Artificial Intelligence for Health and Health Care

Contact: Dolores Derrington — doloresd@mitre.org

December 2017

JSR-17-Task-002

JASON
The MITRE Corporation
7515 Colshire Drive
McLean, VA 22102-7508
(703) 983-6997

Contents

EXECUTIVE SUMMARY

This study centers on how computer-based decision procedures, under the broad umbrella of artificial intelligence (AI), can assist in improving health and health care. Although advanced statistics and machine learning provide the foundation for AI, there are currently revolutionary advances underway in the sub-field of neural networks. This has created tremendous excitement in many fields of science, including in medicine and public health. First demonstrations have already emerged showing that deep neural networks can perform as well as the best human clinicians in well-defined diagnostic tasks. In addition, AI-based tools are already appearing in health-oriented apps that can be employed on handheld, networked devices such as smart phones.

Focus of the Study.

U.S. Department of Health and Human Services (HHS), with support from the Robert Wood Johnson Foundation, asked JASON to consider how AI will shape the future of public health, community health, and health care delivery. We focused on technical capabilities, limitations, and applications that can be realized within the next ten years.

Some questions raised by this study are: Is the recent level of interest in AI just another period of hype within the cycles of excitement that have arisen around AI? Or would different circumstances this time make people more receptive to embracing the promise of AI applications, particularly related to health? AI is primarily exciting to computational sciences researchers throughout academia and industry. Perhaps, the previous advances in AI had no obvious influence on the lives of individuals. The potential influence of AI for health, including health care delivery, may be affected by current societal factors that may make the fate of AI hype different this time. Currently, there is great frustration with the cost and quality of care delivered by the US health care system. To some degree, this has fundamentally eroded patient confidence, opening people's minds to new paradigms, tools, services. Dovetailing with this, there is an explosion in new personal health monitoring technology through smart device platforms and internet-based interactions. This seemingly perfect storm leads to an overarching observation, which defines the environment in which AI applications are now being developed and has helped shape this study:

Overarching Observation: Unlike previous eras of excitement over AI, the potential of AI applications in health may make this era different because the confluence of the following three forces has primed our society to embrace new health centric approaches that may be enabled by advances in AI: 1) frustration with the legacy medical system, 2) ubiquity of networked smart devices in our society, 3) acclimation to convenience and at-home services like those provided through Amazon and others.

Findings and Recommendations:

Overall, JASON finds that AI is beginning to play a growing role in transformative changes now underway in both health and health care, in and out of the clinical setting. At present the extent of the opportunities and limitations is just being explored. However, there are significant

challenges in this field that include: the acceptance of AI applications in clinical practice, initially to support diagnostics; the ability to leverage the confluence of personal networked devices and AI tools; the availability of quality training data from which to build and maintain AI applications in health; executing large-scale data collection to include missing data streams; in building on the success in other domains, creating relevant AI competitions; and understanding the limitations of AI methods in health and health care applications.

Here *we provide the JASON findings and recommendations. Discussion and elaboration on each of these is presented in the text.*

1. AI Applications in Clinical Practice

Findings:

- The process of developing a new technique as an established standard of care uses the robust practice of peer-reviewed R&D, and can provide safeguards against the deceptive or poorly-validated use of AI algorithms. (Section 2.3)
- The use of AI diagnostics as replacements for established steps in medical standards of care will require far more validation than the use of such diagnostics to provide supporting information that aids in decisions. (Section 2.3)

Recommendations:

- Support work to prepare AI results for the rigorous approval procedures needed for acceptance for clinical practice. Create testing and validation approaches for AI algorithms to evaluate performance of the algorithms under conditions that differ from the training set. (Section 2.3)

2. Confluence of AI and Smart Devices for Monitoring Health and Disease

Findings:

- Revolutionary changes in health and health care are already beginning in the use of smart devices to monitor individual health. Many of these developments are taking place outside of traditional diagnostic and clinical settings. (Section 3.1)
- In the future, AI and smart devices will become increasingly interdependent, including in health-related fields. On one hand, AI will be used to power many health-related mobile monitoring devices and apps. On the other hand, mobile devices will create massive datasets that, in theory, could open new possibilities in the development of AI-based health and health care tools. (Section 3.1)

Recommendations:

- Support the development of AI applications that can enhance the performance of new mobile monitoring devices and apps. (Section 3.1)
- Develop data infrastructure to capture and integrate data generated from smart devices to support AI applications. (Section 3.1)
- Require that development include approaches to insure privacy and transparency of data use. (Section 3.1)
- Track developments in foreign health care systems, looking for useful technologies and also technology failures. (Section 3.1)

3. Create Comprehensive Training Databases of Health Data for AI Tool Development

Findings:

- The availability of and access to high quality data are critical in the development and ultimate implementation of AI applications in health care. (Section 4)
    - AI algorithms based on high quality training sets have already demonstrated performance for medical image analysis at the level of the medical capability that is captured in their training data. (Section 2.1)
    - AI algorithms cannot be expected to perform at a higher level than their training data, but should deliver the same standard of performance consistently for data within the training space. (Section 2.1)
- Laudable goals for AI tools include accelerating the discovery of novel disease correlations and helping match people to the best treatments based on their specific health, life-experiences, and genetic profile. Definition and integration of the data sets required to develop such AI tools is a major challenge. (Section 4)
- Extreme care is needed in using electronic health records (EHRs) as training sets for AI, where outputs may be useless or misleading if the training sets contain incorrect information or information with unexpected internal correlations. (Section 6.1)
- Techniques for learning from unlabeled data could be helpful in addressing the issues with using data from a diverse set of sources. (Section 4.2)

Recommendations:

- Support the development of and access to research databases of labeled and unlabeled health data for the development of AI applications in health. (Section 4)
- Support investigations into how to incentivize the sharing of health data, and new paradigms for data ownership. (Section 4)
- Support the assessment of AI algorithms trained with data labeled at levels that significantly exceed standard assessment, for instance the use of outputs from the next stage of diagnostics (e.g., use of biopsy results to label dermatological images). (Section 2.1)
- Support research to characterize the tradeoffs between data quality, information content (complexity and diversity) and sample size, with the goal of enabling quantitative prediction of the quantity and quality of data needed to support a given AI application. (Section 4)
- Identify and develop strategies to fill important data gaps for health. (Section 4)
- Develop automated curation approaches for broadly based data collections to format them for AI tools, e.g., as with well labeled imagery. (Section 4.2)

4. Fill in Critical Missing Data Gaps

Findings:

- AI application development requires training data, and will perform poorly when significant data streams are absent. While DNA is the blueprint for life, health outcomes are highly affected by environmental exposures and social behaviors. There is an imbalance in the effort to capture the diverse data needed for application of AI

techniques to precision medicine, with information on environmental toxicology and exposure particularly suffering: (Section 5.2.2)

- Techniques exist to capture individual environmental exposures, e.g., blood toxin screening, diet questionnaires.
- Techniques exist for environmental pathogen sensing.
- Technologies exist that can capture environmental exposures geographically and create environment tracking systems.

Recommendations:

- Support ambitious and creative collection of environmental exposure data: (Section 5.2.2)
  - Build toxin screening (e.g., dioxin, lead) into routine blood panels, and questions about diet and environmental toxins into health questionnaires.
  - Start urban sensing and tracking programs that align with the geographic areas for the *All of Us* Research Program and similar projects in the future.
  - Support the development of wearable devices for the sensing of environmental toxins.
  - Support the development of broad-based pathogen sensing for rural and urban environments.
  - Develop protocols and IT capabilities to collect and integrate the diverse data.

5. Embrace the Crowdsourcing Movement to Support AI development and Data Generation

Finding: AI competitions have already demonstrated their value in 1) encouraging the creation of large corpuses of data for broad use, and 2) demonstrating the capabilities of AI in health, when provided data that are curated into a well labeled (namely high information content) format. (Section 4.12)

Recommendations:

- Support competitions created to advance our understanding of the nature of health and health care data. (Section 4.12)
- Share data in public forums to engage scientists in finding new discoveries that will benefit health. (Section 4.12)

6. Understand the Limitations of AI Methods in Health and Health care Applications

Findings:

- There is potential for the proliferation of misinformation that could cause harm or impede the adoption of AI applications for health. Websites, Apps, and companies have already emerged that appear questionable based on information available. (Section 3.2)
- Methods to insure transparency in disclosure of large scale computational models and methods in the context of scholarly reproducibility are just beginning to be developed in the scientific community. (Section 6.2)

Recommendations:

- Support the development of critical safeguards that are essential to enable the adoption of AI for public health, community health, and health care delivery:
    - Encourage development and adoption of transparent processes and policies to ensure reproducibility for large scale computational models. (Section 6.2)
    - To guard against the proliferation of misinformation in this emerging field, support the engagement of learned bodies to encourage and endorse best practices for deployment of AI applications in health. (Sections 3.2 and 6.2)

1 INTRODUCTION

Artificial Intelligence (AI), where computers perform tasks that are usually assumed to require human intelligence, is currently being discussed in nearly every domain of science and engineering. Major scientific competitions like ImageNet Large Scale Visual Recognition Challenges [1] are providing evidence that computers can achieve human-like competence in image recognition. AI has also enabled significant progress in speech recognition and natural language processing [2]. All of these advances open questions about how such capabilities can support, or even enhance, human decision making in health and health care. Two recent high-profile research papers have demonstrated that AI can perform clinical diagnostics on medical images at levels equal to experienced clinicians, at least in very specific examples [3,4].

The promise of AI is tightly coupled to the availability of relevant data [5]. In the health domains, there is an abundance of data [6]. However, the quality of, and accessibility to, these resources remain a significant challenge in the United States. On one hand, health data has privacy issues associated with it, making the collection and sharing of health data particularly cumbersome compared to other types of data. In addition, health data are quite expensive to collect, for instance in the case of longitudinal studies and clinical trials, so it tends to be tightly guarded once it is collected. Further, the lack of interoperability of electronic health record systems impedes even the simplest of computational methods [7] and the inability to capture relevant social and environmental information in existing systems leaves a key set of variables out of data streams for individual health [8].

At the same time, there is wide private-sector interest in AI in health data collection and applications as illustrated from the numerous startups related to AI in health and health care (a partial list as of 2016 is captured in Figure 1) [9]. Most (75) of the 106 listed startups are headquartered in the US. There are startups in 15 different countries, with the UK and Israel having the largest number of startups outside the US. The two most popular topics, medical imaging & diagnostics and patient data & risk analytics, are a strong focus in this report. However, another key focus of this report, the importance of environmental factors, is less apparent in the startup activity shown.

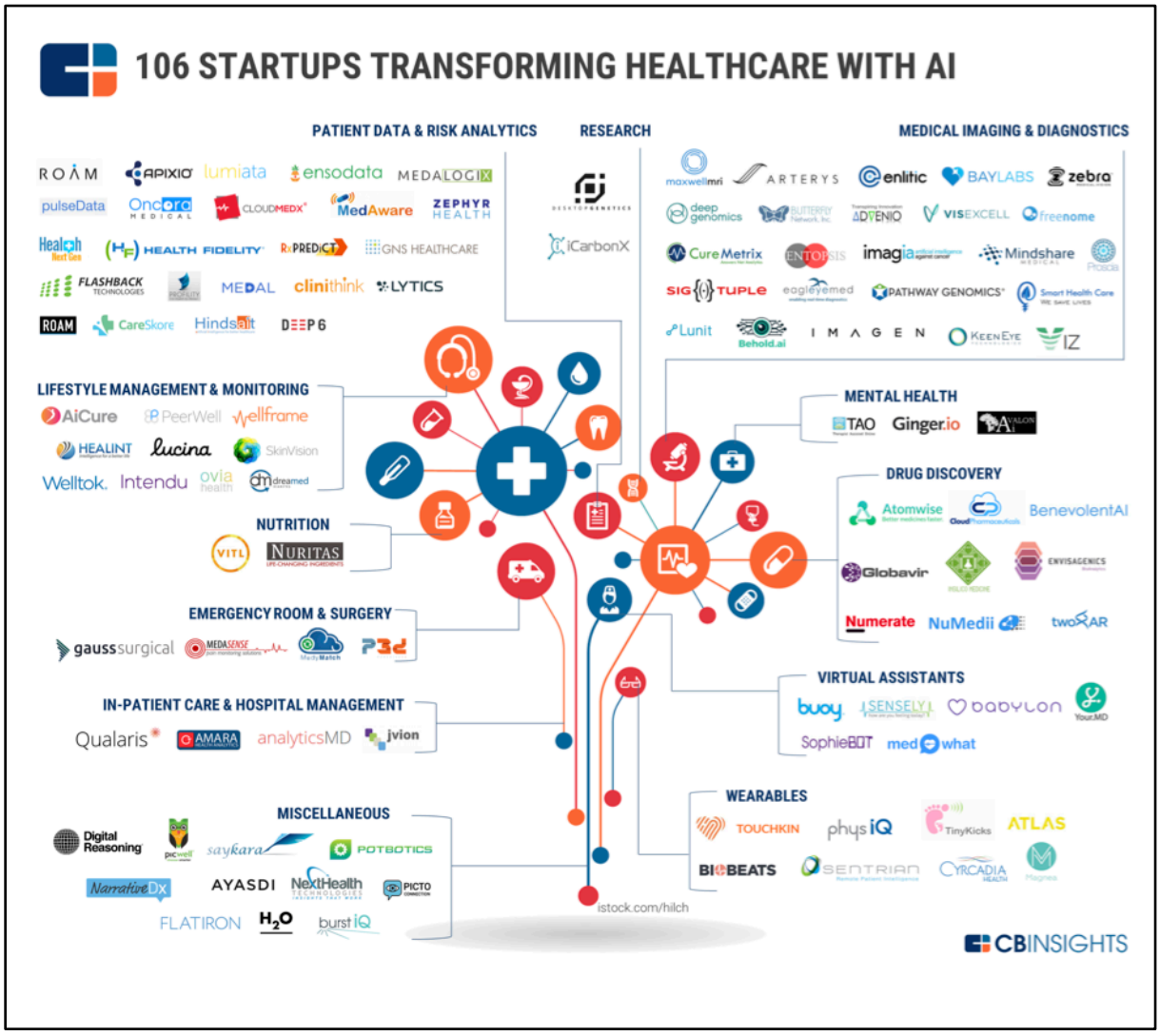

Figure 1: AI in Health Care Startups. From CB Insights (2016) [9].

1.1 Why Now?

AI has been around for decades and its promise to revolutionize our lives has been frequently raised, with many of the promises remaining unfulfilled. Fueled by the growth of capabilities in computational hardware and associated algorithm development, as well as some degree of hype, AI research programs have ebbed and flowed. The JASON 2017 report [10] gives this history and also comments on the current AI revolution stating:

> "Starting around 2010, the field of AI has been jolted by the broad and unforeseen successes of a specific, decades-old technology: multi-layer neural networks (NNs). This phase-change reenergizing of a particular area of AI is the result of two evolutionary developments that together crossed a qualitative threshold: (i) fast hardware Graphics Processor Units (GPUs) allowing the training of much larger—and especially deeper (i.e., more layers)—networks, and (ii) large labeled data sets (images, web queries, social

networks, etc.) that could be used as training testbeds. This combination has given rise to the "data-driven paradigm" of Deep Learning (DL) on deep neural networks (DNNs), especially with an architecture termed Convolutional Neural Networks (CNNs)."

Is the current era just another hype cycle [11]? Or are things different this time that would make people receptive to embracing the promise of AI applications in health and health care? AI is largely exciting to computational sciences researchers throughout academia and industry. Perhaps previously the revolutionary advances in AI had no obvious way to touch the lives of individuals. The opportunities from health, including health care delivery, for AI may today be enhanced by current societal factors that make the fate of AI hype different this time. Currently, there is great frustration in the cost and quality of care delivered by the US health care system [12]. To some degree, this has fundamentally eroded patient confidence, opening people's minds to new paradigms, tools, services. Dovetailing with this, there is an explosion in new personal health monitoring technology through smart device platforms [13,14] and internet-based interactions [15]. This seemingly perfect storm leads to an overarching observation, which defines the environment in which AI applications are now being developed:

Overarching Observation: Unlike previous eras of excitement over AI, the potential of AI applications in health and health care may make this era different because the confluence of the following three forces has primed our society to embrace new health centric approaches that may be enabled by advances in AI: 1) frustration with the legacy medical system, 2) ubiquity of networked smart devices in our society, 3) acclimation to convenience and at-home services like those provided through Amazon and others.

1.2 JASON Study Charge and Process

The U.S. Department of Health and Human Services (HHS), through the Office of the National Coordinator for Health IT (ONC) and the Agency for Healthcare Research and Quality (AHRQ), and with support from the Robert Wood Johnson Foundation requested this JASON study. ONC, reporting directly to the Secretary of HHS, was established by executive order in 2004 and established in statute in 2009 by the Health Information Technology for Economic and Clinical Health Act as the principal federal entity responsible for the coordination and implementation of nationwide efforts for the electronic exchange of health information. AHRQ, an agency within HHS, develops the knowledge, tools, and data needed to improve the quality and safety of the health care system and help Americans, health care professionals, and policymakers make informed health decisions.

HHS asked JASON to assess the full impact that AI can have on health and health care in the context of how AI could shape the future of public health, community health, and health care delivery from a personal level to a system level. Understanding these AI opportunities and considerations can better prepare and inform AI development, policy making, and promote the general welfare of health care consumers and the public.

JASON was introduced to the topic through briefings by various experts, listed in Table 1. Materials recommended by these individuals, together with a wide range of other publically

available materials, were reviewed and discussed by JASON. Most briefers attended the full set of presentations and participated in the accompanying discussions.

Specific mathematical details surrounding current AI applications, including deep learning and convolutional neural networks, will not be discussed here. The reader is referred to JASON 2017 [10] for an excellent exposition of these models and architectures.

Table 1: Briefers

Name	**Affiliation**
Abdul Hamid Halabi	NVIDIA
Kimberly Powell	NVIDIA
Andy Beam	Harvard
Zak Kohane	Harvard
Ziad Obermeyer	Harvard
Eileen Koski	IBM Research
Georgia Tourassi	Oak Ridge National Labs
John Wilbanks	Sage Bionetworks
Kevin Chaney	HHS
Teresa Zayas Cabán	HHS
Lynda Chin	University of Texas
Mark DePristo	Google
Jonathon Shlens	Google
Paul Silvey	MITRE
Russ Altman	Stanford

Focus of the Study.

JASON was asked to consider how AI will shape the future of public health, community health, and health care delivery. We focused on technical capabilities, limitations, and applications that can be realized within the next ten years, in the context of the questions developed with the sponsors (see Appendix).

The organization of this report is as follows. Section 2 focuses on health care and health care delivery. Section 3 reviews the rapid development of smart devices and associated mobile applications in health, and Section 4 argues for the need for good data to drive AI application development generally, and health applications specifically. Section 5 covers issues around large-scale health data collection and missing data streams. Section 6 discusses what is needed for successful adoption of AI in health. The report concludes with a summary of the findings and recommendations in Section 7 and an epilogue in Section 8.

2 AI IN HEALTH DIAGNOSTICS: OPPORTUNITIES AND ISSUES FOR CLINICAL PRACTICE

There have been significant demonstrations of the potential utility of Artificial Intelligence approaches based on Deep Learning [10] for use in medical diagnostics [16]. While continuing basic research on these methods is likely to lead to further advances, we recommend parallel, focused work on creating rigorous testing and validation approaches for the clinical use of AI algorithms. This is needed to identify and ameliorate any problems in implementation [17,18], as soon as possible, in order to develop confidence within the medical community and to provide feedback to the basic research community on areas where continued development is most needed.

We point out a key issue of balance in expectations, which is that AI algorithms, including Deep Learning, should not be expected to perform at higher levels than the training sets. However, where good training sets represent the highest levels of medical expertise, applications of Deep Learning algorithms in clinical settings provide the potential of consistently delivering high quality results. Thus, one aspirational goal for such applications should be to make high quality health care services available to all.

2.1 Advances in AI Applications for Medical Imaging

In the following sections, we review examples in which applications of Deep Learning have been demonstrated, with attention to quantitative understanding of characteristics of the data sets, the problem definition, and the nature of the comparison standard used for labeling the sets. The two examples described are based on medical imaging, specifically diabetic retinopathy and dermatology.

2.1.1 Detection of diabetic retinopathy in retinal fundus images

Many diseases of the eye can be diagnosed through non-invasive imaging of the retina through the pupil [19]. Early screening for diabetic retinopathy is important as early treatment can prevent vision loss and blindness in the rapidly growing population of patients with diabetes. Such screening also provides the opportunity to identify other eye diseases, as well as providing indicators of cardiovascular disease.

The increasing need for such screening, and the demands for expert analysis that it creates, motivates the goal of low cost, quantitative retinal image analysis. Routine imaging for screening uses the specially designed optics of a ‘fundus camera,’ with several images taken at different orientations (fields, see Figure 2) [20] and can be accomplished with (mydriatic) or without (non-mydriatic) dilation of the pupil. Assessment of the image requires skilled readers, and may be performed by remote specialists. With the advent of digital photography, digital recording of retinal images can be carried out routinely through Picture Archiving and Communication Systems (PACS).

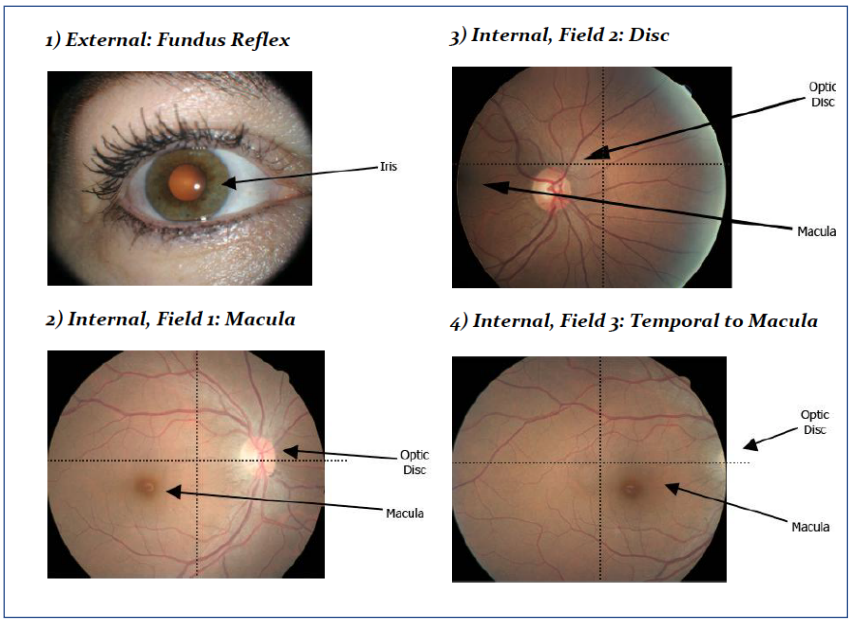

Figure 2: Standard image formats for diabetic retinopathy (right eye). *Source:* taken from EYEPACS LLC 2017.

As a point of reference, the standards for screening [21] for diabetic retinopathy in the UK require at least 80% sensitivity and 95% specificity to determine referral for further evaluation. Screening using fundus photography, followed by manual image analysis, yields sensitivity and specificity rates cited as 96%/89% when two fields (angles of view) are included, and 92%/97% for three fields. (For a single field, cited rates are 78%/86%).

Recently a transformational advance in automated retinal image analysis, using Deep Learning algorithms, has been demonstrated [22]. The algorithm was trained against a data set of over 100,000 images [23], which were recorded with one field (macula-centered). Each image in the training set was evaluated by 3-7 ophthalmologists, thus allowing training with significantly reduced image analysis variability. The results from tests on two validation sets, also involving only one image per eye (fovea centered), are striking. Selecting for high specificity (low false negatives), yielded sensitivities/specificities of 90.3%/98.1% and 87.0%/98.5%). Selecting for high sensitivity yielded values of 97.5%/93.4% and 96.1%/93.9%). These results compare favorably with manual assessments even where those are based on images from multiple fields as noted above. They also are a significant advance over previous automated assessments, which consistently suffered from significantly lower sensitivities [24].

The Deep Learning algorithm shows great promise to provide increased quality of outcomes with increased accessibility. Continued work to establish its use as an approved clinical protocol (see Section 2.3), will be needed. Once validated, its use can be envisioned in a wide range of scenarios, including decision support in existing practice, rapid and reduced cost analysis in place of manual assessment, or enabling diagnostics in non-traditional settings able to reach underserved populations. Greatly expanded accessibility is likely to be aided by deployment of

low cost fundus cameras, which are under rapid development [25,26] and likely to be supported by apps as described in Section 3.

2.1.2 Dermatological classification of skin cancer

Skin cancer represents a challenging diagnostic problem because only a small fraction (3–5% of about ~1.5 million annual US skin cancer cases) are the most serious type, melanoma, which accounts for 75% of the skin cancer deaths. Identifying melanomas early is a critical health issue, and because diagnosis can be performed on photographic images, there are already services that allow individuals to send their smart-phone photos in for analysis by a dermatologist [27]. However, the detection of melanomas in screening exams is limited – sensitivity 40.2% and specificity 86.1% for primary care physicians and 49.0%/ 97.6% for dermatologists [28].

A recent demonstration of automated skin cancer evaluation using a convolutional neural network (CNN) algorithm yielded striking results [29]. The authors drew on a training set of over 125,000 dermatologists labeled images, from 18 different online repositories. Two thousand of the images were also labeled based on biopsies. The algorithm was trained on all the dermatologist labeled images, using 757 disease classes and over 2000 diseases. The top levels of the taxonomy are shown in Figure 3a. In testing the algorithm, the algorithm performed similarly to dermatologists in classifying at the 1st level I (three categories; benign, malignant and non-neoplastic) with 72.1% accuracy vs. 66.0 and 65.56% for two dermatologists. In classifying for the 2d level (9 disease classes), the respective accuracies were 55.4% for the algorithm versus 53.3 and 55.0% for the two dermatologists. The performance levels of the algorithm are almost certainly limited by levels of sensitivity and accuracy for the labeling of images in the training sets.

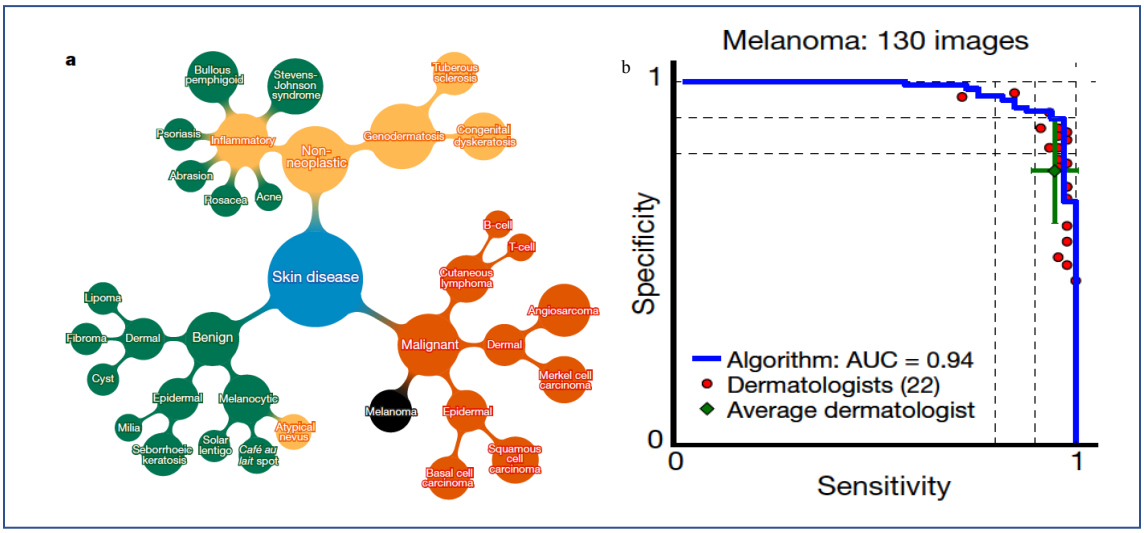

Figure 3: a) (left panel) Illustration of the top levels of the tree-structured taxonomy. The full set of 2032 diseases are leaf nodes and were used for the developing the algorithm. b) (right panel) Classification results for a set of 130 images of melanocytic lesions, blue curve from the algorithm, red dots from individual dermatologists. Images taken from Esteva et al. 2017 [30].

A further classification test was performed drawing only on images that were biopsy-proven to be in a specific disease class. The algorithm then was run to answer only the question of whether the lesion in the image was benign or malignant. The results for analysis of 130 images of melanocytic lesions are shown in Figure 3b, compared with results from assessments by 22 different dermatologists. As with the broader classification tests, the algorithm performs similarly or slightly better than individual dermatologists. The performance for both algorithm and dermatologists is much better for this specific task than for the classification, noted above, of images from a set representing all the different diseases.

As with the retinopathy example, these results indicate that AI algorithms can perform at levels matching their training sets. The poor level of results for the broad screening tests is consistent with the training set, which is based on dermatological characterization. It would be of great interest to understand whether a training set based on a more accurate method of discrimination, for instance biopsies, would allow the algorithm to perform significantly better. The much better results on the narrower classification task, for both the algorithm and the dermatologists, suggests that the clinical decisions that originally led to these cases being selected for biopsies may have removed many less-easily classified images. Overall this is a very promising result, but as the authors note, more work is needed for it to deliver value in a broad clinical setting.

2.1.3 Data issues

Each of the examples above relied on access to large sets of medical images, some of which were available in professionally maintained archives. In addition, however, each also required the development of a labeled training set. In the retinopathy example these were obtained via labor-intensive independent professional assessments of the images. In the dermatology case, the training was also based on clinician assessments of the images. It would be of significant interest to determine whether training against a more rigorous assessment, such as the outcome of a biopsy, improves the performance of the algorithm. The data sets available in the dermatology example included too small a number of biopsy-proven images to serve as the sole training basis. One question to consider is whether mixing images labeled based on a biopsy with those labeled by image reading would improve the performance of the algorithm.

Access to high-quality labeled data is a significant barrier in the development and evaluation of AI algorithms for clinical decision making [30]. The process of developing training labels from existing records requires skilled labor review of patient charts to create meaningful labels. As a result of the time and cost involved many publically accessible labeled sample sets are too small for the best training and testing of AI algorithms. Data sets that are labeled from clinical (as opposed to research-level sources) may be of variable quality that limits the training efficacy, as discussed in Section 6.1.

Findings:

- AI algorithms based on high quality training sets have demonstrated performance for medical image analysis at the levels of the medical capability that is captured in their training data.
- AI algorithms cannot be expected to perform at a higher level than their training data, but should deliver the same standard of performance consistently for images within the training space.

Recommendation:

- Support the assessment of AI algorithms using data labeled at levels that exceed standard assessments, for instance the use of outputs from a further stage of diagnostic testing (e.g., use of biopsy results to label dermatological images).

2.2 Moving Computational Advances into Clinical Practice

Computational methods are recognized as a special type of tool in medical practice and where they impact clinical practice, are subject to different forms of regulation. AI-based tools represent a new set of opportunities, but lack the extensive basis of experience and validation that is needed for acceptance into formal medical practice. In this section and Section 2.3, we address the question of how AI-based computational approaches could become sufficiently trusted to modify protocols for diagnosis and treatment, enabling improved outcomes at lowered costs.

We use the example of a new computational tool that requires large scale computation, but is based on physical properties rather than a trained AI algorithm. The example illustrates the rigorous process of review and development needed to validate new techniques for clinical practice. It also illustrates how important the demonstration of the benefits of a new approach is to its eventual uptake as a clinical tool.

2.2.1 Coronary artery disease – issues driving interest in improved methods

A large number of patients experiencing chest pain undergo invasive testing only to reach a negative diagnosis. The negative impacts on patient experience, access to expensive diagnostic facilities, and cost create a strong incentive to develop alternative approaches.

The diagnosis of whether a patient has coronary artery disease (CAD), and will benefit from cardiac revascularization (bypass surgery or insertion of a stent) is based on two primary factors. The first is structural: the narrowing of the blood vessels (stenosis) around the heart. The second is functional: reduced flow of blood (ischemia) compared with the flow in a normal coronary artery. Direct measurement of stenosis can be accomplished with invasive coronary angiography, in which a cardiac catheter is inserted to deliver a contrast dye to the arteries supplying the heart followed by X-ray imaging. Direct measurement of blood flow is accomplished by the invasive fluid flow reserve (FFR) technique, based on insertion of a pressure sensor through a cardiac catheter.

Unfortunately, the use of invasive coronary angiography for patients with severe chest pain reveals a large fraction (60% for patients who do not have other significant indicators of CAD) who do not have significant arterial narrowing [31]. Furthermore, 50% or more of the cases with significant narrowing do not have significantly impaired blood flow (measured by FFR) and would not benefit from revascularization [32].

2.2.2 Development of new approaches – non-invasive diagnostics

Coronary computed tomographic angiography (CCTA) has been established as a non-invasive technique for screening [33]. However even at its best performance, CCTA-like invasive

coronary angiography (ICA) – has limited ability to discriminate which cases of stenosis are truly causing impaired blood flow. Recently advances in computational fluid dynamics (CFD) and in physical understanding of arterial behavior have enabled computational determination of the fluid flow reserve computationally [34] using CCTA as the input for the structure. An example of such a blood flow assessment is shown in Figure 4.

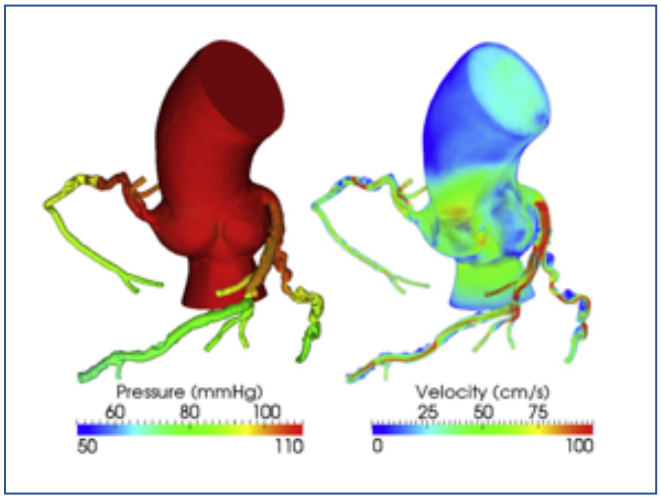

Figure 4: Three-dimensional pressure and velocity fields at one point in the cardiac cycle using FFR- based on CTA imaging (FFR_{CT}). The computation is repeated throughout the cardiac cycle. *Source:* Images have been taken from Taylor et al 2013 [34].

The process of establishing this approach for clinical use included early assessments (ClinicalTrials.gov NCT01189331 and NCT01233518) that compared the performance of computational FFR (FFR_{CT}) and CCTA alone for diagnosing significantly reduced blood flow. The results showed that FFR_{CT} performed dramatically better on per-vessel assessments, and significantly better on per-patient assessments [35]. These positive results, with the potential for greatly reducing the number of invasive tests while maintaining quality in diagnosis, created significant interest in evaluating the technique for use in clinical care.

2.2.3 Development and validation for clinical applications

A private company, Heart Flow, Inc, [36] led development of the FFR_{CT} technology for clinical use. In 2014, the company received FDA approval to market the technology, based on its demonstration of substantial equivalence to predicate devices [37]. In parallel, additional studies were underway to establish whether FFR_{CT} is a feasible alternative to invasive coronary angiography [38] and its potential use as a direct diagnostic for CAD-induced reduced blood flow requiring revascularization [39]. The test of diagnostic performance yielded the results shown in Table 2, showing favorable performance for FFR_{CT}.

Table 2: Per-patient (upper panel) and per-vessel (lower panel) diagnostic performance of CCTA, fractional flow reserve derived from standard acquired coronary computed tomography angiography datasets, and invasive coronary angiography. Ranges represent a 95% confidence interval. Threshold for stenosis was 50% or more reduction in vessel diameter and for FFR was a reduction to 0.8 or less.

	Per-Patient				
	Coronary CTA Stenosis >50%	p Value	FFR_{CT} ≤0.80	p Value	ICA Stenosis >50%
Accuracy	53 (47–57)	<0.001	81 (76–85)	0.09	77 (71–82)
Sensitivity	94 (86–97)	0.058	86 (77–92)	<0.001	64 (53–74)
Specificity	34 (27–41)	<0.001	79 (72–84)	0.29	83 (77–88)
PPV	40 (33–47)	<0.001	65 (56–74)	0.34	63 (52–73)
NPV	92 (83–97)	0.46	93 (87–96)	<0.001	83 (77–89)
	Per-Vessel				
	Coronary CTA Stenosis >50%	p Value	FFR_{CT} ≤0.80	p Value	ICA Stenosis >50%
Accuracy	65 (61–69)	<0.001	86 (83–89)	0.032	82 (79–86)
Sensitivity	83 (74–89)	0.91	84 (75–89)	<0.001	55 (45–65)
Specificity	60 (56–65)	<0.001	86 (82–89)	0.16	90 (86–93)
PPV	33 (27–39)	<0.001	61 (53–69)	0.18	58 (48–68)
NPV	92 (88–95)	0.068	95 (93–97)	<0.001	88 (85–92)

Source: Table extracted from Norgaard et. al. (2014) [41].

These (and other) study results enabled the critical step of evaluation for actual clinical use. Such assessment requires comparison with the established standards of care along with quantification of the impact for patient quality of life and economics [40, 41]. The results confirmed that using FFR_{CT} can correctly predict patients who should not be referred for an ICA, greatly reducing the amount of invasive testing. The reduction in invasive procedures and associated hospitalization and medications during a 90-day follow-up period improved patient quality of life metrics and reduced costs by approximately 30%.

Being established as a standard of care is the final step of approval for new clinical practices. The UK National Institute for Health and Care Excellence (NICE) guidance for FFR_{CT}, illustrates this process [42]. The company presented the case for the FFR_{CT} technology to NICE. NICE carried out an independent literature review and an independent assessment of the cost reductions due to fewer inconclusive or inaccurate diagnostic tests and avoidance of unnecessary staff and procedure costs. The resulting recommendation is that FFR_{CT} should be considered as an option as part of the NICE pathway on chest pain, with a potential cost savings of over £200 per patient. Similar approval processes are required on a country-by-country basis. In the US, almost all health insurance payment and information systems require that a procedure have been awarded an American Medical Association Current Procedural Terminology Code [43].

With approval in place, the challenge becomes the rate of uptake in practice. HeartFlow is continuing with additional clinical trials to demonstrate additional benefits to the medical community. Another pathway to uptake is relationships with equipment providers, which offers the potential for the new diagnostic technique to be marketed as part of an equipment line. The potential to further improve the technology using AI has been recognized and is in development [44].

2.2.4 Summary points for developing clinical applications

The FFR_{CT} example illustrates key issues for a new computational technology to enter into practice:

- The technology should address a significant, identified clinical need.
- The technology must perform at least as well as the existing standard approach.
- Substantial clinical testing is needed to verify the performance of the new technology under the wide range of clinical situations in which it may be used.
- The new technology should provide improvements in patient outcomes, patient quality of life, practicality in use, and reduced medical system costs.

Because the FFR_{CT} technology is based on physical principles, rather than the less-understood correlations of AI, its review and acceptance process likely faced less skepticism than a new AI approach may encounter. For the medical community to develop trust in AI-based tools, assessments at least as rigorous will be needed.

2.3 Evolution of Standards for AI in Medical Applications

In the US, adoption of AI applications for clinical practice will be regulated by a combination of Food and Drug Administration (FDA) regulations and clinical business models. For non-clinical personal smart device uses discussed in the following section, it will require public (user) confidence and may or may not require regulation.

In the example above, the adoption of the FFR_{CT} computational tool required clinical studies. This will also be necessary for AI outcomes to be accepted as definitive decision factors in standard of care. Such decisions may need FDA regulations and the specifics of what those requirements might be are still evolving [45].

FDA has been paying close attention to the international development of both software in medical devices (SiMD) and software as a medical device (SaMD). AI applications can fall into both of these categories. FDA has been participating (actually chairing) an international regulators forum on SiMD and SaMD [46]. The goal of the forum is to develop frameworks, risk categorizations, vocabulary, considerations, and principles that could support regulation around software [47]. Their report on clinical evaluation [48] was put out in the US for public comment [49] by the FDA. FDA plans to use that input as the basis to guide their new development of regulation for SiMD and SaMD and to be responsive to new policies under the 21st Century Cures Act noting [50]

> "... the Act revised FDA's governing statute to, among other things, make clear that certain digital health technologies—such as clinical administrative support software and mobile apps that are intended only for maintaining or encouraging a healthy lifestyle—generally fall outside the scope of FDA regulation. Such technologies tend to pose low risk to patients but can provide great value to the health care system."

FDA's new digital health unit will be using this input to formulate FDA's role in regulation of digital health to include mobile health (mHealth), health information technology (IT), wearable devices, telehealth and telemedicine, and personalized medicine [51].

Clinical trials and regulation are only part of the story for adoption of AI applications. In addition clinicians or health professionals (e.g., physical training, doctor, or public health specialist) must be willing to incorporate these applications into their workflows. The FFR_{CT} example of Section 2.2 and the AMA study [52] discussed in Chapter 3 point out that adoption will depend on how well the mHealth devices and applications fit within existing systems and practices. However, even to support the development of clinical trials and the assurance that the AI applications are legitimate, including for non-regulated applications, the technical soundness of the algorithms need to be confirmed. This point is addressed in more depth in Section 6.

Findings:

- The process of developing a new technique as an established standard of care uses the robust practice of peer-reviewed R&D, and can provide safeguards against the deceptive or poorly-validated use of AI algorithms (see also Section 6).
- The use of AI diagnostics as replacements for established steps in medical standards of care will require far more validation than the use of such diagnostics to provide supporting information that aids in decisions.

Recommendation:

Support work to prepare promising AI applications' results for the rigorous approval procedures needed for acceptance for clinical practice. Create testing and validation approaches for AI algorithms to evaluate performance of the algorithms under conditions that differ from the training set.

3 PROLIFERATION OF DEVICES AND APPS FOR DATA COLLECTION AND ANALYSIS

Smartphones and other smart technologies are already a primary platform for the adoption of health and wellness through mHealth (mobile or digital health) apps and networked devices [53]. These apps and devices support the full spectrum from healthy to sick, e.g., from episodic fitbit users to diabetics who use glucose monitors. Questions arise regarding the use and usefulness of these devices by individuals and the willingness of the medical community to integrate mHealth into health care.

There is active research on the design and testing of the usability of the apps and devices [54]. Additionally, many of the mHealth applications are being included in clinical trials. They are being used as a mechanism to collect information for the trial, to evaluate the specific device or app use, or to evaluate their usefulness in combination with other health behaviors. For example, in 2016 it was reported that Fitbit alone is being used in 21 clinical trials [55].

This growth in the development of more serious health devices that could be used to monitor and communicate health status to a professional has attracted the attention of the American Medical Association (AMA). AMA recently adopted a set of principles to promote safe, effective mHealth applications [56]. AMA is encouraging physicians and others to support and establish patient-physician relationships around the use of apps and associated devices, trackers, and sensors.

An AMA survey released in 2016 reported that 31% of physicians see the potential for digital tools to improve patient care and about half are attracted to digital tools because they believe they will improve current practices with respect to efficiency, patient safety, improved diagnostic ability, and physician-patient relationships [57]. The study goes on to point out that adoption in practice will require these tools fit within existing systems and practices including coverage for liability, data privacy assurance, ability to link to electronic health records, and billing/reimbursements. These same points reflect issues for the adoption of AI applications discussed in Sections 2.2 and 2.3.

These are all promising directions for the development and application of mHealth tools. The focus here is on the types of devices and apps that have the potential to benefit from AI applications either in their individual functions or when the data generated by the device can be integrated with other health information to support wellness versus sickness.

3.1 Personal Networked Devices and Apps

There are many impressive smartphone attachments and apps currently available for monitoring of personal health. These devices 1) empower individuals to monitor and understand their own health, 2) create large corpuses of data that can, in theory, be used for AI applications, and 3) capture health data that can be shared with clinicians and researchers. AI algorithms drive the performance of many of these devices and, reciprocally, these devices are capturing data that

could be used to develop or improve AI algorithms. Here we list some specific examples of modern health-monitoring tools available for use on mobile devices.

- **Personal EKG.** Kardia Mobile [58] has produced an FDA-cleared personal EKG recording device. The platform uses a finger pad and a smartphone app to record an EKG over a 30 second window. The device operates with no wires or gels. The platform claims to use AI-enabled detection of atrial fibrillation.
- **Parkinson's tremors.** CloudUPDRS [59] is a smartphone app that can assess Parkinson's disease symptoms. The app uses the gyroscope found in many mobile devices to analyze and quantify tremors, patterns in gait, and performance in a "finger tapping" test. An AI algorithm differentiates between actual tremors and "bad data," such as a dropped phone or the wrong action in response to the app's question. This tool enables Parkinson's patients to perform in-home testing, providing valuable and quantitative feedback on how their personal lifestyle factors and medications may affect their symptoms.
- **Asthma tracking and control.** AsthmaMD [60] offers a hand-held flow meter which gauges lung performance by assessing peak flow during exhalation. The flow meter pairs with an app that logs data for people with asthma and other respiratory diseases. Users can also record symptoms and medications. An interesting feature of this app is that users may opt-in to a program where their data are uploaded anonymously onto a Google database being assembled for research purposes. AsthmaMD states "anonymous, aggregate data will help correlate asthma with environmental factors, triggers and climate change."

These sorts of technologies can collect information of clear and vital importance to patients and use by clinicians, but we must again emphasize that each new data stream must be evaluated, collected and curated to formats consistent with clinical needs and AI applications.

New devices are surely to come online and new ideas for these devices are being sponsored by government agencies. The National Institutes of Health has recently put out a request for proposals, called Mobile Monitoring of Cognitive Change (U2C) [61]:

> "This Funding Opportunity Announcement (FOA) invites applications to design and implement research infrastructure that will enable the monitoring of cognitive abilities and age, state, context, or health condition-related changes in cognitive abilities on mobile devices. This effort will include the development (or support for development) of apps on the Android and iOS platforms, the validation of tests and items to be used on the two leading smartphone platforms in age groups ranging from 20 to 85, and the norming of successfully validated measures to nationally representative U.S. population samples that will also receive gold standard measures, including the NIH Toolbox® for Assessment of Behavioral and Neurological Function. A goal of this project is to also support data collection efforts from participants enrolled in projects awarded through this FOA as well as other NIH-funded studies through FY2022, and enable the widespread sharing of both the collected data and the test instruments."

There are certain parts of the human body, however, that haven't been successfully scrutinized by cell phone-based attachments. The bloodstream is one example. There are many things that would be desirable to measure in blood, both metrics of health (e.g., vitamins and minerals) and disease (e.g., viruses and cancer biomarkers). However, with the notable exception of glucose,

most things are at concentrations too low to be measured in a finger-prick volume of blood. Although not specific to the immediate report topic of AI applications in Health, the assessment of current and potential ultra-sensitive assays for chemicals and biomarkers in small blood samples should be evaluated. One can imagine a day where people could, for instance, 1) use their cell phone to check their own cancer or heart disease biomarker levels weekly to understand their own personal baseline and trends, or 2) ask a partner to take a cell-phone-based HIV test before a sexual encounter.

3.1.1. Capturing mobile device information – utility and privacy

Mobile devices and apps could provide a rich source of data to leverage broader AI applications. This is already happening as summarized above. But, more could be done to leverage the rapid development of mHealth apps, devices, and sensors. JASON 2014 [8] discussed the potential value in developing a data infrastructure that would ingest the wide range of data that is being generated around individual health, mobile technologies being one of them. Significant steps forward to create such an infrastructure have not been forthcoming, perhaps because the problem is simply too broad. One area of focus should be on data capture and infrastructure to support AI applications that enhance the performance and adoption of mHealth tools and thus give individuals more health autonomy. In addition, providing access to data captured by mHealth apps and devices could enhance the research community's ability to build more insights into public health through AI.

Such data sharing would require informed consent of the participants. One promising example is the combination of a mobile electronic consent application with a mHealth app designed to collect data that will be helpful in managing an individual's Parkinson disease [62]. The app was made available free in the US through the Apple App Store [63]. The app explained the study, who could participate, and gave options for sharing your data. The data sharing options included sharing your data with the specific research team that developed the app and were conducting the study or with a broader research community. The study's "mPower" app was downloaded by 48k people of whom 25% were eligible and participated in the study (12.2k). Of those participants, 78% opted to have their data shared broadly.

This is a nice example because it shows that people are willing to share data collected on their mobile devices for research purposes. In this example, the data being collected could also be a useful addition to an individual's electronic health record. Finally, the implemented informed consent process both educated the individual on what they were consenting to regarding their data, as well as giving an option to consent electronically, thus allowing their data to immediately be captured.

3.1.2 Online plus AI

There is a proliferation of companies developing apps that offer online doctors' appointments. In the U.S., this includes the new company PlushCare [64] but these services currently seem to be more prevalent in the U.K. These apps allow nearly instantaneous access to a live doctor, over mobile devices, anytime of day, every day of the week. One of them, Babylon, claims to use an AI algorithm, along with a series of questions about symptoms, to automatically triage patients [65]. Interestingly, the company is now expanding to Rwanda, where there is a serious shortage of doctors, yet a high penetration of smart phones. Online doctors' appointments are likely to

appeal to many people who are already acclimated to the use of apps to fulfill personal needs (e.g., Amazon, Uber, etc.). However, the potential dangers of sharing personal health information over such networked connections is a concern.

3.1.3 Examples of privacy and transparency

The rapid evolution and adoption of AI applications in remote access health care has a strong presence outside of the US. One such example is deep learning from the company DeepMind Technologies [66]. In 2016, DeepMind launched several initiatives in the health care arena under its DeepMind Health Division [67].

DeepMind is working with UK National Health System (NHS) Hospital Trust Foundation [68] to develop AI applications involving patient electronic health records from multiple London hospitals. It is indicative of the fast moving pace of these AI in health care applications that the most informative information on the plans of DeepMind Health comes from investigative news articles written within the last few months of this study [69]. These have focused on issues of transparency, privacy, and health ethics issues relating to delivery of AI products in health care [70]. NHS patient data was provided without informed consent to Google to test an app designed to help monitor kidney disease. Some argue that this violated UK regulations that state "patient records without explicit consent can only be used for direct care delivery." However, this controversy has not slowed DeepMind's continued and expanding partnership with NHS [71].

From a US government perspective, the work of DeepMind Health using UK NHS data should be viewed as an early large scale "experiment" that will reveal many real-world issues that arise in the application of AI to health care. It should therefore be tracked closely. For example, the problem discussed above with data access transparency may have led DeepMind to an accelerated application of "blockchain-style" technology for securing and tracking data access [72]. Basically, blockchain methodologies use a distributed database consisting of continuously updated (augmented) "blocks" which contain a linked list of all previous transactions [73]. In the case of health care, this encompasses all previous records of access to an individual data record including information about how the data was used and any additions or changes to the data record [74,75].

A second technology application that has emerged from DeepMind Health has many blockchain-like aspects [76,77]. Instead of blockchain, the DeepMind data audit system uses an approach based on Merkle Trees [78], a type of hash tree that allows secure verification of the contents of large data structures. DeepMind hopes to prototype the verifiable data audit system by the end of 2017 for eventual use in its Royal Hospital health care software environment [79]. While the audit system will be prototyped within the confines of the NHS organizations, DeepMind sees the possibility that the audit system itself might become available to individuals in the public so that they can access and verify their actual electronic health records.

Given the rapid growth of similar activities in other countries, the U.S. government should track developments in foreign health care systems, looking for useful technologies and also technology failures.

This leads us to the following findings and recommendations regarding the confluence of AI and personal networked devices and apps.

Findings:

- Revolutionary changes in health and health care are already beginning in the use of smart devices to monitor individual health. Many of these developments are taking place outside of traditional diagnostic and clinical settings.
- In the future, AI and smart devices will become increasingly interdependent, including in health-related fields. On one hand, AI will be used to power many health-related mobile attachments and apps. On the other hand, mobile devices will create massive datasets that may open new possibilities in the development of AI-based health and health care tools.

Recommendations:

- Support the development of AI applications that can enhance the performance of new mobile monitoring devices and apps.
- Develop data infrastructure to capture and integrate data generated from smart devices to support AI applications.
- Require that development include approaches to insure privacy and transparency of data use.
- Track developments in foreign health care systems, looking for useful technologies and also technology failures.

3.2 Concerns about "Snake Oil"

There is enormous money to be made in the inevitable onset of internet-delivered diagnostics and care. This will promote the entry of all sorts of companies into this space, both meritorious and not. For instance, there are already many paid online services available that will help people interpret their Ancestry.com genome data or offer weekly health reports based on 23andMe [80]. How are Americans to know which of these they can trust?

To illustrate the problem, consider the gene for Methylenetetrahydrofolate reductase (MTHFR). MTHFR plays a role in the folate cycle in humans, which is a key step in producing purines, building blocks of DNA. This well studied gene has also been associated with numerous ills including contributing to plaque formation by damaging arterial walls and increasing the risks of clot formation.

This sets up a precarious situation. Here is a genetic variant that millions have been tested for (often inadvertently, via ancestry genetic testing) that sounds pretty scary. Through simple Google search you can find sites that present accessible information about the MTHFR gene and its effects. However, the purpose of some of these websites is to guide visitors to expensive consultations purported to help them determine the best course of treatment for their MTHFR genotype. According to an exposé, [81] in one case the consultation costs $3000, and generally results in a prescription of exotic vitamin combinations only available through the site. If these pills contain B12 and folates, they might be an effective (but unreasonably expensive) remedy for the widespread MTHFR malady. With approximately 10-20% of the population having the

MTHFR mutation, there is a large pool of individuals who could be victimized by such 'services'.

We easily imagine similar situations emerging with AI applications. Consider an example for skin cancer detection. Computer-aided automated skin cancer detection was demonstrated on biopsy-proven clinical images and tested against 21 dermatologists [82]. In parallel, online services already exist for remote dermatologist diagnosis of online-submitted images of skin moles [83]. We could imagine a scam service asking patients to submit self-taken skin mole images along with payment for an automated "quack" diagnosis in return, one that did not actually use any validated classification scheme. More likely, the methods used by any one company may be hidden or obscure, meaning the user has no way to judge the soundness of the company. Skinvision [83] is a new company based in the Netherlands for "skin cancer melanoma detection" where users download an app, take a picture for the app, and receive analysis, diagnosis, and tracking by the app. Very little information is provided about the methods used, other than Skinvision uses the "mathematical theory of 'fractal geometry' for medical imaging to diagnose suspected melanoma" and that "the algorithm has been developed and tested in cooperation with dermatologists and checks for irregularities in color, texture, and shape of the lesion", although elsewhere in fine print the website acknowledges that "our solution is not a diagnostic device".

The threat of bogus web sites is diminished in the presence of trusted, credible resources. For example, information about cancer symptoms, diagnosis, and treatment options are provided by the American Cancer Society [84] and the Mayo Clinic [85]. WebMD was founded in 1996 to be such a resource. It has been tremendously successful as a provider of news and information related to human health and well-being. WebMD has more monthly unique visitors than any other private or government health care website, making it the leading health publisher in the United States. In the fourth quarter of 2016, WebMD recorded an average of 179.5 million unique users per month, and 3.63 billion page views per quarter [86].

Similar trusted web sites, developed and supported by competent experts, will be needed to help consumers navigate the proliferation of AI applications in health at both the individual and public health levels. Addressing this inevitable train wreck before it happens leads us to the following finding and recommendation.

Finding: There is potential for the proliferation of misinformation that could cause harm or impede the adoption of AI applications for health. Websites, apps, and companies have already emerged that appear questionable based on information available.

Recommendation: To guard against the proliferation of misinformation in this emerging field, support the engagement of learned bodies to encourage and endorse best practices and deployment of AI applications in health.

3.3 Concerns about Inequity

One issue with focusing on smart devices as the future gateway to health information, monitoring individual health, and remote health care is equity access to the information and

services. There are two parts to this equity issue. First is access to a smart platform such as a smartphone, and the second is access to the useful apps and devices.

Recent data (2016) from Pew Research Center indicate 95% of Americans own a cellphone of some kind and 77% actually own a smartphone [87]. This is a sharp charge from 2011 where 83% of Americans owned a cellphone of some kind and only 35% owned a smartphone [88].

The distributions of cellphone and smartphone ownership for 2016 are given in Table 3. Ownership rates among men and women and different racial groups are quite similar. There are some differences across income and education levels, but overall rates are higher for every group than in 2011. Of particular note is the variation among age groups. There are differences in age groups with 100% of 18–29 year olds owning some sort of a cell phone and 92% owning a smartphone, compared to 95% and 49% in 2011. The elderly have lower adoption with only 80% of the 65+ age group owning a cellphone and only 42% owning a smartphone in 2016. Nonetheless, adoption is growing for the elderly as only 46% owned any sort of cellphone in 2011 and only 11% owned a smartphone.

A Pew report from 2015 [89] pointed out that a growing number of Americans are using their smartphone as their primary broadband access device, with the largest groups with the behavior being non-white with lower education and income levels. It was also noted that more than half of all smartphone owners used their devices to access health information and do on-line banking. These statistics imply that smartphone adoption could soon be ubiquitous making it a platform of interest for the widespread access to health apps, many of which will be AI enabled.

However, there may be inequitable differences among these demographic groups in the means, education levels, or physical capabilities needed to purchase, understand or use the tools. This may be a role for public health and other government subsidy programs. As various AI applications become useful and also demonstrate a cost savings in the management of health and health care, these products could be provided to qualifying individuals through programs similar to Supplemental Nutrition Assistance (SNAP) [90], Temporary Assistance for Needy Families (TANF) cash assistance [91], and other medical assistance programs.

Table 3: PEW Research on Cellphone and Smartphone Usage 2016.
http://www.pewinternet.org/2011/07/11/smartphone-adoption-and-usage/

% of U.S. adults who own the following devices

	Any cellphone	Smartphone	Cellphone, but not smartphone
Total	95%	77%	18%
Men	96%	78%	18%
Women	94%	75%	19%
White	94%	77%	17%
Black	94%	72%	23%
Hispanic	98%	75%	23%
Ages 18-29	100%	92%	8%
30-49	99%	88%	11%
50-64	97%	74%	23%
65+	80%	42%	38%
Less than high school graduate	92%	54%	39%
High school graduate	92%	69%	23%
Some college	96%	80%	16%
College graduate	97%	89%	8%
Less than $30,000	92%	64%	29%
$30,000-$49,999	95%	74%	21%
$50,000-$74,999	96%	83%	13%
$75,000+	99%	93%	6%
Urban	95%	77%	17%
Suburban	96%	79%	16%
Rural	94%	67%	27%

Source: Survey conducted Sept. 29-Nov. 6, 2016.

PEW RESEARCH CENTER

Source: Reproduced from Pew Research Institute "Mobile Fact Sheet," January 12, 2017 [89].

4 ADVANCING AI ALGORITHM DEVELOPMENT

The examples given earlier in this report demonstrate the value in high quality training data for the development of AI applications. This point needs to be significantly underscored if the potential of AI for health is to be realized. There are many issues of being able to develop and share datasets based on health and health care data. First of all is the cost of creating labeled data of high quality – this may require hiring expert image analysts, or providing additional testing (such as biopsies) to create labels. In the large quantities needed for AI data sets, this can be expensive. Other issues include the fact that the data may contain sensitive information about real people. Additionally, the culture in biology and life sciences, to include health care, is to closely protect one's data that has typically been expensive and time intensive to collect. Few incentives exist to share this data until the primary researchers have squeezed all the results they can from their data. From an industry perspective, protecting intellectual property around AI applications currently does not lend itself to sharing data. These issues will be discussed throughout the remainder of this report. But first, there are overarching findings and recommendations associated with expanding the availability and access to comprehensive databases of health data.

Findings:

- The availability of and access to high quality data is critical in the development and ultimate implementation of AI applications. The existence of some such data has already proven its value in providing opportunities for the development of AI applications in medical imaging.
- Laudable goals for AI tools include accelerating the discovery of novel disease correlations and helping match people to the best treatments based on their specific health, life-experiences, and genetic profile. Definition and integration of the data sets required to develop such AI tools is a major challenge.

Recommendations:

- Support the development of and access to research databases of labeled and unlabeled health data for the development of AI applications in health.
- Support investigations into how to incentivize the sharing of health data, and new paradigms for data ownership.
- Support the assessment of AI algorithms using data labeled at levels that exceed standard assessment, for instance the use of outputs from the next stage of diagnostics (e.g., use of biopsy results to label dermatological images).
- Support research to characterize the tradeoffs between data quality, information content (complexity and diversity) and sample size, with the goal of enabling quantitative prediction of the quantity and quality of data needed to support a given AI application.
- Identify and develop strategies to fill important data gaps for health.

4.1 Crowdsourcing

The examples in the sections above demonstrate the value of large corpuses of labeled data in developing AI applications for health. Several of the examples made use of the highly successful

ImageNet competition [92] results to initiate their deep learning algorithms. These types of successes can be replicated if the broader research community, and the public, becomes more readily engaged in 1) the curation, analysis, and understanding of health and health care data as a novel and important new player in the "Big Data" field, 2) the development of improved and health care-tailored AI algorithms.

The two specific vehicles for this are crowdsourcing via online technical competitions and citizen science via online public engagement activities. The competitions force the creation of good data by the host and bring in people to develop new AI applications. The citizen science approach is a promising path to creating the (labeled) data.

4.1.1 Crowdsourcing competitions

Crowdsourcing is becoming a growing success for AI in Health algorithm development via online competitions. The crowdsourcing competitions are able to engage top data scientists and programmers who are not health care domain experts. Some competitions have seen thousands of participants while others are purposefully limited to a select group of dozens of invited participants chosen from a pool of prior successful competitors. The competition duration can be from a few days up to months. Monetary prizes are often given out to the top winners, and from typical totals as small as $10K through $100K even up to $1M. Leaderboards serve to motivate the competitors and discussion boards promote sharing of information sometimes leading up to collaborations in the competitions themselves. Contributed code is usually made public and serves both as a benchmark and to move the field forward. Crowdsourcing is motivated by the fact that while there are numerous and varied strategies that can be applied to any predictive modeling task, it is impossible to know at the outset which technique will be most effective.

Kaggle [93] is a popular, successful, and leading online data science competition host with over 760,000 members around the world and hundreds of completed competitions in a variety of fields. One of the most successful health competitions to date was held in 2016-2017, run by Kaggle and Data Science Bowl [94] with support from over two dozen organizations. Competitors used anonymized, high-resolution lung scans from hundreds of patients provided by the National Cancer Institute. Competitors applied deep learning AI methods to find solutions that can improve lung cancer screening technology. The participants created algorithms that can accurately determine when lesions in the lungs are cancerous with the motivation to dramatically decrease the false positive rate of current low-dose CT technology.

A competition example which highlights the curated data, is a recent Kaggle competition posted by the Memorial Sloan Kettering Cancer Center (MSKCC) [95]:

> For this contest "we need your help to take personalized medicine to its full potential. Once sequenced, a cancer tumor can have thousands of genetic mutations. But the challenge is distinguishing the mutations that contribute to tumor growth (drivers) from the neutral mutations (passengers). Currently this interpretation of genetic mutations is being done manually. This is a very time-consuming task where a clinical pathologist has to manually review and classify every single genetic mutation based on evidence from text-based clinical literature. For this competition MSKCC is making available an expert-annotated knowledge base where world-class researchers and oncologists have manually

annotated thousands of mutations. We need your help to develop a Machine Learning algorithm that, using this knowledge base as a baseline, automatically classifies genetic variations."

Crowdsourcing competitions for health are actually quite broad in topic and type of data source, even though medical imaging competitions have received the most attention. The competitions have included: assessing cardiac function as a key indicator of heart disease using MRIs [96]; seizure prediction based on intracranial EEG recordings [97]; and predicting actual activity from sensor data fusion for monitoring the elderly [98].

The main challenge for crowdsourcing efforts is the required large, well-labeled public or semi-public datasets in order to get the proper community involvement. Nobody could have predicted the benefits that derived from the careful creation of the ImageNet database [99], and we conjecture that the expanded creation of similar high-quality databases for health data could also lead to unforeseen advances. Additionally, the competitions could help accelerate new AI algorithm development, and an understanding of the biases and errors implicit to health data. There are already websites that are perfectly positioned to host such competitions, such as Kaggle [94].

A second major challenge to online competitions is in how to move competition-resulting software to clinical tools. At least one competition is already addressing this [100]. The Booz Allen Hamilton Data Science Bowl challenges participants to develop clinic-ready software, based on AI models, that can be used to detect early stages of lung cancer. One unique feature of this competition over some of the data-analysis-only competitions is that it will require creating teams of data scientists, software engineers, designers, and clinicians working together.

A third major challenge is that competitions, thus far, are mostly limited to image recognition/computer vision. Complex and heterogeneous datasets, noisy medical datasets are not yet addressed (see Section 6.1). Yet, in contrast, large data-gathering projects (like the *All of Us* Research Program, see Section 5.1) will be gathering large amounts of data of unknown and varying quality. Data fusion, the integration of multiple heterogeneous data sources, is an example suggestion for a new type of contest. Health data has features that make it unique from all other types of data, and contests could be designed to facilitate an understanding of what those features are, and how to correct for them. One example could be giving participants access to both electronic health records and billing data for a collection of patients that would require creative data linkage strategies to develop matched data and corresponding analyses.

4.1.2 Citizen science

In cases where datasets are noisy, limited in number or scope, or otherwise not yet amenable to robust and autonomous computational processing, citizen science may be an approach to develop data sources. Citizen Science is a form of collaboration where members of the public participate in scientific research, a paradigm where the activities of an engaged public are intertwined with professional scientific research. Zooinverse [101] hosts a vast array of citizen science projects.

One recent medical-related citizen science example is on bacterial resistance to antibiotics [102]. The researchers aim is to use a genetics-based approach to improve the treatment of tuberculosis (TB) and reduce the spread of bacterial resistance to antibiotics. Understanding which changes in

the bacterial genome (mutations) lead to antibiotic resistance enables the identification of which antibiotics can be used to treat a particular patient, in a week rather than current practice of a month. The researchers are collecting over 100,000 TB infection samples from around the world. Each sample will have its genome sequenced and for each sample the volunteers will help determine the sample's sensitivity to a range of antibiotics based on images of plates with different antibiotic doses where bacteria have been allowed to grow. Volunteers are simply asked to view images of plates and identify whether or not bacteria are growing. Here the hosts aim to compare and combine inputs by the volunteers, expert opinion, and computer processing of images to get an accurate assessment of each plate.

Public forums are needed to engage citizen scientists in helping find new discoveries that will benefit health and wellness. The creation of discovery-based challenges that build on crowdsourcing and citizen science lessons learned from other areas of science and engineering, e.g., Galaxy Zoo [103] or National Weather Service's Cooperative Observing program [104] leads to the following finding and recommendation.

Finding: AI competitions have already demonstrated their value in 1) encouraging the creation of large corpuses of data for broad use, and 2) demonstrating of the capabilities of AI in health, when provided data that are curated into a well labeled (namely high information content) format.

Recommendation: Embrace the "crowdsourcing" movement that fueled this recent revolution in AI.

- Support competitions created to advance our understanding of the nature of health and health care data.
- Share data in public forums to engage scientists in helping find new discoveries that will benefit health.

4.2 Deep Learning with Unlabeled Data

The most intuitive training process for deep learning involves labeled data as in the examples presented in Section 2. However, various techniques are in fact known for utilizing deep learning in contexts where labeled data are unavailable or impractical. Three worth mentioning are reinforcement learning, auto-encoders, and generative adversarial networks. Each of these in some sense creates a training set.

In reinforcement learning, typically, the input to a deep neural net is a complex image or feature set, while the output is supposed to be a set of policies that maximize a score. To train the net, we must provide an algorithm that evaluates and scores the output policies. In an early successful example, the input images were the continuous screenshots of the Atari game Pong, while the output policies were joystick moves that the computer makes. The score in this case is simply the Atari score displayed on the screen (which, in effect, the computer learns to read). So, in this example, as a substitute for labeled data, the computer simply plays the game many millions of times.

Auto-encoders provide a methodology for autonomously learning to summarize a complex input signal (an image, for example) in terms of a small(er) number of features. An auto-encoder

consists of two deep learning layers, back-to-back. The encoder side gradually necks down (through many layers) to a feature layer. The decoder side starts with these features and attempts to re-synthesize the original complex input. The auto-encode is trained using (as before) a lot of unlabeled – real or simulated – images. The training goal is to maximize the fidelity of the output image, as compared to the input image, after passing through the narrow neck of the feature vector. This is compute-intensive and done off-line. Once it is trained, we can use the auto-encoder in real time in two ways: It can produce feature vector output from complex input. And, it can synthesize a complex output from inputted feature vectors. As a notional example in health care, we might use an auto-encoder to learn how to do differential diagnosis, where the small number of features at the narrow neck are the diagnoses, while the complex input (and synthesized output) are the continuous signals of heart monitors, etc.

In generative adversarial networks (or other similar adversarial networks) we allow two deep neural nets to play against each other, so that each becomes good at countering the behavior of the other. In a symmetric example, the two networks could be players in a two-person game (e.g., Go). Playing against itself (actually an independent copy of itself), the network learns to become a good Go player. In asymmetric examples, one network is often termed generative, the other discriminative. The job of the generative network is to generate images (say) that "fool" the discriminative network, while the job of the discriminative network is not to be fooled. As the two networks play against each other, both networks improve, leading to a trained discriminative network. One new application of the use of these networks in health is to generate simulated or synthetic electronic health records that carry the same properties as the original data, thus purportedly preserving the privacy of the subjects [105].

The use of AI methods for unlabeled data in health applications has had little attention to date.

Finding: Techniques for learning from unlabelled data could be helpful in addressing the issues with using data from a diverse set of sources, such as those discussed in Section 3.

Recommendation: Develop automated curation approaches for broadly based data collections to format them for AI tools—e.g., as with well labeled imagery.

5 LARGE SCALE HEALTH DATA

An aspirational goal for health and health care is to amass large datasets (labeled and unlabeled) and systematically curated health data so that novel disease correlations can be identified, and people can be matched to the best treatments based on their specific health, life-experiences, and genetic profile. AI holds the promise of integrating all of these data sources to develop medical breakthroughs and new insights on individual health and public health. However, major limiting factors will be the availability and accessibility of high quality data, and the ability of AI algorithms to function effectively and reliability on the complex data streams.

It is estimated that 60% of premature deaths [106] are accounted for by social circumstances, environmental exposures, and behavioral patterns [107]. These three areas are a combination of experiences throughout our life based on where we were born, live, learn, work, and play. Frequently coined the social determinants of health [108], these include economic stability, neighborhood and physical environment, education, food, community and social context, and health care system (see Figure 5) [109].

Economic Stability	Neighborhood and Physical Environment	Education	Food	Community and Social Context	Health Care System
Employment	Housing	Literacy	Hunger	Social integration	Health coverage
Income	Transportation	Language	Access to healthy options	Support systems	Provider availability
Expenses	Safety	Early childhood education		Community engagement	Provider linguistic and cultural competency
Debt	Parks	Vocational training		Discrimination	Quality of care
Medical bills	Playgrounds	Higher education			
Support	Walkability				

Figure 5: Social Determinants of Health. *Source:* Figure modified from the Kaiser Family Foundation [110].

Genetic information also must be included in this enterprise. However, it must be recognized that genetic sequencing continues to fall short in explaining many health conditions. In some cases, human diseases are easily tracked to well-characterized mutations in very specific genes. But this seems to be the exception rather than rule. Sometimes, human illnesses result from combinations of genetic mutations, and in these cases, it is much more difficult to track down the genetic underpinnings of disease. But, in addition, the forces of chance are at play, and susceptibilities are being altered by behavior (e.g., exercise, diet, smoking, etc.) and environmental exposures (e.g., environmental toxins, noise pollution, industrial chemicals).

The human genome sequence, comprised of ~3 billion DNA bases, was completely determined in 2003, almost 15 years ago now. A primary goal of the human genome sequencing project was to provide highly-accurate DNA sequence data to researchers who then would identify links

between DNA sequence variation and human disease. This was a visionary goal that was enabled by strides in inexpensive DNA sequencing. The vision was sound, as there already existed several well-proven links between specific DNA mutations and the manifestation of genetic diseases [111]. Given these past findings, many researchers expected that numerous single-nucleotide polymorphisms (SNPs; differences between the genetic code of different human individuals) could be found to strongly correlate with additional human diseases, explaining why some people have them and other people don't. It has been found surprisingly rare that human disease can be linked to specific genetic mutations [112,113]. Ultimately, health is a function of the multiple factors of genetics, behavior, and environmental exposures.

To gain a complete picture of health, data from genetics, health care delivery and outcomes, and the social determinants of health should be integrated. This could include, for example, data on exercise, addictions, diet, the reporting and sharing of family history, treatment experiences, and social consequences associated with a chronic disease, and the widespread collection of health-relevant information from wearable devices and smart technology platform apps.

5.1 Current Effort – *All of Us* Research Program

A major initiative is just beginning in the U.S. to collect a massive amount of individual health data, including social behavioral information. This is a ten year, $1.5B National Institutes of Health (NIH) Precision Medicine Initiative (PMI) project called *All of Us* Research Program [114]. The goal is to develop a 1,000,000 person-plus cohort of individuals across the country willing to share their biology, lifestyle, and environment data for the purpose of research. A soft launch for this data collection has just begun at a collection of locations across the country (see Figure 6). The initial data collected on participants opting into the study will be surveys capturing 1) basic information on medical history and lifestyle (e.g., personal habits and overall health), 2) physical measurements (e.g., blood pressure, pulse, height, weight, and hip and waist circumference), 3) biosample assessments (blood and urine samples) and, in some cases, DNA testing with additional participant consent; and 4) electronic health records data capturing most fields of the common clinical dataset [115] (e.g., demographics, health care visits, diagnoses, procedures, medications, laboratory tests, and vital signs). The electronic health record data will be collected continuously over the 10 years of the study meaning participants recruited now will have 10 years of data associated with them. There are future goals of including all parts of the electronic health record and data from wireless sensor technologies (including data from mobile/wearable devices), and also geospatial/environmental data.

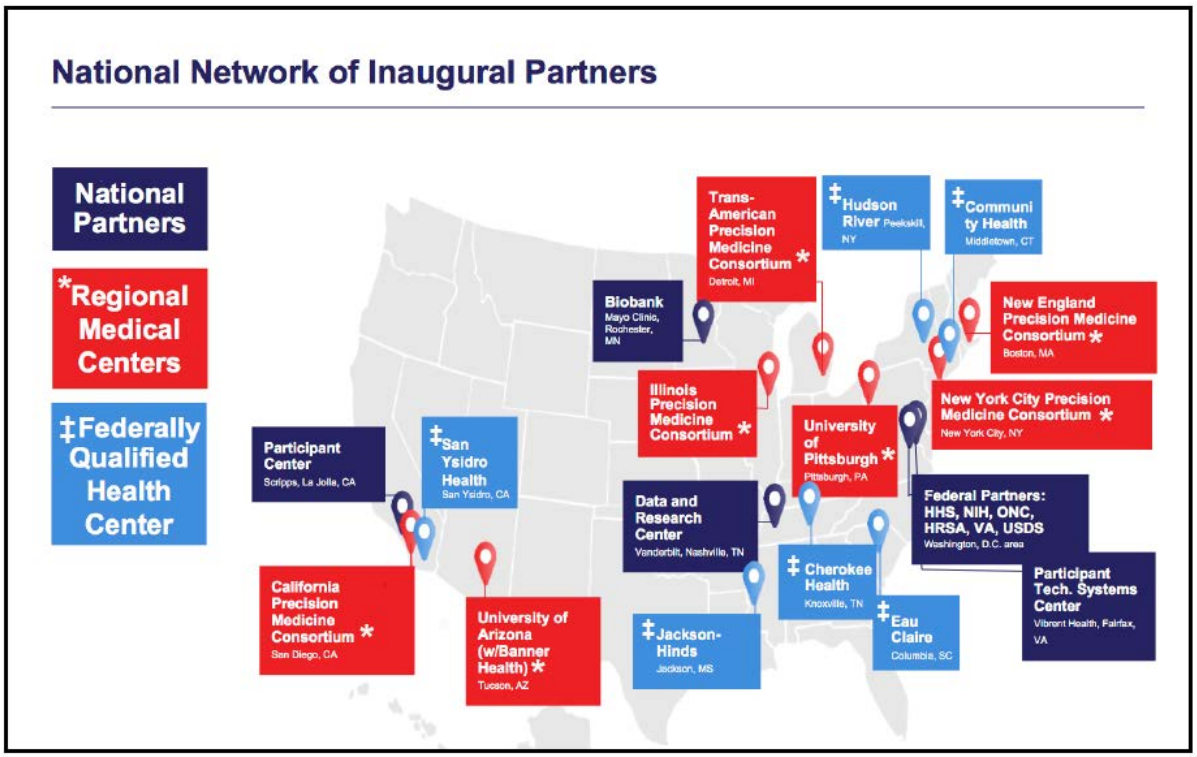

Figure 6: *All of Us* Research Program initial data collection partners. *Source:* Teresa Zayas Cabán, briefing to JASON, June 2017.

The $1.5B decade of funding for this PMI initiative has been authorized by Congress through the 21st Century Cures Act [116]. While this may appear to provide the funding needed to create the data resources envisioned by PMI, it is likely that significantly more funding may be both necessary and warranted to successfully populate the database and distribute this data to the research community and public. We conclude the *All of Us* Research Program discussion with some highlights and cautions regarding participation, privacy, and access.

All of Us Research Program Participation, Privacy, and Access

A major purpose of the *All of Us* Research Program will be to make the data available to participants, researchers, and the public.[154] This raises issues of privacy concerns. PMI has recognized from the start of this initiative that no amount of de-identification (anonymization) of the data will guarantee the privacy protection of the participants, an issue noted by JASON 2014 [7]. To build confidence around privacy, *All of Us* Research Program is laying the foundation for ensuring confidentiality, integrity and availability of the data. This is through the activities of two interagency working groups tasked with developing privacy and trust principles and data security policy principles and framework. [117]

In order to acquire the various pieces of data for the *All of Us* Research Program, privacy guarantees must be provided to the participants. Participants may not want to share all of the required information and so a mechanism must be put into place to enable participants to choose which pieces of information can be released and to whom. On the one hand, novel methods to help participants understand informed consent are being developed based on research that has

been piloted by Sage Bionetworks [118]. These methods have now become part of the *All of Us* Research Program protocol [119]. The informed consent process will allow all participants to opt-in to all or only subsets of their data being included in the study and/or shared with research studies. Participants will also be able to change their preferences and opt-out at any time. On the other hand, parallel consent authorization processes need to be implemented properly.

To pilot technology that will allow individuals to share their health data from EHRs and send it to researchers, NIH and ONC launched Sync for Science (S4S) [154]. S4S is testing several technology innovations. The first is based on the emerging OAuth2.0 standard [120]. OAuth is a standard for delegation of access to online information. It is already in common use. For example, one may want to access a web service without being a locally authorized user of that service. In some cases, the service will offer to authenticate the user through their credentials used for another site typically a large provider. If the user agrees to this, they enter their credentials for the remote site. These credentials are then sent securely via OAuth to the remote site. The remote site then authenticates the user and may also query the use for which authorization is requested. If the user accedes a token is sent to the requesting site indicating the user is valid and has given permission to the site to authenticate them. From then on when the user wants access to the web service the authentication can be mediated through the remote provider.

This approach makes it possible to create web services that can consolidate disparate pieces of data in one place. A common example is a site to let someone see their comprehensive financial position by interrogating all the financial service sites in which a user participates and providing a composite analysis of the data. In the past, this required having the user provide authentication data to the consolidation site so that it could perform remote queries. Using frameworks like OAuth the users need only authenticate themselves to the remote institutions once and establish a kind of digital contract that delegates their permission to acquire the latest financial information from the relevant financial service. Because no authentication information is provided the consolidation site can only query the limited piece of information (e.g., bank balance) and cannot access any other information. The exchanged tokens encode the level of permissions, so if a user does not want to give access to some account data, it is at their discretion not to do so. If the user decides not to share the information the permissions can always be revoked, and the tokens then become invalid.

This approach could work well, in principle, for the sharing of health data. It has many of the attributes of "privacy bundles", a concept that was introduced by JASON in their initial report on electronic health record systems [7]. JASON advocated the development of a fine-grained permission model for EHRs. Because an EHR can be organized in a tree-like directory structure it is possible to assign different levels of privacy depending on the sensitivity of the information. But rather than have patients (participants in the context of *All of Us* Research Program) determine the level of privacy for each data element in an EHR, JASON advocated the development of privacy bundles wherein the patient can elect to share collections of related data but elect not to share others depending on the requestor. For example, one's privacy policies for research projects may differ from those associated with an insurance company. Or one may allow data associated with blood tests to be released but may refuse to share data associated with mental health. The use of OAuth is very much in keeping with this idea as it puts the data owner (e.g., the *All of Us* Research Program participant) in control.

OAuth may not be the ultimate solution, but it is the one being used despite some issues. The first is security. OAuth makes use of a framework called Transport Layer Security (TLS). This is a widely-used set of protocols used to secure e-mail, web browsing etc. The problem is that there is no authoritative implementation of the protocols. There are commercial as well as open source versions. Some of these have suffered security breaches because the protocol was not implemented properly. Today there does not exist a provably secure version of TLS although there is active ongoing work by Amazon, Microsoft and others in creating one. The second issue is phishing. It would not be hard to create fake requests for OAuth tokens that look authentic. It may be necessary to have a secondary verification such as confirmation on one's cell phone to verify that a request is valid. A greater challenge lies in the concept of implementing privacy bundles. Typical EHRs are paged documents. It is not easy to decompose the pages into the atomic data elements associated with a privacy bundle. For example, suppose a user does not want to provide a urine sample. Does this mean that all data on their EHR associated with kidney function should not be accessed? Or perhaps a user declines to share any data related to mental health, this could include many SNPs, blood tests, drug prescriptions etc. How could this be achieved? Issues like these indicate the complexity of the task.

The second S4S technology is the use of blockchains as a way of validating the provenance of data. This is similar to the approach taken by DeepMind, as discussed in Section 3.1.3. This is an excellent idea and may lay the groundwork for the more general use of block chain ledgers for EHRs [121]. The S4S pilot will be using the Health Level 7 (HL7®) Fast Healthcare Interoperability Resource (FHIR®) standard [122] as a way of encoding medical data. JASON has commented positively in the past on the use of FHIR and has endorsed the approach as a way to achieve health care data interoperability [8]. Because FHIR uses a mark-up language to describe the data, it is possible to easily decompose the data so that specific elements can be tracked and reports describing the chronology of the data elements can be readily assembled. Because the *All of Us* Research Program will be highly distributed, the problem arises that various pieces of data may become out of sync with one another. This could have deleterious consequences on the validity of the research data. One way to deal with this is to have a central point of access for the data but this may lead to significant inconvenience on the part of researchers who are distributed around the country. It would also create a single point of failure should the central repository suffer an interruption of service either due to a security incident or perhaps hardware failure.

The idea of the blockchain is to create a distributed ledger where various accesses or changes in the ledger are indicated through the use of a set of interlinking hash codes (known as a Merkle tree [78]). Using this type of data structure, it is possible to verify who made a change and what change was made. It is virtually impossible to corrupt the database because a change in the data causes changes in the hash codes. While it is theoretically possible to forge data with the same hash codes, to date there are no algorithms known that can do this in a reasonable period with current computing resources. In addition, even if one could forge the data, because the database is distributed and available to everyone it is virtually impossible to insert forgeries on all the copies. There is a special concurrence algorithm that allows all users to eventually learn which blockchain is the most recent making it unnecessary for all users of the ledger to perform specific synchronization of their copies.

The use of blockchains would be particularly appropriate for health records in general. It solves the problem of trying to locate the most current records when a patient presents themselves at a

medical facility. By querying the blockchain all interactions would be recorded and the most recent data could be acquired. There are of course privacy issues here. Agreeing to the use of this technology means that a patient is tracked throughout the health care system. While this is clearly beneficial to the medical practitioner the patient must be informed, and may object to this type of EHR tracking.

Beyond the implementations of the S4S technologies for a pilot set of EHRs in the *All of Us* Research Program, little information is available about how the program will meet their privacy and trust principles and their data security policy principles.

5.2 Environmental Data –The Missing Data Stream

If AI is going to be used to build a deeper understanding of disease, one needs to ensure that the training data incorporates all of the relevant data streams including environmental exposures [123]. One type of environmental exposure that is closely monitored by epidemiologists and clinicians is pathogen exposure. However, others such as exposures to UV radiation, industrial chemicals, air pollution, and extreme noise levels, are rarely discussed in the clinical setting. It is interesting to think that someone might live right next to an industrial smokestack, yet this may never make it into their medical record because clinicians don't typically ask questions related to environmental toxicology, and these have not been incorporated into most standard health questionnaires.

Environmental exposures are broadly defined as exposure to chemicals, pathogens, noise, and energy sources (microwave, UV, ionizing radiation). For many diseases, environmental exposures play a bigger role in health outcomes than genetics. Yet, the amount of attention paid to environmental factors is a fraction of the attention that has been devoted to genetics. There are well substantiated environmental risk factors for significant diseases such as many cancers [124] and autism [125,126], yet to our knowledge these known environmental components are not being tracked as part of most health data projects. Data in all of these areas should continue to be developed, but we find that absence of the environmental exposure datastream to be particularly troubling.

5.2.1 Capturing data on toxin exposure

An understanding of variation in toxicity exposure from town to town and even from home to home is needed. Measurements would ideally be taken in each person's home. For instance, one person might have many pieces of furniture containing flame retardants [127], which have been linked to cancer, while another does not. At some level, this information can also be captured by measuring toxins in individuals' bloodstreams. One can imagine that members of some households may have higher levels of certain toxins than members of another household.

Environmental exposure data can fairly easily be captured in health data projects. For instance:

- Blood testing could include assays for common environmental toxins (e.g., lead, dioxin, etc.). Custom and commercial testing kits are available (for example [128]
- Diet-related toxins should be considered (e.g., preservatives, pesticides, plastic residues from packaging, heavy metals). Patient questionnaires could be designed to include

questions about diet-related toxins, for instance the percentage of produce purchased that is organic, number of times per week that seafood [129] is consumed, etc.
- Given the decaying infrastructure in the US [130] it would be worth monitoring water quality in people's homes, and feeding this information into meta-datasets used for AI studies generating human disease correlates. There is some evidence that, in certain places in the country, water contaminants including lead fall outside of allowable bounds [131].

Collection of parallel city-scale environmental data should be incorporated at health study sites (for instance, the cities where the *All of Us* Research Program's data collection centers are located). This can be done by city-wide sensors, as described in the next section, or by having participants wear or carry devices that can take these measurements. Important things to measure would include UV intensity, chemical components of air pollution, allergens, noise, human pathogens, construction/demolition-related pollutants (lead and asbestos), and radiation.

5.2.2 Environmental sensing at different geographic resolutions

National, state, and local (county-based) efforts exist to collect and track environmental data. The Center for Disease Control (CDC), Environmental Protection Agency (EPA), Housing and Urban Development (HUD) and the Census Bureau have environmental data at various levels of geographic aggregation such as population characteristics, air quality, housing quality, climate, asthma, cancer, poisonings, opioid and other drug use and deaths, toxic releases, proximity to transportation infrastructure. In 2009, CDC launched an on-line Environmental Public Health Tracking Network across 26 states that integrates the national data resources with environmental and health data available at the state and local levels [132]. These are important data sources to be used in building an understanding of the social determinants of health and the impact of the environment on health, but a finer level of geographic resolution in the capture of environmental exposure data likely will be needed to unravel the relationship between health status, genetics, environment and behaviors.

To better understand the spatial and temporal variability of air pollution, NASA, in collaboration with the EPA, uses satellite data to provide estimates of pollutants for pixels of the scale of tens of kilometers squared [1331]. These space-based data products are validated via comparison to EPA in-situ monitoring stations [134] and then allow for a broader spatial perspective on the state of air quality in any location. These data products are mature and operationalized, with NASA and EPA presently making a push to train air quality forecasters and managers in their use.

There are numerous academic projects underway to collect environmental data within urban settings. Advances in sensors and data technologies are enabling better measurement of environmental factors, particularly in the urban areas where some 80% of the US population resides. Indeed, there is a flourishing of academic and for-profit efforts to "measure cities" and exploit the data so derived [135]. For instance, the "Array of Things" project is a network of urban sensors placed on telephone poles around Chicago [136] and the "Sounds of New York City" project [137] aims to measure and characterize the urban noise field with high spatial and temporal granularity through stationary *in-situ* sensors.

Similarly, networks of small sensors enable high-resolution measurements of ozone and carbon monoxide in Berkeley California [138] and a citizen science project is capturing particulates in Brooklyn, New York [139] which heretofore are derived only by uncertain modeling of transport from known sources such as roads and emission stacks. Urban metagenomic studies are underway to characterize the space-time variation of the urban microbiome and light pollution in a given housing unit can be measured passively and synoptically by observations of building facades from urban vantage points. Mobile devices offer similar opportunities for environmental sensing. The fusion of such data with individual mobility tracks can provide individual exposures.

Such environmental data should be collected as part of all big-data health and health care projects. Collection of these data streams is particularly practical in projects that are based out of specific medical centers, limiting the environmental data needed to a few particular cities. Studies should also be designed to place sensors within the homes of individuals, with care for privacy concerns, so that the home-to-home variability in environmental exposure can be evaluated. This brings us to the following findings and recommendations.

Finding: AI application development requires training data, and will perform poorly when significant data streams are absent. While DNA is the blueprint for life, health outcomes are highly affected by environmental exposures and social behaviors. There is an imbalance in the effort to capture the diverse data needed for application of AI techniques to personalized medicine, with information on environmental toxicology and exposure particularly suffering.

- Techniques exist to capture individual environmental exposures, e.g., blood toxin screening, diet questionnaires.
- Techniques exist for environmental pathogen sensing.
- Technologies exist that can capture environmental exposures geographically and create environment tracking systems.

Recommendation: Support ambitious and creative collection of environmental exposure data.

- Build toxin screening (e.g., dioxin, lead) into routine blood panels, and questions about diet and environmental toxins into health questionnaires.
- Start urban sensing and tracking programs that align with the geographic areas for *All of Us* Research Program and similar projects in the future.
- Support the development of wearable devices for the sensing of environmental toxins.
- Support the development of broad-based pathogen sensing for rural and urban environments.
- Develop protocols and IT capabilities to collect and integrate the diverse data.

6 ISSUES FOR SUCCESS

If AI applications are to advance beyond supporting specific diagnostics, and significantly enable the broader health and health care activities envisioned in Chapters 3 and 5, there are significant challenges to be addressed. A key issue is the present implicit assumption that the capabilities of AI will automatically overcome the problems of large, complex and imperfect health data. The perils of this assumption, and the associated need to address it with systematic approaches to both data management and transparency in algorithm development is discussed in this section. A striking example of the issues that must be addressed is cited by Ching et al. [140]:

> "A motivating example … can be found in [a case] where a model trained to predict the likelihood of death from pneumonia assigned lower risk to patients with asthma, but only because such patients were treated as higher priority by the hospital. In the context of deep learning, understanding the basis of a model's output is particularly important as deep learning models are unusually susceptible to adversarial examples and can output confidence scores over 99.99% for samples that resemble pure noise."

The broader point is that the "predicted outcome if nothing is done" is a very different thing than "predicted outcome if the usual system does its usual thing." Both could be useful, but to misinterpret one as the other is a deadly mistake.

6.1 Plans for use of Legacy Health Records

The promise of AI is tightly coupled to the availability of relevant data. In the health domains, there is an abundance of data. Electronic health records (EHRs) are part of this. There has been growth in the adoption of "basic EHRs" (demographics, problems lists, medications, discharge summaries, lab reports, radiology tests, and diagnostic tests), but only about 40% of hospitals have comprehensive EHRs that contain clinician notes, full lab orders and reports, and decision guidelines around clinical practice and drug interactions [141]. In addition, the issue of interoperability of electronic health record systems between care settings remains elusive [142]. That has been a topic of prior JASON reports [7,8].

Moreover, the utility of EHR data can be problematic beyond issues of completeness or interoperability, because it was not collected for the purpose or under the controls of use of research studies. This raises the issue of the actual quality of the data in the EHR. If EHR data are to be used to support AI applications, understanding this quality, and how AI algorithms react given the quality issues will be important. To date, very little research has looked at this issue.

A recent striking example of these concerns arises from a study that drew on EHRs from the UK National Health Service [143]. The example drew on a very large data base (over 12 million individual records) to assess the ability to predict cardiovascular disease (CVD). The study addressed a serious issue, which is that standard assessments do poorly in predicting the patients

who eventually do have a cardiovascular event, and generate huge numbers of false positives that stymie effective follow on testing.

The potential to improve risk assessment using machine learning was assessed using electronic health records for the time period between 2005 and 2015. From the 12 million patient records, about 375,000 were suitable for use based on the requirement of complete records on 8 standard diagnostic indicators (i.e., gender, age, smoking, blood pressure, high and low-density cholesterol, body mass index, and diabetes) and no prior history of CVD. About 25,000 patients within the study group suffered an event reported as a CVD event during the 10 years of the data records. The ability of a standard risk assessment tool (ACC/ACA), and four machine learning approaches to predict which patients would have CVD events was evaluated. Three quarters of the records were used as the training case for machine learning, and other quarter were used as the test case. In addition, for the machine learning assessments, 22 more diagnostics available in the health records were added to the input data streams.

The statistical results for the standard risk tool and the two best performing machine learning algorithms are summarized in Table 4. The machine learning algorithms improve the sensitivity (true positives) by almost 5%, but increase the specificity (decrease the false positives) by less than 0.5%. Given the poor baseline, the improvement in sensitivity still leaves much to be desired in correctly identifying patients at risk. The very small improvement in specificity yields an insignificant impact on the serious issue of false positives.

Table 4**:** Comparison of the results using a standard risk assessment tool for cardiovascular disease with 8 diagnostic inputs to the two best performing machine learning algorithms (Gradient Boosting Machines and Neural Networks) using 30 diagnostic inputs, and trained on the EHR record of whether or not the patient had a CVD event during the 10 years of records. *Source:* Adapted from Weng et al 2017 [143].

Algorithms	Total CVD Cases	Cases Correct (True Positive)	Cases Incorrect (False Negative)	Sensitivity (True Positive)	Total Non-Cases	Non-Cases Correct (True Negative)	Non-Cases Incorrect (False Positive)	Specificity (True Negative)
ACC/AHA	7,404	4,643	2,761	62.7%	75,585	53,106	22,479	70.3%
ML: GBL	7,404	4,997	2,407	67.5%	75,585	53,458	22,127	70.7%
ML: NN	7,404	4,998	2,406	67.5%	75,585	53,461	22,124	70.7%

There are many possible reasons for the limited improvements due to the use of machine learning. One is simply that diagnostics used may not represent all the medical links to CVD: there is large individual variability in CVD prognosis that is not necessarily captured in EHRs or indeed for which diagnostics have not yet been identified [144]. Another possibility is that there may be errors in the data used for the training set. Finally, there may even be errors in the diagnoses of patients suffering a CVD event, which is the training standard.

Part of the basis for errors is the nature of the data in the EHRs. An excellent recent study compared cardiovascular risk factors and events from EHRs across a research network of six hospitals to data from a traditional longitudinal cardiovascular cohort study [155]. They point out that most studies looking at the quality of EHR data consider a single health care institution and compare EHR data to clinical care as the reference. In their study, EHR data was from six institutions and they compared data for the same patients derived from research methods in the cohort study, using these as the standard. The study found varying degrees of association between the data sources. Examples of the sensitivity and specificity of the health records relative to the research study measurements include:

Hypertension	sensitivity: 71.2%	specificity: 73.0%
Obesity	sensitivity: 30.9%	specificity: 97.5%
Diabetes	sensitivity: 77.5%	specificity: 95.6%

The impact on the AI algorithms of possible error rates such as these in the training sets should be formally assessed.

In any case, the results indicate the need for extreme care in using EHRs as training sets for AI, where correlations are established that may be meaningless or misleading if the training sets contain incorrect information or information with unexpected internal correlations. The outcomes of the study using UK NHS data highlighted issues with using EHRs as inputs by assessing which factors in the expanded training sets for predicting CVD had the highest weights in the machine learning determinations, as illustrated in Table 5.

The changes in the ranked risk factors for the two best performing machine learning algorithms appear almost idiosyncratic, consistent with the well-known ‘black-box’ nature of machine learning. One observation is that the rankings aren’t readily explained in terms of the relative rates that each of these factors appear in the populations that did and did not experience a CVD event. It seems likely that there are redundant indicators of cardiovascular disease in the EHR and that they are reasonably highly correlated. Transparency in reports on AI algorithm development will be enhanced by such assessment and reporting of weighting factors.

Table 5: Comparison of the top 10 risk factors in the standard ACC/AHA Tool and the most heavily weighted factors in the two best performing machine-learning algorithms. Adapted from Weng et al 2017 [145].

ACC/AHA Algorithm		Machine-learning Algorithms	
Men	**Women**	**ML: Gradient Boosting Machines**	**ML: Neural Networks**
Age	Age	Age	Atrial Fibrillation
Total Cholesterol	HDL Cholesterol	Gender	Ethnicity
HDL Cholesterol	Total Cholesterol	Ethnicity	Oral Corticosteroid Prescribed
Smoking	Smoking	Smoking	Age
Age x Total Cholesterol	Age x *HDL Cholesterol*	*HDL cholesterol*	Severe Mental Illness
Treated Systolic Blood Pressure	Age x Total Cholesterol	Triglycerides	SES: Townsend Deprivation Index
Age x Smoking	Treated Systolic Blood Pressure	Total Cholesterol	Chronic Kidney Disease
Age x *HDL Cholesterol*	Untreated Systolic Blood Pressure	HbA1c	*BMI missing*
Untreated Systolic Blood Pressure	Age x Smoking	Systolic Blood Pressure	Smoking
Diabetes	Diabetes	SES: Townsend Deprivation Index	Gender

The changes in the ranked risk factors for the two best performing machine learning algorithms appear almost idiosyncratic, consistent with the well-known 'black-box' nature of machine learning. One observation is that the rankings aren't readily explained in terms of the relative rates that each of these factors appear in the populations that did and did not experience a CVD event. It seems likely that there are redundant indicators of cardiovascular disease in the EHR and that they are reasonably highly correlated. Transparency in reports on AI algorithm development will be enhanced by such assessment and reporting of weighting factors.

There is a great deal of interest in the potential of using the vast data sets represented in electronic health records (EHRs), in combination with AI algorithms, to draw insights about disease indicators. However, while AI can perform with great accuracy when the relationship between diagnostic data and the diagnosis is well defined, when the relationship between the data and the diagnosis suffers from error, variability or difficulty in discrimination, AI algorithms also perform less well. This creates challenges for developing AI for assessments based on data from EHRs, and foreshadows the opportunities (and challenges) of supplementing EHRs with extended patient reported data (see Section 3) as well as results from new diagnostic tools [146].

Finding: Extreme care is needed in using EHRs as training sets for AI, where outputs may be useless or misleading if the training sets contain incorrect information or information with unexpected internal correlations.

6.2 Evaluation

The clinical trials, regulation, and acceptance by the medical profession, reviewed in Section 2, is only part of the story for adoption of AI applications. However, even to support the development of clinical trials and the assurance that the AI applications are legitimate, even for non-regulated applications, the technical soundness of the algorithms need to be confirmed. While AI algorithms such as deep learning can produce amazing results, work is needed to develop confidence that they will perform as required in situations where health and life are at risk. This is independent of the hope that there will be the kind of continued improvements that have occurred in image recognition or various aspects of natural language processing. The issues here are more pragmatic.

First, no matter how carefully the training data has been assembled, there is the risk that it does not closely enough match what will be encountered in real application – the process of clinical trials outlined in Section 2.2 attempts to address such concerns. Another observation is that not all errors are equally important (or unimportant). As the system is being developed one typically uses error curves or recall/precision statistics, and without special treatment these evaluate all errors as the same. An example of bad errors was with a Google Photos release [147]. It is easy to imagine similarly unexpected, but possibly life-threatening errors in health applications. Assessment of algorithms must include questioning whether the observed error rates are like the expected rates, and identifying what types of errors the algorithm makes and why.

In addition, things change over time. Even diseases change, and the diagnostic aids have to change with them. Sometimes the changes are relatively slow, as with the multi-decadal change in the kinds of pneumonia seen [148]. Sometimes new diseases pop up and require changes to previously sound diagnostic protocols. Thus, even if an application of deep learning were ideally suited to the real world when it is first released, over time the real world may drift and make a static application less and less effective. Assessment of algorithms should include understanding how they will respond, or what indicators may be observed, if the input data characteristics begin to diverge from the original training sets.

There has been recognition that guidance on reproducibility for computational methods is needed. Stodden [149] points out there are actually three important parts to consider: empirical reproducibility, computational reproducibility, and statistical reproducibility. Empirical and statistical are being readily addressed by the research community, however computational is lacking. This would include AI models and applications. A 2016 workshop held by American Association of the Advancement of Sciences (AAAS) stressed that data, code, and workflows should be available and cited [150]. They note:

> "Over the past two decades, computational methods have radically changed the ability of researchers from all areas of scholarship to process and analyze data and to simulate complex systems. But with these advances come challenges that are contributing to broader concerns over irreproducibility in the scholarly literature, among them the lack of transparency in disclosure of computational methods. Current reporting methods are often uneven, incomplete, and still evolving."

They put forward a set of Reproducibility Principles which include:

- Share data, software, workflows, and details of the computational environment that generate published findings in open trusted repositories.
- Persistent links should appear in the published article and include a permanent identifier for data, code, and digital artifacts upon which the results depend.
- To enable credit for shared digital scholarly objects, citation should be standard practice.
- To facilitate reuse, adequately document digital scholarly artifacts.
- Use Open Licensing when publishing digital scholarly objects.
- Journals should conduct a reproducibility check as part of the publication process.
- To better enable reproducibility across the scientific enterprise, funding agencies should instigate new research programs and pilot studies.

They recognize that meeting these principles will be challenging and exceptions will be necessary with human subject research and proprietary codes. This seems to be exactly the situations faced in many of the AI in health application development.

In 2016, the Association for Computing Machinery (ACM), one of the premier publishers of computational research approved a new policy to guide the review of "artifacts." [151] Artifacts are digital objects including computational models. The policy requires transparency around the creation of the models, including experiments, data, and model development. They differentiate repeatability, replication, and reproducibility and give review guidance for each. They even go so far as creating a badging system that could be placed on publications to give their pedigree with respect to these three topics.

The ACM policy and the recommendations from the AAAS workshop have not yet been broadly implemented, but these are an excellent step forward on advising the peer-review process for large scale computational models. More discussion and guidance is needed, particularly in the context of AI applications in health. This discussion should include researchers, technologists, health professionals, industry, regulators, professional societies, and users/patients. This leads to the following finding and recommendation.

Findings: Methods to insure transparency by disclosure of large scale computational models and methods in the context of scholarly reproducibility are just beginning to be developed in the scientific community.

Recommendation: Support the critical research that will ultimately enable the adoption of AI for public health, community health, and health care delivery.

a. Encourage development and adoption of transparent processes and policies to ensure reproducibility for large scale computational models.
b. To guard against the proliferation of misinformation in this emerging field, support the engagement of learned bodies to encourage and endorse best practices for deployment of AI applications in health.

7 FINDINGS AND RECOMMENDATIONS

JASON was asked how AI could shape the future of public health, community health, and health care delivery. This report argues that AI application in health could help clinicians provide the best possible care, thus making high quality health care services available to all, and could increase people's engagement in their own health.

Overall, JASON finds that AI is beginning to play a growing role in transformative changes now underway in both health and health care, in and out of the clinical setting. At present the extent of the opportunities and limitations is just being explored. However, there are significant challenges in this field that include: the acceptance of AI applications in clinical practice, initially to support diagnostics; ability to leverage the confluence of personal networked devices and AI tools; availability of quality training data from which to build and maintain AI applications in health; large-scale data collection to include missing data streams; building on the success in other domains, creating relevant AI competitions; executing and understanding the limitations of AI methods in health and health care applications.

Here we provide the JASON findings and recommendations.

1. AI Applications in Clinical Practice

Findings:

- The process of developing a new technique as an established standard of care uses the robust practice of peer-reviewed R&D, and can provide safeguards against the deceptive or poorly-validated use of AI algorithms. (Section 2.3)
- The use of AI diagnostics as replacements for established steps in medical standards of care will require far more validation than the use of such diagnostics to provide supporting information that aids in decisions. (Section 2.3)

Recommendations:

- Support work to prepare promising AI applications for the rigorous approval procedures needed for acceptance for clinical practice. Create testing and validation approaches for AI algorithms to evaluate performance of the algorithms under conditions that differ from the training set. (Section 2.3)

2. Confluence of AI and Smart Devices for Monitoring Health and Disease

Findings:

- Revolutionary changes in health and health care are already beginning in the use of smart devices to monitor individual health. Many of these developments are taking place outside of traditional diagnostic and clinical settings. (Section 3.1)
- In the future, AI and smart devices will become increasingly interdependent, including in health-related fields. On one hand, AI will be used to power many health-related mobile monitoring devices and apps. On the other hand, mobile devices will create massive datasets that, in theory, could open new possibilities in the development of AI-based health and health care tools. (Section 3.1)

Recommendations:

- Support the development of AI applications that can enhance the performance of new mobile monitoring devices and apps. (Section 3.1)
- Develop data infrastructure to capture and integrate data generated from smart devices to support AI applications. (Section 3.1)
- Require that development include approaches to insure privacy and transparency of data use. (Section 3.1)
- Track developments in foreign health care systems, looking for useful technologies and also technology failures. (Section 3.1)

3. Create Comprehensive Training Databases of Health Data for AI Tool Development

Findings:

- The availability of and access to high quality data is critical in the development and ultimate implementation of AI applications in. (Section 4)
    - AI algorithms based on high quality training sets have already demonstrated performance for medical image analysis at the level of the medical capability that is captured in their training data. (Section 2.1)
    - AI algorithms cannot be expected to perform at a higher level than their training data, but should deliver the same standard of performance consistently for data within the training space. (Section 2.1)
- Laudable goals for AI tools include accelerating the discovery of novel disease correlations and helping match people to the best treatments based on their specific health, life-experiences, and genetic profile. Definition and integration of the data sets required to develop such AI tools is a major challenge. (Section 4)
- Extreme care is needed in using electronic health records (EHRs) as training sets for AI, where outputs may be useless or misleading if the training sets contain incorrect information or information with unexpected internal correlations. (Section 6.1)
- Techniques for learning from unlabelled data could be helpful in addressing the issues with using data from a diverse set of sources. (Section 4.2)

Recommendations:

- Support the development of and access to research databases of labeled and unlabeled health data for the development of AI applications in health. (Section 4)
- Support investigations into how to incentivize the sharing of health data, and new paradigms for data ownership. (Section 4)
- Support the assessment of AI algorithms trained with data labeled at levels that significantly exceed standard assessment, for instance the use of outputs from the next stage of diagnostics (e.g., use of biopsy results to label dermatological images). Section 2.1)
- Support research to characterize the tradeoffs between data quality, information content (complexity and diversity) and sample size, with the goal of enabling quantitative prediction of the quantity and quality of data needed to support a given AI application. (Section 4)
- Identify and develop strategies to fill important data gaps for health. (Section 4)
- Develop automated curation approaches for broadly based data collections to format them for AI tools, e.g., as with well labeled imagery. (Section 4.2)

4. Fill in Critical Missing Data Gaps

Findings:

- AI application development requires training data, and will perform poorly when significant data streams are absent. While DNA is the blueprint for life, health outcomes are highly affected by environmental exposures and social behaviors. There is an imbalance in the effort to capture the diverse data needed for application of AI techniques to precision medicine, with information on environmental toxicology and exposure particularly suffering: (Section 5.2.2)
    - Techniques exist to capture individual environmental exposures, e.g., blood toxin screening, diet questionnaires.
    - Techniques exist for environmental pathogen sensing.
    - Technologies exist that can capture environmental exposures geographically and create environment tracking systems.

Recommendations:

- Support ambitious and creative collection of environmental exposure data: (Section 5.2.2)
    - Build toxin screening (e.g., dioxin, lead) into routine blood panels, and questions about diet and environmental toxins into health questionnaires.
    - Start urban sensing and tracking programs that align with the geographic areas for the *All of Us* Research Program and similar projects in the future.
    - Support the development of wearable devices for the sensing of environmental toxins.
    - Support the development of broad-based pathogen sensing for rural and urban environments.
    - Develop protocols and IT capabilities to collect and integrate the diverse data.

5. Embrace the Crowdsourcing Movement to Support AI development and Data Generation

Finding: AI competitions have already demonstrated their value in 1) encouraging the creation of large corpuses of data for broad use, and 2) demonstrating the capabilities of AI in health, when provided data that are curated into a well labeled (namely high information content) format. (Section 4.12)

Recommendations:

- Support competitions created to advance our understanding of the nature of health and health care data. (Section 4.12)
- Share data in public forums to engage scientists in finding new discoveries that will benefit health. (Section 4.12)

6. Understand the Limitations of AI Methods in Health and Health Care Applications

Findings:

- There is potential for the proliferation of misinformation that could cause harm or impede the adoption of AI applications for health. Websites, Apps, and companies have already emerged that appear questionable based on information available. (Section 3.2)

- Methods to insure transparency in disclosure of large scale computational models and methods in the context of scholarly reproducibility are just beginning to be developed in the scientific community. (Section 6.2)

Recommendations:

- Support the development of critical safeguards that are essential to enable the adoption of AI for public health, community health, and health care delivery:
    - Encourage development and adoption of transparent processes and policies to ensure reproducibility for large scale computational models. (Section 6.2)
    - To guard against the proliferation of misinformation in this emerging field, support the engagement of learned bodies to encourage and endorse best practices for deployment of AI applications in health. (Sections 3.2 and 6.2)

8 EPILOGUE

We conclude with some thoughts about human perception versus digital data in a timeline further out than has been covered in this report. One of the major obstacles to be overcome in making health and health-care information useful is the gap between human cognition and digital data. Information concerning an individual patient is mostly obtained in forms designed to be accessible to medical personnel. Typical data may consist of X-ray or MRI or ultrasound pictures of the patient, visual records of heart or lung function varying with time, or verbal descriptions of the patient as seen by a nurse or a doctor. On the other hand, when data are stored in information systems and used, in medical research or to develop treatment guidelines, it is often reduced to statistical information which is predominantly digital. The conversion of analog input into digital output is a burdensome task, and may result in a loss of significant information that would have been helpful to the user.

When we consider the design of future health care information systems that might be more user-friendly both to providers and to users of the information, two questions need to be answered, one concerned with computer-science and one concerned with fundamental biology. The computer-science question is, whether an entire medical database can be created and used with the data maintained in a form accessible to human cognition, avoiding the cumbersome and costly translation from analog to digital. The fundamental biology question is whether the natural coding of information in a human brain is basically analog and not digital. We do not claim to know the answers to these questions, but we are inclined to guess that the answers to both will be affirmative.

Two known facts support the affirmative answers. One known fact is a mathematical theorem proved by Marian Pour-El and Ian Richards in 1978 [152]. The theorem says that analog computing is in a precise mathematical sense more powerful than digital computing. Pour-El and Richards display a number that is computable with a simple analog device but not computable with any digital device as defined by Alan Turing in his famous paper, ``On Computable Numbers'' in 1937 [153]. Their discovery gives us reason to hope that a new generation of computers operating as analog devices may give us databases more user-friendly to us than our present-day digital databases. The second fact, supporting the view that the human brain operates as an analog device, is our subjective experience of perception and memory. We experience the visual operation of our brains as a rapid and effortless scanning of pictures moving in space and time. To our subjective view, the brain appears to be primarily a device for the direct comparison of images. We see the images as whole scenes with shape and style, not as collections of pixels. Our perception of continuously moving images does not prove that our brain is an analog device, but it makes this a plausible hypothesis.

This JASON report is concerned with practical problems of management of health and health care information with a medium-range perspective, looking ahead only about thirty years. It is unlikely that a fundamental shift of our computing technology from digital to analog could occur within a time as short as thirty years. But if we look ahead further, for fifty or a hundred years into the future, such a fundamental shift becomes possible or even likely. While making plans for the near-term future, it would be wise to keep in mind the more adventurous possibilities that new discoveries in biology and neurology are likely to bring later. One possibility to be taken seriously is a complete change of databases from digital to analog data, making them more user-friendly to expert and non-expert users.

APPENDIX - Statement of Work

The overarching issue is to study how Artificial Intelligence (AI) applied to large sets of complex data could be used to improve decision making and action in health applications. JASON 2016 defined AI as intelligence exhibited by machines and that encompasses areas of R&D practiced by computational Computer Vision, Natural Language Processing (NLP), Robotics (including Human-Robot Interactions), Search and Planning, Multi-agent Systems, Social Media Analysis (including Crowdsourcing), and Knowledge Representation and Reasoning (KRR). Advanced statistics and the field of Machine Learning (ML) are the foundational basis for AI.

Understanding the full impact that artificial intelligence can have on health and health care is important.

The Agency for Healthcare Research and Quality (AHRQ), the Office of the National Coordinator for Health IT (ONC), and the Robert Wood Johnson Foundation are asking JASON to consider the following questions about how artificial intelligence could shape the future of public health, community health, and health care delivery from a personal level to a system level, and provide their insight. Understanding the opportunities and considerations can better prepare and inform developers and policy makers and promote the general welfare of health care consumers and the public.

Scope:
The following questions will be posed to the JASON group for further refinement in collaboration with all stakeholders.

1. AI Opportunities:
    a. Ways to Improve Health and Health Care
        i. In what ways might artificial intelligence advance efforts to improve individual health, health care (care of individuals), and community health (health status of sub-populations)?
        ii. What evidence exists regarding artificial intelligence's relevance for health, health care, and community health? What is the demonstrated state-of-the art in these areas?
        iii. What are the most high-value areas (example, reducing the cost of expensive treatments, prevention of mortality or morbidity in disproportionately affected populations, improvement in productivity due to better health, or focusing on risk mitigation where the impacted population is large) where artificial intelligence could be focused to contribute quickly and efficiently?
        iv. How can the benefits of artificial intelligence applications be defined and assessed?
2. AI Considerations:
    a. Technical
        i. What are the considerations for the data sources needed to support the development of artificial intelligence programs for health and health care?

For example, what is the needed data quality, breadth, and depth necessary to support the deployment of appropriate artificial intelligence technology for health and health care?

ii. How does research in computational, statistical, and data sciences need to advance in order for these technologies to reach their fullest potential?

iii. What technology barriers may arise in the technology adoption associated with artificial intelligence for health and health care? How can these be mitigated?

b. Ethical/Legal

i. What are the potential unintended consequences, including real or perceived dangers, of artificial intelligence focused on improving health and health care?

ii. What are the potential risks of artificial intelligence inadvertently exacerbating health inequalities?

c. Workforce

i. What workforce changes may be needed to ensure effective broad-based adoption of data-rich artificial intelligence applications?

3. AI Implementation

a. Are there relevant projects of AI in individual health, community health and health care that currently demonstrate the potential value of AI and feasibility of scale-up?

b. Depending on relevant projects to learn more, have other industries successfully utilized AI which might translate well to individual health, community health, and health care? Are there similar or dissimilar facets of implementation barriers that exist for AI in individual health, community health, and health care?

REFERENCES

1. http://www.image-net.org/challenges/LSVRC/2017/index.php

2. https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html

3. Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., et al. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. Jama, 316(22), 2402. http://doi.org/10.1001/jama.2016.17216

4. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancerwith deep neural networks. Nature, 542(7639), 115–118. http://doi.org/10.1038/nature21056

5. https://medium.com/machine-intelligence-report/data-not-algorithms-is-key-to-machine-learning-success-69c6c4b79f33

6. http://www.datasciencecentral.com/profiles/blogs/10-great-healthcare-data-sets

7. JASON 2013, *A Robust Health Data Infrastructure.* JSR-13-Task-007.

8. JASON 2014, *Data for Individual Health.* JSR-14-Task-007.

9. https://www.cbinsights.com/research/artificial-intelligence-startups-healthcare/

10. JASON 2017, *Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD*. JSR-16-Task-003.

11. https://contently.com/strategist/2017/05/23/artificial-intelligence-hype-cycle-5-stats/

12. K. Davis, K. Stremikis, C. Schoen, and D. Squires, *Mirror, Mirror on the Wall, 2014 Update: How the U.S. Health Care System Compares Internationally,* The Commonwealth Fund, June 2014.

13. http://www.pewresearch.org/fact-tank/2017/01/12/evolution-of-technology/

14. http://www.pewinternet.org/fact-sheet/mobile/

15. For instance DirectDerm: https://www.directderm.com/, and PlushCare: https://techcrunch.com/2016/11/03/plushcare-nabs-8m-series-a-to-prove-telehealth-can-go-mainstream/

16. Translating Artificial Intelligence into Clinical Care, Andrew L. Beam, Isaac S. Kohane, JAMA 316, 2368, 2016

17. Opportunities and Obstacles for Deep Learning in Biology and Medicine, CS Greene *et al.*, bioRxiv preprint first posted online May. 28, 2017; doi: http://dx.doi.org/10.1101/142760.

18. The Parable of Google Flu: Traps in Big Data Analysis, David Lazer, Ryan Kennedy, Gary King, Alessandro Vespignani, SCIENCE 343, 1203, 2014

19. 2010: Retinal Imaging and Image Analysis, Michael D. Abràmoff, Mona K. Garvin, Milan Sonka, IEEE Trans Med Imaging. 2010 January 1; 3: 169–208. doi:10.1109/RBME.2010.2084567.

20. EYEPACS LLC PHOTOGRAPHER MANUAL Downloaded June 2017: https://www.eyepacs.org/photographer/protocol.jsp#image_right

21. 2016: Retinal Imaging Techniques for Diabetic Retinopathy Screening, James Kang Hao Goh, Carol Y. Cheung, Shaun Sebastian Sim, Pok Chien Tan, Gavin Siew Wei Tan, and Tien Yin Wong, Journal of Diabetes Science and Technology 2016, Vol. 10(2) 282–294

22. 2016: Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs, Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, TomMadams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, Dale R.Webster, JAMA. 2016;316(22):2402-2410. doi:10.1001/jama.2016.17216

23. images were retrospectively obtained from EyePACS in the United States and 3 eye hospitals in India

24. ibid Gulshan et al, Abramoff et al

25. 2017: A Portable, Inexpensive, Nonmydriatic Fundus Camera Based on the Raspberry Pi® Computer, Bailey Y. Shen and Shizuo Mukai, Journal of Ophthalmology, Volume 2017, Article ID 4526243, 5 pages, https://doi.org/10.1155/2017/4526243

26. Viewed June 2017: Peek System Population Vision Screening: https://www.peekvision.org/

27. For instance DirectDerm: https://www.directderm.com/ and PlushCare: https://www.plushcare.com

28. Karen J.Wernli, Nora B. Henrikson, Caitlin C. Morrison, Matthew Nguyen, Gaia Pocobelli, Paula R. Blasi, Screening for Skin Cancer in Adults Updated Evidence Report and Systematic Review for the US Preventive Services Task Force, JAMA. 2016;316(4):436-447. doi:10.1001/jama.2016.5415

29. Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature, 42, 115, 2017, doi:10.1038/nature21056

30. Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones,et al, Opportunities and obstacles for deep learning in biology and medicine, doi: http://dx.doi.org/10.1101/142760.

31. M.R. Patel, et al., Low Diagnostic Yield of Elective Coronary Angiography, The New England Journal of Medicine 362, 886 (2011).

32. N.H. Pijls, P. van Schaardenburgh, G. Manoharan, et al. Percutaneous cornonary intervention of functionally nonsignificant stenosis: 5-year follow-up of the DEFER Study. J. Am Coll. Cardio 49, 2105 (2007).

33. Nørgaard BL et al., Diagnostic performance of noninvasive fractional flow reserve derived from coronary computed tomography angiography in suspected coronary artery disease: the NXT trial, J Am Coll Cardiol. 2014 Apr 1;63(12):1145-55.

34. Charles A. Taylor, Timothy A. Fonte James K. Min, Computational Fluid Dynamics Applied to Cardiac Computed Tomography for Noninvasive Quantification of Fractional Flow Reserve, Journal of the American College of Cardiology, Vol. 61, No. 22, 2013

35. ibid, Taylor, 2013

36. Accessed July 2017, Company Overiew of HeartFlow, Inc. https://www.bloomberg.com/research/stocks/private/snapshot.asp?privcapId=102389812

37. Accessed July 2017, https://www.accessdata.fda.gov/cdrh_docs/pdf16/K161772.pdf

38. Douglas et al. Clinical outcomes of fractional flow reserve by computed tomographic angiography-guided diagnostic strategies vs. usual care in patients with suspected coronary artery disease: the prospective longitudinal trial of FFR_{CT}: outcome and resource impacts study. Eur Heart J. 2015;36(47):3359-67.

39. B.L. Norgaard, J. Leipsic, S. Gaur, et al., Diagnostic performance of Noninvasive Fractional Flow Reserve Derived from Coronary Computed Tomography Angiography in Suspected Coronary Artery Disease, J. Am Coll Cardiology 63, 1145 2014.

40. Hlatky, Mark A., et al. "Quality-of-Life and Economic Outcomes of Assessing Fractional Flow Reserve with Computed Tomography Angiography." *Journal of the American College of Cardiology* 66.21 (2015): 2315-2323.

41. Douglas, Pamela S., et al. "1-Year outcomes of FFR CT-guided care in patients with suspected coronary disease: the PLATFORM study." *Journal of the American College of Cardiology* 68.5 (2016): 435-445.

42. HeartFlow FFR_{CT} for estimating fractional flow reserve from coronary CT angiography, Medical technologies guidance, Published: 13 February 2017, https://www.nice.org.uk/guidance/mtg32

43. https://www.ama-assn.org/practice-management/applying-cpt-codes

44. http://www.heartflow.com/press_release/press-release-example-1

45. https://blogs.fda.gov/fdavoice/index.php/2017/07/fda-announces-new-steps-to-empower-consumers-and-advance-digital healthcare/?source=govdelivery&utm_medium=email&utm_source=govdelivery and https://www.fda.gov/MedicalDevices/DigitalHealth/UCM567265

46. http://www.imdrf.org/workitems/wi-samd.asp

47. http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf

48. http://www.imdrf.org/docs/imdrf/final/consultations/imdrf-cons-samd-ce.pdf

49. https://www.federalregister.gov/documents/2016/10/14/2016-24805/software-as-a-medical-device-clinical-evaluation-international-medical-device-regulators-forum-draft

50. https://blogs.fda.gov/fdavoice/index.php/tag/software-as-a-medical-device-samd/

51. https://www.fda.gov/medicaldevices/digitalhealth/

52. https://www.ama-assn.org/sites/default/files/media-browser/specialty%20group/washington/ama-digital-health-report923.pdf

53. Silva, Bruno MC, et al. "Mobile-health: A review of current state in 2015." *Journal of biomedical informatics* 56 (2015): 265-272.

54. Zapata, Belén Cruz, et al. "Empirical studies on usability of mHealth apps: a systematic literature review." *Journal of medical systems* 39.2 (2015): 1
.

55. http://www.mobihealthnews.com/content/21-clinical-trials-are-using-fitbit-activity-trackers-right-now

56. https://www.ama-assn.org/ama-adopts-principles-promote-safe-effective-mhealth-applications

57. https://www.ama-assn.org/sites/default/files/media-browser/specialty%20group/washington/ama-digital-health-report923.pdf

58. https://www.alivecor.com/

59. http://www.updrs.net/

60. http://www.asthmamd.org/

61. https://grants.nih.gov/grants/guide/rfa-files/RFA-AG-18-012.html

62. Doerr, Megan, et al. "Formative Evaluation of Participant Experience With Mobile eConsent in the App-Mediated Parkinson mPower Study: A Mixed Methods Study." *JMIR mHealth and uHealth* 5.2 (2017).

63. https://itunes.apple.com/us/app/parkinson-mpower-study-app/id972191200?mt=8

64. https://techcrunch.com/2016/11/03/plushcare-nabs-8m-series-a-to-prove-telehealth-can-go-mainstream/

65. https://www.theguardian.com/technology/2016/oct/02/gp-smartphone-apps-threat-to-nhs

66. https://deepmind.com/

67. https://deepmind.com/applied/deepmind-health/

68. https://www.royalfree.nhs.uk/patients-visitors/how-we-use-patient-information/our-work-with-deepmind/

69. http://www.zdnet.com/article/googles-deepmind-and-the-nhs-a-glimpse-of-what-ai-means-for-the-future-of-healthcare/

70. https://www.newscientist.com/article/2131256-google-deepmind-nhs-data-deal-was-legally-inappropriate/

71. https://techcrunch.com/2017/06/22/deepmind-health-inks-another-5-year-nhs-app-deal-in-face-of-ongoing-controversy/

72. https://www.wired.com/2017/03/google-deepminds-untrendy-blockchain-play-make-actually-useful/

73. Peterson, K., Deeduvanu, R., Kanjamala, P., & Boles, K. (2016). A Blockchain-Based Approach to Health Information Exchange Networks. https://www.healthit.gov/sites/default/files/12-55-blockchain-based-approach-final.pdf

74. Ekblaw, A., Azaria, A., Halamka, J. D., & Lippman, A. (2016, August). A Case Study for Blockchain in Healthcare: "MedRec" prototype for electronic health records and medical research data. In *Proceedings of IEEE Open & Big Data Conference*.

75. https://hbr.org/2017/03/the-potential-for-blockchain-to-transform-electronic-health-records

76. https://www.wired.com/2017/03/google-deepminds-untrendy-blockchain-play-make-actually-useful/

77. Hern, A., "Google's DeepMind plans bitcoin-style health record tracking for hospitals", The Guardian, 10 March 2017. https://www.theguardian.com/technology/2017/mar/09/google-deepmind-health-records-tracking-blockchain-nhs-hospitals

78. *Merkle, R. C. (1988). "A Digital Signature Based on a Conventional Encryption Function". Advances in Cryptology — CRYPTO '87. Lecture Notes in Computer Science.* ***293****. p. 369. ISBN 978-3-540-18796-7. doi:10.1007/3-540-48184-2_32.*

79. Lomas, N., "Patient data API pivotal to DeepMinds push into UKs NHS", https://techcrunch.com/, 22 Nov 2016. https://techcrunch.com/2016/11/22/patient-data-api-pivotal-to-deepminds-push-into-uks-nhs/

80. https://www.23andme.com/

81. https://sciencebasedmedicine.org/dubious-mthfr-genetic-mutation-testing/

82. E.g., https://www.directderm.com/ and https://www.firstderm.com/

83. https://skinvision.com/

84. www.cancer.org

85. www.mayoclinic.org

86. http://files.shareholder.com/downloads/WBMD/4075514959x0x929898/89CBE95F-8CF1-4234-B809-D93060391CE5/WBMD_Q4_16_TRANSCRIPT.pdf

87. http://www.pewinternet.org/fact-sheet/mobile/

88. http://www.pewinternet.org/2011/07/11/smartphone-adoption-and-usage/

89. Pew Research Center, April, 2015, “The Smartphone Difference” Available at: http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/

90. https://www.fns.usda.gov/snap/supplemental-nutrition-assistance-program-snap

91. https://www.acf.hhs.gov/ofa/programs/tanf

92. http://www.image-net.org/challenges/LSVRC/

93. https://www.kaggle.com/

94. https://www.kaggle.com/c/data-science-bowl-2017

95. https://www.kaggle.com/c/msk-redefining-cancer-treatment

96. http://www.datasciencebowl.com/competitions/transforming-how-we-diagnose-heart-disease/

97. https://www.kaggle.com/c/seizure-prediction

98. https://www.drivendata.org/competitions/42/senior-data-science-safe-aging-with-sphere/

99. http://www.image-net.org/

100. http://www.datasciencebowl.com/totheclinic/

101. https://www.zooniverse.org/

102. https://www.zooniverse.org/projects/mrniaboc/bash-the-bug/

103. https://www.galaxyzoo.org/

104. http://www.nws.noaa.gov/om/coop/

105. http://www.biorxiv.org/content/biorxiv/early/2017/07/05/159756.full.pdf

106. A premature death is death before a specific expected age that could have been preventable.

107. Marmot, Michael, et al. "Closing the gap in a generation: health equity through action on the social determinants of health." *The lancet* 372.9650 (2008): 1661-1669.

108. Schroeder, Steven A. "We can do better—improving the health of the American people." *New England Journal of Medicine* 357.12 (2007): 1221-1228.

109. http://www.kff.org/disparities-policy/issue-brief/beyond-health-care-the-role-of-social-determinants-in-promoting-health-and-health-equity/

110. Ibid

111. https://www.nature.com/scitable/topicpage/rare-genetic-disorders-learning-about-genetic-disease-979

112. Q. Yang, MJ Khoury, JM Friedman, J Little and WD Flanders, How many genes underlie the occurrence of common complex diseases in the population?, International Journal of Epidemiology 2005;34:1129–1137

113. MI McCarthy, G.R. Abecasis, L.R.Cardon, et al., Genome-wide association studies for complex traits: consensus, uncertainty and challenges, Nature Reviews, 9, 356, 2008.

114. https://allofus.nih.gov/

115. https://www.healthit.gov/sites/default/files/commonclinicaldataset_ml_11-4-15.pdf

116. 21st Century Cures Act, Public Law 114-255, 114th Congress (2015-2016).

117. https://obamawhitehouse.archives.gov/blog/2016/05/25/precision-medicine-initiative-and-data-security

118. Doerr, Megan and Suver, Christine and Wilbanks, John, *Developing a Transparent, Participant-Navigated Electronic Informed Consent for Mobile-Mediated Research (April 22, 2016). Available at SSRN: https://ssrn.com/abstract=2769129*

119. https://allofus.nih.gov/sites/default/files/allofus-initialprotocol-v1_0.pdf

120. https://oauth.net/2/

121. Peterson, K., Deeduvanu, R., Kanjamala, P., & Boles, K. (2016). A Blockchain-Based Approach to Health Information Exchange Networks. https://www.healthit.gov/sites/default/files/12-55-blockchain-based-approach-final.pdf

122. https://www.hl7.org/fhir/

123. Kaput, J., & Rodriguez, R. L. (2004). Nutritional genomics: the next frontier in the postgenomic era. Physiological Genomics, 16(2), 166–177. http://doi.org/10.1152/physiolgenomics.00107.2003

124. Vineis, P., & Wild, C. P. (2014). Global cancer patterns: causes and prevention. Lancet, 383(9916), 549–557. http://doi.org/10.1016/S0140-6736(13)62224-2

125. Landrigan, P. J., Lambertini, L., & Birnbaum, L. S. (2012). A Research Strategy to Discover the Environmental Causes of Autism and Neurodevelopmental Disabilities. Environmental Health Perspectives, 120(7), a258–a260. http://doi.org/10.1289/ehp.1104285

126. Volk, H. E., Hertz-Picciotto, I., Delwiche, L., Lurmann, F., & McConnell, R. (2010). Residential Proximity to Freeways and Autism in the CHARGE Study. Environmental Health Perspectives, 119(6), 873–877. http://doi.org/10.1289/ehp.1002835

127. Castorina, R., Butt, C., Stapleton, H. M., Avery, D., Harley, K. G., Holland, N., et al. (2017). Flame retardants and their metabolites in the homes and urine of pregnant women residing in California (the CHAMACOS cohort). Chemosphere, 179, 159–166. http://doi.org/10.1016/j.chemosphere.2017.03.076

128. https://www.gdx.net/product/toxic-effects-core-test-urine-blood

129. Perl, T. M., Bédard, L., Kosatsky, T., Hockin, J. C., Todd, E. C., & Remis, R. S. (1990). An outbreak of toxic encephalopathy caused by eating mussels contaminated with domoic acid. The New England Journal of Medicine, 322(25), 1775–1780. http://doi.org/10.1056/NEJM199006213222504
130. http://harvardpolitics.com/united-states/drip-drop-americas-crumbling-water-infrastructure/

131. http://www.cnn.com/2016/06/28/us/epa-lead-in-u-s-water-systems/index.html

132. https://ephtracking.cdc.gov/showHome.action

133. https://airquality.gsfc.nasa.gov

134. https://airquality.gsfc.nasa.gov/cities

135. https://metrolabnetwork.org/

136. https://arrayofthings.github.io/

137. https://wp.nyu.edu/sonyc/

138. http://beacon.berkeley.edu/Overview.aspx

139. https://arxiv.org/abs/1609.08780

140. Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones,et al, Opportunities and obstacles for deep learning in biology and medicine, doi: http://dx.doi.org/10.1101/142760.

141. https://dashboard.healthit.gov/evaluations/data-briefs/non-federal-acute-care-hospital-ehr-adoption-2008-2015.php#appendix

142. http://catalyst.nejm.org/ehr-interoperability-blame-game/

143. Stephen F. Weng, Jenna Reps, Joe Kai, Jonathan M. Garibaldi, Nadeem Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?, PLoS ONE 12(4): e0174944. https://doi.org/10.1371/journal.pone.0174944

144. ibid BigData@Heart

145. ibid S.F. Weng et al

146. Press release for BigData@Heart program: http://www.ucl.ac.uk/health-informatics/ihi-news-publication/bigdataatheart-launch

147. https://www.recode.net/2015/6/30/11564016/machine-learning-is-hard-google-photos-has-egregious-facial. Google immediately apologized and fixed it.

148. http://www.sciencedirect.com/science/article/pii/S0954611107000856, 100 years of respiratory medicine: Pneumonia

149. Stodden, Victoria. "Reproducing statistical results." *Annual Review of Statistics and Its Application* 2 (2015): 1-19.

150. Stodden, Victoria, et al. "Enhancing reproducibility for computational methods." *Science* 354.631 (2016): 1240-1241.

151. https://www.acm.org/publications/policies/artifact-review-badging

152. Pour-el, Marian Boylan, and Ian Richards. "A computable ordinary differential equation which possesses no computable solution." *Annals of Mathematical Logic* 17.1-2 (1979): 61-90.

153. Turing, Alan Mathison. "On computable numbers, with an application to the Entscheidungs problem." *Proceedings of the London mathematical society* (1937): 230-265.

154. http://syncfor.science/

155. http://circ.ahajournals.org.ezproxy.lib.vt.edu/content/early/2017/07/07/CIRCULATIONAHA.117.027436