A Path for Translation of Machine Learning Products into Healthcare Delivery

Authors: *Mark P. Sendak,[1] Joshua D'Arcy,[2] Sehj Kashyap,[1,3] Michael Gao,[1] Marshall Nichols,[1] Kristin Corey,[1,2] William Ratliff,[1] Suresh Balu[1,2]

1. Duke Institute for Health Innovation, Durham, North Carolina, USA
2. Duke University School of Medicine, Durham, North Carolina, USA
3. Department of Biomedical Data Science at Stanford Medicine, Palo Alto, California, USA

*Correspondence to mark.sendak@duke.edu

Disclosure: Dr Sendak, Dr Gao, Dr Balu, and Dr Nichols are named inventors of the Sepsis Watch deep-learning model, which was licensed from Duke University, Durham, North Carolina, USA, by Cohere Med Inc., Burlingame, California, USA. Dr Sendak, Dr Gao, Dr Balu, and Dr Nichols do not hold any equity in Cohere Med Inc. The other authors have declared no conflicts of interest.

Received: 25.09.2019

Accepted: 10.12.2019

Keywords: Healthcare, innovation, machine learning, translational research.

Citation: EMJ Innov. 2020; DOI/10.33590/emjinnov/19-00172.

Abstract

Despite enormous enthusiasm, machine learning models are rarely translated into clinical care and there is minimal evidence of clinical or economic impact. New conference venues and academic journals have emerged to promote the proliferating research; however, the translational path remains unclear. This review undertakes the first in-depth study to identify how machine learning models that ingest structured electronic health record data can be applied to clinical decision support tasks and translated into clinical practice. The authors complement their own work with the experience of 21 machine learning products that address problems across clinical domains and across geographic populations. Four phases of translation emerge: design and develop, evaluate and validate, diffuse and scale, and continuing monitoring and maintenance. The review highlights the varying approaches taken across each phase by teams building machine learning products and presents a discussion of challenges and opportunities. The translational path and associated findings are instructive to researchers and developers building machine learning products, policy makers regulating machine learning products, and health system leaders who are considering adopting a machine learning product.

INTRODUCTION

Machine learning is a set of statistical and computational techniques that is becoming increasingly prominent in the lay press and medical research. Outside of healthcare, machine learning was quickly adopted to recommend movies and music, annotate images, and translate language. In healthcare, in which the stakes are high, although the enthusiasm surrounding machine learning is immense,[1] the evidence of clinical impact remains scant. New platforms were created to disseminate machine learning research in healthcare, such as the Machine Learning for Healthcare Conference (MLHC), and certain academic journals have been created to provide a platform for the proliferating research. This enthusiasm for novel technologies unfortunately

overshadows the challenging path to successfully translate machine learning technologies into routine clinical care.

National and international efforts are underway to ensure that appropriate guardrails are in place for machine learning to become part of routine care delivery. The International Medical Device Regulators Forum (IMDRF) has defined Software as a Medical Device as "software intended to be used for medical purposes that performs its objectives without being part of a hardware medical device."[2] European regulatory agencies and the U.S. Food and Drug Administration (FDA) are embracing Software as a Medical Device frameworks to regulate machine learning technologies, and national strategies for machine learning are emerging.[3-6] Regulations are actively being developed and implemented, with new guidance from the FDA in September 2019[7] and upcoming changes in European Union Medical Device Regulation (EU MDR) in March 2020.[8] While regulators and medical professional societies proactively shape the machine learning ecosystem, many challenges remain to achieve the anticipated benefits.

This narrative review proposes a general framework for translating machine learning into healthcare. The framework draws upon the authors' experience building and integrating machine learning products within a local setting as well as 21 case studies of machine learning models that are being integrated into clinical care. This review focusses on machine learning models that input data from electronic health records (EHR) applied to clinical decision support tasks, rather than models applied to automation tasks.[9] Automation tasks are cases in which "a machine operates independently to complete a task," whereas clinical decision support tasks are cases in which "a machine is concerned with providing information or assistance to the primary agent responsible for task completion."[9] Distinct from prior systematic reviews of EHR models,[10,11] the current review focusses on models that have been productised and integrated into clinical care rather than the large body of academic work of published models that are not integrated. The review builds upon related work that highlights how academic and industry partners collaborate to develop machine learning products,[12] as well as the need for engagement from front-line clinicians and standard reporting.[13] Case studies were selected amongst 1,672 presentations at 9 informatics and machine learning conferences between January 2018 and October 2019. The conferences include American Medical Informatics Association (AMIA) Annual Symposia and Summits, MLHC, Health Information and Management Systems Society (HIMSS) Machine Learning and Artificial Intelligence Forum, and the Health AI Deployment Symposium. Machine learning technologies were included as case studies if they met two criteria: 1) they tackle a clinical problem using solely EHR data; and 2) they are evaluated and validated through direct integration with an EHR to demonstrate clinical, statistical, or economic utility. Machine learning technologies that analysed images were excluded. This review also advances prior work to propose best practices for teams building machine learning models within a healthcare setting[14] and for teams conducting quality improvement work following the learning health system framework.[15] However, there is not a unifying translational path to inform teams beyond success within a single setting to diffuse and scale across healthcare. This review fills that gap and highlights how teams building machine learning products approach clinical translation and discusses challenges and opportunities for improvement.

MACHINE LEARNING APPLIED TO CLINICAL DECISION SUPPORT TASKS

Machine learning has been described as "the fundamental technology required to meaningfully process data that exceed the capacity of the human brain to comprehend."[16] Machine learning models are often trained on millions of pieces of information. Existing knowledge about individual data elements and relationships between data elements are not explicitly programmed into the model and are instead learned through repeated iterations of mapping between inputs and outputs. This is in contrast to algorithms that comprise predictors and weights that are agreed upon by medical experts. Collaboration between machine learning and clinical experts is crucial and there are a range of modelling techniques that incorporate varying amounts of clinical expertise into model specifications.[17] Machine learning models can be trained in a supervised and unsupervised fashion. Supervised models assume

that the output labels, for example a disease, are known up front, whereas unsupervised models assume that the output labels are unknown. An example of a supervised model is identifying which patients will develop sepsis, a known entity, whereas an example of an unsupervised model is identifying unknown subgroups of asthma patients. Most models integrated into clinical workflows as clinical decision support are supervised machine learning models.

This review focusses on models applied to clinical decision support tasks rather than models applied to automation tasks. Not only does automation involve heightened regulatory burden,[18] but machine learning is initially expected to impact healthcare through augmenting rather than replacing clinical workflows.[19] The distinction between decision support and automation is critical: "while it may be assumed that decision support is simply a stepping stone on the progression towards full automation, the truth is that decision support systems have fundamentally different considerations that must be accounted for in design and implementation."[9] Recommendations for the design and implementation of machine learning as clinical decision support are only beginning to emerge.[20]

Machine learning can be applied to a wide variety of clinical decision support tasks in the inpatient and outpatient setting. Of the 21 case studies, 14 apply primarily to the inpatient setting and 7 apply primarily to the outpatient setting. Some models, such as a 30-day readmission model, can inform clinical decisions in the inpatient and outpatient setting. Examples of inpatient applications include prediction of intensive care unit transfer, acute kidney injury, sepsis, and *Clostridium difficile* infection, while examples of outpatient applications include prediction of chronic kidney disease progression, death, surgical complications, and colon cancer. Table 1 presents how models from the 21 case studies can be translated into clinical care and provides example clinical workflows. Additional details about each model are provided in the next section. This workflow summary is designed to be illustrative and more comprehensive overviews of models and potential workflows can be found elsewhere.[10,21] Configurations are categorised as either centralised or decentralised. In centralised workflows, the user of the clinical decision support is removed from direct in-person patient interactions. The user may be a physician, nurse, or care manager involved in managing the health of a population or cohort of patients. Centralised workflows are often involved in 'command centres' or 'air-traffic controls'. In decentralised workflows, the user of the clinical decision support is directly involved in in-person patient interactions and is typically a nurse or physician. Decentralised clinical decision support tends to be directly embedded within the EHR. However, there is no single best workflow for a model and in fact many implementations of clinical decision support fail to improve outcomes.

THE TRANSLATIONAL PATH

The pathway for translating machine learning applied to clinical decision support tasks is based on an examination of 21 machine learning products and the authors' own experience integrating machine learning products into clinical care.[22-24] In this section, machine learning technologies are referred to as products rather than models, recognising the significant effort required to productise and operationalise models that are often built primarily for academic purposes. Table 2 summarises the products, provides context on the origin of the team and development effort, and highlights translational milestones. To map between individual products and the translational path, milestones for each product are marked within three phases: 1) design and develop; 2) evaluate and validate; 3) diffuse and scale. These phases are described in more detail below.

Milestone highlights also track the type of source as peer review, online marketing by the product development entity, or industry news. The products are designed to solve a variety of clinical and operational problems ranging from sepsis to escalation of cost and capture experiences from across the globe. The important role of academic research in commercialising products cannot be understated, as 16 of the 21 products were originally developed in academic settings. Many of those products are then externally licensed and are being scaled and diffused via commercial entities. Notably, these products have raised more than $200 million in private venture capital, but several products have also been funded through government grants as well as health systems.

Table 1: Types of workflow configurations for machine learning applied to clinical decision support tasks.

Configuration	Machine learning products	Example user experience
Centralised		
Outpatient	Kidney Failure Risk Equation: Kidney failure model Kensci: End of life model Ayasdi: Escalation of cost model	Interdisciplinary team meets weekly to discuss high-risk patients and support front-line clinicians by identifying gaps in care and providing recommendations.
Inpatient	Advance Alert Monitor: ICU transfer model Sepsis Watch: Sepsis model *Clostridium difficile*: *C. diff* model Pieces™: 30-day readmission model	Nurse continuously reviews model and supports front-line clinicians by identifying gaps in care and providing recommendations.
Decentralised		
Outpatient	Pythia: Surgical complication model Medial EarlySign: Colon cancer model	Front-line clinician receives notification directly for high-risk patient and ensures there are no gaps in patient care.
Inpatient	eCart: Cardiac arrest model eTriage: Emergency department triaging model Rothman Index: Continuous measure of patient state Inpatient fall model Dascena InSight: Sepsis model Jvion Machine: Sepsis model TREW Score: Sepsis model Deep-AISE: Sepsis model EWS 2.0: Sepsis model ePNA: Pneumonia model DeepMind Streams: acute kidney injury model HBI Spotlight: Length of stay model	Front-line clinician receives notification directly for high-risk patient and makes immediate assessment to make diagnosis or change care plan.

AISE: AISepsis Expert; eCART: electronic Cardiac Arrest Risk Triage; EWS: Early Warning Score; ICU: intensive care unit; TREW: targeted real-time early warning.

The translational path focusses on products that are actively being integrated into healthcare delivery settings and are often also being diffused and scaled beyond the original development context. The translational path, depicted in Figure 1, contains four phases discussed in detail below: 1) design and develop; 2) evaluate and validate; 3) diffuse and scale; 4) and continuing monitoring and maintenance. Each step contains a set of activities that teams building machine learning products often complete. The activities do not define requirements for any single product, but are representative of the activities completed across the 21 products in Table 2. The path is not linear and may be highly iterative.

1. Design and Develop

The first step to building a machine learning product is identifying the right problem to solve. Healthcare is a data rich environment and even though a model can be developed to generate an insight, for the insight to support clinical decisions it must be actionable and must have the potential to impact patient care. Many of the products in Table 2 generate insights for conditions that require immediate action in the inpatient setting, such as cardiac arrest, sepsis, and deterioration. Products focussed on the outpatient setting tackle problems associated with high costs to healthcare payers or providers, such as surgical complications, hospital readmissions, and end-stage renal disease.

Table 2: A table examining the experience of 14 machine leaning products actively undergoing translation into clinical care.

Product name	Product description	Origin setting (country)	Translational path milestone (year, type of source)
eTriage	ED triaging algorithm	Academia: Johns Hopkins, Baltimore, Maryland (USA)	**Develop:** Funding from AHRQ and NSF (2018, peer reviewed).[25] **Validate:** Multisite, retrospective, cross-sectional validation study of >170,000 ED visits (2018, peer reviewed).[25]
eCart	Cardiac arrest algorithm	Academia: University of Chicago, Chicago, Illinois (USA)	**Develop:** Retrospective internal data used through University of Chicago's EHR data (2014, peer reviewed).[26] **Validate:** Multicentre dataset (5 total) with 10-fold cross validation for each centre (2016, peer reviewed).[27] **Scale:** $600,000 privately raised to scale via QuantHC Startup (2012, news article).[28] **Scale:** Technology acquired by EarlySense Inc., Waltham, Massachusetts, USA, for undisclosed amount (2018, news article).[29]
PeraHealth Rothman Index	Continuous measure of patient health	Startup (USA)	**Develop and Validate:** Retrospective EHR data from Sarasota Memorial Hospital, Sarasota, Florida (2013, peer reviewed).[30] **Validate:** Model validated using retrospective EHR data from two other hospitals (2013, peer reviewed).[31] **Scale:** Secured funding from Mainsail Partners, San Francisco, California, USA, with $14 million (2017, PeraHealth website).[32] **Scale:** PeraHealth Solutions, Charlotte, North Carolina, USA, used by >80 hospitals (2017, PeraHealth website).[32]
Advance Alert Monitor (AAM)	Intensive care unit transfer and mortality algorithm	Academia: Kaiser Permanente Division of Research, Oakland, California (USA)	**Develop and Validate:** Model trained on data from 14 Kaiser Permanente Northern California hospitals (2012, peer reviewed).[33] **Scale:** Scaling to 21 Kaiser Permanente Northern California hospitals (2018, peer reviewed)[34]
Inpatient Fall Prediction	Inpatient fall algorithm	Academia: Inha University, Incheon (South Korea)	**Develop and Validate**: Model trained and externally validated on data from two hospitals in Seoul, South Korea (2019, peer reviewed).[35] **Validate:** Clinical utility evaluation completed with 12 nursing units (2019, peer reviewed).[36]
Sepsis Watch	Sepsis algorithm	Academia: Duke University, Durham, North Carolina (USA)	**Develop:** Interdisciplinary team at Duke Health working to build sepsis-prediction model (2017, peer reviewed).[22] **Validate:** Prospective internal data using Duke's EHR (2018–2019, ClinicalTrial.gov).[37] **Scale:** Cohere Med Licensing (2019, Duke University website).[38]
Dascena InSight	Sepsis algorithm	Startup (USA)	**Develop:** SBIR/STTR grants from NIH to perform clinical trials (2016–2018, ClinicalTrials.gov).[39] **Validate:** Prospective internal study at University of California, San Francisco, California (2017, peer reviewed).[40] **Validate:** Multicentre retrospective data from six institutions for generalisability (2018, peer reviewed).[41] **Scale:** Approximately $1.9 million from SBIR/STTR (2016–2018, SBIR Website).[39]
Jvion Machine	Example: Sepsis algorithm; ED admission algorithm	Startup (USA)	**Develop:** Model includes a combination of Eigen-based mathematics and a dataset of >16 million patients (2019, Jvion website).[42] **Validate:** Average reductions of 30% in preventable harm incidents/cost savings of $6.3 million a year (2019, Jvion website).[42] **Scale:** $8.9 million in funding (2019, Crunchbase website).[43]

Table 2 continued.

Product name	Product description	Origin setting (country)	Translational path milestone (year, type of source)
TREW Score	Sepsis algorithm	Academia: Johns Hopkins (USA)	**Develop and Validate:** TREWScore developed for early sepsis detection using external retrospective data (MIMIC-II) (2015, peer reviewed).[44] **Develop:** Spinout company Bayesian Health, Wilmington, Delaware, USA, formed (2019, news article).[45] **Scale:** TREWScore has been implemented at two hospitals, with three more planned in 2019 (2019, news article).[45] **Scale:** $15.0 million Series A (2019, Pitchbook website).[46]
Deep-AISE	Sepsis algorithm	Academia: Emory University, Atlanta, Georgia (USA)	**Develop:** Interdisciplinary team working at Emory University to build a sepsis prediction model (2018, peer reviewed).[47] **Validate:** External validation using MIMIC-III (2018, peer reviewed).[47] **Scale:** Approximately $700,000 BARDA grant funding sepsis consortium to scale across 3 sites (2019, Emory University website).[48]
EWS 2.0	Sepsis algorithm	Academia: University of Pennsylvania, Philadelphia, Pennsylvania (USA)	**Develop and Validate:** Model developed using internal EHR data from University of Pennsylvania Health System (2019, peer reviewed).[49] **Validation:** Clinical utility evaluation completed in internal setting (2019, peer reviewed).[49] **Validation:** Clinician perception qualitative evaluation completed in internal setting (2019, peer reviewed).[50]
ePNa	Pneumonia algorithm in the ED	Academia: Intermountain Medical Center, Salt Lake City, Utah (USA)	**Develop and Validate:** Internally developed on EHR data and radiology reports (2013, peer reviewed).[51] **Validate:** Clinical validation performed across four internal ED (2015, peer reviewed).[52] **Validate:** Qualitative evaluation to understand clinician response to CDS completed internally (2019, peer reviewed).[53] **Scale:** Reprogrammed and integrated into Cerner EHR and expanded across Intermountain hospitals (2019, peer reviewed).[54]
Clostridium Difficile	*C. difficile* infection algorithm	Academia: Massachusetts Institute of Technology, Cambridge, Massachusetts (USA)	**Develop and Validate:** Developed on data from a single USA hospital (2012, peer reviewed).[55] **Validate:** Multi-site evaluation of transfer learning methods across three hospitals (2014, peer reviewed).[56] **Validate:** Site-specific model development and evaluation approach across two hospitals (2018, peer reviewed).[57]
DeepMind Streams	Acute kidney injury algorithm	Startup (UK)	**Develop:** Founded in London by two PhD students and an entrepreneur (2010, DeepMind website).[58] **Develop:** Partnered with NHS to develop Streams (2015, news article).[59] **Validate:** Prospective study of workflow tool at NHS hospital (2019, peer reviewed).[60] **Scale:** Acquired by Google for $500.0 million (2014, DeepMind website).[61]
HBI Spotlight platform	Example: Length of stay algorithm	Academia: Stanford University, Stanford, California (USA)	**Develop:** Multidisciplinary Stanford team sought to improve health/reduce care costs (2011, HBI Website).[62] **Validate:** Temporal validation of hospital readmission model (2015, peer reviewed).[63] **Validate:** Temporal validation of inpatient mortality model (2019, peer reviewed).[64] **Scale:** Series A funding of $12.6 million (2015, HBI Website).[65]

Table 2 continued.

Product name	Product description	Origin setting (country)	Translational path milestone (year, type of source)
Pieces™	Example: 30-day readmission algorithm	Academia: University of Texas Southwestern, Dallas, Texas (USA)	**Develop:** Creation of the non-profit named PCCI, funding from various grants (2012, news article).[66] **Validate:** 30-day readmission model with external validation from 7 large hospitals (2015, peer reviewed).[67] **Scale:** Series A funding of $21.6 million (2016, news article).[68]
Pythia	Surgery complication algorithm	Academia: Duke University	**Develop and Validate:** Development and validation on data from a single health system and compared to national benchmark model (2018, peer reviewed).[23] **Validate:** Site-specific and ensemble model development and validation approach across three hospitals (2019, peer reviewed).[69]
Kidney Failure Risk Equation	Kidney failure algorithm	Academia: University of Toronto, Sunnybrook Hospital, Toronto, Ontario (Canada)	**Develop and Validate:** Development and validation at two Independent Canadian hospitals (2011, peer reviewed).[70] **Validate:** Multicentre retrospective data from >30 countries from 1982 to 2014 (2016, peer reviewed).[71] **Scale:** Available as a website calculator (2019, website).[72] **Scale:** Physician lead joins Viewics, Santa Clara, California, USA, advisory board to integrate kidney analytics into products (2016, news article).[73] **Scale:** Viewics acquired by Roche, Basel, Switzerland, for undisclosed amount (2017, Roche website).[74]
Kensci	Example: End-of-life algorithm	Academia: University of Washington, Seattle, Washington (USA)	**Develop and Validate:** 7 years of research and >40 publications in various fields (2012–2019, Kensci website).[75] **Validate and Scale:** Integration in >25 health systems (2017, Kensci website).[75] **Scale:** Partnered with agencies including CDC and various industry entities (2016, Kensci website).[75] **Scale:** Series A funding of $8.5 million (2017, Kensci website).[75] **Scale:** Series B funding of $22.0 million (2019, Kensci website).[75]
Ayasdi platform	Example: Escalation of cost algorithm	Startup (USA)	**Develop:** Machine learning company applying resources and applications to healthcare (2019, Ayasdi website).[76] **Validate:** Multiple peer-reviewed publications (2008–2019, Ayasdi website).[76] **Scale:** $106.3 million in funding through Series C (2015, Crunchbase website).[77]
Medial EarlySign ColonFlag	Colon cancer screening algorithm	Academia: Tel-Aviv University, Tel Aviv (Israel)	**Develop and Validate:** Model trained on data from Israeli insurer and validated on Israeli data and external dataset from the UK (2016, peer reviewed).[78] **Validate:** Temporal validation in Israel to determine clinical utility (2018, peer reviewed).[79] **Validate:** Independent external validations on UK cohort (2017, peer reviewed) 80 and USA cohort (2017, peer reviewed).[81] **Scale:** $30.0 million Series B (2018, Pitchbook website).[82]

The table includes the name, a brief description of the machine learning product, the origin, and the path to translation.

AHRQ: Agency for Healthcare Research and Quality; BARDA: Biomedical Advanced Research and Development Authority; CDC: U.S. Centers for Disease Control and Prevention; CDS: clinical decision support; ED: emergency department; EHR: electronic health record; MIMIC: Medical Information Mart for Intensive Care; NSF: National Science Foundation; PCCI: Parkland Center for Clinical Innovation; SBIR: Small Business Innovation Research; STTR: Small Business Technology Transfer; TREW: targeted real-time early warning.

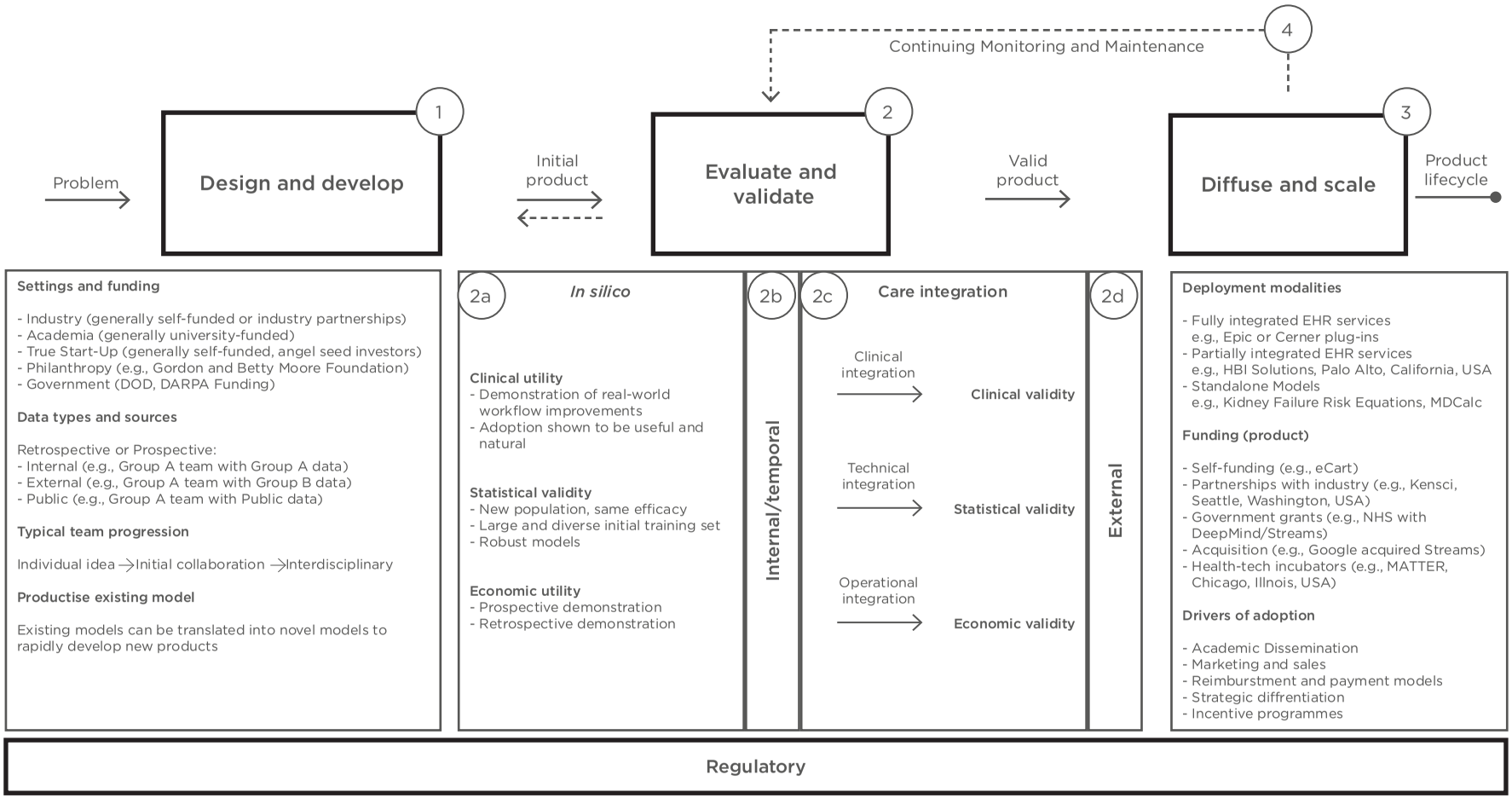

Figure 1: A path for the design, development, evaluation, validation, diffusion, and scaling of machine learning models to be used in clinical care.

The translational path begins on the left with the 'problem' to be addressed by the product and concludes with the 'product lifecycle' on the right. This path is designed to be a comprehensive representation of the current landscape, but is not meant to specify the steps taken by any individual product. Numbers denote the steps of the path.

DARPA: Defense Advanced Research Projects Agency; DOD: Department of Defense; eCART: electronic Cardiac Arrest Risk Triage; EHR: Electronic Health record.

The setting and funding of the team shapes many aspects of how the machine learning product is designed and developed. For example, in an academic setting it may be easier to cultivate collaborations across domains of expertise early on in the process. However, academic settings may have difficulty recruiting and retaining the technical talent required to productise complex technologies. Funding sources can also vary as products proceed through different stages of the translational path. For example, many of the products in Table 2 were initially funded internally or externally through grants and secured private investment once the product was licensed to an outside company. Setting also significantly impacts the data available to develop a machine learning product. In an academic setting, teams may have access to internal data, whereas teams situated outside of a healthcare system need to obtain data through partnerships. Public datasets also play an important role in promoting product development. Three sepsis products (Deep-AISepsis Expert [AISE], targeted real-time early warning [TREW] Score, Insight) used Medical Information Mart for Intensive Care (MIMIC)-III data for training or evaluation.[83]

Finally, there are cases in which existing algorithms are productised while more sophisticated machine learning techniques are developed as product enhancements. A notable example of this is DeepMind Streams, which originally productised a national acute kidney injury algorithm.[84] In parallel, DeepMind implemented and evaluated the workflow solution as well as developed machine learning methods to enhance the product.[60,85] Similarly, the regression-based Kidney Failure Risk Equation (KFRE) used to predict progression of chronic kidney disease, was productised by Viewics inc., Santa Clara, California, USA, while additional technologies were built in parallel.

2. Evaluate and Validate

An initial round of machine learning product evaluation and validation, Step 2a, can be completed entirely on retrospective data. These experiments are called '*in silico*' and demonstrate three dimensions of validity and utility.[21] Clinical utility addresses the question: can the product improve clinical care and patient outcomes? This requires that the team marketing or developing a machine learning product can calculate baseline performance on data relevant to the adopting organisation. Statistical validity addresses the question: can the machine learning product perform well on metrics of accuracy, reliability, and calibration? This requires that there is agreement on important and relevant model performance measures and a sense of what makes a product perform well enough for adoption. Finally, economic utility addresses the question: can there be a net benefit from the investment in the machine learning product? The economic utility can be demonstrated through cost savings, increased reimbursement, increased efficiency, and through brand equity. All forms of utility and validity are ultimately in the eye of the beholder. As such, relationships and communication between the machine learning product team and organisational stakeholders are critical.

The evaluation and validation of machine learning products requires multiple iterations. Demonstrating utility or validity on retrospective, *in silico* settings does not guarantee that the product will perform well in a different setting[57] or in a different time period.[86,87] The utility and validity of the product must be reassessed across time (Step 2b) and space (Step 2d). Evaluating a machine learning product on a hold-out and temporal validation set (Step 2a) is recommended before integrating a product into clinical care.[14] Evaluating a machine learning product on external geographic datasets (Step 2d) can help drive adoption in new settings. Of the products listed in Table 2, Sepsis Watch, Advanced Alert Monitor, ePNa, and Early Warning Score 2.0 were integrated within the healthcare organisation that developed the product without external validation. On the other hand, electronic Cardiac Arrest Risk Triage (eCART),[26,27] KFRE,[71] InSight,[41] Rothman Index,[31] Pieces™ readmission model,[67] Deep-AISE,[47] and eTriage[25] completed peer-reviewed external validations. The two products with the most extensive external validations include the KFRE, which was validated on a multi-national dataset consisting of >30 countries,[71] and ColonFlag, which was validated on cohorts in the USA,[81] UK,[80] and Israel.[88] These external validations evaluate the same model in different geographical contexts. However, several teams are taking a different approach to validating models across settings. The teams working on *C. difficile*

and surgical complication models are building generalisable approaches by which site-specific models are developed and validated.[57,69]

The production environment of a health information technology system often differs dramatically from the environment that stores retrospective or even day-old data. Significant effort and infrastructure investment are required to integrate products into production EHR systems (Step 2c). One study estimated the cost to validate and integrate the KFRE into clinical workflows at a single site at nearly $220,000.[24] Redundant costs across sites, as a result of a lack of interoperability and lack of infrastructure, would require similar investment by institutions following a similar approach. Furthermore, the 'inconvenient truth' of machine learning in healthcare was pointedly described as "at present the algorithms that feature prominently in research literature are in fact not, for the most part, executable at the front lines of clinical practice."[89] Finally, before being integrated into clinical care, a machine learning product needs to be evaluated and validated in a 'silent' mode, "in which predictions are made in real-time and exposed to a group of clinical experts."[14] This period is crucial for finalising workflows and product configurations as well as serving as a temporal validation (Step 2b). An example of a silent mode evaluation is an eCart feasibility study.[90]

Although represented as a single arrow, the 'Clinical Integration' step (Step 2c) can often be the most difficult step in the entire translational path. Most implementations of clinical decision support do not have the intended effect because of the difficulty with clinical integration. What differentiates the products listed in Table 2 from most machine learning models is that these products have undertaken clinical integration. Only one product, InSight (Dascena, Oakland, California, USA),[40] conducted a single-site, randomised control trial with 142 patients and demonstrated positive results. This single study needs to be followed up with larger trials from every team trying to drive adoption of a machine learning product.

3. Diffuse and Scale

There are many machine learning products that focus on solving a local problem. The next challenge is to diffuse and scale across settings, which requires special attention to deployment modalities, funding, and drivers of adoption. As described earlier, machine learning products that ingest structured data from EHR require significant integration effort and infrastructure. This has driven the rapid adoption of models and algorithms sold and distributed by EHR vendors.[91] To scale, machine learning products must be able to ingest data from different EHR and must also support on-premise and cloud deployments. For this reason, many models are also distributed as stand-alone web applications that require manual entry to calculate risk.

During this stage, machine learning product teams seek external investment and financial resources. As shown in Table 2, tens of millions of dollars are often raised by companies trying to scale products. The resources are required for both scaling deployment of the product as well as navigating the drivers of adoption. Several adoption strategies include academic dissemination, marketing and sales, and partnerships with regulators and payers to create reimbursement mechanisms. The "nonadoption, abandomnent, and challenges to the scale-up, spread, and sustainability of health and care technologies" is an example of a technology adoption framework that covers seven domains and has been recently applied to machine learning products.[92,93] To date, no machine learning product ingesting EHR data has successfully diffused and scaled across healthcare. The products listed in Table 2 are as far along as any and will be closely watched over the coming years.

4. Continuing Monitoring and Maintenance

The translational path for a machine learning product does not have a finish line. Data quality, population characteristics, and clinical practice change over time and impact the validity and utility of models. Model reliability and model updating are active fields of research and will be integral to ensure the robustness of machine learning products in clinical care.[94,95] Another example of model maintenance is updating outcome definitions to retrain models as scientific understanding of disease progresses. For example, many of the sepsis products listed in Table 2 use sepsis definitions that pre-date

Sepsis-3, the most recent international consensus definition.[96] Similar to other technology innovations, the product lifecycle continues and will need to adapt to changes in market dynamics and organisational needs. Similarly, the product will evolve over time and will require continued validation and iteration. Throughout the process, teams developing machine learning products need to work closely with regulators and operate within evolving regulatory frameworks.[14]

CHALLENGES AND OPPORTUNITIES

The translational path described above is not well trodden and, for better or for worse, the products listed in Table 2 are establishing norms for the industry. Across the 21 products, there are opportunities to improve how machine learning products are translated into clinical care. Outside of the scope of this review, the path of IDx-DR, a machine learning product used to automatically diagnose diabetic retinopathy, is instructive. IDx-DR has conducted a randomised control trial and has received regulatory approval in both the USA and EU as a medical device.[97,98] IDx-DR is now actively being scaled and diffused. Unfortunately, many products in Table 2 are pursuing 'stealth science' to protect trade secrets and avoiding regulatory or academic scrutiny.[99] While stealth science is not uncommon amongst biomedical innovations, lack of transparency is particularly concerning with machine learning. This narrative review was unable to provide standard metrics of adoption, because many of the figures marketed by product developers have no peer-reviewed evidence. Machine learning products, which often lack inherent interpretability, need evidence that ensures validity as well as safety and efficacy.

There are three opportunities to enhance how all machine learning products that ingest structured EHR data are translated into clinical care. First, data quality systems and frameworks need to be adopted to ensure that machine learning models have face validity. Significant effort and resources are required to transform the raw data extracted from EHR into a usable format for training machine learning models.[24] Distributed research networks that leverage EHR data for clinical research have developed frameworks for assessing quality of EHR data, but these frameworks have not been adopted by machine learning product developers.[100-102] Incorporating high quality data into a model is as important as incorporating that same data into a pharmaceutical clinical trial. However, reporting the results of data quality assessments rarely accompanies reporting of model performance. Second, without interoperability across EHR systems, machine learning products will continue to face significant challenges scaling and diffusing across systems. New regulatory and policy mechanisms need to drive interoperability between EHR systems. Third, products listed in Table 2 that predict the same outcome cannot be easily compared. Reporting of machine learning models often fails to follow establish best practices and model performance measures are not standardised across publications.[21,103,104] Data sources and data transformations also impact model performance across studies. Furthermore, there is no current standard definition of accuracy and patient health outcomes against which to measure the products. There is a head-to-head comparison of the Advance Alert Monitor and eCart on the same dataset,[105] but this practice is exceedingly rare. Benchmark datasets, funding mechanisms, and agreement on model and clinical performance measures must be established to facilitate comparisons across products and settings.

Finally, there are a host of ethical challenges entailed in each step throughout the translation of machine learning in clinical care. First, patients are largely left unaware when personal data is shared with machine learning model developers, whether through a waiver of consent within an academic organisation or through a business agreement with an industry partner. These privacy concerns are prompting legal challenges and revisions of healthcare privacy law across the USA and Europe.[106] Second, the dataset used to train a model[107] or the outcome that a model predicts[108] can have significant implications for how models lessen or worsen disparities in healthcare. Unfortunately, these biases have been discovered in models that run on hundreds of millions of patients. There are additional ethical challenges in machine learning that are described in more detail in related reviews.[107,109] Teams building machine learning products need to consider these challenges early and often and incorporate ethical and legal perspectives into their work.

CONCLUSION

Despite enormous enthusiasm surrounding the potential for machine learning to transform healthcare, the successful translation of machine learning products into clinical care is exceedingly rare. Evidence of clinical impact remains scant. This review examines the experience of 21 machine learning products that integrate with EHR to provide clinical decision support. The steps and activities involved in the design and development, evaluation and validation, and scale and diffusion of the machine learning products are described. This translational path can guide current and future efforts to successfully translate machine learning products into healthcare.

References

1. Hinton G. Deep learning-a technology with the potential to transform health care. JAMA. 2018;320(11):1101-2.
2. Software as a Medical Device Working Group. Software as a Medical Device (SaMD): key Definitions. 2013. Available at: http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-131209-samd-key-definitions-140901.pdf. Last accessed: 11 December 2019.
3. Gordon WJ, Stern AD. Challenges and opportunities in software-driven medical devices. Nature Biomedical Engineering. 2019;3:493-7.
4. National Health Service (NHS). The Topol Review: Preparing the healthcare workforce to deliver the digital future. 2019. Available at: https://topol.hee.nhs.uk/wp-content/uploads/HEE-Topol-Review-2019.pdf. Last accessed: 11 December 2019.
5. National Science Technology Council (NSTC). The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update. 2019. Available at: https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf. Last accessed: 11 December 2019.
6. American Medical Association (AMA). Augmented Intelligence in Health Care. 2018. Available at: https://www.ama-assn.org/system/files/2019-01/augmented-intelligence-policy-report.pdf. Last accessed: 11 December 2019.
7. U.S. Food and Drug Administration (FDA). Clinical Decision Support Software - Draft Guidance for Industry and Food and Drug Administration Staff. 2019. Available at: https://www.fda.gov/media/109618/download. Last accessed: 11 December 2019.
8. European Union (EU). Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017. 2017. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0745. Last accessed: 11 December 2019.
9. Jamieson T, Goldfarb A. Clinical considerations when applying machine learning to decision-support tasks versus automation. BMJ Quality & Safety. 2019;28(10):778-81.
10. Goldstein BA et al. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc. 2016;24(1):198-208.
11. Xiao C et al. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc. 2018;25(10):1419-28.
12. Riemenschneider M et al. Data science for molecular diagnostics applications: From academia to clinic to industry. Systems Medicine. 2018;1(1):13-7.
13. Rawson TM et al. A systematic review of clinical decision support systems for antimicrobial management: are we failing to investigate these interventions appropriately? Clin Microbiol Infect. 2017;23(8):524-32.
14. Wiens J et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337-40.
15. Greene SM et al. Implementing the learning health system: from concept to action. Ann Intern Med. 2012;157(3):207-10.
16. Rajkomar A et al. Machine learning in medicine. N Engl J Med. 2019;380(14):1347-58.
17. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. 2018;319(13):1317-8.
18. Wachter S et al. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law. 2017;7(2):76-99.
19. Obermeyer Z, Lee TH. Lost in thought - the limits of the human mind and the future of medicine. N Engl J Med. 2017;377(13):1209-11.
20. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. JAMA. 2018;320(21):2199-200.
21. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56.
22. Futoma J et al. An improved multi-output gaussian process RNN with real-time validation for early sepsis detection. proceedings of machine learning for healthcare. 2017;Eprint:1708.05894
23. Corey KM et al. Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): a retrospective, single-site study. PLoS Med. 2018;15(11):e1002701.
24. Sendak MP et al. Barriers to achieving economies of scale in analysis of EHR data: a cautionary tale. Applied Clinical Informatics. 2017;8(3):826-31.
25. Levin S et al. Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index. Ann Emerg Med. 2018;71(5):565-74.e2.
26. Churpek MM et al. Multicenter development and validation of a risk stratification tool for ward patients. Am J Respir Crit Care Med. 2014;190(6):649-55.
27. Churpek MM et al. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. CCM. 2016;44(2):368-74.
28. Wolinsky H. Crains Chicago Business. What if you could prevent someone from suffering cardiac arrest? 2018. Available at: https://www.chicagobusiness.com/article/20180201/ISSUE01/180209999/quant-startup-offers-cardiac-arrest-predictive-software-ecart. Last accessed: 1 September 2019.
29. Monegain B. Health IT News. EarlySense acquires predictive analytics to help hospitals assess cardiac risk. Available at: https://www.healthcareitnews.com/news/earlysense-acquires-predictive-analytics-help-hospitals-assess-cardiac-risk. Last accessed: 22 September 2019.

30. Rothman MJ et al. Development and validation of a continuous measure of patient condition using the Electronic Medical Record. J Biomed Inform. 2013;46(5):837-48.
31. Finlay GD et al. Measuring the modified early warning score and the Rothman Index: advantages of utilizing the electronic medical record in an early warning system. J Hosp Med. 2013;9(2):116-9.
32. The Rothman Index. PeraHealth Secures $14 Million in Financing. 2019. Available at: https://www.perahealth.com/press-releases/2017/01/perahealth-secures-14-million-in-financing/. Last accessed: 14 August 2019.
33. Escobar GJ et al. Early detection of impending physiologic deterioration among patients who are not in intensive care: development of predictive models using data from an automated electronic medical record. J Hosp Med. 2012;7(5):388-95.
34. Dummett B et al. Preventing unrecognized deterioration and honoring patients' goals of care by embedding an automated early-warning system in hospital workflows. Perm J. 2018;1-20.
35. Cho I et al. Novel approach to inpatient fall risk prediction and its cross-site validation using time-variant data. J Med Internet Res. 2019;21(2):e1150513.
36. Cho I, Jin I. Changes in nursing activity after implementing a CDS service predicting the risk of falling based on electronic medical records data. Abstract 028. AMIA Annual Symposium, 16-20 November, 2019.
37. Sendak MP et al. Sepsis watch: a real-world integration of deep learning into routine clinical care. JMIR Preprints. 2019:15182.
38. Duke University. "Deep Sepsis" Licensed to Cohere Med. 2019. Available at: https://olv.duke.edu/news/deep-sepsis-licensed-to-cohere-med/. Last accessed: 1 September 2019.
39. SBIR/STTR. America's Seed Fund. Dascena. 2018. Available at: https://www.sbir.gov/sbc/dascena. Last accessed: 1 September 2019.
40. Shimabukuro DW et al. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. BMJ Open Resp Res. 2017;4(1):e000234.
41. Mao Q et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. BMJ Open. 2018;8(1):e017833-11.
42. Jvion. Prescriptive Analytics for Preventable Harm--The Jvion Machine. Available at: https://jvion.com/about. Last accessed: 1 September 2019.
43. Crunchbase. Jvion. Available at: https://www.crunchbase.com/organization/jvion#section-overview. Last accessed: 15 August 2019.
44. Henry KE et al. A targeted real-time early warning score (TREWScore) for septic shock. Science Translational Medicine. 2015;7(299):299ra122.
45. Johns Hopkins Medicine. Early-Warning Algorithm Targeting Sepsis Deployed at Johns Hopkins. 2019. Available at: https://www.hopkinsmedicine.org/news/articles/early-warning-algorithm-targeting-sepsis-deployed-at-johns-hopkins. Last accessed: 1 August 2019.
46. Pitchbook. Bayesian Health. 2018. Available at: https://pitchbook.com/profiles/company/277329-07. Last accessed: 15 September 2019.
47. Nemati S et al. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit Care Med. 2018;46(4):547-53.
48. Emory News Center. DRIVe teams up with academic research consortium to develop deep learning software to provide early warning of sepsis in patients. 2019. Available at: https://news.emory.edu/stories/2019/02/sharma_nemati_drive_academic_corsortium_for_sepsis/index.html. Last accessed: 18 September 2019.
49. Giannini HM et al. A machine learning algorithm to predict severe sepsis and septic shock. Crit Care Med. 2019;47(11):1485-92.
50. Ginestra JC et al. Clinician perception of a machine learning-based early warning system designed to predict severe sepsis and septic shock. Crit Care Med. 2019;47(11):1477-84.
51. Dean NC et al. Performance and utilization of an emergency department electronic screening tool for pneumonia. JAMA Intern Med. 2013;173(8):699-701.
52. Dean NC et al. Impact of an electronic clinical decision support tool for emergency department patients with pneumonia. Ann Emerg Med. 2015;66(5):511-20.
53. Jones B et al. CDS in a learning health care system: Identifying physicians' reasons for rejection of best-practice recommendations in pneumonia through computerized clinical decision support. Appl Clin Inform. 2019;10(01):001-9.
54. Dean NC et al. AMIA. Implementation of real-time electronic clinical decision support for emergency department patients with pneumonia across a healthcare system. 2019. Available at: https://knowledge.amia.org/69862-amia-1.4570936/t004-1.4574923/t004-1.4574924/3195283-1.4575123/3195283-1.4575124?qr=1. Last accessed: 11 December 2019.
55. Wiens J et al. Patient risk stratification for hospital-associated *C.diff* as a time-series classification task. Advances in Neural Information Processing Systems. 2012:467-75.
56. Wiens J et al. A study in transfer learning: Leveraging data from multiple hospitals to enhance hospital-specific predictions. J Am Med Inform Assoc. 2014;21(4):699-706.
57. Oh J et al. A generalizable, data-driven approach to predict daily risk of *Clostridium difficile* infection at two large academic health centers. Infect Control Hosp Epidemiol. 2018;39(4):425-33.
58. DeepMind. About (DeepMind). 2019. Available at: https://deepmind.com/about. Last accessed: 1 September 2019.
59. King D. DeepMind. Why doesn't Streams use AI? 2017. Available at: https://deepmind.com/blog/article/streams-and-ai. Last accessed: 15 August 2019.
60. Connell A et al. Evaluation of a digitally-enabled care pathway for acute kidney injury management in hospital emergency admissions. npj Digital Medicine. 2019;67:1-9.
61. DeepMind. Scaling Streams with Google. 2018. Available at: https://deepmind.com/blog/announcements/scaling-streams-google. Last accessed: 15 August 2019.
62. HBI Solutions. About HBI Solutions. 2019. Available at: https://hbisolutions.com/about-2/. Last accessed: 1 September 2019.
63. Hao S et al. Development, validation and deployment of a real time 30 day hospital readmission risk assessment tool in the Maine Healthcare Information Exchange. PLoS ONE. 2015;10(10):e0140271-15.
64. Ye C et al. A real-time early warning system for monitoring inpatient mortality risk: prospective study using electronic medical record data. J Med Internet Res. 2019;21(7):e13719-3.
65. HBI Solutions. Silicon Valley's HBI Solutions Secures Series A Funding to Expand Real-time Healthcare Analytics Services in the US and China. 2015. Available at: https://hbisolutions.com/silicon-valleys-hbi-solutions-secures-series-a-funding-to-expand-real-time-healthcare-analytics-services-in-the-us-and-china/. Last accessed: 15 August 2019.
66. MedCity News. Founder of PCCI talks about path to Pieces Tech launch. 2016. Available at: https://medcitynews.com/2016/03/founder-pcci-talks-path-pieces-tech-launch/. Last accessed: 15 August 2019.
67. Amarasingham R et al. Electronic medical record-based multicondition models to predict the risk of 30 day readmission or death among adult medicine patients: validation and comparison to existing models. BMC

Med Inform Decis Mak. 2015;15:39.

68. MedCity News. Clinical decision support startup launches, raises $21.6M. 2016. Available at: https://medcitynews.com/2016/03/clinical-decision-support-startup/. Last accessed: 15 August 2019.
69. Corey K et al. Model ensembling vs data pooling: Alternative ways to merge hospital information across sites. Proceedings of Machine Learning for Healthcare. 2019. Available at: https://static1.squarespace.com/static/59d5ac1780bd5ef9c396eda6/t/5d473e91b0f598000 1a24186/1564950161664/Corey.pdf. Last accessed: 11 December 2019.
70. Tangri N et al. A predictive model for progression of chronic kidney disease to kidney failure. JAMA. 2011;305(15):1553-9.
71. Tangri N et al. Multinational assessment of accuracy of equations for predicting risk of kidney failure: A meta-analysis. JAMA. 2016;315(2):164-74.
72. The Kidney Failure Risk. The Kidney Failure Risk Equation. Available at: https://kidneyfailurerisk.com/. Last accessed: 15 September 2019.
73. Business Wire. Medical Researcher Dr. Navdeep Tangri Joins Advisory Board for Healthcare Analytics Company Viewics. 2016. Available at: https://www.businesswire.com/news/home/20160503006400/en/Medical-Researcher-Dr.-Navdeep-Tangri-Joins-Advisory. Last accessed: 15 September 2019.
74. Roche. Roche to acquire Viewics, Inc. to provide data-driven lab business analytics and add further digital capabilities along the laboratory value chain. 2017. Available at: https://www.roche.com/media/releases/med-cor-2017-11-17b.htm. Last accessed: 15 September 2019.
75. KenSci. Death Vs. Data Science. 2017. Available at: https://www.kensci.com/company/about/. Last accessed: 15 August 2019.
76. Ayasdi. Transform your business with machine 1 intelligence and big data. 2019. Available at: https://www.ayasdi.com/resources/publications/. Last accessed: 15 August 2019.
77. Crunchbase. Ayasdi. 2019. Available at: https://www.crunchbase.com/organization/ayasdi. Last accessed: 1 September 2019.
78. Kinar Y et al. Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study. J Am Med Inform Assoc. 2016;23(5):879-90.
79. Goshen R et al. Computer-assisted flagging of individuals at high risk of colorectal cancer in a large health maintenance organization using the ColonFlag test. JCO Clin Cancer Inform. 2018;2(2):1-8.
80. Birks J et al. Evaluation of a prediction model for colorectal cancer: Retrospective analysis of 2.5 million patient records. Cancer Med. 2017;6(10):2453-60.
81. Hornbrook MC et al. Early colorectal cancer detected by machine learning model using gender, age, and complete blood count data. Digestive Diseases and Sciences. 2017;62(10):2719-27.
82. Pitchbook. Medial EarlySign. 2019. Available at: https://pitchbook.com/profiles/company/162221-32. Last accessed: 6 December 2019.
83. Johnson AEW et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035-9.
84. Selby NM et al. Standardizing the early identification of acute kidney injury: the NHS England National patient safety alert. Nephron. 2015;131(2):113-7.
85. Tomašev N et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature. 2019;572(7767):116-9.
86. Davis SE et al. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc. 2017;24(6):1052-61.
87. Davis SE et al. A nonparametric updating method to correct clinical prediction model drift. J Am Med Inform Assoc. 2019;26(12):1148-57.
88. Kinar Y et al. Performance analysis of a machine learning flagging system used to identify a group of individuals at a high risk for colorectal cancer. PLoS ONE. 2017;12(2):e0171759-8.
89. Panch T et al. The "inconvenient truth" about AI in healthcare. npj Digital Medicine. 2019:2(77):1-3.
90. Kang MA et al. Real-time risk prediction on the wards: A feasibility study. Crit Care Med. 2016;44(8):1468-73.
91. Amland RC, Sutariya BB. An investigation of sepsis surveillance and emergency treatment on patient mortality outcomes: an observational cohort study. JAMIA Open. 2018;1(1):107-14.
92. Greenhalgh T et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res. 2017;19(11):e367.
93. Shaw J et al. Artificial intelligence and the implementation challenge. J Med Internet Res. 2019;21(7):e13659-11.
94. Lenert MC et al. Prognostic models will be victims of their own success, unless... J Am Med Inform Assoc. 2019;26(12):1645-50.
95. Saria S, Subbaswamy A. Tutorial: Safe and reliable machine learning. 2019. Available at: https://arxiv.org/abs/1904.07204. Last accessed: 11 December 2019.
96. Singer M et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). 2016;315(8):801-10.
97. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. npj Digital Medicine. 2018;1(40):1-3.
98. Abràmoff MD et al. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. npj Digital Medicine. 2018;1(39):1-8.
99. Cristea IA et al. Stealth research: Lack of peer-reviewed evidence from healthcare unicorns. Eur J Clin Invest. 2019;49(4):e13072-8.
100. Callahan TJ et al. A comparison of data quality assessment checks in six data sharing networks. eGEMs (Generating Evidence and Methods to Improve Patient Outcomes). 2017;5(1):8.
101. Khare R et al. A longitudinal analysis of data quality in a large pediatric data research network. J Am Med Inform Assoc. 2017;24(6):1072-9.
102. Kahn MG et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. eGEMs (Generating Evidence and Methods to Improve Patient Outcomes). 2016;4(1):1-18.
103. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. The Lancet. 2019;393(10181):1577-9.
104. He J et al. The practical implementation of artificial intelligence technologies in medicine. Nat Med. 2019;25(1):30-6.
105. Kipnis P et al. Development and validation of an electronic medical record-based alert score for detection of inpatient deterioration outside the ICU. J Biomed Inform. 2016;64:10-9.
106. Cohen IG, Mello MM. Big data, big tech, and protecting patient privacy. JAMA. 2019; doi: 10.1001/jama.2019.11365. [Epub ahead of print].
107. Rajkomar A et al. Ensuring fairness in machine learning to advance health equity. Ann Intern Med. 2018;169(12):866-8.
108. Obermeyer Z et al. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-53.
109. Gianfrancesco MA et al. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med. 2018;178(11):1544–7.