing which distributive justice principle should be incorporated in an ML application, and therefore providing a clear definition of the goal of the application (eg, maximize overall accuracy while equalizing benefits), is a common recommendation and a step forward in bias reduction. However, this is a complex question that has technical, clinical, and ethical dimensions, and in fact another common recommendation to achieve fairness in ML applications is the use of interdisciplinary teams and the involvement of a diverse group of stakeholders, starting at the early stages of the project.[263,269]

The design of the algorithm used in the ML application is not the only element for which one can intervene to achieve fairness, though. The data used to develop an ML application also play a crucial role,[270] and steps can be taken to minimize the perpetuation of biases already present in the data or to ensure that the sample is sufficiently representative of protected and nonprotected groups. For example, both proper annotation of the data,[271] which documents how they were collected and labeled, and improved design of data collection methods,[256,269] which take in account the uneven distribution of health care access, have been recommended as strategies to reduce bias.[270]

Finally, there are also actions that can be taken after an application has been designed and that can contribute to improve fairness—for example, monitoring predetermined metrics across protected and nonprotected groups while deploying the application can help catch potential issues at early stages. In particular, combining monitoring with stepped-wedge trials,[272] in which the intervention is deployed sequentially on different clusters of subjects, has been recommended as best practice.[263]

Interpretability

A common criticism of AI and ML applications is that they sometimes behave like a "black box" or an oracle: when queried they provide an answer but may not reveal how they arrived at that specific answer.[31] Depending on the function and the users of the application, this issue can become a barrier to adoption. For example, when the function of the application is risk stratification and the user is a health care professional, interpretability is important because one needs to know both the level and the determinants of the risk to decide which actions to take. However, when an ML-powered tool is used as a device or a component of a larger system, or is used in the context of resource optimization, a user may simply trust the tool, as in the case of many other devices or software tools employed in clinical contexts.

The degree of interpretability of the application is a function of the specific type of ML used.[273] At the end of the noninterpretable spectrum is neural networks, such as DL architectures, whose predictions depend on the input variables in a complicated way and tend to have a large number of parameters. Interpretable models tend to be somewhat simpler and have fewer parameters, and they include the classic statistical logistic and multinomial models as well as tree-based models, such as CART, and similarity-based models. Though it is true that more complex (and less interpretable) models often achieve higher accuracy, if the size of the data set is sufficiently large, this does not necessarily imply that there is a trade-off between accuracy and interpretability and that one need give up interpretability to gain accuracy. It is entirely possible that a simple and interpretable model performs better than more complex and less interpretable models,[274] and in general one does not know a priori which method will perform better on a specific data set.

Given that interpretability of an ML application is an important attribute, we considered making it an explicit part of the framework we used to categorize applications (summarized in Table 2.2). However, interpretability is a function of the detailed type of AI that was used. Because we have shown in Table 4.5 that, for 61% of the applications, the AI type is unknown, it follows that it is currently not possible to make definitive statements about the level of interpretability among AI applications in health.

Security

The increased use of individual-level data for the purpose of both developing and making use of AI applications raises concerns about protecting individual privacy, and all the stakeholders interviewed expressed this concern. An issue that was discussed to a much lesser extent was security—that is, the safeguarding of data and devices from theft, corruption, and malicious attacks.[275] Across the 109 applications we reviewed, we did not find any evidence of how these issues are tackled by the developers of these applications, other than adhering to current regulations. Although privacy is mentioned in some of the documents we reviewed, cybersecurity is never acknowledged. This omission could be partly explained by the fact that only recently has cybersecurity in health been recognized as a serious issue that needs to be addressed. The Health Care Industry Cybersecurity Task Force was initiated only in 2015 as part of the Cybersecurity Act, and its first report was published in 2017.[276]

7. Discussion

Before discussing lessons learned from this analysis, it is important to reiterate some important features and limitations of this study:

- We have intentionally focused on applications that are either already in use or are likely to be in use in the near future, and therefore we paint a picture that is inevitably different from the picture someone would get from the vast existing literature about the *potential* of AI in health care. Researchers have been studying applications of AI to health care for more than 20 years, but the time lag between research and translation into clinical practice is notoriously long,[277] and therefore the health sector is only in the early days of implementation.
- We focused on applications that directly affect clinical care, and therefore we are not considering here applications, such as fraud detection, that may have a significant impact on the health care system.
- We have intentionally excluded the entire area of medical imaging from the scope of the report; AI and ML applications in this area are at a more advanced stage of adoption and are already quite well understood.[3,123,124,278-283]
- Due to the breadth of the review scope and range of document types examined, we did not formally assess the quality of the evidence presented in evaluation studies; rather, we focused on the type of evidence presented.
- We examined only publicly available evaluation studies for evidence surrounding AI applications. Developers often evaluate applications without publishing the results. One stakeholder suggested that this may be especially the case for established businesses whose applications do not require FDA approval and that do not need to use evaluations to attract funding.

With these points in mind, the first conclusion we draw from the results of our narrative review and evidence map is that **some evidence regarding the benefits of current applications of ML in health is available**. Though the evidence is scattered across peer-reviewed journals and different types of gray literature, and its quality varies widely, we found that some evidence is available for about 80% of the applications considered, and in that group about 40% of the applications are the subject of at least one peer-reviewed paper. Among the peer-reviewed papers, only 21% are RCTs, and another 18% are other studies that report health-related outcomes, implying that 61% of peer-reviewed publications report only accuracy or usability. Similarly, only 50% of all evaluation studies are field evaluations with reported health outcomes. When health outcomes are reported, the application is shown to lead to some health benefit. Therefore, although the evidence base is not very broad, we do find that the lives of patients, especially those with cardiovascular disease[143,145,172,284] or diabetes,[37,201,215] are improved by some AI applications. Patients with priority conditions other than

cardiovascular disease or diabetes have fewer applications specific to them; however, such patients still benefit from a wide range of applications that are not specific to a health condition but rather target service users (such as early sepsis detection models[170]) or individuals who may be in need of care (such as symptom checkers[161,163,164]). Only a single evaluation study—which found that use of a symptom checker led to greater patient anxiety—reported direct evidence of harm associated with any of the AI applications considered in this report.[248]

A second conclusion we can draw is that, when studying the evidence regarding current AI applications in health, **the notion of evidence is complex**. When analyzing the effectiveness of pharmaceutical drugs, devices, or health interventions, evidence is usually reported in terms of health outcomes or utilization measures (or both in cost-effectiveness studies). However, in the case of AI applications, evidence tends to be collected about a broader range of outcomes. As described in the "Methods and Conceptual Frameworks" section, in addition to health-related outcomes and resource utilization, we also found evidence about outcomes related to accuracy and user satisfaction. Ideally, all 3 of these types of outcomes are needed to evaluate an application, but we found that, in many cases only one of these outcomes is reported, providing a limited view of the application. For example, it is common for a study to report only the accuracy of an application. Although this certainly counts as evidence, it is not very useful, especially to patients, since it does not necessarily translate into a health benefit. Similarly, we found that health-related outcomes are reported in only 50% of all studies, and user satisfaction, a key determinant of adoption and continued use, is reported in only 9% of the studies.

Though we have not performed a rigorous assessment of the quality of the evidence, we can safely observe that it is extremely variable. Evidence is sometimes found in peer-reviewed papers documenting the results of an RCT but can also be found on the website of a vendor, without any reference to an actual study. In addition, even for studies that reported evidence with rigorous reporting standards, we found that the ML algorithms behind the applications are not disclosed in most cases. In fact, for 62% of the applications considered, we were not able to determine which type of AI was involved; during our search of the gray literature it often proved very difficult to understand whether an application included an AI component at all.

Part of the variation in the quality and type of evidence that we observed is likely related to access to resources and incentives: large, multisite institutions or well-established businesses seem more likely to sponsor a field study to evaluate the effectiveness of an application and have an incentive to show an FDA clearance and to present themselves as transparent. But younger and smaller innovative companies do not necessarily have the means to perform proper studies (which may or may not prove favorable to them), and transparency may not be a high priority. In addition, another factor contributing to the uneven quality in evidence reporting is the fact that there is no well-established framework for designing and reporting on this type of study. This situation may change soon, since, as of mid-September 2020, several papers were published by the SPIRIT-AI and CONSORT-AI Working Groups.[285-287] These texts offer recommended guidelines for reporting on clinical trials interventions involving AI as well as for designing the protocols of those studies.

Potential Areas of Future Work

One difficulty in understanding the state of the evidence on the topic of AI and health is that gathering information about applications currently in use is challenging. Several stakeholders pointed out in their interviews that much of the information we were looking for would not be found in the academic literature. Indeed, to collect the list of applications that form the basis of this report, we searched, in addition to the traditional literature, hundreds of industry magazines and websites. This process is inefficient and difficult to update with new results, but this does not have to be the case. For example, in the area of AI and medical imaging, the Data Science Institute of the American College of Radiology already maintains a website called "FDA Cleared AI Algorithms," which provides for each algorithm a short summary, model manufacturer, FDA product code, predicate devices, product evaluations, clinical validation, and other useful information.[106]

Borrowing from this idea, one can envision a curated website that allows both developers and researchers to maintain an up-to-date description of AI applications in use and any evidence attached to them. At a minimum, each entry would contain information similar to what we provide in the appendices of this report, but one could build on the newly released guidelines on reporting on clinical trials[285,286] to design a new reporting standard. The effort on the part of developers and researchers to keep this information current would be minimal and would be rewarded by increased visibility and a reputation for transparency. If a sufficiently large number of developers joined initially, there would be a strong incentive for other developers to join so that they are not "left behind." The curation effort would be minimal, as some of the information is already available and would be focused on guaranteeing that the application is indeed currently in use, which is one feature that is often difficult to verify.

Although an effort of this type would increase transparency and would be a starting point in building a searchable evidence base, deeper issues must be addressed. For example, an important problem that we have not seen addressed in the literature is a lack of clarity on the meaning of a *sufficiently accurate algorithm*—that is, an algorithm that achieves a level of accuracy with which the field of health care is comfortable. Consider a case in which a predictive algorithm achieves a better false-positive rate than do humans but a slightly worse false-negative rate. Under which conditions does one decide that the algorithm performs better or worse compared with humans? And what if the AI solution performs slightly worse than do humans but saves a significant amount of resources that can be allocated to another task? The answers to these questions depend in a complicated way on the clinical, economic, legal, and ethical *consequences* of the type of error being committed, including the safeguards put in place to catch those errors, which are typically unknown.

Although this is not necessarily a new problem—and it arises when building any sort of device—a framework in which the notion of accuracy of AI algorithms applied to health care can be properly discussed seems to be missing. Clearly, providing an answer to such questions as "what are the costs of a false positive/negative?" is not possible in general, as it requires the simulation of the future consequences of those errors and an evaluation of their health and cost implications, which are specific

to the AI application considered. However, answers could be provided in some cases by taking advantage of the existing body of work in health services research and health economics. Work along these lines could go a long way to clarify the implications of the introduction of AI algorithms in the health care domain.

That said, accuracy looks only at the number and type of errors. Concerns raised by stakeholders and frequently cited in the literature show that the distribution of errors is also important, especially when an AI application is applied to a population to which it has never been applied before (the value misalignment problem). There are steps that could be taken to address these concerns. For example, the Data Science Institute of the American College of Radiology already offers a service, called Certify-AI,[288] which independently evaluates and validates an AI algorithm against a large data set that spans known sources of variability. This comparison is possible because medical imaging data sets are more standardized and have a higher degree of interoperability than do most data sets in other areas of clinical care. Nevertheless, despite the variety of data types used by AI applications, it would be possible, to some extent, to establish similar services for applications built on data sets such as EHRs or that contain specific types of data (say, ECG and EEG). Again, this problem is not new, and it also affects developers of medical devices.[289,290] However, the range of variables that could introduce bias is potentially wider in the case of AI applications, which depend on a large number of inputs. In addition, the potential scale at which some of these AI applications could operate is enormous, and therefore the potential for harm to specific subpopulations is large if this issue is not addressed. Any step in this direction would give confidence to both developers and the public that the algorithm will perform well when applied to a population that has a different composition than the one for which it was originally developed.

Studies of this type, as well as much needed rigorous prospective field evaluations, require resources. This is particularly problematic when developers of new AI application are start-ups or small and medium-size enterprises. However, novel funding schemes could be developed. The Australian Cooperative Research Centers (CRCs)[291] for applied research stands as an interesting model. In this model, industries and universities contribute cash and in-kind resources toward the formation of a CRC, and the government provides matching funds. This setup allows industry partners of the CRC to leverage both the matching funds and the in-kind personnel and expertise provided by universities. The result is that industries use the funds to pay a fraction of the cost of PhD students or postdocs and at the same time gain free access to senior university researchers. Because the objective of the CRC program is to support businesses and stimulate the economy, all the CRC-sponsored projects belong to the category of applied research. For CRCs in the health domain, that often means that the sponsored projects are field evaluation studies, as exemplified by some of the ongoing projects of the Digital Health CRC,[292] the largest CRC dedicated to health currently in existence. Understanding how such a program could be adapted to the US environment might inform the discussion on what other alternative funding schemes could be implemented to build a broader evidence base for AI in health.

Conclusions

In this report we have focused on AI applications that are currently in use or are in near-future use, and we have intentionally excluded applications of AI to medical imaging, an area that is in a more advanced adoption stage. As a result, the picture that emerges from the analysis of the evidence we have collected is quite different from the picture one obtains from the vast literature about the *potential* of AI in health. The large amount of publications on the subject of AI and health and the attention that this topic has received, fueled by the undisputed success of AI in the specific area of medical imaging, may have created high expectations and a sense that the field is in a more advanced stage than it actually is.[293,294]

In fact, we found that, when it comes to AI applications to clinical care, the health sector is still in the early days of implementation. This finding should not be surprising: researchers have been working in this area for more than 20 years now, but the time lag between research and translation into clinical practice is notoriously long.[277]

By searching both the academic and gray literature, we have found 109 applications that fit in the scope of the report. For 94 of them we were able to find some evidence along the dimensions of accuracy, health-related outcomes, and user satisfaction. In 84% of the cases in which a comparison with the status quo is performed, the evidence points to an improvement in some of these outcomes. Though the quality of the evidence varies, and in many cases health outcomes are not reported, overall we did find evidence that some of these AI applications are delivering health benefits to the population of patients. About a quarter of the applications do not target a specific health condition and benefit anyone who uses health services. Among the applications that are disease specific, the most prevalent are those targeting patients with cardiovascular disease, accounting for about 30% of all applications, and diabetes, accounting for about 15% of all applications. By contrast, patients with Alzheimer's or dementia, kidney disease, or substance abuse issues have very few applications dedicated to their needs.

The evidence varies in quality and ranges from peer-reviewed journal publications describing results of RCTs to short, often unauthored reports on the website of a commercial application. In addition, the number of applications for which it is possible to establish whether their safety has been evaluated, or will be evaluated, is slightly less than 50%, and FDA-cleared applications account for only 20% of the total.

The fact that there are problems that need to be addressed before AI applications in clinical care can be confidently evaluated and used does not mean that the future of this area is not bright, and it is important to reiterate that, especially in the area of clinical care, we are still in the early stages of research translation. When viewed through this lens, it is not unexpected that the state of the evidence is incomplete—but improving rapidly.

Appendix A. Current and Near-Future ML Applications

Details on each of the 109 in-scope ML applications that we identified as in current or potential near-future use are provided in the table below, in alphabetical order.

Details on the individual studies that evaluate these applications can be found in appendix B, also presented in alphabetic order by application.

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
Ada, Ada Health[163] Smartphone-based chatbot and symptom-checking app for broad range of health conditions In use, not FDA cleared	Health recommendations General: All conditions Potential presence of condition; need for clinical visit	Conversational AI Patient-entered text Unspecified training data	All patients Patient Home Smartphone	Peer-reviewed article, pre-post trial[248] Peer-reviewed article, performance test[295] Peer-reviewed article, implementation study[296] Academic preprint, performance test[297] Gray literature, implementation study[298] Gray literature, performance test[299]
Advanced Electronic Safety of Prescriptions Model, Taipei Medical University[300] App that flags potentially inappropriate prescriptions for physician review Development: Field evaluation: Not yet published	Patient evaluation General: All conditions Presence of inappropriate prescription	Parametric Structured variables: EHR Unspecified training data	All patients Health care professionals Outpatient	Clinical trial record, RCT[300]
Advisor Pro/MD-Logic, DreaMed[201] App that provides diet and insulin dosage recommendations for diabetic patients FDA cleared: Class II, 510(k), following prior De Novo approval	Health recommendations Diabetes Optimal insulin dosage	Unspecified ML type Nonimagery sensor: Continuous glucose monitor Unspecified training data	Children and adults (aged 6-65) with type 1 diabetes receiving insulin treatment Health care professionals Home Wearable	Peer-reviewed article, RCT[301] Peer-reviewed article, RCT[302] Peer-reviewed article, implementation study[44] Peer-reviewed article, performance test[303]

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
Ahead, Brainscope[149] EEG device and algorithm to evaluate possible brain injury FDA cleared: Class II, 510(k), following prior De Novo approval	Patient evaluation General: Traumatic brain injury Potential presence of traumatic brain injury; cognitive function	Unspecified ML type Nonimagery sensor: EEG Unspecified training data	Adults (aged 18-85) with mild traumatic brain injury in past 72 hours Health care professionals Outpatient	Peer-reviewed article, performance test[304] FDA summary, performance test[305] FDA summary, performance test[306]
AI-Assisted Insulin Titration System, Shanghai Zhongshan Hospital[204] App that uses AI to determine insulin dose to provide patient Development: Field evaluation: Not yet published	Health recommendations Diabetes Optimal insulin dosage	Unspecified ML type Unspecified input Unspecified training data	Adults (aged ≥ 18) with type 2 diabetes receiving insulin treatment Patient Home	Clinical trial record, RCT[204] Clinical trial record, RCT[307]
AI-ECG Tracker, Lepu Medical/Carewell[144] ECG processing and analysis software for arrythmia detection FDA cleared: Class II, 510(k)	Patient evaluation Cardiovascular: Arrythmia Presence of arrythmia; measurement of heart rhythm	Parametric Nonimagery sensor: ECG Unspecified training data	Adults (aged ≥ 22) without pacemakers Health care professionals Inpatient; outpatient	No studies
AIM@BP, University of Michigan[196] Text messaging app that provides medication adherence reminders to patients with hypertension Development: Field evaluation: Results published	Health recommendations Cardiovascular: Hypertension Optimal medication reminder message type	Parametric Nonimagery sensor: Medication use monitor 48 patients with hypertension	Patients with hypertension Patient Home Smartphone	Gray literature, RCT[308]

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
Anemia Control Model, Fresenius Medical Care[203] ML algorithm that generates optimal erythropoiesis-stimulating agents dosage for patients with end-stage renal disease Development: Field evaluation: Results published, further trials ongoing	Health recommendations Kidney disease; anemia Patient response to treatment dosage; optimal dosage	Parametric Structured variables: HER 101 918 records from 4135 patients (29% with diabetes) at Italian, Spanish, and Portuguese clinics	Adults (aged 19-90) with chronic kidney disease undergoing hemodialysis Health care professionals Outpatient	Peer-reviewed article, pre-post trial[40] Peer-reviewed article, pre-post trial[309] Peer-reviewed article, performance test[310] Clinical trial record, pre-post trial[203]
Anticoagulation Management Service, Brigham and Women's Hospital[184] App to identify patients with AF at high risk for stroke and not on anticoagulants Development: Field evaluation: Results published	Patient evaluation Cardiovascular: Arrythmia: AF Cerebrovascular: stroke Presence of AF and stroke risk factors	Unspecified ML type Structured variables: EHR Text: EHR Unspecified training data	Adults (aged ≥ 18) Health care professionals Outpatient	Peer-reviewed article, RCT[311]
Apple Watch 4: Fall Detection App, Apple[217] Smartwatch app for detecting user falls; calls emergency services if user remains immobile In use, not FDA cleared	Treatment delivery General: Fall Detection of fall and continued patient immobility	Unspecified ML type Nonimagery sensor: Accelerometer, gyroscope More than 250 000 recorded people-days from more than 2500 people	All patients (function enabled by default for adults aged >55) Patient Home Wearable	No studies

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
ASSIST, Case Western Reserve University[212] Wearable app that evaluates patient behavior to determine how often to deliver mindfulness meditation training Development: Field evaluation: Not yet published	Treatment delivery Mental health: Depression/ anxiety Optimal type and frequency of meditation mindfulness training	Unspecified ML type Nonimagery sensor: Blood pressure, heart rate, movement Unspecified training data	Adults (aged ≥ 18) who are providing care for a critically ill family member Patient Home Wearable	Clinical trial record, RCT[212]
Assisted Rehabilitation Care, Camlin[211] Physical rehabilitation device that provides ML-based personalized physical therapy Development: Field evaluation: Not yet published	Health recommendations Cerebrovascular: Poststroke rehabilitation Quality of patient movement following stroke	Unspecified ML type Nonimagery sensor: Accelerometer/ inertial sensor Unspecified training data	Adults (aged ≥ 18) who have had a stroke in the previous 6 months Patient Home Wearable	Clinical trial record, pre-post trial[211]
Atrial Fibrillation Risk Prediction, Bristol-Myers Squibb[146] App that analyzes ECG and EHR to predict patient risk of AF Development: Field evaluation: Not yet published	Patient evaluation Cardiovascular: Arrythmia: AF Presence of AF	Unspecified ML type Structured variables: EHR Nonimagery sensor: ECG Unspecified training data	Adults (aged ≥ 30) without prior diagnosis of AF Health care professionals Outpatient	Clinical trial record, RCT[146]
Babylon Health, Babylon[167] Smartphone-based chatbot and symptom-checking app for broad range of health conditions In use, not FDA cleared	Health recommendations General Unspecified	Unspecified ML type Patient-entered text Structured variables: EHR Unspecified training data	All patients Patient Home Smartphone	Gray literature, implementation study[246] Gray literature, implementation study[312] Gray literature, performance test[313] Gray literature, performance test[314]

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
Bionic Pancreas, Beta Bionics[216] App that uses AI to determine insulin dose to provide patient Development: Field evaluation: Not yet published	Treatment delivery Diabetes Optimal insulin dosage	Unspecified ML type Nonimagery sensor: Continuous glucose monitor Structured variables: Patient weight Unspecified training data	Adults (aged ≥ 18) with diabetes receiving insulin treatment Patient Home Wearable	Peer-reviewed article, RCT[315] Clinical trial record, RCT[316] Clinical trial record, RCT[317]
Biovitals Analytics Engine/Biovitals HF, Biofourmis[141] Cloud-based app that identifies significant changes in patient vital signs FDA cleared: Class II, 510(k)	Patient evaluation General: Adverse events Vital sign health index score reflecting physiological change	Regression Nonimagery sensors measuring vital signs (heart rate, respiratory rate, activity) Unspecified training data	Adults Health care professionals Home; outpatient	FDA summary, performance test[318] FDA summary, performance test[319]
BlueStar, Welldoc[37] App that analyzes blood glucose levels to provide health recommendations to patients with diabetes FDA cleared: Class II, 510(k)	Health recommendations Diabetes Unspecified	Unspecified ML type Nonimagery sensor: Blood glucose Unspecified training data	Adults (aged ≥ 18) with diabetes Patient; health care professionals Home Smartphone	Peer-reviewed article, RCT[320] Peer-reviewed article, implementation study[46] Gray literature, implementation study[321]
BQ Device, BrainQ Technologies[220] App that uses AI and EEG data to deliver personalized brain electromagnetic field therapy to patients with recent stroke Development: Field evaluation: Not yet published	Treatment delivery Cerebrovascular: Stroke Optimal electromagnetic treatment	Unspecified ML typeNonimagery sensor: EEG Unspecified training data	Adults (aged 18-80) with recent ischemic stroke Health care professionals Unspecified	Clinical trial record, RCT[322]

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
BrightArm Compact, Bright Cloud[209] Physical and cognitive rehabilitation device that provides AI-based personalized physical therapy Development: Field evaluation: Not yet published	Health recommendations Cerebrovascular: Stroke Appropriate level for virtual reality game–based rehabilitation therapy	Unspecified ML type Prior patient performance using VR game device Unspecified training data	Patients who recently suffered a stroke Patient Inpatient; outpatient Wearable; VR	Clinical trial record, pre-post trial[209]
Buoy Health, Buoy[164] Smartphone-based chatbot and symptom-checking app for broad range of health conditions In use, not FDA cleared	Patient evaluation General Presence of condition	Conversational AI Text Unspecified training data	General Patient Home Smartphone	Peer-reviewed article, pre-post trial[323]
Cardiologs Platform, CardioLogs[142] App that analyzes ECG data to detect arrythmia FDA cleared: Class II, 510(k)	Patient evaluation Cardiovascular: Arrythmia Presence of arrythmia; intermediate health measures	Parametric Nonimagery sensor: ECG Approximately 130 000 ECGs with expert annotation	Adults (aged ≥ 18) Health care professionals Unspecified	Peer-reviewed article, performance test[324] Peer-reviewed article, performance test[325]
Cardiomatics, Cardiomatics[137] Cloud-based system that uses ML for analysis and interpretation of ECG signals In use, not FDA cleared	Patient evaluation Cardiovascular Presence of condition	Parametric Nonimagery sensor: ECG Unspecified training data	Unspecified Health care professionals Unspecified	No studies
CLEWICU, Clew Medical[131] App that predicts risk of respiratory failure or hemodynamic instability in ICU patients FDA: Emergency use authorization for COVID-19	Patient evaluation Respiratory: COVID-1 General: Adverse events Risk of respiratory failure; risk of hemodynamic instability	Unspecified ML type Structured variables: HER Records from more than 100 000 hospitalized patients	Adults (aged >18) in ICU Health care professionals Hospital ICU	FDA summary, performance test[326]

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
Companion, CompanionMx[226] Smartphone app that monitors user's voice and usage to predict changes in mental health Development: Field evaluation: Not yet published	Patient evaluation Mental health: Depression/ anxiety Composite measures of patient mental health	Unspecified ML type Audio: Human speech Structured variables: Smartphone usage patterns Unspecified training data	Health status Patient; health care professionals Home Smartphone	Peer-reviewed article, performance test[327] Clinical trial record, RCT[328]
Control Tower, Mayo Clinic[48] App that uses EHR data to identify patients who may benefit from early palliative care review Development: Field evaluation: Not yet published	Health recommendations General: Palliative care Patient health status; need for palliative care	Unspecified ML type Structured variables: EHR Unspecified training data	Hospitalized adults (aged ≥ 18) Health care professionals Inpatient	Clinical trial record, RCT[329]
Corti, Corti[330] Triage platform that detects cases of cardiac arrest using real-time audio from phone calls to emergency medical services In use, not FDA cleared	Patient evaluation Cardiovascular: Cardiac arrest Presence of cardiac arrest	Parametric Audio: Human speech Unspecified training data	All emergency service callers Health care professionals Home	Peer-reviewed article, performance test[331] Clinical trial record, RCT[332]
COVID-19 Alert System, Beijing Tsinghua Chang Gung Hospital[333] App that assesses COVID-19 risk and makes health recommendations based on patient-entered information Development: Field evaluation: Not yet published	Health recommendations Respiratory: COVID-19 Presence of COVID-19	Unspecified ML type Structured variables: Patient-entered information Unspecified training data	All patients Patient Home Mobile phone	Clinical trial record, pre-post trial[333]

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
Current Platform, Current Health[181] App that identifies patients at risk of deterioration FDA cleared: Class II, 510(k)	Patient evaluation General: Adverse events Risk of deterioration	Unspecified ML type Nonimagery sensors: Temperature, oxygen saturation, heart rate, movement Unspecified training data	Adult patients Health care professionals Hospital inpatient (not ICU); skilled nursing facilities; home Wearable	Gray literature, Implementation study[334]
Dashboard for Diabetes Care, University of Utah[206] Dashboard app that uses EHR data to predict patient response to diabetes treatment options Development: Field evaluation: Not yet published	Health recommendations Diabetes Likely impact of various treatment options	Unspecified ML type Structured variables: EHR Unspecified training data	Adults (aged ≥ 18) with diabetes Health care professionals; patient; caregiver Outpatient	Clinical trial record, RCT[206]
DayTwo, DayTwo[193] Smartphone app that provides personally tailored dietary suggestions to diabetic patients In use, not FDA cleared	Health recommendations Diabetes Optimal diet	Unspecified ML type Biomarkers; text Unspecified training data	Health status Patient Home Smartphone	Peer-reviewed article, RCT[335] Peer-reviewed article, performance test[336] Conference paper, pre-post trial[337]
Diabetes Prevention App, New York University Langone Health[194] Smartphone app for prediabetic patients that provides personalized diet recommendations based on prediction of glycemic response Development: Field evaluation: Not yet published	Health recommendation Diabetes Glycemic response to future meal	Unspecified ML type Nonimagery sensor: Continuous glucose monitor Structured variables: Patient-entered food type, food weight, exercise, sleep, and activity 46 898 meal records from 800 patients	Adults (aged 18-80) who are overweight or obese with prediabetes Patient Home Smartphone	Clinical trial record, RCT[194,338]

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
Diagnostic AI for Pediatric Diseases, Guangzhou Women and Children's Medical Center[339] ML app that analyzes free text in EHR to diagnose common pediatric diseases Development: Field evaluation: Not yet published	Patient evaluation General: Pediatric diseases Presence of condition	Parametric Text: HER 101.6 million data points from 1 362 559 pediatric patient visits from 567 498 patients	Pediatric patients (aged 0-18) Health care professionals Outpatient	Peer-reviewed article, performance test[158]
eCART (electronic cardiac arrest triage), AgileMD[171] App that identifies hospital patients at risk of adverse events who may need transfer to ICU within next 8 hours In use, not FDA cleared	Patient evaluation Cardiovascular: Cardiac arrest General: Adverse events 8-hour risks of cardiac arrest, transfer to ICU, or death	Regression Structured variables: HER Data from more than 250 000 patients at 5 hospitals Unspecified training data	Hospitalized patients Health care professionals Inpatient	Peer-reviewed article, performance test[340] Peer-reviewed article, performance test[341] Peer-reviewed article, performance test[342] Peer-reviewed article, performance test[343]
ECG AI-Guided Screening, Mayo Clinic[344] App that analyzes ECG data to detect low ejection fraction Development: Field evaluation	Patient evaluation Cardiovascular: Low ejection fraction Presence of low ejection fraction	Parametric Nonimagery sensor: Electrocardiogram ECG and transthoracic echocardiogram data from 35 970 adult patients (≥ 18)	Adults (aged ≥ 18) Health care professionals Outpatient: Primary care	Peer-reviewed article, performance test[344] Peer-reviewed article, performance test[242]
Eko Analysis Software, Eko Health[145] App that analyzes patient heart sounds and ECG to detect heart murmurs and AF FDA cleared: Class II, 510(k)	Patient evaluation Cardiovascular: Arrythmia: AF Cardiovascular: Heart murmur Presence of AF or heart murmur; intermediate health measures	Parametric Nonimagery sensor: ECG Audio: Heart sounds 6 publicly available data sets plus 2 proprietary data sets including 375 patients	Adults (aged ≥18) Health care professionals Unspecified	FDA summary, performance test[345]

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
eMurmur AI, eMurmur[147] App that analyzes heart sounds to detect pathologic heart murmur FDA cleared: 510(k) (class information not available, FDA summary document blank)	Patient evaluation Cardiovascular: Heart failure Cardiovascular: Heart murmur Presence of pathologic heart murmur vs innocent or no murmur	Parametric Nonimagery sensor: Electronic stethoscope-recorded heart sounds Unspecified training data	Patients of all ages Health care professionals Inpatient; outpatient; home	Peer-reviewed article, performance test[346] Peer-reviewed article, performance test[347]
EnsoSleep, EnsoData[150] App that uses EEG-based neurological phenotypes to rapidly produce sleep scores FDA cleared: Class II, 510(k)	Patient evaluation Respiratory: Sleep apnea Presence of condition	Unspecified ML type Nonimagery sensor: EEG 4650 adult patients	Health status Health care professionals Outpatient	FDA summary, performance test[348]
EPIC Deterioration Index, EPIC Systems[177] Apps that identifies hospital patients at risk of adverse events who may need transfer to ICU In use, not FDA cleared	Patient evaluation Cerebrovascular: Cardiac arrest General: Adverse events Risks of cardiac arrest, transfer to ICU, severe sepsis, or death	Unspecified ML type Structured variables: HER Data from more than 130 000 patient encounters	Hospitalized patients not in the ICU Health care professional Inpatient	Academic preprint, performance test[349] Academic preprint, performance test[350] Gray literature, pre-post trial[351] Gray literature, pre-post trial[352]
FibriCheck, FibriCheck[138] Smartphone self-monitoring app to analyze heart rhythms and detect arrythmia FDA cleared: Class II, 510(k)	Patient evaluation Cardiovascular: Arrythmia, AF Presence of arrythmia or AF	Unspecified ML type Nonimagery sensor: Photoplesmography Unspecified training data	Adults diagnosed with or at risk of AF Health care professionals; patient Home Smartphone	Peer-reviewed article, implementation study[353] Peer-reviewed article, performance test[354,355]

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
FIND FH Algorithm, FIND FH (Family Hypercholesterolemia) Foundation[356] App that uses genetic samples to identify individuals with familial hypercholesterolemia (FH) Development: Field evaluation: Results published	Patient evaluation Cardiovascular: High cholesterol Presence of condition	Decision trees Structured variables: HER 939 individuals with FH and 83 136 individuals without FH	Adults with at least one cardiovascular comorbidity Health care professionals Outpatient	Peer-reviewed article, implementation study[284]
Ginger.io, Ginger[227] App that assesses patient mental health and provides alerts to mental health coach/therapist In use, not FDA cleared	Patient evaluation Mental health Need for mental health intervention	Unspecified ML type Patient-entered text and structured variables Unspecified training data	Adults (aged ≥ 18) Health care professionals Home Smartphone	No studies
HealthTap AI, HealthTap[165] Smartphone-based chatbot and symptom-checking app for broad range of health conditions In use, not FDA cleared	Health recommendations General: All conditions Presence of condition; need for clinical consultation	Conversational AI Text HealthTap repository of data on patient conditions and clinician-patient interactions	Adults and older adolescents (aged ≥ 16) Patient Home Smartphone	Gray literature, implementation study[298]
Heart Failure Medication Reminder App, Washington State University[42] Text messaging app that provides medication adherence reminders to patients with heart failure Development: Field evaluation: Not yet published	Health recommendations Cardiovascular Daily activities and activity transitions; medication reminder success/failure	Unspecified ML type Structured variables: Smartphone usage patterns Unspecified training data	Adults (aged ≥ 21) recently hospitalized for heart failure Patient Home Smartphone	Clinical trial record, pre-post trial[42]

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
Heart Failure Risk Calculator, MAGGIC[357] Online calculator to assess patient risk of heart failure In use, not FDA cleared	Patient evaluation Cardiovascular: Heart failure Risk of death	Regression Structured variables 39 372 patients with heart failure	Adults (aged ≥ 18) Health care professionals Online tool	Peer-reviewed article, performance test[358]
Heart Failure Treatment Gap Model, Geisinger Clinic[207] App that analyzes EHR data to identify gaps in care/treatment options for patients with heart failure Development: Field evaluation: Not yet published	Health recommendations Cardiovascular: Heart failure Optimal treatment	Unspecified ML type Structured variables: EHR Unspecified training data	Adults (aged ≥ 18) with heart failure Health care professionals Outpatient	Clinical trial record, RCT[207]
HeartHero AED, HeartHero[218] Portable AED that uses ML to detect cardiac arrest prior to delivering shock Development: Described as submitting for FDA clearance/approval in 2020	Treatment delivery Cardiovascular: Cardiac arrest Presence of cardiac arrest	Unspecified ML type Nonimagery sensor: ECG Unspecified training data	Unspecified Caregiver Home Portable AED device	No studies
Hypotension Prediction, University of Amsterdam[359] Early warning system for pending intraoperative hypotension Development: Field evaluation: Results published	Patient evaluation General Presence of condition	Unspecified ML type Nonimagery sensor Unspecified training data	General Health care professionals Inpatient	Peer-reviewed article, RCT[359]
Ibis, Senscio Systems[360] Tablet for at-home self-management of complex chronic conditions In use, not FDA cleared	Patient evaluation General Unspecified	Unspecified ML type Text; nonimagery sensor Unspecified training data	Multiple chronic conditions Patient Home Smartphone	No studies

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
iDEFCO, Centre Hospitalier Universitaire de Saint Etienne[186] App that uses patient-submitted questionnaire to identify cancer patients with fragile social support to prioritize social worker engagement Development: Field evaluation: Not yet published	Patient evaluation Cancer Composite measure of patient social fragility	Parametric Text and/or structured variables Unspecified training data	Adults (aged ≥ 18) with cancer Health care professionals Home Tablet	Clinical trial record, pre-post trial[186]
Intermountain Healthcare Readmission/Mortality Prediction, Intermountain Healthcare[172] ML algorithm for early identification of heart failure and readmission risk In use, not FDA cleared	Patient evaluation Cardiovascular Heart failure; hospital readmission; mortality	Unspecified ML type Structured variables: HER 16 971 hospitalization records	Hospitalized patients Health care professionals Inpatient	Peer-reviewed article, pre-post trial[172]
Jumpstart, University of Washington[185] App that uses EHR to identify seriously ill patients who may benefit from goals-of-care discussion; ML is used to detect prior goals-of-care discussions in HER Development: Field evaluation: Not yet published	Patient evaluation General: Serious illness Presence of documented goals-of-care discussions/advance directives	Unspecified ML type Structured variables: EHR; Text Unspecified training data	Seriously ill hospitalized patients Health care professionals Inpatient	Clinical trial record, RCT[185]

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
JVION Machine, JVION[174] App that analyzes EHR and patient socioeconomic data to predict risk of adverse events and recommend interventions In use, not FDA cleared	Patient evaluation General: Adverse events Risk of adverse events (readmission, sepsis, etc)	Decision trees Structured variables: EHR, patient socioeconomic data Data from 3 health systems including 138 115 adult and pediatric patients	All patients Health care professionals Inpatient; outpatient	Peer-reviewed article, pre-post trial[361] Conference paper, pre-post trial[41] Conference paper, pre-post trial[43] Conference paper, performance test[362]
K Health, K Health[166] Smartphone-based chatbot and symptom-checking app for broad range of health conditions In use, not FDA cleared	Patient evaluation General: All conditions Presence of condition	Conversational AI Patient-entered text; text and structured variables from HER More than 400 million clinical notes	All adults Patient Home Smartphone	Peer-reviewed articles, performance test[363]
Karantis360, Karantis360[236] App that uses multiple sensors to detect abnormal behavior among elderly adults, including those with dementia In use, not FDA cleared	Patient evaluation Dementia General: Elderly Presence of abnormal patient behavior	Unspecified ML type Nonimagery sensors: Temperature, movement, pressure Unspecified training data	Elderly adults, including with dementia or in assisted living facilities Health care professionals Home; inpatient	No studies
KardiaAI/Kardia Mobile, AliveCor[143] ECG processing and analysis software for arrythmia and AF detection FDA cleared: Class II, 510(k)	Patient evaluation Cardiovascular: Arrythmia: AF Presence of AF	Parametric Nonimagery sensor: Heart rate monitor Unspecified training data	Adults (aged ≥ 18) Patient Home Smartphone; other	Peer-reviewed article, RCT[364] Peer-reviewed article, implementation study[365] Peer-reviewed article, implementation study[366] Peer-reviewed article, performance test[367] Peer-reviewed article, performance test[368]

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
KDPI-EPTS Survival Benefit Estimator, Johns Hopkins School of Medicine[235,369] Online calculator to assess survival of kidney transplant recipients based on patient and donor profile In use, not FDA cleared	Patient evaluation Kidney disease 5-year survival of patient	Decision trees Structured variables 120 818 kidney transplant recipients	Kidney transplant candidates Health care professionals; patient Online tool	Peer-reviewed article, performance test[235]
KelaHealth, KelaHealth[175] App to assess risk of postsurgery complications Development: Field evaluation: Not yet published	Patient evaluation General: Adverse events 30-day postsurgery risk of complications including deep vein thrombosis and sepsis; length of hospital stay postsurgery	Parametric Structured variables: HER 4 million patient health records	Adults (aged ≥ 18) undergoing elective, high-risk colon, rectal, pancreas, or gastric surgery with predicted length of hospital stay ≥ 3 days Health care professionals Inpatient	Gray literature, performance test[370] Gray literature, performance test[371] Clinical trial record, pre-post trial[372]
Lark Diabetes Care, Lark[187] Diabetes management program that uses input from connected devices and from the patients themselves In use, not FDA cleared	Health recommendations Diabetes Unspecified	Unspecified ML type Text; nonimagery sensor Unspecified training data	Patients with diabetes Patient Home Smartphone	Gray literature, pre-post trial[373]
Lark DPP, Lark[192] Smartphone diabetes prevention program In use, not FDA cleared	Health recommendations Diabetes Unspecified	Unspecified ML type Text; nonimagery sensor Unspecified training data	Patients at risk of developing type 2 diabetes Patient Home Smartphone	Peer-reviewed article, pre-post trial[374] Gray literature, pre-post trial[375]

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
Lark for Hypertension, Lark[195] Hypertension management program that uses input from connected devices and from the patients themselves In use, not FDA cleared	Health recommendations Cardiovascular: Hypertension Unspecified	Unspecified ML type Text; nonimagery sensor Unspecified training data	Adults (aged ≥ 18) with hypertension Patient Home Smartphone	Peer-reviewed article, RCT[376] Gray literature, pre-post trial[377]
Loop System, Spry[182] App to assess risk of patient deterioration using wearable monitor in the home environment FDA cleared: Class II, 510(k)	Patient evaluation Cardiovascular: Congestive heart failure Respiratory: Chronic obstructive pulmonary disease Respiratory: COVID-19 General: Adverse events Risk of patient deterioration	Unspecified ML type Nonimagery sensor: Accelerometer Nonimagery sensor: Photoplesmography Unspecified training data	Adults Health care professionals; patient Home Wearable	No studies
MATRx Plus, Zephyr[219] Oral testing appliance to treat sleep apnea that uses ML for optimal self-adjustment FDA Approved: Class II, De Novo	Treatment delivery Respiratory: Sleep apnea Likelihood of patient benefit from therapy; optimal position of device	Decision trees Nonimagery sensor: Respiratory airflow, oxygen saturation 149 patients	Adult patients with obstructive sleep apnea Patient; health care professionals Home Wearable	Peer-reviewed article, performance test[378] Conference paper, performance test[379] Clinical trial record, implementation study[380]
Medical Early Warning Score ++, Icahn School of Medicine at Mount Sinai[178] App that identifies hospital patients at risk of adverse events who may require escalation of care within next 6 hours Development: Field evaluation: Not yet published	Patient evaluation General 6-hour risk of patient deterioration	Unspecified ML type Structured variables: HER 157 984 hospital encounters and 244 343 bed movements from 96 645 patients	Hospitalized adults (aged ≥ 18) not in ICU Health care professionals Inpatient	Clinical trial record, pre-post trial[178] Clinical trial record, performance test[178]

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
Minimed 780G/MD-Logic Artificial Pancreas, Medtronic[215] Artificial pancreas Development: Field evaluation: Results published, further trials ongoing	Treatment delivery Diabetes Treatment: Dosage	Unspecified ML type Nonimagery sensor: Continuous glucose monitor Unspecified training data	Children and adults (aged ≥ 7) with type 1 diabetes receiving insulin treatment Patient Diabetes camp Wearable	Peer-reviewed article, RCT[301] Peer-reviewed article, RCT[381] Peer-reviewed article, RCT[382] Peer-reviewed article, pre-post trial[44]
Nectarine Health, Nectarine Health[383] Cloud-supported wearable for the detection of emergencies and behavioral changes in frail population In use, not FDA cleared	Patient evaluation Dementia General: Elderly Unspecified	Unspecified ML type Nonimagery Sensor Unspecified training data	Elderly adults Health care professionals Assisted living Wearable	No studies
Neuro Motor Index, Altoida[157] App that analyzes wide variety of patient data to detect Alzheimer's disease Development: Field evaluation: Results published	Patient evaluation Dementia Potential presence of Alzheimer's disease	Unspecified ML type Structured variables: Cognitive exam Nonimagery sensor: EEG Audio: Human voice Genetic Imagery 215 adults	Adults (aged 55-90) Health care professionals Outpatient; home Smartphone; Tablet	Peer-reviewed article, performance test[384]
Omada, Omada Health[199] Coaching platform for management and prevention of chronic conditions In use, not FDA cleared	Health recommendations Diabetes; hypertension; mental health Unspecified	Unspecified ML type Text Unspecified training data	Health status Patient Home Smartphone	No studies

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
One Drop, One Drop[189] Smartphone app that predicts blood glucose levels in next 12 hours and suggests preventive actions In use, not FDA cleared (while ML algorithm is not FDA cleared, it is used in conjunction with an FDA-cleared sensor device)	Health recommendations Diabetes Patient blood glucose levels in the next 12 hours; suggested actions to avoid glycemic event	Decision trees Nonimagery sensor: Glucose monitor Structured variables More than 1.1 billion data points collected by more than 860 000 users	Patients with diabetes Patient Home Smartphone	Peer-reviewed article, pre-post trial[385] Conference paper, performance test[386] Conference paper, performance test[387]
Optima 4 Blood Pressure, Optima Integrated Health[205] App that analyzes EHR and blood pressure readings to recommend optimal treatment for hypertension Development: Field evaluation: Not yet published	Health recommendations Cardiovascular: Hypertension Optimal treatment type and dosage	Unspecified ML type Structured variables: HER Nonimagery sensor: Blood pressure Unspecified training data	Adults (aged 21-80) taking medication to treat hypertension Health care professionals Home N/A	Clinical trial record, RCT[205]
Owlytics, Owlytics Healthcare[234] Wearable app that assesses risk of patient falling and risk of patient deterioration, and detects falls In use, not FDA cleared	Patient evaluation Dementia General: Elderly Risk of patient deterioration; risk of fall; presence of patient fall	Unspecified ML type Nonimagery sensor: Heart rate, movement Unspecified training data	Elderly adults Health care professionals; patient Home; inpatient Wearable	No studies
Pathwork Tissue of Origin Test, Pathwork Diagnostics[155] App that identifies tumor type based on RNA in tissue sample FDA cleared: Class II, 510(k), no longer in use	Patient evaluation Cancer Type of tumor	Parametric Genetic: Biopsy tissue Data from 2039 tumor specimens	Unspecified Health care professionals Unspecified N/A	FDA summary, performance test[155] FDA summary, performance test[388]

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
PD_Manager, PD_Manager[389] Wearable monitoring device and app that uses ML to detect change in motor control and cognitive symptoms of Parkinson's disease Development: Field evaluation: Not yet published	Patient evaluation Dementia Change in motor control or cognitive symptoms	Unspecified ML type Patient-entered text and/or structure variables Nonimagery sensor: Accelerometer, insole pressure sensors, smart pillbox Audio: Human speech Unspecified training data	Adults (aged ≥ 18) with Parkinson's disease Health care professionals; patient; caregiver Home Smartphone; wearable	Peer-reviewed article, implementation study[245] Clinical trial record, implementation study[390]
Pediatric Symptom Checker, Shanghai Jiao Tong University School of Medicine[391] Smartphone app to check pediatric patient symptoms, recommend tests, and suggest diagnosis Development: Field evaluation: Not yet published	Health recommendations General: All conditions Presence of condition	Unspecified ML type Structured variables; patient-entered text Unspecified training data	Pediatric patients Patient Outpatient Smartphone	Clinical trial record, RCT[391]
PhysIQ Personalized Physiology Engine, PhysIQ[138,139] App that analyzes ECG data to detect arrythmia and identify changes to patient vital signs FDA cleared: Class II, 510(k)	Patient evaluation Cardiovascular: Arrythmia: AF General: Adverse events Measures of patient heart rate, respiratory rate; presence of AF; composite measure of physiological change	Unspecified ML type Nonimagery sensor: ECG, accelerometer Unspecified training data	Adult patients Health care professionals Home; inpatient; outpatient N/A	No studies

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
Preventice BeatLogic Platform, Preventice Solutions[392] DL model that interprets and classifies ECG In use, not FDA cleared	Patient evaluation Cardiovascular: Arrythmia, AF Presence of arrythmia, including AF	Parametric Nonimagery sensor: Electrocardiogram Various sized data sets (5000-21 000 data records)	Health status Health care professionals Outpatient Wearable	Peer-reviewed article, performance test[393] Conference paper, performance test[394] Conference paper, performance test[395]
Qventus, Qventus[396] Complex platform for the automation of hospital patient flow In use, not FDA cleared	Patient evaluation General Deterioration; resource utilization	Unspecified ML type EHR Unspecified training data	Hospital patients Health care professionals Inpatient N/A	Gray literature, pre-post trial[397,398]
Radiation Therapy Risk Algorithm, Duke University[399] App that identifies patients receiving radiation therapy who are at high risk for readmission Development: Field evaluation: Results published	Patient evaluation Cancer General: Adverse events Risk of hospital readmission	Unspecified ML type Structured variables: HER 8134 records of radiation therapy courses for 6879 patients	Adults (aged ≥ 18) undergoing radiation therapy for cancer Health care professionals Outpatient N/A	Academic preprint, RCT[400]
rapid Whole Genome Sequencing, Rady's Children's Institute for Genomic Medicine[153] App that uses ML and clinical natural language processing to diagnose rare genetic diseases In use, not FDA cleared	Patient evaluation General: Genetic condition Presence of condition	Unspecified ML type Genetic: Blood Text: EHR Unspecified training data	Children with suspected genetic diseases Health care professionals Inpatient N/A	Peer-reviewed article, implementation study[401]
Rare Disease Auxiliary Diagnosis System, Center of Bioinformatics in East China Normal University[152] Online tool that uses list of clinical phenotypes to diagnose rare disease In use, not FDA cleared	Patient evaluation General: Rare diseases Presence of potential rare disease	Unspecified ML type Structured variables Between 44 000 and 420 000 disease-phenotype associations	Patients with potential rare diseases Health care professionals Online tool N/A	Peer-reviewed article, performance test[402]

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
REACH VET, Department of Veterans Affairs (VA)[183] App that identifies veterans at high risk of suicide by using EHR records In use, not FDA cleared	Patient evaluation Mental health: Suicide General: Adverse events Risk of suicide; risk of hospitalization or other adverse event	Unspecified ML type Structured variables: EHR Unspecified training data	Veterans in VA health care system Health care professionals Unspecified N/A	Clinical trial record, implementation study[403]
ResApp, ResApp Health[151] App that diagnoses and measures the severity of respiratory diseases by analyzing cough and breathing sounds In use, not FDA cleared	Patient evaluation Respiratory: Asthma, COPD, COVID-19 Presence or severity of respiratory condition	Unspecified ML type Audio: Cough sound Unspecified training data	Patients with potential respiratory condition Health care professionals Clinical N/A	Peer-reviewed article, performance test[404] Conference paper, performance test[405] Conference paper, performance test[406]
Rhythm Express RX-1 MDSP Technology, VivaQuant[228] App that analyzes ECG data to detect various types of arrythmia FDA cleared: Class II, 510(k)	Patient evaluation Cardiovascular: Arrythmia: AF Cardiovascular: Heart rhythm Classification of ECG noise vs ECG signal; presence and type of arrythmia	Unspecified ML type Nonimagery sensor: ECG 80 000 hours of ECG data	Adult patients at risk of heart disease Health care professionals Home Wearable	No studies
RhythmAnalytics, Biofourmis[141] App that analyzes ECG to detect arrythmia FDA cleared: Class II, 510(k)	Patient evaluation Cardiovascular: Arrythmia Presence of arrythmia	Parametric Nonimagery sensor: ECG 121 346 ECG records	Adults (aged ≥ 18) Health care professionals Unspecified N/A	Gray literature, performance test[407]
Rose platform, Rose[197] Smartphone app that assesses patient mental health and suggests therapy Development: Field evaluation: Results published	Health recommendations Mental health: Depression, anxiety Presence of depression, anxiety, or other mental health condition	Unspecified ML type Patient-entered text Unspecified training data	All adults Patient; health care professionals Home Smartphone	Peer-reviewed article, implementation study[408] Clinical trial record, RCT[197]

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
Seattle Heart Failure Model, University of Washington[173] Online calculator to predict heart failure patient survival at baseline and for various interventions In use, not FDA cleared	Patient evaluation Cardiovascular: Heart failure Likelihood of 1-, 2-, and 3-year survival at baseline and for various interventions	Regression Structured variables 1125 heart failure patients	Adults (aged ≥ 18) with heart failure Health care professionals; patient General Online tool	Peer-reviewed article, performance test[409]
Sepsis Prediction Algorithm, Dascena[230] Early sepsis prediction algorithm Development: Field evaluation	Patient evaluation General: Sepsis Presence of condition	Unspecified ML type Structured variables: EHR Unspecified training data	General Health care professionals Inpatient N/A	Peer-reviewed article, RCT[240] Peer-reviewed article, pre-post trial[238] Peer-reviewed article, pre-post trial[239] Peer-reviewed article, performance test[241] Clinical trial record, RCT[410] Clinical trial record, RCT[247]
Sepsis Watch, Duke University[229] App that monitors EHR data to detect early signs of sepsis In use, not FDA cleared	Patient evaluation General: Sepsis Risk of sepsis	Parametric Structured variables: EHR 50 000 patient records	Adults (aged ≥ 18) receiving emergency department care Health care professionals Emergency department N/A	Peer-reviewed article, implementation study[262] Peer-reviewed article, performance test[411] Clinical trial record, pre-post trial[412]
Short Arm Human Centrifuge Rehab, Greek Aerospace Medical Association and Space Research[210] App that uses AI to analyze patient EEG data, to determine optimal centrifuge training Development: Field evaluation: Not yet published	Health recommendations Cerebrovascular: Poststroke Respiratory: COPD General: Elderly Optimal centrifuge setting	Parametric Nonimagery sensor: EEG Unspecified training data	Adults (aged 17-90), including with recent stroke, COPD, or elderly Health care professionals Outpatient N/A	Clinical trial record, RCT[210]

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
Sinedie, Universidad Politécnica de Madrid[190] App that provides diet and insulin recommendations for diabetic patients Development: Field evaluation: Not yet published	Health recommendations Diabetes Patient metabolic condition (supervised ML); patient mealtimes (unsupervised ML)	Decision trees Nonimagery sensor: Glucose monitor Structured variables 7113 glycemia measurements from 42 patients	Patients with diabetes Patient; health care professionals Home Online tool	Peer-reviewed article, RCT[190]
Smart Angel, Evolucare[413] App that analyzes patient-entered and sensor data to assess risk of deterioration following surgery Development: Field evaluation: Not yet published	Patient evaluation General: Adverse events Risk of patient deterioration	Unspecified ML type Patient-entered text and/or structured variables Nonimagery sensors: Heart rate, blood pressure, oxygen saturation	Adults (aged 18-80) who have undergone outpatient surgery Health care professionals Home Tablet	Clinical trial record, RCT[414]
Smoking Cessation App, University of Hong Kong[198] Text messaging chatbot that makes behavioral recommendations as part of smoking cessation treatment Development: Field evaluation: Not yet published	Health recommendations Substance abuse: Smoking Unspecified	Conversational AI Patient-entered Text Unspecified training data	Adults (aged ≥ 18) who smoke daily Patient Home Smartphone	Clinical trial record, RCT[198]
SOPHiA GENETICS, SOPHiA GENETICS[415] Genomic platform to detect and characterize genomic variants associated with cancers and hereditary disorder In use, not FDA cleared	Patient evaluation Cancer Presence of condition: Tumor type	Unspecified ML type Genetic: Biopsy tissue Unspecified training data	Health status Health care professionals Clinical N/A	No studies

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
Steth IO Software, StethIO[148] Smartphone app that uses heart sounds to detect heart murmur In use, not FDA cleared (while ML algorithm is not FDA cleared, it is used in conjunction with an FDA-cleared sensor device)	Patient evaluation Cardiovascular: Heart murmur Presence of heart murmur	Parametric Audio: Heart sounds Unspecified training data	Adults Health care professionals Inpatient; outpatient Smartphone	No studies
Sugar.IQ, Medtronic[237] App that helps patients manage their glucose levels In use, not FDA cleared (while ML algorithm is not FDA cleared, it is used in conjunction with an FDA-cleared sensor device)	Patient evaluation Diabetes Biomarker: Glucose level	Unspecified ML type Nonimagery sensor: Continuous glucose monitor Unspecified training data	Health status Patient Home Smartphone	Conference paper, pre-post trial[416] Conference paper, pre-post trial[417] Gray literature, pre-post trial[418]
Symptomate, Infermedica[159] Online symptom checker In use, not FDA cleared	Patient evaluation General Presence of condition	Parametric Patient-entered information Unspecified training data	General Patient Home Web	Peer-reviewed article, performance test[160] Gray literature, implementation study[298]
t2.coach, Columbia University[191] Smartphone app that uses AI to provide diet and other treatment recommendations to patients with type 2 diabetes Development: Field evaluation: Not yet published	Health recommendations Diabetes Optimal patient health recommendations	Unspecified ML type Patient-entered text and structured variables, including self-reported blood glucose Nonimagery sensor: Heart rate, movement Unspecified training data	Adults (aged 18-65) with type 2 diabetes Patient Home Smartphone	Clinical trial record, RCT[191]

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
Targeted Real-time Early Warning Score, Bayesian Health[169] App that identifies ICU patients at risk of septic shock In use, not FDA cleared	Patient evaluation General: Sepsis Risk of developing septic shock	Unspecified ML type Structured variables: EHR 13 014 ICU patients	ICU patients Health care professionals ICU N/A	Peer-reviewed article, performance test[169]
Tempus Oncology Testing, Tempus Labs[156] App that performs genomic profiling and prediction of cancer origin In use, not FDA cleared	Patient evaluation Cancer Presence of condition: Tumor type	Unspecified ML type Genetic: Biopsy tissue Various data sets (10 000-25 000 records)	Health status Health care professionals Clinical N/A	Peer-reviewed article, performance test[419] Conference paper, performance test[419,420]
Tess, X2[38] Text messaging chatbot that delivers mental health therapy In use, not FDA cleared	Treatment delivery Mental health: Depression, anxiety Unspecified	Conversational AI Text Unspecified training data	All patients Patient Home Smartphone	Peer-reviewed article, RCT[421] Peer-reviewed article, implementation study[422]
Virta, Virta Health[39] App that provides personalized coaching for diabetes management In use, not FDA cleared	Health recommendations Diabetes Unspecified	Unspecified ML type Biomarkers; text Unspecified training data	Health status Patient Home Smartphone	Peer-reviewed article, pre-post trial[244] Peer-reviewed article, pre-post trial[423]
VITEK MS, bioMerieux[135] Mass spectrometer that identifies microorganisms cultured from human specimens FDA approved: Class II; De Novo	Patient evaluation General: Bacterial and fungal infections Identity of microorganism	Unspecified ML type Nonimagery sensor: Mass spectrometry Unspecified training data	Unspecified Health care professionals Inpatient; outpatient N/A	Peer-reviewed article, performance test[424] FDA summary, performance test[425] FDA summary, performance test[426]

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
Warfarin Dosage App, Wuhan Asia Heart Hospital[200] Chatbot that aids in managing optimal warfarin dosage Development: Field evaluation: Not yet published	Health recommendations Cerebrovascular; cardiovascular Optimal dosage of warfarin	Conversational AI Unspecified data input Unspecified training data	Adults (aged 18-65) with AF and mechanical valve replacement who are receiving warfarin therapy Patient Home Mobile phone	Clinical trial record, RCT[200]
Watson for Oncology and Genomics, IBM[208] App that uses ML to extract therapeutic information from peer-reviewed studies relevant to specific genetic markers identified in patient tests In use, not FDA cleared	Health recommendations Cancer Relevance of therapeutic information	Unspecified ML type Geneti Text: Peer-reviewed research Unspecified training data	Patients with cancer Health care professionals Inpatient; outpatient N/A	Peer-reviewed article, implementation study[427] Peer-reviewed article, performance test[428]
WAVE Clinical Platform: Visensia, the Safety Index, Excel Medical[176] App that identifies hospitalized patients at risk of deterioration FDA cleared: Class II, 510(k)	Patient evaluation General: Adverse events; cardiovascular Risk of patient deterioration	Unspecified ML typeStructured variables: EHR; nonimagery sensors 3500 hours of high-risk, in-hospital patient monitoring	All hospitalized patients Health care professionals Inpatient N/A	Peer-reviewed article, pre-post trial[429] Peer-reviewed article, performance test[430]
Wellthy Diabetes, Wellthy Therapeutics[188] Smartphone chatbot app that evaluates diabetic patient health and provides behavioral coaching Development: Field evaluation	Health recommendations Diabetes Unspecified	Conversational AI Structured variables: Self-reported blood glucose, weight, meals, and physical activity Unspecified training data	Patients with type 2 diabetes Patient Home Smartphone	Gray literature, pre-post trial[243]

Application, developer Summary Current status	Function Health condition Predicted variable	ML type Data input Training data	Target population User Setting Platform	Evaluation studies
Woebot, Woebot Health[214] Smartphone chatbot app that delivers mental health therapy In use, not FDA cleared	Treatment delivery Mental health: Depression, anxiety Unspecified	Conversational AI Patient-entered Text Unspecified training data	Adults and adolescents (aged ≥ 13) Patient Home Smartphone	Peer-reviewed article, RCT[431]
Wysa, Wysa[213] Smartphone chatbot app that delivers mental health therapy In use, not FDA cleared	Treatment delivery Mental health: Depression, anxiety Unspecified	Conversational AI Text Unspecified training data	Adolescents and adults (aged ≥13) Patient Home Smartphone	Peer-reviewed article, pre-post trial[432]
Your.MD, Healthily[161] Online symptom checker In use, not FDA cleared	Patient evaluation General Presence of condition	Unspecified ML type Text Unspecified training data	General Patient Home Smartphone	Peer-reviewed article, performance test[162] Gray literature, implementation study[298]
Zio AT ECG Monitoring System, iRhythm Technologies[140] App that analyzes ECG data to detect arrythmia FDA cleared: Class II, 510(k)	Patient evaluation Cardiovascular: Arrythmia Presence of arrythmia	Parametric Nonimagery sensor: ECG Unspecified training data	Adults (aged ≥ 18) Health care professionals Home; inpatient Wearable	Peer-reviewed article, RCT[433] Peer-reviewed article, RCT[434] Peer-reviewed article, performance test[435]

Appendix B. Full List of Evaluation Studies

We identified 173 evaluation studies providing information on current and near-future ML applications in clinical care. Details on these studies are provided in the table below. These studies are organized alphabetically by the application they discuss, beginning with Ada and ending with Zio ATG.

Multiple applications are listed for the very small number of studies that discussed multiple applications. These receive a single table entry just like other studies and are grouped with the first application in the list of those covered.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Miller et al, 2020[296] Ada Health recommendations General: All conditions	Peer-reviewed article Field evaluation: Implementation study 522 patients in primary care setting (n = 522)	No comparator Patient perceptions of app usability and acceptability	Patients rated the app as highly usable and acceptable in a primary care setting.
UserTesting, 2019[298] Ada, HealthTap, Symptomate, Your.MD Health recommendations General: All conditions	Gray literature Field evaluation: Implementation study 500 health care consumers (n = 500)	5 apps were compared: Ada Health, HealthTap, Mediktor, Symptomate, Your.MD User perceptions of apps' ease of use, credibility, aesthetics, efficiency, delight	Ada scored highest across all categories, followed closely by Mediktor, Your.MD, and Symptomate. HealthTap scored much lower than all other apps.
Jungman et al, 2020[248] Ada Health recommendations General: All conditions	Peer-reviewed article Field evaluation: Pre-post trial 141 students participated (n = 141)	Google search; no symptom search (control) 1-day intervention period; patient affect; health anxiety; perceived need to see a doctor	Ada Search was significantly associated with greater negative affect, health anxiety, and perceived need to see a doctor compared with not searching for symptoms. Google Search was significantly associated with greater negative affect, health anxiety, and perceived need to see a doctor compared with the control, to a greater extent than Ada was.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Gilbert et al, 2020[297] Ada, Babylon, Buoy, K Health, Symptomate, Your.MD Health recommendations General: All conditions	Academic preprint Performance test: Simulated 200 clinical vignettes and 7 clinicians	Apps were compared with each other and with clinicians Accuracy as measured by percentage of true diagnoses included in the top 3 diagnoses identified by the app; safety of urgency advice; number of conditions covered by an app	Top-3 accuracy was highest for clinicians (82.1%), followed by Ada, Buoy, Mediktor, K Health, Your.MD, Symptomate, and Babylon. Most apps provided accurate or slightly conservative advice on clinical urgency. Some apps (Babylon, Symptomate, Your.MD) covered only 50% to 65% of conditions.
Burgess, 2017[299] Ada Health recommendations General: All conditions	Gray literature Performance test: Simulated 2 researchers conducted an ad hoc test of 3 apps	Your.MD; Babylon Accuracy	Ada's accuracy was found to be better than that of Your.MD and Babylon in an ad hoc test.
Jungmann et al, 2019[295] Ada Health recommendations General: All conditions	Peer-reviewed article Performance test: Simulated 120 evaluations (6 users independently evaluated 20 case vignettes)	No comparator Accuracy	Average kappa was 0.64 for adult cases and 0.40 for child and adolescent cases. Kappa was higher when Ada was used by health professionals and lower when used by health students or laypersons.
Taipei Medical University, 2018[300] Advanced Electronic Safety of Prescriptions model Patient evaluation General: All conditions	Clinical trial record Field evaluation: RCT 37 physicians, unspecified number of patients	Standard care 3-month intervention period; proportion of inappropriate prescription reminders that are acted on; number of inappropriate prescriptions	Study ongoing/not yet published
Ziegler et al, 2015[44] Advisor Pro/MD-Logic Health recommendations Diabetes	Peer-reviewed article Field evaluation: Implementation study 75 patients with diabetes at home (n = 75)	Preintervention 4-day intervention period; patient fear of hypoglycemia, patient satisfaction, and ease of use	After 4 days of using MD-LOGIC as a closed-loop artificial pancreas at home at night, patients reported lower fears of hypoglycemia. Patients reported a high level of satisfaction with the device and increased ease of use.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Nimri et al, 2018[303] Advisor Pro/MD-Logic Health recommendations Diabetes	Peer-reviewed article Performance test: Retrospective Glucose monitoring and health data from 15 pediatric patients with diabetes taken over a 3-week period (n = 15)	Insulin dosage recommendations from 26 physicians Agreement on insulin dosage between physicians and Advisor Pro	Insulin dosage recommended by Advisor Pro did not significantly differ from recommendations provided by physicians.
Nimri et al, 2014[302] Advisor Pro/MD-Logic Health recommendations Diabetes	Peer-reviewed article Field evaluation: RCT 15 patients (both children and adults) with diabetes (n = 15)	Standard Sensor-Augmented Pump therapy 8-day intervention period; time spent with glucose levels below 70 mg/dL (hypoglycemia); percentage of nights with average glucose levels between 90 and 140 mg/dL	Use of MD-Logic as a closed-loop artificial pancreas at night was associated with less time spent with glucose below 70 mg/dL (hypoglycemia). No difference was found in the percentage of nights with average glucose between 90 and 140 mg/dL.
Nimri et al, 2014[301] Advisor Pro/MD-Logic Health recommendations Diabetes	Peer-reviewed article Field evaluation: RCT 24 patients (aged 12-43) with type 1 diabetes (n = 24)	Standard sensor-augmented pump therapy 6-week intervention period; time spent in hypoglycemia (<70 mg/dL), within target range (70-140 mg/dL), and in substantial hyperglycemia (>240 mg/dL)	Use of MD-Logic as a closed-loop artificial pancreas at home at night was associated with lower hypoglycemia, more time in target range, and 52% less time spent in substantial hyperglycemia.
Naunheim et al, 2011[304] Ahead Patient evaluation General: Traumatic brain injury	Peer-reviewed article Performance test: Prospective 153 patients who presented to a tertiary referral hospital with headache or altered mental status (n = 153)	No comparator Accuracy	Ahead prototype had a sensitivity of 96% and specificity of 87%.
US FDA, DEN140025, 2014[305] Ahead Patient evaluation General: Traumatic brain injury	FDA summary document Performance test: Prospective 552 adults aged 18-80 years (n = 552)	No comparator Accuracy	Ahead 100 had a sensitivity of 78.5% and specificity of 48.6%.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
US FDA, K161068, 2016[306] Ahead Patient evaluation General: Traumatic brain injury	FDA summary document Performance test: Prospective 720 adults; 60.7% males, mean age of 43.12 years, mean Glasgow Coma Scale score of 14.97, and an average time since injury of 13.9 hours (n = 720)	No comparator Accuracy	Ahead 300 accuracy, including sensitivity and specificity, exceeded goals (no specific numbers reported).
Shanghai Zhongshan Hospital, 2020[204] AI-Assisted Insulin Titration System Health recommendations Diabetes	Clinical trial record Field evaluation: RCT 46 adults (aged 18-75) with type 2 diabetes receiving insulin treatment for at least 3 months (n = 46)	Insulin regimen decided by physicians 7-day intervention period; percentage of time of sensor glucose measurements in target range (3.9-10 mmol/L); total insulin dose	Study ongoing/not yet published
Shanghai Zhongshan Hospital, 2020[307] AI-Assisted Insulin Titration System Health recommendations Diabetes	Clinical trial record Field evaluation: RCT 120 adults (aged 18-99) with type 2 diabetes receiving insulin treatment for at least 3 months (n = 120)	Insulin regimen decided by physicians 7-day intervention period; patient fasting plasma glucose levels; proportion of time glucose measurements are in targeted range; total insulin dose	Study ongoing/not yet published
Farris, 2017[308] AIM@BP Health recommendations Cardiovascular: Hypertension	Gray literature Field evaluation: RCT 45 patients with low medication possession ratio for their antihypertensive medication (n = 45)	Standard care 6-month intervention period; self-reported medication adherence; proportion of days covered; attitudes toward medication	The app led to an increase in self-reported adherence at 3 months; no significant changes in proportion of days covered or attitudes toward medication were found.
Barbieri et al, 2016[40] Anemia Control Model Health recommendations Kidney disease; anemia	Peer-reviewed article Field evaluation: Pre-post trial 752 patients with kidney disease on hemodialysis (n = 752)	Preintervention health status 12-month control followed by 12-month intervention period; proportion of patients with hemoglobin in target range; hemoglobin fluctuations; medication dosage	During the intervention period, patients had significant reductions in medication dosage, significant reductions in hemoglobin variability, and significant increases in on-target hemoglobin values.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Vifor Fresenius Medical Care Renal Pharma, 2020[203] Anemia Control Model Health recommendations Kidney disease; anemia	Clinical trial record Field evaluation: Pre-post trial 88 adult patients with end-stage renal disease (n = 88)	Preintervention health status 6-month intervention period; proportion of patients with hemoglobin in target range; hemoglobin fluctuations; medication dosage	Study ongoing/not yet published
Bucalo et al, 2018[309] Anemia Control Model Health recommendations Kidney disease; anemia	Peer-reviewed article Field evaluation: Pre-post trial 219 patients with kidney disease on hemodialysis (n = 219)	Patient health status during 6 months of control period between two 6-month intervention periods 6-month intervention, then 6-month control, then 6-month intervention period; proportion of patients with hemoglobin in target range; hemoglobin fluctuations; medication dosage	The interventional periods saw patient improvement in all measures: increased hemoglobin in range, reduced hemoglobin, and fewer adverse events (transfusion, hospitalization, cardiac arrest, death).
Barbieri et al, 2015[310] Anemia Control Model Health recommendations Kidney disease; anemia	Peer-reviewed article Performance test: Retrospective 101 918 records from 4135 patients (29% with diabetes) at Italian, Spanish, and Portuguese clinics (n = 4135)	Application model compared with a different neural network ML model and with a linear ML model Accuracy	The new model outperformed other models with a mean absolute error of hemoglobin prediction of about 0.6 g per dl.
Wang et al, 2019[311] Anticoagulation Management Service Patient evaluation Cardiovascular: Arrythmia: AF Cerebrovascular: Stroke	Peer-reviewed article Field evaluation: RCT 432 patients identified as high risk without evidence of anticoagulant prescription in the prior year (n = 432)	Usual care Proportion of potentially undertreated patients with AF started on anticoagulant therapy within 28 days	No increases in new anticoagulation prescriptions due to the intervention were found.
Case Western Reserve University, 2018[212] ASSIST Treatment delivery Mental health: Depression, anxiety	Clinical trial record Field evaluation: RCT 19 adults providing care for critically ill family member (n = 19)	Same wearable device without algorithm-generated reminders Various measures of patient well-being (sleep, stress, anxiety, depression, caregiving burden, quality of life)	Study ongoing/not yet published

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Camlin, Ltd, 2020[211] Assisted Rehabilitation Care Health recommendations Cerebrovascular: Poststroke rehabilitation	Clinical trial record Field evaluation: Pre-post trial 41 adults diagnosed with stroke (n = 41)	Preintervention Patient adherence, resource utilization (unscheduled visits to clinician); patient quality of life; patient mobility; device-related adverse events	Study ongoing/not yet published
Bristol-Myers Squibb, 2019[146] Atrial Fibrillation Risk Prediction Patient evaluation Cardiovascular: Arrythmia: AF	Clinical trial record Field evaluation: RCT 18 000 adults older than 30 years of age (n = 18 000)	Routine clinical practice Percentage of patients with diagnosed AF; expected health care resource utilization; quality-adjusted life-years; life-years	Study ongoing/not yet published
Ipsos MORI, 2019[246] Babylon Health Health recommendations General	Gray literature Field evaluation: Implementation study 49 000 patients in Babylon GP at Hand system (n = 49 000)	Video and telephone consultation with a clinician Patient use of app, patient perceptions of app	The symptom checker app was used by 55% of patients. Use of the app declined over time since registration. The app was viewed less positively than were video and telephone appointments with a clinician.
Babylon Health, 2017[312] Babylon Health Health recommendations General	Gray literature Field evaluation: Implementation study 9700 patient triage recommendations (n = 9700)	Prior care Estimated cost savings; number of app triage recommendation outputs in each category	Britain's National Health Service (NHS) saves an estimated £10 each time a patient opts to share app triage outcomes with the NHS.
Middleton et al, 2016[313] Babylon Health Health recommendations General	Gray literature Performance test: Simulated 102 patient vignettes	Triage recommendations from 12 clinicians and 7 nurses Accuracy; time to make triage recommendation	The app provided more accurate triage recommendations than did clinicians or nurses in both emergency and nonemergency cases. The app also provided triage recommendations more quickly.
Razzaki et al, 2018[314] Babylon Health Health recommendations General	Gray literature Performance test: Simulated Clinicians role-playing as patients in 100 clinical vignette scenarios	Clinicians Accuracy of diagnosis and triage recommendations	The app provided comparably accurate diagnoses and provided safer (equal or more urgent) triage recommendations than did clinicians.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
El-Khatib et al, 2017[315] Bionic Pancreas Treatment delivery Diabetes	Peer-reviewed article Field evaluation: RCT 39 adults with type 1 diabetes (n = 39)	Conventional and sensor-augmented pump insulin therapy 2-month intervention period; mean patient glucose concentration; mean time with glucose < 3.3 mmol/L; mean nausea score on the Visual Analogue Scale	Patient glycemic regulation was improved with use of the bionic pancreas.
Massachusetts General Hospital, 2019[316] Bionic Pancreas Treatment delivery Diabetes	Clinical trial record Field evaluation: RCT 129 adults (aged ≥ 18) with cystic fibrosis–related diabetes who are receiving insulin treatment (n = 129)	Patient's usual insulin therapy 7-day intervention period; average patient glucose; time with patient glucose < 54 ml/dL; time with patient glucose in multiple numerical ranges; number of self-reported experiences of hypoglycemia symptoms	Study ongoing/not yet published
Jaeb Center for Health Research, 2020[317] Bionic Pancreas Treatment delivery Diabetes	Clinical trial record Field evaluation: RCT 440 adults and children (aged ≥ 6) with type 1 diabetes who have been receiving insulin treatment for at least 1 year (n = 440)	Patient's usual insulin therapy 13-week intervention period; superiority for patient HbA_{1c}; time with patient glucose < 54 mg/dL	Study ongoing/not yet published
US FDA, K183282, 2019[318] Biovitals Analytics Engine/Biovitals HF Patient evaluation General: Adverse events	FDA summary document Performance test: Prospective 50 patients released from emergency department; testing was performed on a total of 50 subjects presenting at an emergency department who were deemed appropriate for home monitoring (n = 50)	No comparator Accuracy	The lower bound of the 95% confidence interval of the positive percentage agreement (PPA) was greater than 0.7.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Chen et al, 2019[319] Biovitals Analytics Engine/Biovitals HF Patient evaluation General: Adverse events	FDA summary document Performance test: Retrospective 50 healthy adults (n = 50)	No comparator Accuracy in identifying period of physiological change (subjects in hypobaric chamber to simulate activity at 3000 m altitude)	When using a specific classification threshold, AUC was 0.99. sensitivity was 0.91, and specificity was 0.98.
IBM Watson Health, 2018[321] BlueStar Health recommendations Diabetes	Gray literature Field evaluation: Implementation Data from more than 3000 patients using BlueStar (n = 3000)	Prior care Estimated cost savings associated with reductions in patient HbA_{1c}	Use of BlueStar was estimated to save between $1392 and $3672 annually per Medicare patient and between $1824 and $5244 annually per commercial sector patient.
Desveaux et al, 2018[46] BlueStar Health recommendations Diabetes	Peer-reviewed article Field evaluation: Implementation study 26 interviews conducted with 16 patients (n = 16)	Prior care Patient experience of using app; barriers and facilitators to using app; patient HbA_{1c}	Patient self-efficacy and willingness to engage with the app were associated with improvement in HbA_{1c}. Competing patient priorities and psychosocial issues were barriers to effective use of the app and improved HbA_{1c}.
Agarwal et al, 2019[320] BlueStar Health recommendations Diabetes	Peer-reviewed article Field evaluation: RCT 223 adults (aged ≥ 18) with type 2 diabetes (n = 223)	Usual care 6-month intervention period; patient HbA_{1c} levels; patient-reported self-efficacy, experience of care, health care utilization	BlueStar had no significant impact on HbA_{1c} or patient-reported measures of self-efficacy, quality of life, or health care utilization.
BrainQ Technologies, Ltd, 2019[322] BQ Device Treatment delivery Cerebrovascular: Stroke	Clinical trial record Field evaluation: RCT 50 adults (aged 18-80) with recent ischemic stroke (n = 50)	Sham device Various neurological and motor function test scores	Study ongoing/not yet published
Bright Cloud International Corp, 2020[209] BrightArm Compact Health recommendations Cerebrovascular: Stroke	Clinical trial record Field evaluation: Pre-post trial 6 patients who had prior stroke (n = 6)	Preintervention Patient motor function, strength, range of motion, independence in daily living, cognitive function, mood, and acceptance of technology	Study ongoing/not yet published

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Winn et al, 2019[323] Buoy Health Patient evaluation General	Peer-reviewed article Field evaluation: Pre-post trial 158 083 patient encounters with the application Users of Buoy Health	Preintervention Patient intentions to seek care at different levels of urgency	Patients reported after using the app that they intended to seek a less urgent level of care in more than 25% of cases, a more urgent level in about 5% of cases, and the same level of care in the remaining cases.
Smith et al, 2019[324] Cardiologs Platform Patient evaluation Cardiovascular: Arrythmia	Peer-reviewed article Performance test: Retrospective 1500 ECGs	Veritas classification algorithm used in Mortara ECG machines Accuracy	In identifying major ECG abnormalities, Cardiologs had a sensitivity of 88.7% and specificity of 94.0%. Veritas had a sensitivity of 92.0% and specificity of 84.7%.
Smith et al, 2019[325] Cardiologs Platform Patient evaluation Cardiovascular: Arrythmia	Peer-reviewed article Performance test: Retrospective 24 123 ECG readings recorded over 6 months; second database of 1473 ECGs for measuring sensitivity	Veritas classification algorithm used in Mortara ECG machines; physician interpretation using Veritas (Veritas + Physician) Accuracy	The first version of the Cardiologs DL algorithm was more accurate (91.2%) than was Veritas (80.2%) and similar to Veritas + Physician (90.0%) in classifying atrial fibrillation. Cardiologs also had higher sensitivity (92%) than did Veritas (87%).
US FDA, 2020[326] CLEWICU Patient evaluation Respiratory: COVID-19 General: Adverse events	FDA summary document Performance test Data from 7 adult ICUs at 2 hospitals over 11 years	No comparator Accuracy	When used to predict respiratory failure, CLEWICU's sensitivity was 53.7% and positive predictive value (PPV) was 4.7%, with 3.9 hours median lead time provided for true positive alerts. When used to predict hemodynamic instability, CLEWICU's sensitivity was 56.9% and PPV was 18.5%, with 3.7 hours median lead time for true positive alert.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Place et al, 2017[327] Companion Patient evaluation Mental health: Depression, anxiety	Peer-reviewed article Performance test: Prospective 73 adults (aged ≥ 18) with at least one symptom of post-traumatic stress disorder (PTSD) or depression; 1217 audio recordings and 51 080 131 "digital trace points" (n = 73)	No comparator Accuracy	Companion's AUC was 0.74 for depressed mood, 0.56 for fatigue, 0.75 for interest in activities, and 0.83 for social connectedness.
Department of Veterans Affairs (VA) Eastern Colorado Health Care System, 2019[328] Companion Patient evaluation Mental health: Depression, anxiety	Clinical trial record Field evaluation: RCT 750 adults (aged 18-55) in the US Navy (n = 750)	Access to noninteractive mental health app 3-month intervention period; patient psychological distress, time until treatment, utilization of treatment, depression, PTSD symptoms, suicidal ideation	Study ongoing/not yet published
Mayo Clinic, 2020[329] Control Tower Health recommendations General: Palliative care	Clinical trial record Field evaluation: RCT 20 000 adults (aged ≥ 18) admitted to hospital (n = 20 000)	Standard care 1-year intervention period; time until palliative care team consultation; number of palliative care consults; transition time to hospice; emergency department visit within 30 days of discharge; hospital readmission within 30 days of discharge; ICU transfers; ratio of hospice to hospital deaths; inpatient length of stay	Study ongoing/not yet published
Blomberg et al, 2019[331] Corti Patient evaluation Cardiovascular: Cardiac arrest	Peer-reviewed article Performance test: Retrospective 108607 calls to Copenhagen emergency services (n = 108 670)	Dispatcher interpretation of call Accuracy	The ML algorithm had a sensitivity of 84.1% and specificity of 97.3% in recognizing out-of-hospital cardiac arrest, compared with dispatcher sensitivity of 72.% and specificity of 98.8%. The ML algorithm also recognized cardiac arrest faster (median 44 seconds) than did dispatchers (median 54 seconds).

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Emergency Medical Services, Capital Region, Denmark, 2020[332] Corti Patient evaluation Cardiovascular: Cardiac arrest	Clinical trial record Field evaluation: RCT 5242 calls to Copenhagen emergency services (n = 5242)	Dispatcher response with alert compared with usual dispatcher response without automated alert Dispatcher identification of caller cardiac arrest; time to identification; dispatcher request to caller for cardiopulmonary resuscitation (CPR) initiation; time until dispatcher starts guiding caller in CPR	Study ongoing/not yet published
Beijing Tsinghua Chang Gung Hospital, 2020[333] COVID-19 Alert System Health recommendations Respiratory: COVID-19	Clinical trial record Field evaluation: Pre-post trial 102 456 patients (n = 102 456)	Preintervention Accuracy; avoidance of unnecessary outpatient visits; patient relief from anxiety	Study ongoing/not yet published
Current Health, 2019[334] Current Platform Patient evaluation General: Adverse events	Gray literature Field evaluation: Implementation study Unspecified	Prior care Time to train clinicians and deploy app, patient adherence, number of clinician home visits, number of emergency department visits	The clinical deployment team was trained in one hour and their time to deployment was less than 24 hours after training. Patient adherence rate was 92% with a 22% reduction in unnecessary home visits.
University of Utah, 2019[206] Dashboard for Diabetes Care Health recommendations Diabetes	Clinical trial record Field evaluation: RCT 17 000 primary care patients with diabetes (n = 17 000)	Standard care without dashboard 1-year intervention period; patient HbA_{1c}; body mass index (BMI); cost of diabetes medications prescribed; use of the diabetes dashboard; user opinions of the diabetes dashboard	Study ongoing/not yet published
Shlomo et al, 2019[337] DayTwo Health recommendations Diabetes	Conference paper Field evaluation: Pre-post trial 28 patients with diabetes not on insulin (n = 28)	Prior to intervention 4- to 20-month intervention period; patient HbA_{1c}, time in target glucose range (70-140), mean glucose	Average patient HbA_{1c} fell from 7.2% to 6.5%. Time in reference range increased from 69.1% to 79.6%. Mean glucose levels decreased from 125.6 to 114.6 mg/dL.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Zeevi et al, 2015[335] DayTwo Health recommendations Diabetes	Peer-reviewed article Field evaluation: RCT 26 patients (n = 26)	Clinician dietary recommendations 2-week intervention period; patient post-prandial glycemic response	Ten of 12 patients had lower postprandial glycemic response to app dietary recommendations compared with 8 of 14 using clinician dietary recommendations.
Mendes-Soares et al, 2019[336] DayTwo Health recommendations Diabetes	Peer-reviewed article Performance test: Prospective 327 individuals without diabetes (n = 327)	Model trained using different data Performance in predicting postprandial glycemic response	The model performed best when trained on both previous and current patient data (R = 0.618).
New York University (NYU) Langone Health, 2020[194] Hu et al, 2020[338] Diabetes Prevention App, NYU Langone Health Health recommendations Diabetes	Clinical trial record Field evaluation: RCT 200 adults 18 or older who are overweight or obese and prediabetic (n = 200)	One-size-fits all dietary recommendations Body weight at 6 and 12 months; body composition at 6 and 12 months; glycemic variability; self-efficacy for weight loss	Preliminary findings found no difference for self-efficacy at 3 months.
Liang et al, 2019[158] Diagnostic AI for Pediatric Diseases, Guangzhou Women and Children's Medical Center Patient evaluation General: Pediatric diseases	Peer-reviewed article Performance test: Retrospective Unspecified	Clinician review of data Accuracy	The model accuracy was "comparable to experienced pediatricians in diagnosing common childhood diseases."
Kang et al, 2016[340] eCART (electronic cardiac arrest triage) Patient evaluation Cardiovascular: Cardiac arrest General: Adverse events	Peer-reviewed article Performance test: Prospective 5751 admissions of 3889 distinct adult inpatients (n = 3889)	Standard care Accuracy in predicting cardiac arrest or ICU transfer; time of prediction prior to cardiac arrest or ICU transfer	eCART identified 52% of ICU transfers early compared with 34% by the current system. eCART identified high-risk patients 30 hours prior to ICU transfer or cardiac arrest compared with 1.7 hours for the current system. eCART's AUC was 0.80 for ICU transfer and 0.88 for cardiac arrest.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Hirai et al, 2018[341] eCART (electronic cardiac arrest triage) Patient evaluation Cardiovascular: Cardiac arrest General: Adverse events	Peer-reviewed article Performance test: Retrospective 61 adults following intervention for pulmonary embolism (n = 61)	Cardiac Arrest Triage (CART) score; Pulmonary Embolism Severity Index (PESI) Accuracy in predicting 30-day mortality	eCART was most accurate (AUC 0.84), followed by CART (AUC 0.72) and PESI (AUC 0.69).
Green et al, 2018[342] eCART (electronic cardiac arrest triage) Patient evaluation Cardiovascular: Cardiac arrest General: Adverse events	Peer-reviewed article Performance test: Retrospective 107 868 patient admissions (n = 107 868)	Modified Early Warning Score (MEWS); National Early Warning Score (NEWS); Between the Flags (BTF) system Accuracy in predicting adverse event (ICU transfer, cardiac arrest, death) within next 24 hours	eCart accuracy was highest, with an AUC of 0.801, followed by NEWS (AUC 0.718), MEWS (AUC 0.698), and BTF (AUC 0.663).
Bartkowiak et al, 2019[343] eCART (electronic cardiac arrest triage) Patient evaluation Cardiovascular: Cardiac arrest General: Adverse events	Peer-reviewed article Performance test: Retrospective 32 537 patient admissions (all postoperative adults admitted to hospital inpatient; n = 32 537)	MEWS; NEWS Accuracy in predicting adverse event (ICU transfer, cardiac arrest, death) in postoperative period (duration unspecified)	eCART was most accurate (AUC 0.79), followed by NEWS (AUC 0.76) and MEWS (AUC 0.75).
Attia et al, 2019[344] ECG AI-Guided Screening Patient evaluation Cardiovascular: Low ejection fraction	Peer-reviewed article Performance test: Retrospective 52 870 patients (n = 52 870)	No comparator Accuracy	AUC was 0.93; sensitivity was 86.3%; specificity was 85.7%.
Attia et al, 2019[242] ECG AI-Guided Screening Patient evaluation Cardiovascular: Low ejection fraction	Peer-reviewed article Performance test: Prospective 16 056 patients who received routine ECGs (n = 16 056)	No comparator Accuracy	Sensitivity was 82.5%; specificity was 86.8%. Performance was similar across different ages, sexes, and races.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
US FDA, K192004, 2020[345] Eko Analysis Software Patient evaluation Cardiovascular: Arrythmia: AF Cardiovascular: Heart murmur	FDA summary document Performance test: Retrospective 2177 recordings from 375 patients (n = 375)	No comparator Accuracy	For AF: Sensitivity was 100%; specificity was 96.2%. For heart murmur: Sensitivity was 87.6%; specificity was 87.8%.
Lai et al, 2016[346] eMurmur AI Patient evaluation Cardiovascular: Heart Failure Cardiovascular: Heart murmur	Peer-reviewed article Performance test: Prospective 106 pediatric patients (n = 106)	Clinician review of data Accuracy of detecting pathologic heart murmurs	Sensitivity was 87%; specificity was 100%. Accuracy levels were similar to those of pediatric cardiologists and neonatologists.
Thompson et al, 2019[347] eMurmur AI Patient evaluation Cardiovascular: Heart Failure Cardiovascular: Heart murmur	Peer-reviewed article Performance test: Retrospective 3180 hear sound recordings from 603 outpatient visits (n = 603)	No comparator Accuracy in detection of pathologic heart murmurs	Sensitivity was 93%; specificity was 81%. Accuracy was similar to that of cardiologists.
US FDA, K162627, 2017[348] EnsoSleep Patient evaluation Respiratory: Sleep apnea	FDA summary document Performance test: Retrospective 823 records from 72 patients (n = 72)	Clinician review of data Accuracy in classifying obstructive sleep apnea severity	Sensitivity was 86.9%; specificity was 99.5%.
Ochsner Health, 2018[351] EPIC Deterioration Index Patient evaluation Cerebrovascular: Cardiac arrest General: Adverse events	Gray literature Field evaluation: Pre-post trial Unspecified	Prior care without use of predictive app 90-day intervention period; number of adverse patient events outside the ICU	There was a 44% reduction in adverse patient events outside of the ICU.
Potolsky, 2020[352] EPIC Deterioration Index Patient evaluation Cerebrovascular: Cardiac arrest General: Adverse events	Gray literature Field evaluation: Pre-post trial Unspecified	Prior care without rapid response team use of predictive app Number of in-hospital cardiac arrests; number of patient transfers to higher level of care	Study ongoing/not yet published

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Singh et al, 2020[349] EPIC Deterioration Index Patient evaluation Cerebrovascular: Cardiac arrest General: Adverse events Respiratory: COVID-19	Academic preprint Performance test: Retrospective 392 patient hospitalizations due to COVID-19 (n = 392)	No comparator Composite outcome measure of adverse events (ICU transfer, ventilator use, death)	The model AUC was 0.79. Model accuracy was better for the highest- and lowest-risk patients. No prediction bias was found by patient race or sex.
Cummings et al, 2020[350] EPIC Deterioration Index Patient evaluation Cerebrovascular: Cardiac arrest General: Adverse events Respiratory: COVID-19	Academic preprint Performance test: Retrospective Unspecified	PICTURE (Predicting Intensive Care Transfers and other UnfoReseen Events) algorithm, which is not in current use Accuracy of predicting patient adverse event (ICU transfer, ventilator use, death) for COVID-19 and non-COVID-19 patients	EPIC deterioration index had an AUC of 0.762 compared with 0.819 for PICTURE for non-COVID-19 patients. EPIC had an AUC of 0.792 vs 0.828 for PICTURE in COVID-19 patients.
Pluymaekers et al, 2020[353] FibriCheck Patient evaluation Cardiovascular: Arrythmia, AF	Peer-reviewed article Field evaluation: Implementation study 38 patients (n = 38)	No comparator Number of patients agreeing to use application; number of measurements recorded of sufficient quality for analysis	Thirty out of 38 patients had a smartphone/tablet and agreed to use the application; 587 out of 651 recorded measurements were of sufficient quality for analysis.
Proesmans et al, 2019[354] US FDA K173872, 2018[355] FibriCheck, KardiaAI Patient evaluation Cardiovascular: Arrythmia, AF	Peer-reviewed article Performance test: Prospective 223 adults (aged ≥ 65; n = 223)	Fibricheck compared with KardiaAI Accuracy of detecting AF	FibriCheck had a sensitivity of 95.6% and specificity of 96.55%, compared with KardiaAI sensitivity of 94.09% and specificity of 97.47%.
Myers et al, 2019[284] FIND FH Algorithm Patient evaluation Cardiovascular: High cholesterol	Peer-reviewed article Performance test: Prospective Field evaluation: Implementation study 170 589 934 patients in implementation portion of study, 148 patients in performance test (n = 148)	Clinician review of 148 patients to established gold standard for performance test Number of patients identified with potential family hypercholesterolemia (FH); accuracy	The model identified 1 332 625 patients with potential FH. Model accuracy was AUC 0.89, sensitivity was 0.45, and specificity was 0.85.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Washington State University, 2019[42] Heart Failure Medication Reminder App Health recommendations Cardiovascular	Clinical trial record Field evaluation: Pre-post trial 40 Adults 21 or older with diagnosis of heart failure (n = 40)	Preintervention Medication adherence rate (up to 12 months)	Study ongoing/not yet published
Geisinger Clinic, 2020[207] Heart Failure Treatment Gap Model Health recommendations Cardiovascular: Heart failure	Clinical trial record Field evaluation: RCT 600 adults (aged ≥ 18) with heart failure (n = 600)	Standard care Patient mortality; hospital admission; health care utilization	Study ongoing/not yet published
Wijnberge et al, 2020[359] Hypotension Prediction Patient evaluation General	Peer-reviewed article Field evaluation: RCT 68 adult patients (aged ≥18 years old) scheduled to undergo an elective noncardiac surgical procedure under general anesthesia (n = 68)	Standard care Time-weighted average of patient hypotension during surgery	The median time of hypotension per patient was reduced by 16.7 minutes. No serious adverse events occurred in the intervention group (n = 34); 2 events leading to death occurred in the control group (n = 34).
Centre Hospitalier Universitaire de Saint Etienne, 2020[186] iDEFCO Patient evaluation Cancer	Clinical trial record Field evaluation: Pre-post trial 850 patients with cancer (n = 850)	Prior to intervention 12-month intervention period; number of people who have used iDEFCO; proportion of patients identified as socially fragile; patient self-reported depression, anxiety, and quality of life	Study ongoing/not yet published
Evans et al, 2016[172] Intermountain Healthcare Readmission/Mortality Prediction Patient evaluation Cardiovascular	Peer-reviewed article Field evaluation: Pre-post trial 175 adult hospitalized patients (n = 175)	Standard care 5-month intervention period; 30-day readmission rate; 30-day mortality rate; patient discharge to home care	Implementation of the prediction model led to significant reduction in 30-day mortality, significant increase in patient discharge, and no change in readmission.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
University of Washington, 2020[185] Jumpstart Patient evaluation General: Serious illness	Clinical trial record Field evaluation: RCT 150 adults (aged ≥ 18) with serious illness (n = 150)	Usual care (no use of app) Number of patients with documented goals-of-care discussions	Study ongoing/not yet published
Gajra et al, 2020[41] JVION Machine Patient evaluation General: Adverse events	Conference paper Field evaluation: Pre-post trial 21 providers managing an average of 4329 patients with cancer per month (n = 4329)	Standard care preintervention 17-month intervention period; rate of palliative care consultations; rate of hospice referrals	The average rate of palliative care consults increased from 17.3 to 29.1 per 1000 patients per month. The average rate of hospice referrals increased from 0.2 to 1.6 per 1000 patients per month.
Frownfelter et al, 2020[43] JVION Machine Patient evaluation General: Adverse events	Conference paper Field evaluation: Pre-post trial 21 providers managing an average of 4329 patients with cancer per month (n = 4329)	Standard care preintervention 17-month intervention period; rate of depression screenings; rate of antidepressant prescriptions	The average rate of depression screenings increased from 6 to 16.2 per 1000 patients per month. The average rate of antidepressant prescriptions increased from 9.2 to 15.5 per 1000 patients per month.
Romero-Brufau et al, 2020[361] JVION Machine Patient evaluation General: Adverse events	Peer-reviewed article Field evaluation: Pre-post trial 2460 hospitalizations (n = 2460)	Standard care at same hospital preintervention and at matched control hospitals 1-year intervention period; accuracy in predicting hospital acquired preventable infections, 30-day patient readmission rate	Risk assignment accuracy: Sensitivity was 65%; specificity was 89%. Readmission rates decreased 3.3% (from 11.4% to 8.1%) following implementation, as compared with a reduction of 0.5% observed at control hospitals.
Ravi et al, 2019[362] JVION Machine Patient evaluation General: Adverse events	Conference paper Performance test: Retrospective 63 476 hospitalizations (n = 63 476)	Braden scale Accuracy in predicting hospital-acquired preventable infections	The JVION machine AUC was 0.84 compared with the Braden AUC of 0.72.
Koren et al, 2019[363] K Health Patient evaluation General: All conditions	Peer-reviewed article Performance test 1215 patients at 2 health systems (n = 1215)	Clinician review of data Patient perceptions of application accuracy	In the trial, 82.4% of Maccabi patients and 85.4% of Integrity Health patients reported that K Health diagnoses agreed with their doctors' final diagnosis.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Duarte et al, 2020[366] KardiaAI/Kardia Mobile Patient evaluation Cardiovascular: Arrythmia: AF	Peer-reviewed article Field evaluation: Implementation Various (multiple studies included in meta-analysis)	imPulse; MyDiagnostick; RhythmPad GP; Zenicor ECG; Generic lead-I ECG device Cost per quality-adjusted life-year	Kardia Mobile is "the most cost-effective option in a full incremental analysis" compared with the standard diagnostic pathway and other lead-I ECG devices.
Godin et al, 2019[365] KardiaAI/Kardia Mobile Patient evaluation Cardiovascular: Arrythmia: AF	Peer-reviewed article Field evaluation: Implementation study 133 physicians; 7585 patients (n = 7585)	Standard prior physician practice Patient screenings per physician; detection of AF; initiation of anticoagulation therapy; physician-perceived clinical value, satisfaction, and feasibility of use	AF was detected in 6.2% of patients; 270 patients were initiated on anticoagulation therapy. Physician perceptions of clinical value and feasibility of use were favorable, including ease of explaining application results to patients.
Wegner et al, 2020[367] KardiaAI/Kardia Mobile Patient evaluation Cardiovascular: Arrythmia: AF	Peer-reviewed article Performance test: Prospective 296 ECGs from 99 patients (n = 99)	Electrophysiologist interpretation Accuracy of detecting AF	The app had a sensitivity of 55% to 70% and specificity of 60% to 69%, compared with electrophysiologist sensitivity of 96% and specificity of 97%.
Selder et al, 2019[368] KardiaAI/Kardia Mobile Patient evaluation Cardiovascular: Arrythmia: AF	Peer-reviewed article Performance test: Retrospective 982 ECGs from 233 patients (n = 233)	No comparator Accuracy of detecting AF; number of ECG recordings unclassified or unreadable	Sensitivity was 0.92; specificity was 0.95; 17% of recordings were unclassified, and 2% were unreadable.
Reed et al, 2019[364] KardiaAI/Kardia Mobile Patient evaluation Cardiovascular: Arrythmia: AF	Peer-reviewed article Field evaluation: RCT 243 patients aged 16 years or older (n = 243)	Standard care without application Number of symptomatic arrythmias detected at 90 days; days until detection	Arrythmia was detected at an average of 9.5 days by the application compared with 42.9 days in the control group. The application detected arrythmia in 8.9% of patients compared with 0.9% in standard care.
Bae et al, 2019[235] KDPI-EPTS Survival Benefit Estimator Patient evaluation Kidney disease	Peer-reviewed article Performance test: Retrospective 120 818 patients who received kidney transplants (n = 120 818)	No comparator Accuracy	Application C statistic was 0.637.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Duke University, 2018[372] Kelahealth Patient evaluation General: Adverse events	Clinical trial record Field evaluation: Pre-post trial 200 adults (aged ≥ 18) postsurgery (n = 200)	Pre-adoption patient care (clinician judgment alone) 30-day surgical complications, including mortality, wound, cardiac, respiratory, thrombotic, renal, urinary tract infection, cerebrovascular, sepsis, bleeding, and readmission	Study ongoing/not yet published
Kelahealth, 2019[371] Kelahealth Patient evaluation General: Adverse events	Gray literature Performance test: Retrospective 2506 patients (n = 2506)	Actual care received in which all 2503 patients received 3 days of ICU care and 1 day of floor care postsurgery Accuracy in predicting whether patient would stay over 4 days in hospital postsurgery; retrospective/simulated cost	The AUC was 0.75; 2003 low-risk patients were assigned to 4 days of floor care postsurgery, with estimated $3 million savings per year.
Kelahealth, 2019[370] Kelahealth Patient evaluation General: Adverse events	Gray literature Performance test: Retrospective 370 patients following vascular surgery (n = 370)	Actual care received in which surgeons assessed patient need for negative pressure therapy vs sterile dressings Retrospective/simulated cost and infection rate	The app could have led to an estimated 41.3% reduction in surgical site infections and 26.0% reduction in costs, amounting to $148 458 across the 370 patient cases.
Stein, 2019[373] Lark Diabetes Care Health recommendations Diabetes	Gray literature Field evaluation: Pre-post trial 22 patients enrolled in Lark DMP (n = 22)	Preintervention 6-month intervention period; patient Hba_{1c}	There was a significant reduction in patient HbA_{1c} from 8.4% to 7.4%.
Stein, 2020[375] Lark DPP Health recommendations Diabetes	Gray literature Field evaluation: Pre-post trial 610 patients enrolled in Lark DPP (n = 610)	Preintervention 1-year intervention period; patient weight loss and BMI	The average patient weight loss was 4.2%. Of participants, 31% decreased by at least one BMI category.
Stein and Brooks, 2017[374] Lark DPP Health recommendations Diabetes	Peer-reviewed article Field evaluation: Pre-post trial 70 overweight and obese patients (n = 70)	Preimplementation 6-month intervention period; patient weight loss, proportion of healthy meals, duration of app use	Patients lost an average of 2.4 kg weight and increased their percentage of healthy meals by 31%. Average duration of app use was 15 weeks.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Lark, 2018[377] Lark for Hypertension Health recommendations Cardiovascular: Hypertension	Gray literature Field evaluation: Pre-post trial 76 adults (aged ≥ 18) with hypertension (n = 76)	Preintervention Change in patient blood pressure	There was a significant reduction in patient blood pressure.
Persell et al, 2020[376] Lark for Hypertension Health recommendations Cardiovascular: Hypertension	Peer-reviewed article Field evaluation: RCT 297 patients with uncontrolled hypertension (n = 297)	Preimplementation 6-month intervention period; patient blood pressure, antihypertensive medication adherence, self-monitoring practices, self-efficacy	There was a significant increase in patient self-confidence at controlling blood pressure and some reduction in systolic blood pressure (–2.0 mm Hg; p = .16) for patients using the app.
Zephyr Sleep Technologies, 2019[380] MATRx Plus Treatment delivery Respiratory: Sleep apnea	Clinical trial record Field evaluation: Implementation study 10 adult patients (n = 10)	No comparator Patient ability to successfully use application; patient satisfaction with application	Study ongoing/not yet published
Mosca et al, 2018[379] MATRx Plus Treatment delivery Respiratory: Sleep apnea	Conference paper Performance test: Prospective 60 patients (n = 60)	No comparator Accuracy of prediction that patient would respond to therapy	Sensitivity was 83%; specificity was 86%.
Remmers et al, 2017[378] MATRx Plus Treatment delivery Respiratory: Sleep apnea	Peer-reviewed article Performance test: Prospective 53 patients with obstructive sleep apnea (n = 53)	No comparator Accuracy of prediction that patient would respond to therapy; proportion of recommended device positions that resulted in oxygen desaturation index value of less than 10 events per hour (efficacy)	Patient response to therapy predictive accuracy: Sensitivity was 85%; specificity was 93%. The predicted position was efficacious in 86% of cases in which patients were correctly predicted to respond to therapy.
Icahn School of Medicine at Mount Sinai, 2020[178] Medical Early Warning Score ++ (MEWS ++) Patient evaluation General	Clinical trial record Performance test 96 645 patients with 157 984 hospital encounters and 244 343 bed movements (n = 96 645)	Standard MEWS score Accuracy in predicting patient deterioration	MEWS++ sensitivity was 81.46% and specificity was 75.5% as compared with MEWS sensitivity of 44.6% and specificity of 64.5%.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Icahn School of Medicine at Mount Sinai, 2020[178] Medical Early Warning Score ++ Patient evaluation General	Clinical trial record Field evaluation: Pre-post trial (nonrandomized trial, 2 arms) 2915 hospitalized adults (aged ≥ 18; n = 2915)	Standard care Rate of care escalation; number of patient cardiac arrests, deaths; number of alerts; accuracy of alert	Study ongoing/not yet published
Ziegler et al, 2015[44] Minimed 780G/MD-Logic Artificial Pancreas Treatment delivery Diabetes	Peer-reviewed article Field evaluation: Pre-post trial 58 children and adults (aged 10-65) with type 1 diabetes, HbA_{1c} between 7% and 10% (n = 58)	Standard care Fear of hypoglycemia; acceptance of the artificial pancreas	This analysis of the psychological impact of using the automated closed-loop MD-Logic system under real-life conditions in the patients' home demonstrated reduced worries of hypoglycemia with the artificial pancreas.
Nimri et al, 2014[301] Minimed 780G/MD-Logic Artificial Pancreas Treatment delivery Diabetes	Peer-reviewed article Field evaluation: RCT 19 children and adults (aged 12-65) with type 1 diabetes at least 1 year since diagnosis, use of an insulin pump for at least 3 months, experience using continuous glucose monitor, HbA_{1c} between 6.5% and 10% (n = 19)	Standard care Nocturnal hypoglycemia; time within range	Glycemic control with MD-Logic showed significantly reduced nocturnal hypoglycemia, with increased time within range and lower mean glucose levels.
Nimri et al, 2013[381] Minimed 780G/MD-Logic Artificial Pancreas Treatment delivery Diabetes	Peer-reviewed article Field evaluation: RCT 12 children and adults (aged ≥ 10) with type 1 diabetes diagnosed for more than 1 year; HbA_{1c} at inclusion between 7% and 10% (n = 12)	Standard care Nocturnal hypoglycemia; time within range; adverse events	Fewer hypoglycemic events occurred. Patients experienced increased time within range. No application-related severe adverse events occurred.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Phillip et al, 2013[382] Minimed 780G/MD-Logic Artificial Pancreas Treatment delivery Diabetes	Peer-reviewed article Field evaluation: RCT 54 children (aged 10-18) with at least a 1-year history of type 1 diabetes, HbA_{1c} of 7% to 10% (n = 54)	Standard care Nocturnal hypoglycemia; time within range; adverse events	Patients at a diabetes camp who were treated with MDLAP had less nocturnal hypoglycemia and tighter glucose control than when they were treated with a sensor-augmented insulin pump. No adverse events were reported.
Buegler et al, 2020[384] Neuro Motor Index, Altoida Patient evaluation Dementia	Peer-reviewed article Performance test: Prospective 548 adults (aged 55-90) including those with normal cognition, minor cognitive impairment, or early dementia (n = 548)	No comparator Accuracy in predicting progress from minor cognitive impairment to dementia within 3 years	AUC was 0.91; sensitivity was 0.81; specificity was 0.90.
Osborn et al, 2017[385] One Drop Health recommendations Diabetes	Peer-reviewed article Field evaluation: Pre-post trial 1288 patients with diabetes using One Drop (n = 1288)	No comparator 60- to 365-day intervention period; patient HbA_{1c}	Patient HbA_{1c} significantly decreased by absolute 1.07%, with greater decreases for users with type 2 diabetes than for those with type 1.
Goldner et al, 2018[386] One Drop Health recommendations Diabetes	Conference paper Performance test: Retrospective 1 923 416 blood glucose measurements from 14 706 people with type 2 diabetes who were not receiving insulin treatment (n = 14 706)	No comparator Accuracy	Application predictions were within 50 mg/dL of observed values 91% of the time. Mean error was 21.3 mg/dL; median error was 14.2 mg/dL.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Wexler et al, 2020[387] One Drop Health recommendations Diabetes	Conference paper Performance test: Pre-post evaluation 10 million hours of monitoring data from 3000 patients with diabetes (n = 3000)	ML algorithm developed by IBM Accuracy in predicting glucose level below 70 mg/dl or above 180 mg/dl in 30 minutes, 1 hour, and 4 hours	One Drop was more accurate than was the IBM algorithm. One Drop predictions of hypoglycemia in 30min had 93.2% sensitivity and 89.4% specificity, compared to 1 hour predictions which had 83.2% sensitivity and 74.1% specificity as well as 4 hour predictions which had an AUC of 91.9%. Predictions of hyperglycemia in 30min had 98.9% sensitivity and 97.6% specificity compared to 1 hour predictions which had 95.0% sensitivity and 92.6% specificity.
Optima Integrated Health, 2019[205] Optima 4 Blood Pressure Health recommendations Cardiovascular: Hypertension	Clinical trial record Field evaluation: RCT 18 adults (aged 21-90) with hypertension (n = 18)	Standard care Number of patients with change in treatment	Study ongoing/not yet published
US FDA K080896, 2008[155] Pathwork Tissue of Origin Test Patient evaluation Cancer	FDA summary document Performance test: Prospective 192 samples from 60 individual tumors divided among 4 laboratories for concordance and physician agreement measures; 545 specimens for accuracy (n = 192)	No comparator Concordance between laboratories using application; physician agreement with application diagnosis; accuracy	Pair-wise regression concordance analysis found slopes 0.87 = 0.92 and *R* of 0.84 to 0.90. There was a 91.2% agreement with physician diagnosis, 2.1% disagreement, and 6.9% indeterminate. Across cancer types, accuracy measures ranged from 76.0% to 94.9% positive percent agreement, 98.5% to 100.0% negative percent agreement, and 0.953 to 0.999 AUC.
US FDA K120489, 2012[388] Pathwork Tissue of Origin Test Patient evaluation Cancer	FDA summary document Performance test: Prospective 45 tumor samples (n = 45)	No comparator Accuracy	Application classifications were 97.7% correct and 2.3% incorrect.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Gage, 2018[390] PD_manager Patient evaluation Dementia	Clinical trial record Field evaluation: Implementation study 200 patients with Parkinson's disease (n = 200)	Standard care Perceived ease of use; perceived utility	Study ongoing/not yet published
Gatsios et al, 2020[245] PD_manager Patient evaluation Dementia	Peer-reviewed article Field evaluation: Implementation study 75 patients with Parkinson's disease (n = 75)	No comparator 2-week intervention period; number of patients completing study protocol, use of app	The application protocol was completed by 87% of participants. Patients used the app for a median 11.57 days.
Shanghai Jiao Tong University School of Medicine, 2019[391] Pediatric Symptom Checker Health recommendations General: All conditions	Clinical trial record Field evaluation: RCT 2000 pediatric patients aged 2 months to 18 years (n = 2000)	Standard care Patient waiting time; patient visit time; patient satisfaction with care; cost to patient; patients seen per hour by health care professionals; accuracy of diagnosis; accuracy of health recommendations	Study ongoing/not yet published
Teplitzky et al, undated[394] Preventice BeatLogic Platform Patient evaluation Cardiovascular: Arrythmia, AF	Conference Paper Performance test: Retrospective 2500 ECG records from 512 patients (n = 512)	No comparator Accuracy in identifying AF	Sensitivity was 96.7%; specificity was 96.7%.
Teplitzky and McRoberts, 2018[395] Preventice BeatLogic platform Patient evaluation Cardiovascular: Arrythmia, AF	Conference paper Performance test: Retrospective More than 950 patients (n = 950)	6 previously published algorithms Accuracy in identifying ventricular ectopic beats	Sensitivity was 97%; specificity was 99%. Accuracy was better than that of previously published algorithms.
Teplitzky et al, 2020[393] Preventice BeatLogic platform Patient evaluation Cardiovascular: Arrythmia, AF	Peer-reviewed article Performance test: Retrospective 3000 ECG recordings	A commercial algorithm Accuracy in identifying ventricular ectopic beats	Sensitivity was 99.84%; PPV was 99.78%. BeatLogic accuracy was better than that of the commercial algorithm.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Qventus, 2017[397] Qventus, undated[398] Qventus Patient evaluation General	Gray literature Field evaluation: Pre-post trial Patients in hospitals using Qventus	Standard care Patient time to receiving care; length of stay; cost of care	Hospitals implementing Qventus have seen reduced time from patient arrival to receiving care, reduced length of patient stay, and substantial estimated cost savings.
Hong et al, 2020[400] Radiation Therapy Risk Algorithm Patient evaluation Cancer General: Adverse events	Academic preprint Field evaluation: RCT 963 adult courses of radiation and chemoradiation therapy	Standard care Rates of acute care during treatment; accuracy in identifying high-risk patients	For identifying high-risk patients, the AUC was 0.851. Identified high-risk patient rates of acute care during treatment decreased from 22.3% to 12.3%.
Clark et al, 2019[401] rapid Whole Genome Sequencing Patient evaluation General: Genetic condition	Peer-reviewed article Field evaluation: Implementation Performance test 95 children (performance test component); 7 infants in ICU (field evaluation component; n = 102)	Standard care Accuracy in extracting children's phenomes from EHRs; accuracy of diagnosis; speed of diagnosis	Accuracy in extracting children's phenomes from EHRs: Sensitivity was 93%, and specificity was 80%. Accuracy of diagnosis: Sensitivity was 97%; specificity was 99%. The application correctly diagnosed 3 of 7 seriously ill ICU infants, with average time savings compared with standard care of 22 hours.
Jia et al, 2018[402] Rare Disease Auxiliary Diagnosis System Patient evaluation General: Rare diseases	Peer-reviewed article Performance test: Retrospective 818 patient records (n = 818)	4 model versions tested; no other comparators Accuracy	All 4 of the models had specificity at or above 98%. The highest sensitivity achieved by 1 of the 4 models was 95%; the lowest was 67%.
VA Office of Research and Development, 2020[403] REACH VET Patient evaluation Mental Health: Suicide General: Adverse events	Clinical trial record Field evaluation: Implementation study Veterans in 28 VA medical facilities	No comparator Number of veterans receiving care due to application; application adoption by mental health and primary care providers; fidelity of implementation; cost of application implementation	Study ongoing/not yet published

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Claxton et al, 2020[405] ResApp Patient evaluation Respiratory: Asthma, chronic obstructive pulmonary disease, COVID-19	Conference paper Performance test: Retrospective 187 patients with controlled or exacerbated asthma (n = 187)	Clinician review of data used as gold standard Accuracy of diagnosis of asthma exacerbation	Sensitivity was 89%; specificity was 84%.
Swarnkar et al, 2019[404] ResApp Patient evaluation Respiratory: Asthma, COPD, COVID-19	Peer-reviewed article Performance test: Prospective 224 children with and without acute asthma (n = 224)	No comparator Accuracy of asthma severity classification	The model distinguished high severity from no/mild severity asthma groups with sensitivity of 82.61% and specificity of 78.38%.
Porter et al, 2019[406] ResApp Patient evaluation Respiratory: Asthma, COPD, COVID-19	Conference paper Performance test: Retrospective 366 adults (aged ≥ 40 years old; n = 366)	Clinician review of data used as gold standard Accuracy of diagnosis of COPD	Sensitivity was 86%; specificity was 73%.
Rajput et al, 2019[407] RhythmAnalytics Patient evaluation Cardiovascular: Arrythmia	Gray literature Performance test: Retrospective 600 ECG records (n = 600)	2 other DL models; cardiologist annotation Accuracy	The Biofourmis model was most accurate (sensitivity 0.908, specificity 0.82) compared with the other models or the cardiologists.
Adam et al, 2020[408] Rose platform Health recommendations Mental health: Depression, anxiety	Peer-reviewed article Field evaluation: Implementation study 30 adults (aged ≥ 18) experiencing symptoms of depression and/or anxiety (n = 30)	Standard care 5-week intervention period; patient use of application; patient perceptions of application	After 5 weeks, 70% of patients completed daily check-ins with the app, 97% completed anxiety tests using the app, and 70% found the app function and quality of information to be excellent.
Johns Hopkins University, 2020[197] Rose platform Health recommendations Mental health: Depression, anxiety	Clinical trial record Field evaluation: RCT 45 adults (aged 21-28) experiencing symptoms of depression and/or anxiety (n = 45)	Standard care 5-week intervention period; patient use of application; change in patient anxiety	Study ongoing/not yet published

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Levy et al, 2006[409] Seattle Heart Failure Model Patient evaluation Cardiovascular: Heart failure	Peer-reviewed article Performance test: Retrospective 9942 heart failure patients with 17 307 person-year records (n = 9942)	No comparator Accuracy	Overall AUC was 0.729.
McCoy and Das, 2017[238] Sepsis Prediction Algorithm Patient evaluation General: Sepsis	Peer-reviewed article Field evaluation: Pre-post trial 992 hospitalized patients with indications of sepsis (n = 992)	Preimplementation 3-month preimplementation period; 3 month- to 4-month postimplementation period; sepsis-related in-hospital mortality, length of stay, 30-day readmission	Sepsis-related in-hospital mortality fell by 60.24%, sepsis-related hospital length of stay fell by 9.55%, and sepsis-related 30-day readmission rate decreased by 50.14%.
Burdick et al, 2020[239] Sepsis Prediction Algorithm Patient evaluation General: Sepsis	Peer-reviewed article Field evaluation: Pre-post trial 17 757 adults hospitalized patients with indications of sepsis (n = 17 757)	Preintervention standard care 12 month- to 16-month intervention period; sepsis-related in-hospital mortality, length of stay, and 30-day readmission rates	Implementation was associated with an average 39.5% decrease in in-hospital mortality, 32.3% decrease in hospital length of stay, and 22.7% decrease in 30-day readmission rates.
Calvert et al, 2016[241] Sepsis Prediction Algorithm Patient evaluation General: Sepsis	Peer-reviewed article Performance test: Retrospective 32 000 hospitalized adults (aged ≥ 18; n = 32 000)	No comparator Accuracy	The AUC was 0.83; sensitivity was 0.90; sensitivity was 0.81.
Dascena, 2019[247] Sepsis Prediction Algorithm Patient evaluation General: Sepsis	Clinical trial record Field evaluation: RCT 51 645 adults presenting to emergency department or admitted to hospital (n = 51 645)	Standard care Sepsis-related mortality	Study ongoing/not yet published
Dascena, 2019[410] Sepsis Prediction Algorithm Patient evaluation General: Sepsis	Clinical trial record Field evaluation: RCT 51 645 adults presenting to emergency department or admitted to hospital (n = 51 645)	Locked version of algorithm (InSight) vs continuously learning version of algorithm (HindSight) Reduction in number of false alerts	Study ongoing/not yet published

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Shimabukuro et al, 2017[240] Sepsis Prediction Algorithm Patient evaluation General: Sepsis	Peer-reviewed article Field evaluation: RCT 142 hospitalized patients (n = 142)	Standard care (existing EHR-based sepsis alert) 3-month intervention period; patient in-hospital mortality; length of stay	The algorithm intervention saw reduced patient length of stay from 13.0 days in control to 10.3 days and saw in-hospital mortality reduced by absolute 12.4% (58.0% lower than in control).
Sendak et al, 2020[262] Sepsis Watch Patient evaluation General: Sepsis	Peer-reviewed article Field evaluation: Implementation study An academic health system	No comparator Qualitative measures of implementation success; health system staff involvement	Implementation was successful, clinicians were highly involved in design and implementation of Sepsis Watch, and personnel time was the largest resource requirement for implementation.
Duke University, 2019[412] Sepsis Watch Patient evaluation General: Sepsis	Clinical trial record Field evaluation: Pre-post trial 32 003 adults aged 18 or older presenting to emergency department (ED) (n = 32 003)	No comparator Proportion of patients with sepsis who receive recommended treatment; mean time from ED arrival to sepsis; number of patients who develop sepsis and are not treated; mean ED and hospital lengths of stay for patients with sepsis; mean inpatient mortality for patients with sepsis; mean ICU requirement rate for patients with sepsis; number of sepsis diagnosis codes; use of antibiotics and other recommended care resources	Study ongoing/not yet published
Bedoya et al, 2020[411] Sepsis Watch Patient evaluation General: Sepsis	Peer-reviewed article Performance test: Retrospective 42 000 hospitalized patients (n = 42 000)	7 other clinical risk score and ML models False alarm rate; sensitivity	False alarm rate was 1.4 false alarms per true alarm; sensitivity was 80%. The algorithm outperformed all 7 other clinical risk score and ML models.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Greek Aerospace Medical Association and Space Research, 2020[210] Short Arm Human Centrifuge Rehab Health recommendations Cerebrovascular: Poststroke Respiratory: COPD General: Elderly	Clinical trial record Field evaluation: RCT 105 patients (n = 105)	No treatment Multiple cardiovascular, neurological, and muscle-control measures	Study ongoing/not yet published
Caballero-Ruiz et al, 2017[190] Sinedie Health recommendations Diabetes	Peer-reviewed article Field evaluation: RCT 90 patients with gestational diabetes (n = 90)	No comparator Clinician time spent per patient; number of patient visits to clinician; patient adherence to self-monitoring; patient satisfaction; safety of app recommendations	Sinedie implementation led to reduction of 27.4% in clinician time spent on patient evaluation and reduced in-person clinician visits by 88.6%. Patients reported high satisfaction with the system and adherence was high. All app recommendations were safe.
Centre Hospitalier Universitaire de Nīmes, 2020[414] Smart Angel Patient evaluation General: Adverse events	Clinical trial record Field evaluation: RCT 1260 adult patients undergoing ambulatory surgery (n = 1260)	App without AI; no app (control) Unscheduled patient hospitalization; other measures of patient care utilization	Study ongoing/not yet published
University of Hong Kong, 2019[198] Smoking Cessation App Health recommendations Substance Abuse: Smoking	Clinical trial record Field evaluation: RCT 664 adults (aged ≥ 18) who smoke daily (n = 664)	Chatbot app with nicotine replacement therapy compared with non-chatbot SMS messaging without nicotine replacement therapy Exhaled carbon monoxide (as biochemical measure of smoking cessation); self-reported smoking cessation	Study ongoing/not yet published

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Medtronic, 2019[418] Sugar.IQ Patient evaluation Diabetes	Gray literature Field evaluation: Pre-post trial 3100 patients with diabetes (n = 3100)	Patients using the Guardian Connect system without the Sugar.IQ app Time in range	Patients using the Guardian Connect system with the Sugar.IQ app experienced 4.1% more time in range (63.4%) compared with Guardian Connect alone (59.3%).
Arunachalam et al, 2019[417] Sugar.IQ Patient evaluation Diabetes	Conference paper Field evaluation: Pre-post trial 1765 users of Guardian Connect GCMk (n = 1765)	Patients using the Guardian Connect system using the Sugar.IQ app compared with those not using the Sugar.IQ app 5-month observation period; patient time in target glucose range; glucose management indicator (GMI)	Users of Sugar.IQ had 2.7% higher time in range and GMI of 6.8% compared with 6.9% among nonusers.
Zhong et al, 2018[416] Sugar.IQ Patient evaluation Diabetes	Conference paper Field evaluation: Pre-post trial 256 users of Medtronic MiniMed Connect, with 11 356 user-days recorded (n = 256)	Preintervention standard care 90-day intervention period; patient time in target glucose range; number of hypoglycemic events; number of hyperglycemic events	Users of Sugar.IQ had an average of 33 minutes longer time in range, 1.0 fewer hypoglycemic events per month, and 1.3 fewer hyperglycemic events per month.
Berry et al, 2019[160] Symptomate Patient evaluation General	Peer-reviewed article Performance test: Retospective 168 patients with HIV and/or hepatitis C presenting to ED (n = 168)	ED clinician-determined diagnosis Accuracy	Symptomate was far less accurate than were clinicians. Symptomate correctly identified hepatitus C or HIV as its most likely diagnosis 7.1% of the time and provided the correct diagnosis among its top 10 most likely diagnoses 8.9% of the time.
Columbia University, 2020[191] t2.coach Health recommendations Diabetes	Clinical trial record Field evaluation: RCT 280 adults (18-65) with type 2 diabetes (n = 280)	Usual care 6-month intervention period; patient HbA_{1c} value	Study ongoing/not yet published

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Henry et al, 2015[169] Targeted Real-time Early Warning Score (TREWS) Patient evaluation General: Sepsis	Peer-reviewed article Performance test: Retrospective Model development study 13 014 patients (n = 13 014)	MEWS; routine screening Accuracy	TREWS was more accurate (AUC 0.83) than was MEWS (AUC 0.73). TREWS was also more accurate (specificity 0.67, sensitivity 0.85) than was routine screening (specificity 0.64, sensitivity 0.74).
Michuda et al, 2019[419] Tempus Oncology Testing Patient evaluation Cancer	Peer-reviewed article Performance test: Retrospective 10 000 tissue samples	No comparator Accuracy of tumor type classification	Application sensitivity varied from 98.1% to 99.9% for detecting different tumor characteristics.
Michuda et al, 2020[419,420] Tempus Oncology Testing Patient evaluation Cancer	Conference paper Performance test: Retrospective 25 000 tissue samples	No comparator Accuracy of tumor type classification	ML model accuracy increased with additional data inputs.
Stephens et al, 2019[422] Tess Treatment delivery Mental Health: Depression, anxiety	Peer-reviewed article Field evaluation: Implementation study 23 adolescent patients (aged 9-18) coping with weight management and prediabetes symptoms (n = 23)	No comparator Perceived usefulness of app messages	Patients rated the app as useful 96% of the time.
Fulmer et al, 2018[421] Tess Treatment delivery Mental Health: Depression, anxiety	Peer-reviewed article Field evaluation: RCT 74 participants (n = 74)	Suggested reading material with no use of app Change in self-reported depression using health questionnaires	Four weeks of app access led to significant reduction in depression and anxiety compared with control; 2 weeks of app access led to significant reduction in anxiety compared with control.
McKenzie et al, 2017[244] Virta Health recommendations Diabetes	Peer-reviewed article Field evaluation: Pre-post trial 238 adults (aged 21-65) with type 2 diabetes (n = 238)	Preintervention Patient HbA_{1c}; diabetes medication amount (number and dosage)	HbA_{1c} was reduced by 1.0%, and percentage of patients with HbA_{1c} < 6.5% increased from 19.8% to 56.1%. Diabetes medications were reduced by 56.8% of participants.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Athinarayanan et al, 2019[423] Virta Health recommendations Diabetes	Peer-reviewed article Field evaluation: Pre-post trial 349 adults (aged 21-65) with type 2 diabetes and BMI > 25kg/m2 (n = 349)	Preintervention health status; patients opting for usual care Patient HbA_{1c}; diabetes medication amount (number and dosage); BMI; glucose level; blood pressure	Patients using Virta saw sustained improvements: reduced HbA_{1c}, lower weight, lower blood pressure, and reduced use of glycemic control medications.
US FDA K124067, 2012[425] VITEK MS Patient evaluation General: Bacterial and fungal infections	FDA summary document Performance test: Prospective 3 panels of 100 microorganism stains	No comparator Accuracy	Raw accuracy was 96.1% correct identification, 0.2% incorrect, and 3.7% no identification.
US FDA K181412, 2018[426] VITEK MS Patient evaluation General: Bacterial and fungal infections	FDA summary document Performance test: Prospective 4241 test result records	No comparator Accuracy	Accuracy was 98.8% correct identification, 0.3% incorrect, and 0.9% no identification.
Fan et al, 2017[424] VITEK MS Patient evaluation General: Bacterial and fungal infections	Peer-reviewed article Systematic review of performance tests 27 studies covering 3540 *streptococci* strains	MALDI Biotyper system Accuracy	Correct identification was 98% compared with 94% by the MALDI Biotyper.
Wuhan Asia Heart Hospital, 2019[200] Warfarin Dosage App Health recommendations Cerebrovascular; cardiovascular	Clinical trial record Field evaluation: RCT 500 adults (aged 18-65) on Warfarin (n = 500)	Dosage recommendations from clinicians not using application Time in therapeutic range; bleeding events; thrombotic events	Study ongoing/not yet published
Somashekhar et al, 2019[427] Watson for Oncology and Genomics Health recommendations Cancer	Peer-reviewed article Field evaluation: Implementation study 1000 adult patients with breast, lung, or colorectal cancer (n = 1000)	No comparator Rate at which multidisciplinary tumor board changed decision following application use; agreement between the board and the app	The tumor board changed its treatment recommendation due to Watson assessment in 13.6% of cases. The tumor board and app agreed on treatment 92% of the time.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Patel et al, 2018[428] Watson for Oncology and Genomics Health recommendations Cancer	Peer-reviewed article Performance test: Retrospective 1018 patients with cancer (n = 1018)	Standard care (review by traditional multidisciplinary tumor board) Number of new genomic events detected	Use of Watson led to identification of additional genomic events of potential significance in 32% of patients. Most of these additional genomic events were considered actionable in that they could qualify patients for biomarker-selected clinical trials.
Hravnak et al, 2011[429] WAVE Clinical Platform: Visensia, the Safety Index Patient evaluation General: Adverse events; cardiovascular	Peer-reviewed article Field evaluation: Pre-post trial 642 monitored patients (n = 642)	Alerts generated by individual vital sign parameters and thresholds (single-channel) Measures of patient cardiorespiratory instability (heart rate, respiratory rate, blood pressure, oxygen saturation); number of unanticipated patient deaths	Use of the application was associated with decrease in unanticipated patient deaths and with decrease in duration and number of occurrences of patient instability episodes.
Tarassenko et al, 2006[430] WAVE Clinical Platform: Visensia, the Safety Index Patient evaluation General: Adverse events; cardiovascular	Peer-reviewed article Performance test 168 monitored patients (n = 168)	Alerts generated by individual vital sign parameters and thresholds (single channel) Accuracy of predicting patient deterioration	The app alerts were deemed true positives by clinical experts 95% of the time. App PPV was 0.32 compared with 0.03 for single-channel alerts. The app produced true alerts in advance of single-channel alerts.
Sosale et al, 2018[243] Wellthy Diabetes Health recommendations Diabetes	Gray literature Field evaluation: Pre-post trial 108 adults with diabetes (n = 108)	Preintervention standard care HbA_{1c}	Hemoglobin A_{1c} dropped from a mean of 8.51% preintervention to 8.02% postintervention.
Fitzpatrick et al, 2017[431] Woebot Treatment delivery Mental health: Depression, anxiety	Peer-reviewed article Field evaluation: RCT 70 adults (aged 18-28; n = 70)	No app use 2 week- to 3-week intervention period; patient depression anxiety	Woebot users reported significant reductions in depression and anxiety.

Evaluation study Application(s) Function Health condition	Publication type Study design Study population	Comparator(s) Outcomes measured	Author conclusions
Inkster et al, 2018[432] Wysa Treatment delivery Mental health: Depression, anxiety	Peer-reviewed article Field evaluation: Pre-post trial 139 Wysa users (n = 139)	High users vs low users of app Self-reported depression	High users of the app had significant reduction in depression compared with low users of the app.
Fleming and Jeannon, 2020[162] Your.MD Patient evaluation General	Peer-reviewed article Performance test: Simulated Cancer symptom vignettes	Ada; Babylon Accuracy in diagnosing cancer	Babylon was most accurate: sensitivity was 45%. Ada sensitivity was 32%; Your.MD sensitivity was 23%.
Steinhubl et al, 2018[433] Zio AT ECG Monitoring System Patient evaluation Cardiovascular: Arrythmia	Peer-reviewed article Field evaluation: RCT 2659 adult patients (n = 2659)	No active monitoring Identification of new AF; initiation of anticoagulants; number of patient visits to health facilities	The study found that 3.9% of patients received new AF diagnoses; anticoagulant initiation, cardiologist visits, and primary care visits all increased. No difference was found in number of ED visits or hospitalizations.
Kaura et al, 2019[434] Zio AT ECG Monitoring System Patient evaluation Cardiovascular: Arrythmia	Peer-reviewed article Field evaluation: RCT 90 adult patients (n = 90)	Standard care (short-duration Holter monitoring) Detection of arrythmia; initiation of anticoagulants	Rate of detection of arrythmia increased. Patient use of anticoagulants increased.
Yenikomshian et al, 2019[435] Zio AT ECG Monitoring System Patient evaluation Cardiovascular: Arrythmia	Peer-reviewed article Systematic review of performance tests 23 studies	No comparator Rate of detecting arrythmia	Mean detection of AF was 12.2%, supraventricular tachycardia or ectopy 45.5%, and ventricular tachycardia 17.3%.

Appendix C. Stakeholder Interview Protocol

We conducted interviews with 9 stakeholder representatives prior to finalizing our review scope and analytic design. In the next 2 sections we include, verbatim, the written study description and informed consent protocol that was emailed to interviewees prior to the interview. The final section describes the interview guide and questionnaire that we followed for all the interviews.

Study Description Provided to Interviewees

"This project is funded by the Patient-Centered Outcomes Research Institute (PCORI) and is being conducted by the RAND Corporation. The goal of our study is to better understand the use of artificial intelligence in health care, focusing especially on the use of machine-learning predictive analytics in clinical care. We are interested in understanding the types of machine-learning applications being used in clinical care, the functions that they are being used for, and the conditions they are being used to treat. We are also interested in assessing the evidence surrounding these applications, in terms of both their benefits as well as their potential risks or unintended impacts, with a particular goal of identifying any gaps in the evidence base for applications currently in use or expected to be in use in the near future.

To do this, we will be producing a narrative review and evidence map on these subjects. We are also conducting key informant interviews, in order to ensure that we take into account a range of stakeholder perspectives and priorities as we determine the review scope and guiding questions. We also hope these interviews will help us better understand relevant regulatory issues as well as barriers and facilitators to evaluating machine-learning-based applications in clinical care."

Informed Consent Protocol

"We are conducting interviews with a range of stakeholders, including patients and patient advocates, clinicians, hospitals and health systems, health care purchasers, health care payers and insurers, health policymakers, industry, and researchers. We would like to interview you for this study. You have been selected because of your perspective, interest, and experience with issues relevant to this research.

***Risks:** We do not expect that you would face any risks related to your participation in this interview.*

***Confidentiality:** We will keep your responses during this interview confidential, and our notes from this interview will only be accessible to the study team. We will not include your name in our interview notes, and we will store all interview notes securely, and separate from the list of interviewee names and identifying information. We will be reporting themes and variation in responses evident across all of our interviews. We may refer to information or opinions you express in the interview in our report,*

but we will be attributing these generically to a 'key informant interviewee,' and will not be attributing comments to anyone by name, position, affiliation, or in any way that could be used to identify you.

Duration: *Your participation in this interview will last about an hour.*

Participation and Withdrawal*: Participation in this interview is entirely voluntary. Deciding not to participate will have no negative consequences. If you decide to participate, you are free to end the interview at any point or decline to answer any question for any reason.*

Questions for Interviewer: *Do you have any questions about this study or about participation in this interview that you would like answered now?*

Informed Consent: *Do you consent to participate in this interview?"*

Interview Guide

Below we report the questions and prompts we used to guide the interviews.

A. Background

A1. Tell us a little about your background on this topic. In what ways have you been involved in or affected by the use of artificial intelligence in health care?

B. Key Questions

Our study is focused in particular on the use of machine-learning-based predictive analytics in clinical care. This includes health information technology tools that are used to diagnose conditions, evaluate patient prognosis, and assess likely treatment benefits and harms. We plan to exclude diagnostic imaging tools since these are relatively well-studied compared to other types of machine-learning applications.

We are in the process of refining the study questions that will be used to guide our literature review. Our current list includes five key questions. We are interested in both your answers to these questions as well as your thoughts on whether these are the right questions to be asking in our study.

We will first ask you to answer each of these questions directly. We know these questions cover a broad range of issues and we will not have time to address them in complete detail. So please just answer with whatever first comes to mind.

B1. What kinds of machine-learning applications are you familiar with that are currently being used in clinical care?

B2. What types of evidence are you aware of that is available on the potential benefits, harms, and other impacts of these applications?

B3. What machine-learning applications are you aware of that are currently under evaluation or in development, and might be adopted into clinical care in the next 5 years? (note: please do not provide us with any proprietary or otherwise sensitive information)

B4. From your perspective, what health conditions, health care functions, and patient populations are being addressed by these machine-learning applications?

B5. What evidence is missing from our understanding of the potential harms, benefits, and other impacts of using machine-learning-based applications in clinical care? What should be prioritized in future research to address these evidence gaps?

B6. In your view, what are the main barriers to developing and fielding machine-learning-based predictive tools in clinical care? What are the main facilitators? Please include consideration of any regulatory or system-level barriers or facilitators.

B7. In your view, what are the main barriers to evaluating the potential benefits, harms, and other impacts associated with using machine-learning-based predictive tools in clinical care? What are the main facilitators? Please include any regulatory or system-level considerations.

C. Research Scope and Key Questions

We appreciate your answers to those questions, as this is important information for our study. We'd like to continue to talk about these same topics, though taking a step back to think about them a bit differently, since we'd like to get your opinion on which aspects of these topics you think we should focus our research on.

As I mentioned earlier, PCORI has asked us to produce a literature review on AI in clinical care and the evidence surrounding its use.

This literature review has two goals. First, it is intended to help them, and other health care stakeholders such as yourself, to understand the current state of the field. Secondly, this report should identify gaps in the evidence surrounding the use of AI in clinical care, in order to help PCORI set priorities for future research.

We are just starting this project, so are very interested in getting your perspective as we determine our study scope and key research questions to help us address these goals.

(refer to draft scope and key questions in interview guide if helpful)

C1. What questions about AI in clinical care, and the evidence surrounding its use, should we seek to answer in our literature review?

C2. Are there specific types of AI applications that we should make sure to include in our review?

C3. Are there specific types of AI that are already well-studied, or have less impact on clinical care, that we should exclude from our review?

C4. What types of potential benefits, harms, and other impacts should we make sure to examine in our review?

C5. Are there any other particular types of documents or data sources that you think would be especially relevant and useful for us to examine in our study?

Appendix D. Literature Search and Screening

We conducted systematic searches of PubMed, Web of Science, the Institute of Electrical and IEEE Xplore Digital Library, the ClinicalTrials.gov database, and the FDA CDRH document library. We then applied 2 screening steps to determine if the documents resulting from these searches were within the scope of our review. This appendix documents the terms used in each of these searches as well as the literature screening process that followed. We selected the search terms by leveraging the researchers' experience in the area of both health care and AI, by drawing on RAND past experience on similar projects, and by choosing terms used in similar systematic reviews.

PubMed Searches for Reviews of AI in Health Care

We collected results from 2 searches of this database.

Search 1: May 4, 2020

Dates: January 1, 2019–May 4, 2020
"machine learning"[tiab] OR "artificial intelligence"[tiab] OR "data mining"[tiab] OR "big data"[tiab] OR "deep learning"[tiab] OR neural net*[tiab] OR support vector machine*[tiab] OR SVM[tiab] OR random forest*[tiab] OR "supervised learning"[tiab] OR "unsupervised learning"[tiab] OR "reinforcement learning"[tiab] OR "unsupervised clustering"[tiab] OR "unsupervised classification"[tiab] OR "supervised classification"[tiab] OR "natural language processing"[tiab] OR "NLP"[tiab] OR "gradient boosting"[tiab] OR expert system*[tiab] OR "rules engine"[tiab] OR "Fuzzy logic"[tiab] OR "knowledge graph*"[tiab]
AND
"health*"[tiab] OR "clinic*"[tiab] OR "hospital"[tiab] OR hospitals[tiab] OR patient*[tiab] OR therap*[tiab] OR medic*[tiab] OR care[tiab] OR drug*[tiab]
AND
review[ti]

Search 2: May 4, 2020

Dates: January 1, 2019–May 4, 2020
"machine learning"[tiab] OR "artificial intelligence"[tiab] OR "data mining"[tiab] OR "big data"[tiab] OR "deep learning"[tiab] OR neural net*[tiab] OR support vector machine*[tiab] OR SVM[tiab] OR random forest*[tiab] OR "supervised learning"[tiab] OR "unsupervised learning"[tiab] OR "reinforcement learning"[tiab] OR "unsupervised clustering"[tiab] OR "unsupervised classification"[tiab] OR "supervised classification"[tiab] OR "natural language processing"[tiab] OR "NLP"[tiab] OR "gradient boosting"[tiab] OR expert system*[tiab] OR "rules engine"[tiab] OR "Fuzzy logic"[tiab] OR "knowledge graph*"[tiab]
AND

"health*"[tiab] OR "clinic*"[tiab] OR "hospital"[tiab] OR hospitals[tiab] OR patient*[tiab] OR therap*[tiab] OR medic*[tiab] OR care[tiab] OR drug*[tiab]
AND
systematic review[ti]

Web of Science Search for Reviews of AI in Health Care

We collected results from 2 searches of this database.

Search 1: June 1, 2020

Dates: January 1, 2019–June 1, 2020
TS=("machine learning" OR "artificial intelligence" OR "data mining" OR "big data" OR "deep learning" OR "neural net*" OR support vector machine* OR SVM OR "random forest*" OR "supervised learning" OR "unsupervised learning" OR "reinforcement learning" OR "unsupervised clustering" OR "unsupervised classification" OR "supervised classification" OR "natural language processing" OR NLP OR "gradient boosting" OR "expert system*" OR "rules engine" OR "Fuzzy logic" OR "knowledge graph*")
AND
TS=(health* OR clinic* OR hospital OR hospitals OR patient* OR therap* OR medic* OR care OR drug*)
AND
TI=(review)

Search 2: June 1, 2020

Dates: January 1, 2019–June 1, 2020
TS=("machine learning" OR "artificial intelligence" OR "data mining" OR "big data" OR "deep learning" OR "neural net*" OR support vector machine* OR SVM OR "random forest*" OR "supervised learning" OR "unsupervised learning" OR "reinforcement learning" OR "unsupervised clustering" OR "unsupervised classification" OR "supervised classification" OR "natural language processing" OR NLP OR "gradient boosting" OR "expert system*" OR "rules engine" OR "Fuzzy logic" OR "knowledge graph*")
AND
TS=(health* OR clinic* OR hospital OR hospitals OR patient* OR therap* OR medic* OR care OR drug*)
AND
TI=(systematic review)

IEEE Xplore Digital Library Search for Reviews of AI in Health Care

We collected results from a single search of this database.

Search: June 1, 2020

Dates: January 1, 2019–June 1, 2020
"machine learning" OR "artificial intelligence" OR "data mining" OR "big data" OR "deep learning" OR "neural net*" OR "support vector machine*" OR SVM OR "random forest*"OR "supervised learning" OR "unsupervised learning" OR "reinforcement learning" OR "unsupervised clustering" OR "unsupervised classification" OR "supervised classification" OR "natural language processing" OR NLP OR "gradient boosting" OR "expert system*" OR "rules engine" OR "Fuzzy logic" OR "knowledge graph*"
AND
health* OR clinic* OR hospital OR hospitals OR patient* OR therap* OR medic* OR care OR drug*
AND
review

ClinicalTrials.gov Search for Clinical Trials of ML Applications

We collected results from 3 searches of this database.

Search 1: May 4, 2020

Dates: January 1, 2012–May 4, 2020
Other terms:
"machine learning" OR "artificial intelligence" OR "data mining" OR "big data" OR "deep learning" OR "neural net" OR "neural network" OR "neural networks" OR "support vector machine" OR "support vector machines" OR "SVM" OR "random forest"

Search 2: May 4, 2020

Dates: January 1, 2012–May 4, 2020
Other terms:
"random forests" OR "supervised learning" OR "unsupervised learning" OR "reinforcement learning" OR "unsupervised clustering" OR "unsupervised classification" OR "supervised classification" OR "natural language processing" OR "NLP"

Search 3: May 4, 2020

Dates: January 1, 2012–May 4, 2020
Other terms:
"gradient boosting" OR "expert system" OR "expert systems" OR "rules engine" OR "Fuzzy logic" OR "knowledge graph" OR "knowledge graphs"

FDA CDRH Document Library Search for Approved ML Applications

We collected results from a single Google search of this database.

Search: June 24, 2020

Dates: January 1, 2000–June 24, 2020

Google Search string: “machine learning” OR “artificial intelligence” OR “neural net*” OR “deep learning” site: https://www.accessdata.fda.gov/cdrh_docs

References

1. Topol E. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again.* Basic Books; 2019.
2. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med.* 2019;380(14):1347-1358.
3. Ting DS, Liu Y, Burlina P, Xu X, Bressler NM, Wong TY. AI for medical imaging goes deep. *Nat Med.* 2018;24(5):539-540.
4. Triantafyllidis AK, Tsanas A. Applications of machine learning in real-life digital health interventions: review of the literature. *J Med Internet Res.* 2019;21(4):e12286. DOI: 10.2196/12286
5. Miake-Lye IM, Hempel S, Shanman R, Shekelle PG. What is an evidence map? A systematic review of published evidence maps and their definitions, methods, and products. *Syst Rev.* 2016;5(1):28.
6. National Science and Technology Council. *Preparing for the Future of Artificial Intelligence.* Executive Office of the President; October 2016.
7. US Food and Drug Administration. *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD): Discussion Paper and Request for Feedback.* 2019. https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf
8. National Institute of Standards and Technology. *US Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools.* US Department of Commerce; August 9, 2019.
9. Joudaki H, Rashidian A, Minaei-Bidgoli B et al. Using data mining to detect health care fraud and abuse: a review of literature. *Glob J Health Sci.* 2015;7(1):194–202.
10. Li J, Huang K-Y, Jin J, Shi J. A survey on statistical methods for health care fraud detection. *Health Care Manag Sci.* 2008;11(3):275-287.
11. Baesens B, Van Vlasselaer V, Verbeke W. *Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection.* John Wiley & Sons; 2015.
12. van Capelleveen G, Poel M, Mueller RM, Thornton D, van Hillegersberg J. Outlier detection in healthcare fraud: a case study in the Medicaid dental domain. *Int J Account Inf Syst.* 2016;21:18-31.
13. Bauder R, Khoshgoftaar T. Medicare fraud detection using machine learning methods. Paper presented at: Machine Learning and Applications (ICMLA), 16th IEEE International Conference on IEEE; December 18, 2017; Cancun, Mexico.
14. Bauder RA, Khoshgoftaar TM, Richter A, Herland M. Predicting medical provider specialties to detect anomalous insurance claims. Paper presented at: 2016 IEEE 28th International Conference on Tools With Artificial Intelligence (ICTAI); November 6-8, 2016; San Jose, CA.
15. How to determine if your product is a medical device. US Food and Drug Administration. Published 2019. Accessed December, 21, 2020. https://www.fda.gov/medical-devices/classify-your-medical-device/how-determine-if-your-product-medical-device
16. Changes to existing medical software policies resulting from Section 3060 of the 21st Century Cures Act: final guidance. US Food and Drug Administration. Published 2019. Updated September 27, 2019. Accessed December 21, 2020.https://www.fda.gov/medical-devices/webinar-changes-existing-medical-software-policies-resulting-section-3060-21st-century-cures-act
17. Learn if a medical device has been cleared by FDA for marketing. US Food and Drug Administration. Published 2017. Accessed December 21, 2020. https://www.fda.gov/medical-devices/consumers-medical-devices/learn-if-medical-device-has-been-cleared-fda-marketing

18. Class I/II exemptions. US Food and Drug Administration. Published 2019. Accessed December 21, 2020. https://www.fda.gov/medical-devices/classify-your-medical-device/class-i-ii-exemptions
19. De Novo classification request. US Food and Drug Administration. Published 2019. Accessed December 21, 2020. https://www.fda.gov/medical-devices/premarket-submissions/de-novo-classification-request
20. Premarket Notification 510(k). US Food and Drug Administration. Published 2020. Accessed December 21, 2020. https://www.fda.gov/medical-devices/premarket-submissions/premarket-notification-510k
21. Premarket approval (PMA). US Food and Drug Administration. Published 2019. Accessed December 21, 2020. https://www.fda.gov/medical-devices/premarket-submissions/premarket-approval-pma
22. Emergency use authorization. US Food and Drug Administration. Published 2020. Accessed December 21, 2020. https://www.fda.gov/emergency-preparedness-and-response/mcm-legal-regulatory-and-policy-framework/emergency-use-authorization#abouteuas
23. PCORI's stakeholders. PCORI. Updated October 9, 2018. Accessed December 21, 2020. https://www.pcori.org/about-us/our-programs/engagement/public-and-patient-engagement/pcoris-stakeholders
24. Products and medical procedures. US Food and Drug Administration. Published 2020. Accessed December 21, 2020. https://www.fda.gov/medical-devices/products-and-medical-procedures
25. US National Library of Medicine. https://clinicaltrials.gov/. Accessed December 21, 2020.
26. Sittig DF, Singh H. A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. *Qual Saf Health Care.* 2010;19(suppl 3):i68-i74.
27. Rippen HE, Pan EC, Russell C, Byrne CM, Swift EK. Organizational framework for health information technology. *Int J Med Inform.* 2013;82(4):e1-e13. doi: 10.1016/j.ijmedinf.2012.01.012.
28. Livingstone DJ. *Artificial Neural Networks: Methods and Applications.* Springer; 2008.
29. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521(7553):436-444.
30. Steinwart I, Christmann A. *Support Vector Machines.* Springer Science & Business Media; 2008.
31. Chakraborty S, Tomsett R, Raghavendra R, et al. Interpretability of deep learning models: a survey of results. Paper presented at: 2017 IEEE SmartWorld; August 4-8, 2017; San Francisco, CA.
32. Che D, Liu Q, Rasheed K, Tao X. Decision tree and ensemble learning algorithms with their applications in bioinformatics. In: Arabnia H, Tran Q, eds. *Software Tools and Algorithms for Biological Systems.* Springer; 2011:191-199.
33. Rokach L, Maimon OZ. *Data Mining With Decision Trees: Theory and Applications.* Vol 69. World Scientific; 2008.
34. Hara S, Hayashi K. Making tree ensembles interpretable. *arXiv:160605390.* Preprint posted online June 17, 2016. https://arxiv.org/abs/1606.05390
35. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Science & Business Media; 2009.
36. Gaffney H, Mansell W, Tai S. Conversational agents in the treatment of mental health problems: mixed-method systematic review. *JMIR Ment Health.* 2019;6(10):e14166. https://mental.jmir.org/2019/10/e14166
37. BlueStar. WellDoc. Accessed December 21, 2020. https://www.welldoc.com/users/
38. X2: Tess. Accessed December 21, 2020. https://www.x2ai.com/individuals
39. Virta. Accessed December 21, 2020. https://www.virtahealth.com/

40. Barbieri C, Molina M, Ponce P, et al. An international observational study suggests that artificial intelligence for clinical decision support optimizes anemia management in hemodialysis patients. *Kidney Int.* 2016;90(2):422-429.

41. Gajra A, Zettler M, Kish J, et al. Impact of augmented intelligence (AI) on utilization of palliative care (PC) services in oncology. *J Clin Oncol.* 2020;38(suppl 15):12015-12015.

42. Activity-aware prompting to improve medication adherence in heart failure patients. Washington State University. ClinicalTrials.gov identifier: NCT04152031. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT04152031

43. Frownfelter J, Blau S, Zettler M, et al. Impact of augmented intelligence (AI) on identification and management of depression in oncology. *J Clin Oncol.* 2020;38(suppl 15):e14059-e14059. 10.1200/JCO.2020.38.15_suppl.e14059

44. Ziegler C, Liberman A, Nimri R, et al. Reduced worries of hypoglycaemia, high satisfaction, and increased perceived ease of use after experiencing four nights of MD-Logic artificial pancreas at home (DREAM4). *J Diabetes Res.* 2015;2015. https://doi.org/10.1155/2015/590308

45. Oliver D. David Oliver: lessons from the Babylon Health saga. *BMJ.* 2019;365:l2387.

46. Desveaux L, Shaw J, Saragosa M, et al. A mobile app to improve self-management of individuals with type 2 diabetes: qualitative realist evaluation. *J Med Internet Res.* 2018;20(3):e81. https://doi.org/10.2196/jmir.8712

47. Machine learning assisted recognition of out-of-hospital cardiac arrest during emergency calls. ClinicalTrials.gov identifier: NCT04219306. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT04219306

48. Artificial intelligence/machine learning modeling on time to palliative care review in an inpatient hospital population. ClinicalTrials.gov identifier: NCT03976297. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT03976297

49. Hripcsak G, Vawdrey DK, Fred MR, Bostwick SB. Use of electronic clinical documentation: time spent and team interactions. *J Am Med Inf Assoc.* 2011;18(2):112-117.

50. Blackley SV, Huynh J, Wang L, Korach Z, Zhou L. Speech recognition for clinical documentation from 1990 to 2018: a systematic review. *J Am Med Inform Assoc.* 2019;26(4):324-338.

51. Straub L, Gagne JJ, Maro JC, et al. Evaluation of use of technologies to facilitate medical chart review. *Drug Saf.* 2019;42(9):1071-1080.

52. Suominen H, Johnson M, Zhou L, et al. Capturing patient information at nursing shift changes: methodological evaluation of speech recognition and information extraction. *J Am Med Inf Assoc.* 2015;22(e1):e48-e66. https://doi.org/10.1136/amiajnl-2014-002868

53. 3M. 3M CodeAssist system. Accessed December 21, 2020. https://www.3m.com/3M/en_US/health-information-systems-us/improve-revenue-cycle/coding/professional/code-assist/

54. Mirchi N, Bissonnette V, Yilmaz R, Ledwos N, Winkler-Schwartz A, Del Maestro RF. The virtual operative assistant: an explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLoS One.* 2020;15(2):e0229596. https://doi.org/10.1371/journal.pone.0229596

55. Shorey S, Ang E, Yap J, Ng ED, Lau ST, Chui CK. A virtual counseling application using artificial intelligence for communication skills training in nursing education: development study. *J Med Internet Res.* 2019;21(10):e14658. https://doi.org/10.2196/14658

56. Levin M, McKechnie T, Khalid S, Grantcharov TP, Goldenberg M. Automated methods of technical skill assessment in surgery: a systematic review. *J Surg Educ.* 2019;76(6):1629-1639.

57. Ryan P, Luz S, Albert P, Vogel C, Normand C, Elwyn G. Using artificial intelligence to assess clinicians' communication skills. *BMJ.* 2019;364:l161.

58. Joudaki H, Rashidian A, Minaei-Bidgoli B, et al. Using data mining to detect health care fraud and abuse: a review of literature. *Glob J Health Sci.* 2014;7(1):194-202.

59. Dua P, Bais S. Supervised learning methods for fraud detection in healthcare insurance. In: Dua S, Acharya U, Dua P (eds) *Machine Learning in Healthcare Informatics.* Springer; 2014:261-285.

60. Waghade SS, Karandikar AM. A comprehensive study of healthcare fraud detection based on machine learning. *Int J Appl Eng Res.* 2018;13(6):4175-4178.

61. Herland M, Khoshgoftaar TM, Bauder RA. Big data fraud detection using multiple medicare data sources. *J Big Data.* 2018;5(1):29.

62. Shin H, Park H, Lee J, Jhee WC. A scoring model to detect abusive billing patterns in health insurance claims. *Expert Syst Appl.* 2012;39(8):7441-7450.

63. Bauder RA, Khoshgoftaar TM. Medicare fraud detection using machine learning methods. Paper presented at: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA); January 18, 2018; Cancun, Mexico.

64. Hehner S, Kors B, Martin M. Artificial intelligence in health insurance. McKinsey and Company. Published September 2017. Accessed January 30, 2021. https://healthcare.mckinsey.com/wp-content/uploads/2020/02/Artificial-intelligence-in-Health-Insurance.pdf.

65. Centaur system. WhiteHatAI. Accessed December 21, 2020. https://whitehatai.com/solutions/centaur/

66. Suleiman M, Demirhan H, Boyd L, Girosi F, Aksakalli V. Incorporation of expert knowledge in the statistical detection of diagnosis related group misclassification. *Int J Med Inform.* 2020;136:104086.

67. Tang M, Mendis BSU, Murray DW, Hu Y, Sutinen A. Unsupervised fraud detection in Medicare Australia. Paper presented at: Proceedings of the Ninth Australasian Data Mining Conference. December 2011;(121):103-110. Ballarat, Australia.

68. Musal RM. Two models to investigate Medicare fraud within unsupervised databases. *Expert Syst Appl.* 2010;37(12):8628-8633.

69. Menger V, Spruit M, van Est R, Nap E, Scheepers F. Machine learning approach to inpatient violence risk assessment using routinely collected clinical notes in electronic health records. *JAMA Netw Open.* 2019;2(7):e196709-e196709. doi:10.1001/jamanetworkopen.2019.6709

70. Wang KZ, Bani-Fatemi A, Adanty C, et al. Prediction of physical violence in schizophrenia with machine learning algorithms. *Psychiatry Res.* 2020;289:112960.

71. Rosen T, Zhang Y, Bao Y, et al. Can artificial intelligence help identify elder abuse and neglect? *J Elder Abuse Negl.* 2020;32(1):97-103.

72. Mowery J, Andrei A, Le E, Jian J, Ward M. Assessing quality of care and elder abuse in nursing homes via Google reviews. *Online J Public Health Inform.* 2016;8(3):e201. https://doi.org/10.5210/ojphi.v8i3.6906

73. Hurley D. Can an algorithm tell when kids are in danger? *The New York Times Magazine*. 2018. Accessed December 21, 2020. https://www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html

74. Yuan M, Boston-Fisher N, Luo Y, Verma A, Buckeridge DL. A systematic review of aberration detection algorithms used in public health surveillance. *J Biomed Inform.* 2019;94:103181.

75. Thiébaut R, Cossin S. Artificial intelligence for surveillance in public health. *Yearb Med Inform.* 2019;28(1):232.

76. Neill DB. New directions in artificial intelligence for public health surveillance. *IEEE Intell Syst.* 2012;27(1):56-59.

77. Gupta A, Katarya R. Social media based surveillance systems for healthcare using machine learning: a systematic review. *J Biomed Inform.* 2020;108:103500.

78. Rastgoo MN, Nakisa B, Rakotonirainy A, Chandran V, Tjondronegoro D. A critical review of proactive detection of driver stress levels based on multimodal measurements. *ACM Comput Surv.* 2019;51(5):35.

79. Poh CQX, Ubeynarayana CU, Goh YM. Safety leading indicators for construction sites: a machine learning approach. *Autom Constr.* 2018;93:375-386.

80. Sarkar S, Vinay S, Raj R, Maiti J, Mitra P. Application of optimized machine learning techniques for prediction of occupational accidents. *Comput Oper Res.* 2019;106:210-224.

81. de Naurois CJ, Bourdin C, Stratulat A, Diaz E, Vercher J-L. Detection and prediction of driver drowsiness using artificial neural network models. *Accid Anal Prev.* 2019;126:95-104.

82. Iranitalab A, Khattak A. Comparison of four statistical and machine learning methods for crash severity prediction. *Accid Anal Prev.* 2017;108:27-36.

83. Müller M, Botsch M, Böhmländer D, Utschick W. Machine learning based prediction of crash severity distributions for mitigation strategies. *J Advan Inform Tech.* 2018;9(1):15-25. doi:10.12720/jait.9.1.15-24

84. Ma R, Ban J, Wang Q, Li T. Statistical spatial-temporal modeling of ambient ozone exposure for environmental epidemiology studies: a review. *Sci Total Environ.* 2020;701:134463.

85. Lary D, Lary T, Sattler B. Using machine learning to estimate global PM2. 5 for environmental health studies. *Environ Health Insights.* 2015;9:(Suppl 1):41-52. doi: 10.4137/EHI.S15664

86. Bellinger C, Mohomed Jabbar MS, Zaïane O, Osornio-Vargas A. A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health.* 2017;17(1):907.

87. Schmidt CW. Into the black box: what can machine learning offer environmental health research? *Environ Health Perspect.* 2020;128(2):22001-22001.

88. Luechtefeld T, Marsh D, Rowlands C, Hartung T. Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility. *Toxicol Sci.* 2018;165(1):198-212.

89. Miller TH, Gallidabino MD, MacRae JI, et al. Machine learning for environmental toxicology: a call for integration and innovation. *Environ Sci Technol.* 2018;52(22):12953-12955.

90. Goswami S, Pal S, Goldsworthy S, Basu T. An effective machine learning framework for data elements extraction from the literature of anxiety outcome measures to build systematic review. In: Abramowicz W, Corchuelo R, eds. *Business Information Systems, Pt I.* Vol 353. Springer-Verlag; 2019:247-258.

91. Thomas J, Askie LM, Berlin JA, et al. Prospective approaches to accumulating evidence. In: Higgins JP, Thomas J, Chandler M, et al. eds. *Cochrane Handbook for Systematic Reviews of Interventions.* 2nd ed. Wiley-Blackwell; 2019:547-568.

92. Evidence pipeline. Cochrane. Accessed December 21, 2020. https://community.cochrane.org/help/tools-and-software/evidence-pipeline

93. Zhang T, Leng J, Liu Y. Deep learning for drug-drug interaction extraction from the literature: a review. *Brief Bioinform.* 2019;21(5): 1609–1627.

94. Zhang Y, Lin H, Yang Z, et al. Neural network-based approaches for biomedical relation classification: a review. *J Biomed Inform.* 2019;99:103294.

95. Chang C, Hung C, Tang CY. A review of deep learning in computer-aided drug design. Paper presented at: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); November 18-21, 2019. San Diego, CA.

96. Amasyali MF, Ersoy O. A comparative review of regression ensembles on drug design datasets. *Turk J Electr Eng Comput Sci.* 2013;21(2):586-602.

97. Luo H, Li M, Yang M, Wu FX, Li Y, Wang J. Biomedical data and computational models for drug repositioning: a comprehensive review. *Brief Bioinform.* 2020. https://doi.org/10.1093/bib/bbz176

98. Carpenter KA, Huang X. Machine learning-based virtual screening and its applications to alzheimer's drug discovery: a review. *Curr Pharm Des.* 2018;24(28):3347-3358.

99. Lin X, Li X, Lin X. A review on applications of computational methods in drug screening and design. *Molecules.* 2020;25(6):1375. https://doi.org/10.3390/molecules25061375

100. Wesolowski M, Suchacz B. Artificial neural networks: theoretical background and pharmaceutical applications: a review. *J AOAC Int.* 2012;95(3):652-668.

101. Ellingson SR, Davis B, Allen J. Machine learning and ligand binding predictions: a review of data, methods, and obstacles. *Biochim Biophys Acta Gen Subj.* 2020;1864(6):129545.

102. Graves J, Byerly J, Priego E, et al. A review of deep learning methods for antibodies. *Antibodies (Basel).* 2020;9(2):12. https://doi.org/10.3390/antib9020012

103. Hu Y, Zhao T, Zhang N, Zhang Y, Cheng L. A review of recent advances and research on drug target identification methods. *Curr Drug Metab.* 2019;20(3):209-216.

104. Ding H, Takigawa I, Mamitsuka H, Zhu S. Similarity-based machine learning methods for predicting drug-target interactions: a brief review. *Brief Bioinform.* 2014;15(5):734-747.

105. Huang G, Yan F, Tan D. A review of computational methods for predicting drug targets. *Curr Protein Pept Sci.* 2018;19(6):562-572.

106. FDA cleared AI algorithms. American College of Radiology Data Science Institute. Accessed December 21, 2020. https://www.acrdsi.org/DSI-Services/FDA-Cleared-AI-Algorithms

107. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med.* 2020;3(1):118.

108. TrueFidelity. GE Healthcare. Accessed December 21, 2020. https://www.gehealthcare.com/products/truefidelity

109. AiCE. Canon. Accessed December 21, 2020. https://global.medical.canon/products/computed-tomography/aice

110. SubtlePET. Subtle Medical. Accessed December 21, 2020. https://subtlemedical.com/subtlepet/

111. iQMR (intelligent Quick Magnetic Resonance). MedicVision. Accessed December 21, 2020. https://medicvision.com/en/iqmr

112. Caption AI. Caption Health. Accessed December 21, 2020. https://captionhealth.com/products/caption-ai/

113. Enlitic. Accessed December 21, 2020. https://www.enlitic.com/

114. Aidoc. Accessed December 21, 2020. https://www.aidoc.com/

115. Viz.ai. Accessed December 21, 2020. https://viz.ai/

116. Zebra-Med. Zebra Medical Vision. Accessed December 21, 2020.

117. AI-Rad companion. Siemens. Accessed December 21, 2020. https://www.siemens-healthineers.com/en-us/digital-health-solutions/digital-solutions-overview/clinical-decision-support/ai-rad-companion

118. OsteoDetect. Imagen. Accessed December 21, 2020.

119. FDA permits marketing of artificial intelligence algorithm for aiding providers in detecting wrist fractures. News release. US Food and Drug Administration; 2018. Accessed December 21, 2020.

https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-algorithm-aiding-providers-detecting-wrist-fractures

120. Icometrix. Accessed December 21, 2020. https://icometrix.com/
121. Cardio AI. Arterys. Accessed December 21, 2020. https://arterys.com/
122. Enterprise imaging solutions. IBM Watson Health. Accessed December 21, 2020. https://www.ibm.com/watson-health/solutions/enterprise-imaging
123. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology*. 2016;278(2):563-577.
124. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ. Artificial intelligence in radiology. *Nat Rev Cancer*. 2018;18(8):500-510.
125. Recht MP, Dewey M, Dreyer K, et al. Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations. *Eur Radiol.* 2020;30(6):3576–3584.
126. Chockley K, Emanuel E. The end of radiology? Three threats to the future practice of radiology. *J Am Coll Radiol.* 2016;13(12):1415-1420.
127. Reardon S. Rise of robot radiologists. *Nature.* 2019;576(7787):S54.
128. Jha S. Will computers replace radiologists?. *Medscape Radiology*. May 12, 2016. https://www.medscape.com/viewarticle/863127
129. Kalmet PHS, Sanduleanu S, Primakov S, et al. Deep learning in fracture detection: a narrative review. *Acta Orthop.* 2020;91(2):215-220.
130. Langerhuizen DWG, Janssen SJ, Mallee WH, et al. What are the applications and limitations of artificial intelligence for fracture detection and classification in orthopaedic trauma imaging? A systematic review. *Clin Orthop Relat Res.* 2019;477(11):2482-2491.
131. US Food and Drug Administration. Emergency Use Authorization Letter to CLEW Medical Ltd. May 26, 2020.
132. *K191370 510(k) Summary*. US Food and Drug Administation; 2019.
133. *K143643 510(k) Summary*. US Food and Drug Administation; 2014.
134. Classify your medical device. US Food and Drug Administration. Published 2020. Accessed December 21, 2020. https://www.fda.gov/medical-devices/overview-device-regulation/classify-your-medical-device.
135. VITEK MS: healthcare. bioMerieux. Accessed December 21, 2020. https://www.biomerieux-usa.com/clinical/vitek-ms-healthcare.
136. Yao X, McCoy RG, Friedman PA, et al. ECG AI-guided screening for low ejection fraction (EAGLE): rationale and design of a pragmatic cluster randomized trial. *Am Heart J.* 2020;219:31-36.
137. Cardiomatics. Accessed December 21, 2020. http://cardiomatics.com/
138. FibriCheck. Accessed December 21, 2020. https://www.fibricheck.com/
139. PhysIQ. Accessed September 14, 2020.
140. Zio AT. iRhythm Technologies. Accessed December 21, 2020. https://www.irhythmtech.com/products-services/zio-at
141. Portfolio. Biofourmis. Accessed December 21, 2020. https://www.biofourmis.com/solutions/
142. Product. Cardiologs. Accessed December 21, 2020. https://cardiologs.com/#product
143. Kardia. Alivecor. Accessed December 21, 2020. https://clinicians.alivecor.com/

144. AI-ECG Tracker. Carewell. Accessed December 21, 2020. https://www.carewellhealth.com/products_aiecg.html
145. CORE digital attachment. Eko. Accessed December 21, 2020. https://shop.ekohealth.com/products/core-digital-attachment
146. A study to assess the effectiveness of an atrial fibrillation (AF) risk prediction algorithm and diagnostic test in identifying patients with AF (PULsE AI). ClinicalTrials.gov identifier: NCT04045639. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT04045639
147. eMurmur connect. eMurmur. Accessed December 21, 2020. https://emurmur.com/
148. Steth IO. Accessed December 21, 2020. https://stethio.com/clinicians/
149. BrainScope. Accessed December 21, 2020. https://brainscope.com/
150. EnsoSleep. Accessed December 21, 2020. https://www.ensodata.com/
151. ResApp. Accessed December 21, 2020. https://www.resapphealth.com.au/
152. RDAD (Rare Disease Auxiliary Diagnosis System). Accessed December 21, 2020. http://119.3.41.228:8080/RDAD/index.php
153. Rady Children's Institute for Genomic Medicine (RCIGM). Accessed December 21, 2020. https://www.radygenomics.org/2019/04/24/rady-childrens-institute-for-genomic-medicine-uses-artificial-intelligence-to-diagnose-genetic-diseases/
154. Brasil S, Pascoal C, Francisco R, et al. Artificial intelligence (AI) in rare diseases: is the future brighter? *Genes (Basel).* 2019;10(12):978.
155. K080896 510(k) summary. US Food and Drug Administration. Published 2008. Accessed December 21, 2020. https://www.accessdata.fda.gov/cdrh_docs/reviews/K080896.pdf
156. Tempus. Accessed December 21, 2020. https://www.tempus.com/
157. Altoida. Accessed December 21, 2020. https://altoida.com/
158. Liang H, Tsui BY, Ni H, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat Med.* 2019;25(3):433-438.
159. Symptomate. Accessed December 21, 2020. https://symptomate.com/.
160. Berry AC, Cash BD, Wang B, et al. Online symptom checker diagnostic and triage accuracy for HIV and hepatitis C. *Epidemiol Infect.* 2019;147:e104-e104. doi:10.1017/S0950268819000268
161. Healthily. Your.MD. Accessed December 21, 2020. https://www.livehealthily.com/
162. Fleming J, Jeannon J-P. Head and neck cancer in the digital age: an evaluation of mobile health applications. *BMJ Innov.* 2020;6(1):13-17.
163. Ada. Accessed December 21, 2020. https://ada.com/
164. Buoy. Accessed December 21, 2020. https://www.buoyhealth.com/
165. HealthTap. Accessed December 21, 2020.
166. K Health. Accessed December 21, 2020. https://www.khealth.ai/
167. Babylon. Accessed December 21, 2020. https://www.babylonhealth.com/us/what-we-offer/chatbot
168. Strickland E. Hospitals fight sepsis with AI: by predicting cases, sepsis watch could save lives. *IEEE Spectr.* 2018;55(11):9-10.
169. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. *Sci Transl Med.* 2015;7(299):299ra122. DOI: 10.1126/scitranslmed.aab3719
170. Dascena. Accessed December 21, 2020. https://www.dascena.com/
171. eCART. AgileMD. Accessed December 21, 2020. https://www.agilemd.com/#our-solutions

172. Evans RS, Benuzillo J, Horne BD, et al. Automated identification and predictive tools to help identify high-risk heart failure patients: pilot evaluation. *J Am Med Inf Assoc.* 2016;23(5):872-878.

173. Seattle heart failure model. University of Washington. Accessed December 21, 2020. https://depts.washington.edu/shfm/

174. Jvion CORE. Jvion. Accessed December 21, 2020.

175. KelaHealth. Accessed December 21, 2020. https://www.kelahealth.com/new-index.

176. WAVE Clinical Platform. Accessed December 21, 2020. https://www.excel-medical.com/predictive-algorithms.

177. Singh K, Valley TS, Tang S, et al. Validating a widely implemented deterioration index model among hospitalized COVID-19 patients. *medRxiv.* Preprint posted online April 29, 2020. https://doi.org/10.1101/2020.04.24.20079012

178. Realtime streaming clinical use engine for medical escalation (ReSCUE-ME). ClinicalTrials.gov identifier: NCT04026555. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT04026555

179. CLEW. Accessed December 21, 2020. https://clewmed.com/

180. phisIQ. Accessed December 21, 2020. https://www.physiq.com/

181. current health. Accessed December 21, 2020. https://currenthealth.com/

182. Loop system. Spry. Accessed December 21, 2020.

183. VA REACH VET initiative helps save veterans lives. News release, April 3, 2017. US Department of Veteran Affairs. Accessed January 31, 2021. https://www.va.gov/opa/pressrel/pressrelease.cfm?id=2878

184. Improving quality of care—managing atrial fibrillation through care teams and health information technology (IQ-MATCH). ClinicalTrials.gov identifier: NCT02734875. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT02734875

185. Project to improve communication about serious illness—pilot study (PICSI-P). ClinicalTrials.gov identifier: NCT03746392. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT03746392

186. Implementation of a tool to identify social weaknesses in the cancer care pathway (iDEFCO). ClinicalTrials.gov identifier: NCT04015895. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT04015895

187. Lark Diabetes Care. Accessed December 21, 2020. https://www.lark.com/diabetes/

188. Wellthy. Accessed December 21, 2020. https://www.wellthytherapeutics.com/product/

189. One Drop. Accessed December 21, 2020. https://onedrop.today/

190. Caballero-Ruiz E, Garcia-Saez G, Rigla M, Villaplana M, Pons B, Hernando ME. A web-based clinical decision support system for gestational diabetes: automatic diet prescription and detection of insulin needs. *Int J Med Inform.* 2017;102:35-49.

191. Dynamically tailored behavioral interventions in diabetes. ClinicalTrials.gov identifier: NCT04226027. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT04226027

192. Lark DPP. Accessed December 21, 2020. https://www.lark.com/dpp-diabetes-prevention-program/

193. Patients. Day Two. Accessed December 21, 2020. https://www.daytwo.com/patients/

194. A personalized diet study to reduce glycemic exposure. ClinicalTrials.gov identifier: NCT03336411. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT03336411

195. Lark for Hypertension. Accessed December 21, 2020. https://www.lark.com/hypertension/

196. Improving adherence and outcomes by artificial intelligence-adapted etxt messages (AIM@BP). ClinicalTrials.gov identifier: NCT02454660. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT02454660

197. Examining the feasibility of the Ask RoSE mobile mental health application. ClinicalTrials.gov identifier: NCT03909685. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT03909685

198. Interactive communication technologies and nicotine replacement therapy sampling for smokers. ClinicalTrials.gov identifier: NCT04001972. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT04001972

199. Omada. Accessed December 21, 2020. https://www.omadahealth.com/

200. AI-based social software to manage warfarin therapy (AI-SMART). ClinicalTrials.gov identifier: NCT03870581. Published 2019. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT03870581

201. Advisor Pro. DreaMed. Accessed December 21, 2020.

202. Predictive accuracy of MATRx Plus in identifying favorable candidates for oral appliance therapy. ClinicalTrials.gov identifier: NCT03217383. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT03217383

203. Artificial intelligence for optimal anemia management in end-stage renal disease: the Anemia Control Model (ACM) trial (ANEMEX). ClinicalTrials.gov identifier: NCT03214627. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT03214627

204. Artificial intelligence assisted insulin titration system (iNCDSS). ClinicalTrials.gov identifier: NCT04053959. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT04053959

205. Piloting healthcare coordination in hypertension (PRECISION). ClinicalTrials.gov identifier: NCT02988193. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT02988193

206. Evaluation of a dashboard for diabetes care integrated with the electronic health record. ClinicalTrials.gov identifier: NCT03826290. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT03826290

207. Risk and benefit informed MTM pharmacist intervention in heart failure. ClinicalTrials.gov identifier: NCT03804606. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT03804606

208. Watson for genomics. IBM. Accessed December 21, 2020. https://www.ibm.com/products/watson-for-genomics/details

209. Arm motor rehabilitation, entertainment and cognition system for the elderly. ClinicalTrials.gov identifier: NCT04252170. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT04252170

210. Short arm human centrifuge therapeutic training and rehabilitation (GRACER1). ClinicalTrials.gov identifier: NCT04369976. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT04369976

211. Assisted rehabilitation care during post-stroke mANaGement: fEasibiLity assessment (ARCANGEL). ClinicalTrials.gov identifier: NCT03787433. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT03787433

212. Self-management behaviors of caregivers of the chronically critically ill (ASSIST). ClinicalTrials.gov identifier: NCT03065829. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT03065829

213. Wysa. Accessed December 21, 2020. https://www.wysa.io/

214. Woebot. Accessed December 21, 2020. https://woebothealth.com/
215. MiniMed 780G. Medtronic. Accessed December 21, 2020. https://www.medtronic-diabetes.co.uk/insulin-pump-therapy/minimed-780g-system
216. Our technology. Beta Bionics. Accessed December 21, 2020. https://www.betabionics.com/technology
217. Fall detection. Apple Watch. Accessed December 21, 2020. https://support.apple.com/en-us/HT208944
218. HeartHero. Accessed December 21, 2020. https://hearthero.com/
219. MATRx plus. Zephyr. Accessed December 21, 2020. https://zephyrsleep.com/for-professionals/products/
220. BrainQ. Accessed December 21, 2020. https://brainqtech.com/
221. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J Am Med Inform Assoc.* 2016;23(5):1007-1015.
222. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Med Inform.* 2019;7(2):e12239 https://dx.doi.org/10.2196/12239
223. Burckhardt P, Padman R. deidentify. In: *AMIA Annual Symposium Proceedings AMIA Symposium.* April 16, 2018;2017:485-494.
224. Friedlin FJ, McDonald CJ. A software tool for removing patient identifying information from clinical documents. *J Am Med Inform Assoc.* 2008;15(5):601-610.
225. Meystre S, Friedlin F, South B, Shen S, Samore M. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Med Res Methodol.* 2010;10:70.
226. How the companion system works. CompanionMx. Accessed December 21, 2020. https://companionmx.com/product/
227. Ginger. Accessed December 21, 2020. https://www.ginger.io/
228. RX-1 mini. Rhythm Express. Accessed December 21, 2020. https://rhythmexpressecg.com/
229. Sepsis watch: the implementation of a Duke-specific early warning system for sepsis. Sepsis Watch. Accessed December 21, 2020. https://dihi.org/project/sepsiswatch/
230. Insight by Dascena. Accessed December 21, 2020. https://www.dascena.com/insight
231. Poon CC, Liu Q, Gao H, Lin W-H, Zhang Y-T. Wearable intelligent systems for e-health. *J Comput Sci Eng.* 2011;5(3):246-256.
232. Baig MM, Afifi S, GholamHosseini H, Mirza F. A systematic review of wearable sensors and IoT-based monitoring applications for older adults: a focus on ageing population and independent living. *J Med Syst.* 2019;43(8):11.
233. Loncar-Turukalo T, Zdravevski E, Machado da Silva J, Chouvarda I, Trajkovik V. Literature on wearable technology for connected health: scoping review of research trends, advances, and barriers. *J Med Internet Res.* 2019;21(9):e14017. https://doi.org/10.2196/14017
234. How it works. Owlytics. Accessed December 21, 2020.
235. Bae S, Massie AB, Thomas AG, et al. Who can tolerate a marginal kidney? Predicting survival after deceased donor kidney transplant by donor-recipient combination. *Am J Transplant.* 2019;19(2):425-433.
236. About Karantis360. Karantis. Accessed December 21, 2020. https://karantis360.com/about-karantis360/
237. Sugar.IQ. Medtronic. Accessed December 21, 2020. https://www.medtronicdiabetes.com/products/sugar.iq-diabetes-assistant

238. McCoy A, Das R. Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units. *BMJ Open Qual.* 2017;6(2):e000158. doi:10.1136/bmjoq-2017-000158

239. Burdick H, Pino E, Gabel-Comeau D, et al. Effect of a sepsis prediction algorithm on patient mortality, length of stay and readmission: a prospective multicentre clinical outcomes evaluation of real-world patient data from US hospitals. *BMJ Health Care Inform.* 2020;27(1):e100109. doi:10.1136/bmjhci-2019-100109

240. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ Open Respir Res.* 2017;4(1):e000234. doi: 10.1136/bmjresp-2017-000234

241. Calvert JS, Price DA, Chettipally UK, et al. A computational approach to early sepsis detection. *Comput Biol Med.* 2016;74:69-73.

242. Attia ZI, Kapa S, Yao X, et al. Prospective validation of a deep learning electrocardiogram algorithm for the detection of left ventricular systolic dysfunction. *J Cardiovasc Electrophysiol.* 2019;30(5):668-674.

243. Sosale AR, Shaikh M, Shah A, et al. Real-world effectiveness of a digital therapeutic in improving glycaemic control in South Asians living with type 2 diabetes. *Diabetes.* 2018;67(suppl 1):866-P.

244. McKenzie AL, Hallberg SJ, Creighton BC, et al. A novel intervention including individualized nutritional recommendations reduces hemoglobin A1c level, medication use, and weight in type 2 diabetes. *JMIR Diabetes.* 2017;2(1):e5. doi: 10.2196/diabetes.6981

245. Gatsios D, Antonini A, Gentile G, et al. Feasibility and utility of mHealth for the remote monitoring of parkinson disease: ancillary study of the PD_manager randomized controlled trial. *JMIR Mhealth Uhealth.* 2020;8(6):e16414. doi:10.2196/16414

246. Ipsos MORI and York Health Economics Consortium. Evaluation of Babylon GP at hand: final evaluation report. NHS Hammersmith and Fulham Clinical Commissioning Group website. Published May 2019. Accessed January 31, 2021. https://www.hammersmithfulhamccg.nhs.uk/media/156123/Evaluation-of-Babylon-GP-at-Hand-Final-Report.pdf

247. RCT of sepsis machine learning algorithm. ClinicalTrials.gov identifier: NCT03882476. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT03882476

248. Jungmann SM, Brand S, Kolb J, Witthöft M. Do Dr. Google and health apps have (comparable) side effects? An experimental study. *Clin Psychol Sci.* 2020;8(2):306-317.

249. The clinical advisory board. Healthily. Accessed December 21, 2020. https://www.livehealthily.com/blog/meet-your-mds-advisory-board

250. Patient safety. Babylon. Accessed December 21, 2020. https://www.babylonhealth.com/responsibility/patient-safety-and-accountability

251. Goldstein MM, Bowers DG. The patient as consumer: empowerment or commodification? Currents in contemporary bioethics. *J Law Med Ethics.* 2015;43(1):162-165.

252. Mun S, Park JH, Baek SM, Lee M, Choi SM, Lee S. Self-care use patterns in the UK, US, Australia, and Japan: a multinational web-based survey. *Integr Med Res.* 2016;5(2):151-160.

253. Lober WB, Flowers JL. Consumer empowerment in health care amid the internet and social media. *Semin Oncol Nurs.* 2011;27(3):169-182. https://doi.org/10.1016/j.soncn.2011.04.002

254. Arnold T, Kasenberg D, Scheutz M. Value alignment or misalignment: what will keep systems accountable? Paper presented at: Thirty-First AAAI Conference on Artificial Intelligence; February 4-9, 2017; San Francisco, CA.

255. Osoba OA, Boudreaux B, Yeung D. Steps towards value-aligned systems. Paper presented at: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society; February 7-8, 2020; New York, NY. https://doi.org/10.1145/3375627.3375872

256. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. *PLoS Med.* 2018;15(11):e1002689. https://doi.org/10.1371/journal.pmed.1002689

257. Healthy people 2020: framework. US Department of Health and Human Services. Accessed December 21, 2020. https://www.healthypeople.gov/sites/default/files/HP2020Framework.pdf

258. Conitzer V, Sinnott-Armstrong W, Borg JS, Deng Y, Kramer M. Moral decision making frameworks for artificial intelligence. Paper presented at: Thirty-First AAAI Conference on Artificial Intelligence; February 4-9, 2017; San Francisco, CA. https://ojs.aaai.org/index.php/AAAI/article/view/11140

259. Osoba OA, Boudreaux B, Saunders J, Irwin JL, Mueller PA, Cherney S. *Algorithmic Equity: A Framework for Social Applications.* RAND Corporation; 2019.

260. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366(6464):447-453.

261. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA.* 2017;318(6):517-518.

262. Sendak MP, Ratliff W, Sarro D, et al. Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. *JMIR Med Inform.* 2020;8(7):e15182. doi:10.2196/15182

263. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med.* 2018;169(12):866-872.

264. Chouldechova A, Roth A. The frontiers of fairness in machine learning. *arXiv:181008810.* Preprint posted online October 20, 2018. https://arxiv.org/abs/1810.08810

265. Corbett-Davies S, Goel S. The measure and mismeasure of fairness: a critical review of fair machine learning. *arXiv:180800023.* Preprint posted online August 14, 2018. https://arxiv.org/abs/1808.00023

266. Holstein K, Wortman Vaughan J, Daumé H III, Dudik M, Wallach H. Improving fairness in machine learning systems: what do industry practitioners need? Paper presented at: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; May 4-9, 2019. Galsgow, Scotland. https://doi.org/10.1145/3290605.3300830

267. Ahmad MA, Patel A, Eckert C, Kumar V, Teredesai A. Fairness in machine learning for healthcare. Paper presented at: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; August 23-27, 2020; Virtual Event, USA. https://doi.org/10.1145/3394486.3406461

268. Pfohl SR, Foryciarz A, Shah NH. An empirical characterization of fair machine learning for clinical risk prediction. *arXiv:200710306.* Preprint posted online November 20, 2020. https://arxiv.org/abs/2007.10306

269. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in health. *arXiv:200910576.* Preprint posted online October 8, 2020. https://arxiv.org/abs/2009.10576

270. Cahan EM, Hernandez-Boussard T, Thadaney-Israni S, Rubin DL. Putting the data before the algorithm in big data addressing personalized healthcare. *NPJ Digit Med.* 2019;2(1):78.

271. Zou J, Schiebinger L. AI can be sexist and racist—it's time to make it fair. *Nature.* 2018;559(7714):324-326.

272. Portela MC, Pronovost PJ, Woodcock T, Carter P, Dixon-Woods M. How to study improvement interventions: a brief overview of possible study types. *BMJ Qual Saf.* 2015;24(5):325-336.

273. Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: a survey on methods and metrics. *Electronics.* 2019;8(8):832.

274. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol.* 2019;110:12-22.

275. Jarrett MP. Cybersecurity—a serious patient care concern. *JAMA.* 2017;318(14):1319-1320.

276. Health Care Industry Cybersecurity Task Force. Report on improving cybersecurity in the health care industry. Public Health Emergency website. Published June 2017. Accessed January 30, 2021. https://www.phe.gov/Preparedness/planning/CyberTF/Documents/report2017.pdf

277. 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;104(12):510-520.

278. Bernal J, Kushibar K, Asfaw DS, et al. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. *Artif Intell Med.* 2019;95:64-81.

279. Biswas M, Kuppili V, Saba L, et al. State-of-the-art review on deep learning in medical imaging. *Front Biosci (Landmark Ed).* 2019;24:392-426.

280. Iezzi R, Goldberg SN, Merlino B, Posa A, Valentini V, Manfredi R. Artificial intelligence in interventional radiology: a literature review and future perspectives. *J Oncol.* 2019;2019:6153041. https://doi.org/10.1155/2019/6153041

281. Lenchik L, Heacock L, Weaver AA, et al. Automated segmentation of tissues using CT and MRI: a systematic review. *Acad Radiol.* 2019;26(12):1695-1706.

282. Liu SF, Wang Y, Yang X, et al. Deep learning in medical ultrasound analysis: a review. *Engineering.* 2019;5(2):261-275.

283. Song J, Yin Y, Wang H, Chang Z, Liu Z, Cui L. A review of original articles published in the emerging field of radiomics. *Eur J Radiol.* 2020;127:108991.

284. Myers KD, Knowles JW, Staszak D, et al. Precision screening for familial hypercholesterolaemia: a machine learning study applied to electronic health encounter data. *Lancet Digit Health.* 2019;1(8):e393-e402. https://doi.org/10.1016/S2589-7500(19)30150-5

285. Cruz Rivera S, Liu X, Chan A-W, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *BMJ* 2020;370:m3210. https://doi.org/10.1136/bmj.m3210

286. Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *BMJ* 2020; 370:m3164. https://doi.org/10.1136/bmj.m3164

287. Topol EJ. Welcoming new guidelines for AI clinical research. *Nat Med.* 2020;26(9):1318-1320.

288. Certify-AI. American College of Radiology Data Science Institute. Accessed December 21, 2020. https://www.acrdsi.org/DSI-Services/Certify-AI

289. Dhruva SS, Mazure CM, Ross JS, Redberg RF. Inclusion of demographic-specific information in studies supporting US Food & Drug Administration approval of high-risk medical devices. *JAMA Intern Med.* 2017;177(9):1390-1391.

290. Fox-Rawlings SR, Gottschalk LB, Doamekpor LA, Zuckerman DM. Diversity in medical device clinical trials: do we know what works for which patients? *Milbank Q.* 2018;96(3):499-529.

291. Cooperative Research Centres (CRC). Accessed December 21, 2020. https://www.industry.gov.au/funding-and-incentives/cooperative-research-centres

292. Active projects. Digital Health CRC. Accessed December 21, 2020. https://www.digitalhealthcrc.com/active-projects/

293. Wilkinson J, Arnold KF, Murray EJ, et al. Viewpoint time to reality check the promises of machine learning- powered precision medicine. *The Lancet.* 2020;7500(20):1-4.

294. Panch T, Mattie H, Celi LA. The "inconvenient truth" about AI in healthcare. *NPJ Digit Med.* 2019;2(1):77-77.

295. Jungmann SM, Klan T, Kuhn S, Jungmann F. Accuracy of a chatbot (Ada) in the diagnosis of mental disorders: comparative case study with lay and expert users. *JMIR Form Res.* 2019;3(4):e13863. https://doi.org/10.2196/13863

296. Miller S, Gilbert S, Virani V, Wicks P. Patients' utilization and perception of an artificial intelligence–based symptom assessment and advice technology in a British primary care waiting room: exploratory pilot study. *JMIR Hum Factors.* 2020;7(3). https://doi.org/10.2196/19713

297. Gilbert S, Mehl A, Baluch A, et al. Original research: how accurate are digital symptom assessment apps for suggesting conditions and urgency advice? a clinical vignettes comparison to GPs. *medRxiv.* Preprint posted online May 13, 2020. https://doi.org/10.1101/2020.05.07.20093872

298. Derome J. Healthcare chatbot diagnosis: will consumers trust them with their health? UserTesting website. Published January 16, 2019. Accessed January 30, 2021. https://www.usertesting.com/blog/healthcare-chatbot-cx-index

299. Burgess M. Can you really trust the medical apps on your phone? Wired. Published 2017. Accessed December 21, 2020. https://www.wired.co.uk/article/health-apps-test-ada-yourmd-babylon-accuracy

300. Reduce medication errors by translating AESOP model into CPOE systems. ClinicalTrials.gov identifier: NCT03484793. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT03484793

301. Nimri R, Muller I, Atlas E, et al. MD-Logic overnight control for 6 weeks of home use in patients with type 1 diabetes: randomized crossover trial. *Diabetes Care.* 2014;37(11):3025-3032.

302. Nimri R, Muller I, Atlas E, et al. Night glucose control with MD-Logic artificial pancreas in home setting: a single blind, randomized crossover trial—interim analysis. *Pediatr Diabetes.* 2014;15(2):91-99.

303. Nimri R, Dassau E, Segall T, et al. Adjusting insulin doses in patients with type 1 diabetes who use insulin pump and continuous glucose monitoring: variations among countries and physicians. *Diabetes Obes Metab.* 2018;20(10):2458-2466.

304. Naunheim RS, Treaster M, English J, Casner T. Automated electroencephalogram identifies abnormalities in the ED. *Am J Emerg Med.* 2011;29(8):845-848.

305. *De Novo Classification Request for Brainscope Ahead 100*. US Food and Drug Administation; 2014.

306. *K161068 510(k) Summary*. US Food and Drug Administation; 2016.

307. AI-assisted insulin titration system on inpatients glucose control. ClinicalTrials.gov Identifier: NCT04517201. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/record/NCT04517201

308. Farris K. Improving adherence and outcomes by artificial intelligence-adapted text messages. AHRQ website. 2017. Accessed December 21, 2020. https://digital.ahrq.gov/ahrq-funded-projects/improving-adherence-and-outcomes-artificial-intelligence-adapted-text-messages/final-report

309. Bucalo ML, Barbieri C, Roca S, et al. The anaemia control model: does it help nephrologists in therapeutic decision-making in the management of anaemia? *Nefrología (English Edition).* 2018;38(5):491-502.

310. Barbieri C, Mari F, Stopper A, et al. A new machine learning approach for predicting the response to anemia treatment in a large cohort of end stage renal disease patients undergoing dialysis. *Comput Biol Med.* 2015;61:56-61.

311. Wang SV, Rogers JR, Jin Y, et al. Stepped-wedge randomised trial to evaluate population health intervention designed to increase appropriate anticoagulation in patients with atrial fibrillation. *BMJ Qual Saf.* 2019;28(10):835-842.

312. Babylon. *NHS 111 Powered by Babylon: Outcomes Evaluation.* Babylon Health website. Published October 2017. Accesses January 30, 2021. https://assets.babylonhealth.com/nhs/NHS-111-Evaluation-of-outcomes.pdf

313. Middleton K, Butt M, Hammeria N, Hambklin S, Mehta K, Parsa A. Sorting out symptoms: design and evaluation of the 'babylon check' automated triage system. *arXiv preprint:160602041*. Preprint posted online June 7, 2016.

314. Razzaki S, Baker A, Perov Y, Middleton K, Baxter J, al. e. A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis. *arXiv preprint:180610698.* Preprint posted online June 27, 2018. https://arxiv.org/abs/1806.10698

315. El-Khatib FH, Balliro C, Hillard MA, et al. Home use of a bihormonal bionic pancreas versus insulin pump therapy in adults with type 1 diabetes: a multicentre randomised crossover trial. *Lancet.* 2017;389(10067):369-380.

316. Feasibility of outpatient closed loop control with the bionic pancreas in cystic fibrosis related diabetes. ClinicalTrials.gov identifier: NCT03258853. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT03258853

317. The insulin-only bionic pancreas pivotal trial. ClinicalTrials.gov identifier: NCT04200313. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT04200313

318. *K183282 510(k) Summary*. US Food and Drug Administation; 2019.

319. Chen J, Yan M, Chin Howe RL, et al. Biovitals™: a personalized multivariate physiology analytics using continuous mobile biosensors. *Conf Proc IEEE Eng Med Biol Soc.* July 23-27,2019; Berlin, Germany.

320. Agarwal P, Mukerji G, Desveaux L, et al. Mobile app for improved self-management of type 2 diabetes: multicenter pragmatic randomized controlled trial. *JMIR Mhealth Uhealth.* 2019;7(1):e10321. https://doi.org/10.2196/10321

321. *Estimating the Economic Impact of a Digital Therapeutic in Type 2 Diabetes.* IBM Watson Health; 2018.

322. Efficacy of EMF BCI based device on acute stroke. ClinicalTrials.gov identifier: NCT04039178. Published 2019. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT04039178

323. Winn AN, Somai M, Fergestrom N, Crotty BH. Association of use of online symptom checkers with patients' plans for seeking care. *JAMA Netw Open.* 2019;2(12):e1918561-e1918561. doi:10.1001/jamanetworkopen.2019.18561

324. Smith SW, Walsh B, Grauer K, et al. A deep neural network learning algorithm outperforms a conventional algorithm for emergency department electrocardiogram interpretation. *J Electrocardiol.* 2019;52:88-95.

325. Smith SW, Rapin J, Li J, et al. A deep neural network for 12-lead electrocardiogram interpretation outperforms a conventional algorithm, and its physician overread, in the diagnosis of atrial fibrillation. *Int J Cardiol Heart Vasc.* 2019;25:100423.

326. *Fact Sheet for Healthcare Providers: Emergency Use of the CLEWICU System During the COVID-19 Pandemic*. US Food and Drug Administation; 2020.

327. Place S, Blanch-Hartigan D, Rubin C, et al. Behavioral indicators on a mobile sensing platform predict clinically validated psychiatric symptoms of mood and anxiety disorders. *J Med Internet Res.* 2017;19(3):e75. https://doi.org/10.2196/jmir.6678

328. Facilitating assessment of at-risk sailors using technology (FAAST). ClinicalTrials.gov identifier: NCT04159480. Accessed December 21, 2020. https://www.clinicaltrials.gov/ct2/show/study/NCT04159480

329. Artificial intelligence/machine learning modeling on time to palliative care review in an inpatient hospital population. ClinicalTrials.gov identifier: NCT03976297. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT03976297

330. Corti. Accessed December 21, 2020. https://www.corti.ai/

331. Blomberg SN, Folke F, Ersbøll AK, et al. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. *Resuscitation.* 2019;138:322-329.

332. Machine learning assisted recognition of out-of-hospital cardiac arrest during emergency calls. ClinicalTrials.gov identifier: NCT04219306. Published 2020. Accessed September 21, 2020. https://clinicaltrials.gov/ct2/show/NCT04219306

333. Efficacy of a self-test and self-alert mobile applet in detecting susceptible infection of COVID-19 (COVID-19). ClinicalTrials.gov identifier: NCT04256395. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT04256395

334. Dartford and Gravesham NHS home health and current partner to reduce hospital readmissions and emergency department visits. Current Health website. Published April 11, 2019. Accessed January 30, 2020. https://currenthealth.com/dartford-and-gravesham-nhs-home-health-and-current-partner-to-reduce-hospital-readmissions-and-emergency-department-visits

335. Zeevi D, Korem T, Zmora N, et al. Personalized nutrition by prediction of glycemic responses. *Cell.* 2015;163(5):1079-1094.

336. Mendes-Soares H, Raveh-Sadka T, Azulay S, et al. Model of personalized postprandial glycemic response to food developed for an Israeli cohort predicts responses in Midwestern American individuals. *Am J Clin Nutr.* 2019;110(1):63-75.

337. Shlomo YB, Azulay S, Raveh-Sadka T, Cohen Y, Hanemann A. 782-P: personalized, machine learning-based nutrition reduces diabetes markers in type 2 diabetic patients. *Diabetes.* 2019;68(suppl 1). https://doi.org/10.2337/db19-782-P

338. Hu L, Wang C, Li H, et al. Does personalized nutrition increase weight loss self-efficacy? *Curr Dev Nutr.* 2020;4(suppl 2):1310-1310. https://doi.org/10.1093/cdn/nzaa059_027

339. Effects of expert arbitration on clinical outcomes when disputes over diagnosis arise between physicians and their artificial intelligence counterparts: a randomized, multicenter trial in pediatric outpatients. ClinicalTrials.gov identifier: NCT04011761. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT04011761

340. Kang MA, Churpek MM, Zadravecz FJ, Adhikari R, Twu NM, Edelson DP. Real-time risk prediction on the wards: a feasibility study. *Crit Care Med.* 2016;44(8):1468-1473.

341. Hirai T, Tate S, Dryer K, et al. Electronic cardiac arrest triage score best predicts mortality after intervention in patients with massive and submassive pulmonary embolism. *Catheter Cardiovasc Interv.* 2018;92(2):366-371.

342. Green M, Lander H, Snyder A, Hudson P, Churpek M, Edelson D. Comparison of the between the flags calling criteria to the MEWS, NEWS and the electronic cardiac arrest risk triage (eCART) score for the identification of deteriorating ward patients. *Resuscitation.* 2018;123:86-91.

343. Bartkowiak B, Snyder AM, Benjamin A, et al. Validating the electronic cardiac arrest risk triage (eCART) score for risk stratification of surgical inpatients in the postoperative setting: retrospective cohort study. *Ann Surg.* 2019;269(6):1059-1063.

344. Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. *Nat Med.* 2019;25(1):70-74.

345. *K192004 510(k) Summary*. US Food and Drug Administation; 2020.

346. Lai LS, Redington AN, Reinisch AJ, Unterberger MJ, Schriefl AJ. Computerized automatic diagnosis of innocent and pathologic murmurs in pediatrics: a pilot study. *Congenit Heart Dis.* 2016;11(5):386-395.

347. Thompson WR, Reinisch AJ, Unterberger MJ, Schriefl AJ. Artificial intelligence-assisted auscultation of heart murmurs: validation by virtual clinical trial. *Pediatr Cardiol.* 2019;40(3):623-629.

348. *K162627 510(k) Summary*. US Food and Drug Administation; 2017.

349. Singh K, Valley TS, Tang S, et al. Evaluating a widely implemented proprietary deterioration index model among hospitalized COVID-19 patients. *medRxiv.* Preprint posted online June 20, 2020. https://doi.org/10.1101/2020.04.24.20079012

350. Cummings BC, Ansari S, Motyka JR, et al. Validation and comparison of PICTURE analytic and Epic deterioration index for COVID-19. *medRxiv.* Preprint posted online June 20, 2020. https://doi.org/10.1101/2020.07.08.20145078

351. Ochsner Health adopts new AI technology to save lives in real-time. News release. Ochsner Health; 2018. Accessed December 21, 2020. https://news.ochsner.org/news-releases/ochsner-health-system-adopts-new-ai-technology-to-save-lives-in-real-time

352. Potolsky A. Implementation of artificial intelligence initiated rapid responses to reduce in-hospital cardiac arrest. Doctor of Nursing Practice (DNP) Projects. 2020. Accessed December 21, 2020. https://repository.usfca.edu/dnp/233/

353. Pluymaekers N, Hermans ANL, van der Velden RMJ, et al. On-demand app-based rate and rhythm monitoring to manage atrial fibrillation through teleconsultations during COVID-19. *Int J Cardiol Heart Vasc.* 2020;28:100533.

354. Proesmans T, Mortelmans C, Van Haelst R, Verbrugge F, Vandervoort P, Vaes B. Mobile phone-based use of the photoplethysmography technique to detect atrial fibrillation in primary care: diagnostic accuracy study of the FibriCheck app. *JMIR Mhealth Uhealth.* 2019;7(3):e12284. https://doi.org/10.2196/12284

355. *K173872 510(k) Summary*. US Food and Drug Administation; 2018.

356. IN-TANDEM familial hypercholesterolemia pilot study. ClinicalTrials.gov identifier: NCT03253432. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT03253432

357. Heart Failure Risk Calculator. Accessed December 21, 2020. http://www.heartfailurerisk.org/

358. Pocock SJ, Ariti CA, McMurray JJ, et al. Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. *Eur Heart J.* 2013;34(19):1404-1413.

359. Wijnberge M, Geerts BF, Hol L, et al. Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. *JAMA.* 2020;323(11):1052-1060.

360. The Ibis health management solution. Senscio Systems. Accessed December 21, 2020. https://www.sensciosystems.com/

361. Romero-Brufau S, Wyatt KD, Boyum P, Mickelson M, Moore M, Cognetta-Rieke C. Implementation of artificial intelligence-based clinical decision support to reduce hospital readmissions at a regional hospital. *Appl Clin Inform.* 2020;11(4):570-577.

362. Ravi V, Zheng J, Subramaniam A, et al. Artificial intelligence (AI) and machine learning (ML) in risk prediction of hospital acquired pressure injuries (HAPIs) among oncology inpatients. *J Clin Oncol.* 2019;37(suppl 15):e18095-e18095. DOI: 10.1200/JCO.2019.37.15_suppl.e18095

363. Koren G, Souroujon D, Shaul R, et al. "A patient like me": an algorithm-based program to inform patients on the likely conditions people with symptoms like theirs have. *Medicine (Baltimore).* 2019;98(42):e17596. doi:10.1097/MD.0000000000017596

364. Reed MJ, Grubb NR, Lang CC, et al. Multi-centre randomised controlled trial of a smartphone-based event recorder alongside standard care versus standard care for patients presenting to the emergency department with palpitations and pre-syncope: the IPED (investigation of palpitations in the ED) study. *EClinicalMedicine.* 2019;8:37-46.

365. Godin R, Yeung C, Baranchuk A, Guerra P, Healey JS. Screening for atrial fibrillation using a mobile, single-lead electrocardiogram in Canadian primary care clinics. *Can J Cardiol.* 2019;35(7):840-845.

366. Duarte R, Stainthorpe A, Greenhalgh J, et al. Lead-I ECG for detecting atrial fibrillation in patients with an irregular pulse using single time point testing: a systematic review and economic evaluation. *Health Technol Assess.* 2020;24(3):1-164.

367. Wegner FK, Kochhauser S, Ellermann C, et al. Prospective blinded evaluation of the smartphone-based AliveCor Kardia ECG monitor for atrial fibrillation detection: the PEAK-AF study. *Eur J Intern Med.* 2020;73:72-75.

368. Selder JL, Breukel L, Blok S, van Rossum AC, Tulevski I, Allaart CP. A mobile one-lead ECG device incorporated in a symptom-driven remote arrhythmia monitoring program. The first 5,982 Hartwacht ECGs. *Neth Heart J.* 2019;27(1):38-45.

369. KDPI-EPTS survival benefit estimator. Transplant Models. Accessed December 21, 2020. http://www.transplantmodels.com/kdpi-epts/

370. Vascular case study: prediction of surgical site infection and risk-based use of negative pressure therapy on closed incisions for vascular surgery. Kelahealth. Published 2019. Accessed December 21, 2020. https://www.kelahealth.com/new-index

371. Cardiac case study: prediction of hospital length of stay and risk-based recommendation of postoperative care setting for patients receiving an aortic valve replacement. Kelahealth. Published 2019. Accessed December 21, 2020. https://www.kelahealth.com/new-index

372. CALYPSO pilot study: machine learning based predictions of surgical complications (CALYPSO). ClinicalTrials.gov identifier: NCT02828475. Accessed December 21, 2020. https://www.clinicaltrials.gov/ct2/show/study/NCT02828475

373. Stein N. *Lark for Diabetes Study Shows Positive Outcomes for People With Type 2 Diabetes.* Lark; 2019.

374. Stein N, Brooks K. A fully automated conversational artificial intelligence for weight loss: longitudinal observational study among overweight and obese adults. *JMIR Diabetes.* 2017;2(2):e28. https://doi.org/10.2196/diabetes.8590

375. Stein N. *One-Year Clinical Outcomes of an Artificial Intelligence-Based Digital Diabetes Prevention Program*. Lark; 2020. Accessed December 21, 2020. https://lark.com/wp-content/uploads/2020/05/Lark-1-year-Outcomes-of-AI-based-DPP.pdf

376. Persell SD, Peprah YA, Lipiszko D, et al. Effect of home blood pressure monitoring via a smartphone hypertension coaching application or tracking application on adults with uncontrolled hypertension: a randomized clinical trial. *JAMA Netw Open.* 2020;3(3):e200255. doi:10.1001/jamanetworkopen.2020.0255

377. Lark & Omron. Analysis shows personalized health coaching, blood pressure monitoring leads to control. 2018. Accessed December 21, 2020. https://lark.com/wp-content/uploads/2020/05/OmronLark-Health-Blood-Pressure-Study.pdf

378. Remmers JE, Topor Z, Grosse J, et al. A feedback-controlled mandibular positioner identifies individuals with sleep apnea who will respond to oral appliance therapy. *J Clin Sleep Med.* 2017;13(7):871-880.

379. Mosca EV, Topor Z, Grosse J, Bruehlmann S, Jahromi SAZ. Prediction of outcome with oral appliance therapy for obstructive sleep apnea using a feedback controlled mandibular positioner: validation on a new population of obstructive sleep apneics. In: American Thoracic Society, eds. *C77. Predictors of Sleep Disordered Breathing and Response to Treatment.* May 1, 2018:A5892-A5892.

380. Workflow validation of an in-home feedback controlled mandibular positioner. ClinicalTrials.gov identifier: NCT03616327. Published 2019. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT03616327

381. Nimri R, Danne T, Kordonouri O, et al. The "Glucositter" overnight automated closed loop system for type 1 diabetes: a randomized crossover trial. *Pediatr Diabetes.* 2013;14(3):159-167.

382. Phillip M, Battelino T, Atlas E, et al. Nocturnal glucose control with an artificial pancreas at a diabetes camp. *N Engl J Med.* 2013;368(9):824-833.

383. Intelligent care. Nectarine Health. Accessed December 21, 2020. https://nectarinehealth.com/

384. Buegler M, Harms RL, Balasa M, et al. Digital biomarker-based individualized prognosis for people at risk of dementia. *Alzheimers Dement (Amst).* 2020;12(1):e12073. https://doi.org/10.1002/dad2.12073

385. Osborn CY, van Ginkel JR, Rodbard D, et al. One drop mobile: an evaluation of hemoglobin A1c improvement linked to app engagement. *JMIR Diabetes.* 2017;2(2):e21. https://doi.org/10.2196/diabetes.8039

386. Goldner DR, Osborn CY, Sears LE, Huddleston B, Dachis J. A machine-learning model accurately predicts projected blood glucose. *Diabetes.* 2018;67(suppl 1). https://doi.org/10.2337/db18-46-LB

387. Wexler Y, Goldner D, Osborn C, Huddleston B, Dachis J. The Official Journal of ATTD Advanced Technologies & Treatments for Diabetes Conference Madrid, Spain—February 19–22, 2020. *Diabetes Technology & Therapeutics.* 2020;22(suppl 1):A-1-A-250.

388. *K120489 510(k) Summary.* US Food and Drug Administration; 2012.

389. PD Manager. Accessed December 21, 2020. http://www.parkinson-manager.eu/

390. Gage H. A pilot trial of devices monitoring symptoms of Parkinson's disease. ISRCTN identifier 17396879. Accessed January 30, 2021. http://www.isrctn.com/ISRCTN17396879

391. Artificial intelligence in children's clinic. ClinicalTrials.gov identifier: NCT04186104. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT04186104

392. Deep learning for detection of AF/AFLUTTER. Preventice Solutions. Accessed December 21, 2020. https://www.preventicesolutions.com/machine-learning

393. Teplitzky BA, McRoberts M, Ghanbari H. Deep learning for comprehensive ECG annotation. *Heart Rhythm.* 2020;17(5, pt B):881-888.

394. Teplitzky B, McRoberts M, Mehta P, Ghanbari H. S-PO02-197: Real-world performance of atrial fibrillation detection from wearable patch ecg monitoring using deep learning. Poster presented at Heart Rhythm 40th Scientific Sessions May 8-11, 2019, San Francisco, CA. Accessed January 30, 2021. https://cslide-us.ctimeetingtech.com/hrs19/attendee/eposter/poster/652

395. Teplitzky BA, McRoberts M. Fully-automated ventricular ectopic beat classification for use with mobile cardiac telemetry. Paper presented at: 2018 IEEE 15th International Conference on Wearable and Implantable Body Sensor Networks (BSN); March 4-7, 2018; Las Vegas, NV.

396. Qventus. Accessed December 21, 2020. https://qventus.com/

397. How Mercy Fort Smith reduced emergency department LWBS rate by 30%. Qventus. Published 2017. Accessed January 31, 2021. https://qventus.com/wp-content/uploads/2017/02/Qventus-Mercy-4Page-Case-Study-WEB_30.pdf

398. How two leading health systems reduced length of stay with a system of action. Qventus. Accessed December 21, 2020.

399. Hong JC, Niedzwiecki D, Palta M, Tenenbaum JD. Predicting emergency visits and hospital admissions during radiation and chemoradiation: an internally validated pretreatment machine learning algorithm. *JCO Clinical Cancer Informatics.* 2018(2):1-11.

400. Hong JC, Eclov NCW, Dalal NH, et al. System for high-intensity evaluation during radiation therapy (SHIELD-RT): a prospective randomized study of machine learning-directed clinical evaluations during radiation and chemoradiation. *J Clin Oncol.* 2020;38(31)3652-3661. https://doi.org/10.1200/jco.20.01688

401. Clark MM, Hildreth A, Batalov S, et al. Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. *Sci Transl Med.* 2019;11(489) DOI:10.1126/scitranslmed.aat6177.

402. Jia J, Wang R, An Z, Guo Y, Ni X, Shi T. RDAD: a machine learning system to support phenotype-based rare disease diagnosis. *Front Genet.* 2018;9:587.

403. REACH VET program evaluation (REACH VET). ClinicalTrials.gov identifier: NCT03280225. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT03280225

404. Swarnkar V, Abeyratne U, Tan J, et al. Stratifying asthma severity in children using cough sound analytic technology. *J Asthma.* 2019. https://doi.org/10.1080/02770903.2019.1684516

405. Claxton S, Porter P, Brisbanne J, et al. Detection of asthma exacerbation in adolescent and adult subjects with chronic asthma using a cough-centred, smartphone-based algorithm. *Respirology.* 2020;25(suppl 1):111-230.

406. Porter P, Claxton S, Brisbane J, et al. Diagnosis of chronic obstructive pulmonary disease (COPD) using a smartphone-based cough-centred algorithm in a mixed disease acute-care cohort. *Respirology.* 2019;24(suppl 2):168-168.

407. Rajput KS, Wibowo S, Hao C, Majmuda M. On arrhythmia detection by deep learning and multidimensional representation. *arxiv:190400138.* Preprint posted online April 11, 2019. https://arxiv.org/abs/1904.00138

408. Adam A, Jain A, Pletnikova A, et al. Use of a mobile app to augment psychotherapy in a community psychiatric clinic: feasibility and fidelity trial. *JMIR Form Res.* 2020;4(7):e17722. https://doi.org/10.2196/17722

409. Levy WC, Mozaffarian D, Linker DT, et al. The Seattle heart failure model: prediction of survival in heart failure. *Circulation.* 2006;113(11):1424-1433.

410. HindSight phase II. ClinicalTrials.gov identifier: NCT04005001. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT04005001

411. Bedoya AD, Futoma J, Clement ME, et al. Machine learning for early detection of sepsis: an internal and temporal validation study. *JAMIA Open.* 2020;3(2):252-260.

412. Implementation and evaluations of sepsis watch. ClinicalTrials.gov identifier: NCT03655626. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT03655626

413. Personalized medical supervision by an evolving expert system and a connected medical device. Evolucare. Accessed December 21, 2020. https://www.evolucare.com/en/support-with-patient-monitoring

414. Evaluation of the impact of the SMART ANGEL™ device on follow-up at home following major or intermediate outpatient surgery (SMART ANGEL 2). ClinicalTrials.gov identifier: NCT04068584. Accessed December 21, 2020. https://clinicaltrials.gov/ct2/show/NCT04068584

415. SOPHiA. Accessed December 21, 2020. https://www.sophiagenetics.com/en_US/home.html

416. Zhong Y, Arunachalam S, Agrawal P, Neemuchwala H, Cordero TL, Kaufman FR. Real-world assessment of Sugar.IQ with Watson—a cognitive computing-based diabetes management solution. *Diabetes.* 2018;67(suppl 1):16-OR. https://doi.org/10.2337/db18-16-OR

417. Arunachalam S, Zhong Y, Abraham SB, et al. 939-P: real-world performance of the guardian connect system with Sugar.IQ. *Diabetes.* 2019;68(suppl 1):939-P. https://doi.org/10.2337/db19-939-P

418. Real-world data from Guardian Connect and Sugar.IQ reveal improved diabetes outcomes. News release. Medtronic; 2019. Accessed December 21, 2020. https://newsroom.medtronic.com/news-releases/news-release-details/real-world-data-guardiantm-connect-and-sugariqtm-reveal-improved

419. Michuda J, Igartua C, Taxter T, Bell JS, Pelossof R, White K. Transcriptome-based cancer type prediction for tumors of unknown origin. *J Clin Oncol.* 2019;37(suppl 15):3081-3081.

420. Michuda J, Leibowitz B, Amar-Farkash S, et al. Multimodal prediction of diagnosis for cancers of unknown primary. *AACR Annual Meeting 2020 Online Proceedings.* April 27-28, 2020;Philadelphia, PA.

421. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. *JMIR Ment Health.* 2018;5(4):e64. https://doi.org/10.2196/mental.9782

422. Stephens TN, Joerin A, Rauws M, Werk LN. Feasibility of pediatric obesity and prediabetes treatment support through Tess, the AI behavioral coaching chatbot. *Transl Behav Med.* 2019;9(3):440-447.

423. Athinarayanan SJ, Adams RN, Hallberg SJ, et al. Long-term effects of a novel continuous remote care intervention including nutritional ketosis for the management of type 2 diabetes: a 2-year non-randomized clinical trial. *Front Endocrinol (Lausanne).* 2019;10:348-348. https://doi.org/10.3389/fendo.2019.00348

424. Fan WT, Qin TT, Bi RR, Kang HQ, Ma P, Gu B. Performance of the matrix-assisted laser desorption ionization time-of-flight mass spectrometry system for rapid identification of streptococci: a review. *Eur J Clin Microbiol Infect Dis.* 2017;36(6):1005-1012.

425. *K124067 510(k) Summary*. US Food and Drug Administation; 2012.

426. *K181412 510(k) Summary*. US Food and Drug Administation; 2018.

427. Somashekhar SP, Sepúlveda M-J, Shortliffe EH, et al. A prospective blinded study of 1000 cases analyzing the role of artificial intelligence: Watson for oncology and change in decision making of a Multidisciplinary Tumor Board (MDT) from a tertiary care cancer center. *J Clin Oncol.* 2019;37(suppl 15):6533-6533.

428. Patel NM, Michelini VV, Snell JM, et al. Enhancing next-generation sequencing-guided cancer care through cognitive computing. *Oncologist.* 2018;23(2):179-185.

429. Hravnak M, Devita MA, Clontz A, Edwards L, Valenta C, Pinsky MR. Cardiorespiratory instability before and after implementing an integrated monitoring system. *Crit Care Med.* 2011;39(1):65-72.

430. Tarassenko L, Hann A, Young D. Integrated monitoring and analysis for early warning of patient deterioration. *Br J Anaesth.* 2006;97(1):64-68.

431. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR Ment Health.* 2017;4(2):e19. https://doi.org/10.2196/mental.7785

432. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR Mhealth Uhealth.* 2018;6(11):e12106. https://doi.org/10.2196/12106

433. Steinhubl SR, Waalen J, Edwards AM, et al. Effect of a home-based wearable continuous ecg monitoring patch on detection of undiagnosed atrial fibrillation: the mSToPS randomized clinical trial. *JAMA.* 2018;320(2):146-155.

434. Kaura A, Sztriha L, Chan FK, et al. Early prolonged ambulatory cardiac monitoring in stroke (EPACS): an open-label randomised controlled trial. *Eur J Med Res.* 2019;24(1):25.

435. Yenikomshian M, Jarvis J, Patton C, et al. Cardiac arrhythmia detection outcomes among patients monitored with the Zio patch system: a systematic literature review. *Curr Med Res Opin.* 2019;35(10):1659-1670.