HHS Public Access

Author manuscript

Phys Med. Author manuscript; available in PMC 2022 March 11.

Published in final edited form as:

Phys Med. 2021 March ; 83: 72–78. doi:10.1016/j.ejmp.2021.02.024.

Requirements and reliability of AI in the medical context

Yoganand Balagurunathan[1,*], **Ross Mitchel**[1,2], **Issam El Naqa**[1,*]

[1]Department of Machine Learning, Health Data Services, Tampa, FL

[2]H. Lee. Moffitt Cancer Center, Tampa, FL

Abstract

The digital information age has been a catalyst in creating a renewed interest in Artificial Intelligence (AI) approaches, especially the subclass of computer algorithms that are popularly grouped into Machine Learning (ML). These methods have allowed one to go beyond limited human cognitive ability into understanding the complexity in the high dimensional data. Medical sciences have seen a steady use of these methods but have been slow in adoption to improve patient care. There are some significant impediments that have diluted this effort, which include availability of curated diverse data sets for model building, reliable human-level interpretation of these models, and reliable reproducibility of these methods for routine clinical use. Each of these aspects has several limiting conditions that need to be balanced out, considering the data/model building efforts, clinical implementation, integration cost to translational effort with minimal patient level harm, which may directly impact future clinical adoption. In this review paper, we will assess each aspect of the problem in the context of reliable use of the ML methods in oncology, as a representative study case, with the goal to safeguard utility and improve patient care in medicine in general.

Keywords

artificial intelligence; machine learning; reliability; medical application

1. Introduction.

The digital information era has seen a surge in the ways to generate and collect data, in recent times, data has been equated to be the new '*oil*' of the future that will fuel technological innovation [1]. Every enterprise wants to have a '*Big data*' resource that follows five popular characteristics: volume, velocity, variety, veracity and value. Availability of such resources and a motivation to aid human cognitive computing capability have led to resurgence in the development of advanced data analytics techniques to cope

*Corresponding Authors: Yoganand Balagurunathan, Issam El Naqa, Ph.D., Department of Machine Learning, H. Lee. Moffitt Cancer Center, 12902 USF Magnolia Ave, Tampa, FL 33612. yogab@moffitt.org, Issam.ElNaqa@moffitt.org.

Conflict of Interest.

YB: He is funded by National Institutes of Health grants and pharmaceutical research grant (Kite pharma). No conflicts related to the present work.

RM: No conflicts to report.

IE-N: Board member of Endectra, LLC and receives National Institute of Health (NIH) grants.

with such big data with artificial intelligence (AI) being the driving force [2]. It has been widely perceived that data will enable wide ranges of enterprises from finance, marketing, entertainment to medicine to better understand their customer behavior and cater future services that can be directly linked to their economic and well-being of future interests [3].

The medical radiological and oncological sciences have been traditionally collecting data in silos for a few decades that are stored in various forms. However, making inferences using the vast resource has been a challenge. Nevertheless, there were some noted successes in the field that had made use of the data repositories, that include radiological sciences, radiation oncology and large patient records [4–8].

In the recent past there has been a convergence of multiple factors that enabled the resurgence of AI, that includes availability of Big data resources, open-source machine learning (ML) tools and cheaper computational units, especially the development of parallel computing technologies such graphical processing units (GPUs) and tensor processing units (TPU) by the high-tech commercial information technology (IT) vendors. The ‘Big data’ wave has given the private sector an edge to forge ahead with technological advancements to make machine-based adaptive learning possible. These developments had several phases, certainly tumultuous iterations, most vibrant media display of these technological advancement can be traced back to initial success starting with the IBM’s *Deep Blue* computer that won the Grand chess master game against chess champion, Kasparov [9], The *Watson* computer that won the Jeopardy game [10], Deep Mind’s computers that won over the Chinese game of GO against its known human champions, which was significant due to the fact that GO game had more potential positions than number of atoms in the universe [11]. Relentless efforts by the IT industry to adopt open-source tool developments have accelerated the Deep learning wave in recent years. The pioneering open source efforts from industry and academic leaders was unheard-of in recent history, where a massive code base has been made open source that has allowed widespread adoption from varied sectors [12, 13] [14]. Specifically, the current advancements in convolutional deep neural network architecture followed an evolutionary process in the research community, one recent milestone in this effort was when machine intelligence was shown to be at par or beyond human perception that showed to identify image groups in a database of over 10,000 categories in natural images [15].

The success of recent deep networks represented a paradigm shift from conventional ML algorithms that relied on human extracted features. In deep neural network learning, the use of convolutional layers, prior to which use of fully connected layers for traditional detection or classification layers have fully automated the learning process. In addition, algorithmic development of better optimization methods for back propagation based adaptive learning has been fundamental to improved network based learning [16, 17]. Although these ML methods provided the ability to solve complex problems, there has been a checkered history for their usage. That included the backbox stigma, requirement of contextual data and complexity in implementation. These challenges have led to instances of missed expectations, sometimes referred to in the literature as *AI winters* [18, 19].

There has been a tremendous growth in medical data that is regularly collected at clinical centers across the world. Resurgence of AI methods and its application in medicine has opened many possibilities of improved diagnosis. Patients' privacy and data security are legitimate concerns when dealing with medical data. Hence, clinical records have been traditionally restricted due to Health Insurance Portability and Accountability Act (HIPPA) guidelines in the USA and the General Data Protection Regulation (GDPR) in Europe. Most often patient records are non-accessible for research, though de-identified access is possible with an institutional protocol, resulting in minimal availability for data mining, increased overhead in data processing and limiting development of ML applications by the general research community compared to other data resources.

In the USA, most of the human genomic data efforts have been initiated through the National Institute of Health (NIH)'s centers, specifically, the National Center of Biomedical Information (NCBI). This center maintains about 39 distinct databases with over 2.9 billion records. The genome database, GeneBank is over 6.25 trillion base pairs from over 1.6 billion nucleotide sequences for 450,000 formally described species, occupying 1057GB of uncompressed disk storage, all of which are open for public access [20, 21]. Similarly, The Cancer Imaging Archive (TCIA) houses open imaging data about 126 imaging studies (CT, PET, MRI, etc.) datasets available to be public [22–24].

In this review, we highlight the evolution of AI/ML methods, outline the challenges in using them for medical applications and focus on factors that contribute to their reliable use in medicine in general and oncological science, to be specific. We then provide recommendations for appropriate AI/ML model development to improve reliability, reproducibility, and user's confidence for clinical research and care adoption.

2. Evolution of Machine Learning Methods in Oncology

Learning patterns and dependency trends derived from the data has been a fundamental discipline of ML methods in AI [25, 26]. Clinical sciences have traditionally been following a conservative quasi-quantitative approach to understand the disease physiology, biological processes; while use of quantitative methods has shown utility in the last decade [27–29]. The first notable application of AI in medicine was the development of MYCIN, as a rule-based system in the 1970s at Stanford University. The system would use patient information and lab measurement for diagnosis of potential bacterial infections and recommend treatment options accordingly [30]. This initial success led to the development of initial AI methods in medicine and specifically in radiology, which saw initial use of computer-aided diagnosis/detection using image analysis in the 1980s and 1990s [31]. However, this optimism[32] went through a dormant period, while computer technology went through a steep developmental period, coupled with the development of the information age, this mindset has transformed into possibility of advancement in most and skepticism in a few others [33].

The recent advancement of AI methods and its application in medicine, especially the use of machine intelligence has rejuvenated excitement. Despite many, it has also raised various concerns among clinical practitioners, patients and public on the issues related to

data protection, ethical use of the technology, biases in future care and more importantly algorithmic limitation and its influence [34] We outline some of the common concerns among the medical practitioners and thought leaders in AI/ML :

Can ML replace a physician?

There is a level of uncertainty that exists in accepting technological advancement related to ML applications and its implications for clinical care. One primary concern among clinical practitioners is being sidelined or replaced by the machines. While in a broader context, most clinical decisions are arrived at by consensus, in oncology tumor boards are classic working example. The role of ML will be to provide a decision aid to the practitioner, like having a personal *smart* assistant that can swiftly run through enormous numbers of records, make comparisons in real time and provide recommendations. The clinical expert may still need to interact with the human patient, understand their personal events and choose an option in consultation with the patients accommodating personal, cultural, social needs and preferences. Machine decisions to most extent will be complementary, most often deferring the final clinical decision to the experts [35]. False decisions may have serious ramifications, which could be overriding (non-action) an AI finding (false negative) or acting on a spurious AI finding (false positives). Recent consensus reports reviewed the use of these technologies in medicine in general and oncology in particular and have emphasized the trustworthy (ethical) use of AI methods [36]. It is expected that new technologies may need the right context of use, which is based on appropriate training cohorts and unbiased algorithms, to mitigate errors, it may still be helpful to have a human expert in the loop. This expert can understand the broad aspects of the model (as a non AI expert) and take a role to interact with the patients, most importantly take responsibility for the outcome [36, 37].

Human oversight in AI decision support.

It is widely believed that human experts could possibly identify spurious findings (false positives) better than an AI system, this is to offset wrong decisions due to contextual differences in the test population or other biases in training [36]. It is also argued that human knowledge is subjective and limited to visible traits in the data, that may lead to biased decisions. Machines follow a set rules learnt from observable and unobservable (non-linear) patterns, complete automation in the decision process would provide best utilization of an AI system [8, 38, 39]. Most neural architectural based AI systems still to date are heavily based on corporate developed codebase, that would ideally require varied train/test and validation across medical centers to reduce over treatment (false negatives), till then human experts in the loop may be unavoidable.

Examples of AI Systems in Oncology and Limitations.

The image concentric sciences (radiology/pathology) were pioneers to use AI methods with the advent of deep learning networks [15, 40, 41]. One example of a translational application of deep learning convolutional neural network (CNN) was to detect skin lesions (melanoma) using pictures [42], this application is projected to attract/benefit billions of users with the deployment of mobile applications [43]. These applications are stated to be a frontline virtual diagnostic care and a screening aid. Another notable application of AI method is the ability to detect diabetic retinopathy using retinal fundus photographs, which has shown to

be useful to scale up in regions with lower medical resources [44, 45]. Deep networks have been successfully shown to be useful in detecting polyps in colonoscopy images [46]. There are several applications in pathology, one recent work shows detection of breast mitosis in whole slides[47], in an another application investigators have shown the ability to detect tumor-infiltrating lymphocytes (TIL) in different cancer types [48]. HistoQC and DeepFocus are some practical use cases of AI methods to standardize and improve the quality of whole slide imaging that could improve the detection ability [49, 50]. There have been many useful applications in radiology [51, 52], one recent work shows ability to detect malignant nodules in screening CT images [53]. Though much of the promise to use these methods to improve detection and patient care has impeded by the lack of “high quality” curated datasets, limited ability to share data across institutions due to understandable privacy compliance regulations and in some cases it has practical implementational limitations [54].

3. Validation and Reproducibility of ML Methods

Recent advancement of deep learning analysis has shown promise in clinical research, but actual impacts on routine clinical care may still be anticipatory. To date most of the studies that claim to have clinical translation are based on retrospective data, likely to belong to prior technologies and medical research continues to evolve at a rapid phase[55]. It may be necessary that ML methods follow a series of internal and external validation for ML related clinical applications to improve chances of being useful for routine patient care with some extent of prospective (or live) training [56, 57]. There are several recommendations that outline best practices to develop and implement AI in medicine. One such recommendation that has relevance provides reliability score for multivariate analysis, called the Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) [58]. More specific recommendations for radiological application have been proposed, such as the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) [59]. Additional guidelines relevant to oncological applications were proposed in MI-CLAIM (Minimum information about clinical artificial intelligence modeling) [60]. There are various other domain specific applications proposed in respective societies, such as radiation oncology/medical physics and others [61, 62]. it is likely the conformance of these recommendation will apply to AI/ML in oncology.

Clinical Implementation and Economic Impact.

Clinical validation of ML methods would require close partnership with the clinicians, who need to understand the system, and a functional overview of the method both as a non-expert, which will allow them to spot and identify possible spurious results, i.e., results that may be technically accurate but clinically meaningless. Recent review of AI related publications shows no or limited economic clinical assessment, most published work has no inclusion criteria or consistency that allows comparison between studies [63]. In a clinical context the core element of the implementation process will involve initial investment cost, operational cost for an AI system and a tangible benefit in terms of patient care which may involve reduction in medical errors etc. Analysis of these characteristics will be critical and plays a role in clinical adoption. Additionally, comparisons need to be made with existing

technologies, like finer genomic analysis down to the cell level and other similar methods with contrasting technologies to achieve the desired improvement in patient care.

In a clinical context, adoption of AI systems to current care will need to be seamlessly integrated into the current Electronic Medical Record (EMR) workflow. Any small addition to clinician's time that may involve system review or additional staff time to the current workflow will become a real challenge and most often requires a clinical cost-benefit analysis at the institutional level. As an example, Breast and Genitourinary Oncology suites perform several targeted biopsies during an office visit. Additional AI system interaction will result in an overhead that causes lost time to accommodate fewer patients during the day, roughly this may trigger 3 to 4 lost appointments in a day (assuming a workload of 10 visits/day) and such decisions typically trigger hospital management evaluation. Certainly, AI systems will come with renewed hope of improved care, reduced or enhanced clinical decisions based on adaptive learning will play a role in disease management.

Interpretability and Governance of AI Models.

There has been renewed debate about governance and interpretability of AI models in medical sciences, these concerns have been exacerbated due to unexplainable models using deep networks [64, 65].Although it is expected that medically relevant models need to be interpretable, open to investigate and expected to provide a level of confidence to the practitioner that allows informed decision. Contrarily, most practitioners are not expected to understand the mechanics of the system. As an analogy, a driver is expected to operate a vehicle, with just operational experience. While a mechanic is considered a specialist, who understands the intricacies of the system. It is common to have users operate a vehicle (mechanical or AI driven) without needing to have any knowledge about the system.

Reliability of ML Methods.

A reliable model needs to perform a stated function that will improve clinical decisions and to reduce unknown spurious consequences with a goal to protect human life. In recent consensus statements that provide criteria to evaluate multivariate statistical methods have been well documented, below are a list of widely accepted contenders TRIPOD, CLAIM, MI-CLAIM [58–60], which were developed by a large survey of scholars and industrial experts' opinions. The ML methods are yet to follow suit, in a recent review of ML methods in oncology provides insights and best practices [34]. However, to be a reliable method performance accuracy along with added clinical value are considered, these aspects are well emphasized in the recent Food and Drug Administration's (FDA) regulations on AI systems [66, 67].

It is widely accepted that models need to be developed using a diverse population and the methods to be validated in an independent dataset that allows developing an unbiased generalizable model. While in ML, complexity of these models, especially deep learning networks are non-interpretable and may pose a challenge to model transparency in oncology. The conventional wisdom of training and validation are traditionally used to develop a model that works in the context of a disease or a molecular subtype. This conventional wisdom is in contention with the new suite of AI based methods. Where-in models trained

on common object images (cats, dogs, toilets, etc.) that are from mixed context have been used to pre-trained deep network models and successfully claimed to provide disease risk assessment in oncology [41, 68]. It has been broadly accepted that the data requirements to train a large deep network are certainly high, successful AlexNet [15] had over 650,000 neurons and about 60 million parameters, which certainly poses a challenge. Due to limited curated public datasets in oncology has led to use of disparate disease types to train the network and used to discriminate against other disease types follows an approach called the transfer learning [40]. These approaches may have an acceptable technical basis to train large deep networks, but this direction is certainly a deviation from conventional wisdom.

Transparency of Models.

It is an implicit requirement that any clinical model needs to be interpretable or transparent and has an ability to describe its functioning, describe its structure, underlying parameters and explicitly state its assumptions. Recent consensus statement describes the expectations of a decision making model in healthcare [69], which states that transparency can be achieved by explicit model definition followed by a model validation. The black box approach is often associated with the ML model algorithms; there is a conscious effort to identify features and characteristics that may be responsible for a machine decision [70]. It may be necessary for the models to document the shortcoming, quantify risk of false detection or uncertainty. Several methods have been developed under the umbrella of explainable (XAI), which is a higher order of expectation than being transparent and interpretable. Efforts have been originally spearheaded by the Defense Advanced Research Agency (DARPA), some of these methods employ approximations methods such as LIME (Local Interpretable Model-Agnostic Explanations), others have focused on visualization efforts such as saliency or activation maps[38, 71]. An interesting approach that gets traction to mitigate bias is the utilization of SHAP values, this concept originated in game theory that tends to assign an importance value for a feature, in a specific prediction. In this approach both local and global interpretability requirements are being satisfied [72]. It has to be emphasized that interpretability is just one step towards better explaining the model and does not necessarily imply causality, although it may be needed towards an optimal clinical decision support system [39].

Repeatability and Reproducibility of findings.

Reproducibility of clinical models by an independent research group either by re-implementing the method or using the same code base to achieve (close to) similar results would be a critical need in clinical sciences and would fulfill at the highest level (level 4) as assessed by the TRIPOD statement [73]. Repeatability of a method in a technically repeated cohort may be the first step in the process of quantifying variability to gain confidence in the use of a method. In most methods, such as imaging, scanning an individual patient twice will result in some differences in extracted metrics due to variations in the patient placement, culminated by reconstruction methods and delineations procedure. In case of imaging, estimating reproducible metrics in repeated patient scans was most fundamental to evaluate the volume metrics [74] and other quantitative imaging features (radiomic) [75, 76]. It is well recognized that quantitative features may be altered due to patient factors (motion, breathing) and other physiological causes. It is a clinical requirement that features

obtained at the lesion level can withstand a certain level of variation and that does not alter the clinical decision, detection or risk assessment. In comparison, a biomarker developed in basic sciences, omics-based do have a level of experimental variability in the lab. The larger development phase has allowed assay refinement, that has improved the variability to a small acceptable percentage of coefficient of variation [77, 78]. Thus, it is necessary that the ML methods follow an evolution to obtain a level of acceptable reproducibility and repeatability [79].

Central versus distributed model training.

Model training from a diverse data source has been traditionally used to sample different populational distributions [25]. Need for diverse data sources for Machine/Deep learning models has been restricted in the medical clinical domain, some are due to various patient privacy related concerns, institutional restrictions and unavailability of standardized data, all of this has limited clinical translation. Recently, there has been an extensive effort to learn localized data at multiple sources in a federated fashion that encouraged decentralized learning [80]. There has been a larger effort to find a compromise solution that would enable usage in the clinical context that has allowed use of distributed or federated learning in oncology [81] and has shown initial success [82] and holds promise for the future.

Model Reliability and Evaluation.

The use of AI in medicine has brought various challenges starting with postulation of the evaluation criteria, sufficient datasets to test varying approaches and establish comparable tools to create standards for clinical translation [83, 84]. Model reliability in a clinical setting involves accuracy of the system, relevance for clinical translation, ethical fairness in its decision and expert/patient trust or acceptance of these systems [85]. Challenges for translation of AI systems in health care include logical difficulties in the implementation, barrier to adoption, sociocultural related issues and non-amenable for workflow changes [86]. On the other hand, it is estimated that human error related cost of over/under diagnosis and missed medical procedures accounts for a large portion of incurred medical costs [87]. AI based medical support system would improve better diagnosis, treatment and have a greater potential to mitigate wrong inferences in highly stressful environments like intensive care facilities (ICU) and clinical trials [88, 89]. To have a reliable AI system in medicine, we could divide the lifecycle of the application into three phases. First early phase would start with a concept, leading to a discovery where an algorithm is established with an initial cohort study. In the mid phase level will allow the system to be established and tested in a larger population (clinical trial) with a limited product test. At every step in the process the methods are refined based on response and algorithms altered with an intention to reduce biases. In the last phase (late) will be when the AI based product is ready to be launched for adoption in a population. Figure 1 describes our broad workflow.

Ethical Use of AI methods.

Machine intelligence and widespread use of AI methods have evoked a larger discussion on its ethical use of the technology. It has been argued that AI methods are subjected to inaccurate and discriminatory outcomes that lead to biases as reported in recent UNESCO review[90]. Most prevalent use of AI technology has been in surveillance of the population,

where AI tools are widely used to identify potential social disruptors and provide leads on wanted individual. Most often, like in the US, certain groups are vulnerable to be identified as potential matches. There have been instances that have resulted in false identification that has resulted in social and human distrust on the use of technology [91, 92].

4. Discussion

The complexity of human biology and heterogeneous nature of oncological disease has allowed multifold investigation, starting with physiology of the diseases, deep dive exploration at the cellular level using genetic, epigenetics to epidemiological analysis at the population level, all of which has allowed generating enormous amounts of information. Most of the learning methods that are used in most studies falls under supervised and unsupervised [93, 94]. The recent advancement of deep networks [95] has brought about new advancements in machine intelligence that shows promise to go beyond human perception. Early applications of these methods were in radiological sciences [53, 96, 97] and has made inroads in other oncological -omics datasets[98, 99]. Most of the deep learning networks have been trained using small data sets, randomly iterate multiple times to train the weights of the network or in other cases the network is trained using unrelated context and primed for the context of interest using a smaller cohort. Recently advancement of Few-Shot learning methods provide promise to use these AI methods in small data applications[100].

There is a significant challenge among users and clinical practitioners to assess the reliability of these network approaches for clinical use. Implementational challenges for AI technology does exist, recent retinopathy detection system had several ground level problems that include missed expectations due to quality and user level issues [54]. Certainly, common for any technological implementations, to date automotive recalls are accepted as proper correction procedure and other technologies enhance any shortfall [101]. Though human perception, future adoption will be at stake, but reliability of the technology depends on consequential steps taken to mitigate the current and future problems.

This decade has seen infiltration of the high-tech industry led methods that includes use of enormous code based that are painstakingly modified for oncological data, with unknown extent of pre-processing with a goal to discern and go beyond human levels of perception, some examples include ability to identify disease types, subtle traits in image sub-class categorization. These approaches have shown enormous promise but pose a grand challenge to reproduce findings across unseen diverse datasets. One such concern has been widely referenced in a recent effort to improve transparency and reliability of these deep network findings [102], which has outlined the need for open disclosure of codebase, datasets used for training and preprocessing steps to achieve desired outcome. It may be necessary to allow free and fair assessment of these method’s reliability and reproducibility.

In oncology, ML methods have brought us to a critical junction that has resulted in various contentions among the oncologists that the AI systems are a) unexplainable and b) it would replace their role as machines that can do a better job than humans. It may be true that these networks are unexplainable but certainly have utility once we provide extensive

training samples. Some successful examples include product support recommendation based on prior purchases and picture identification to find close family members. It is essential that clinical systems are transparent or at least interpretable and that would allow adaptable multi-expert involvement to be considered in the AI decisions. It has become apparent that AI technologies would have widespread impact on human society in many ways, ethical use of technology with some oversight becomes necessary for adaptation in our societies. AI technologies may have influence across borders with widespread implications, this has led to a recent report on ethical use of AI technologies under the auspices of UNESCO [90].

Model training using silos of data sources that could update the model in a federated model has shown promise in medicine. Recent use of distributed learning across multiple centers, spread across three continents shows enormous hope to build a robust clinical model.

Medicine in general and oncology in specific, due to its complexities; machines may not completely replace human's role in diagnosis and treatment, certainly needs to follow a multi-phased approach with early invention, first user application to a period for technological adoption. There is a greater role for AI systems to support the oncologist and to provide decision support to enhance human understanding with enormous hope to discover cures for diseases and improve patient care.

5. Conclusion

We believe that AI methods will provide an ample opportunity to break the barrier to understand complex human cognitive ability to make decisions and to automate processes. These methods along with ML approaches are here in stay and will support the expert medical professions, specifically more in oncology due to disease complexity [103]. To improve reliability and transparency of ML and deep methods in medical sciences, we summarize the following checklist based on literature survey and research experiences discussed in this report.

- Diverse cohort of patient records for model training, achieved either through centralized or using federated/distributed learning models that uses silos of different data sources.
- Use of independent data cohort for testing, preferably in a distributed setting with diverse patient types.
- Transparency of deep network model architecture with confidence levels in its decisions.
- Ethically appropriate use of AI methods with some level of oversight.
- Assessment of reproducibility of AI models with test-retest type studies.
- Model transparent that discloses the architecture, data sets and trained weights for the network.
- Quality assurance program for implementation and continuous performance monitoring.

Once the AI/ML systems are trained with large, diverse cohorts and the methods transparently reproduced, these models would be an asset to aid/train future medical professionals to advance knowledge that may be uncommon to human perception. As Heraclitus, a Greek philosopher would say, "*Change is the only constant in life*", this era will see a renewed focus in building reliable AI/ML networks that will allow humans to learn from the machines. As Niels Bohr (*Nobel Laureate*), says '*An expert is a man who has made all the mistakes which can be made in a very narrow field*', here we expect the network to be a specialist that has learned from examples and could help humans to discern facts. This approach is relevant in healthcare, especially oncology that shows great promise, and will have greater impact to improve quality of medicine and provide scalable medical access to the world population.

Acknowledgements

Authors like to acknowledge institutional support and grant funding received through various agencies to support their research time. YB (U01-CA200464, U01CA143062, Cohen's family donation), IE (R37-CA222215, R01-CA233487, R41 CA243722, and subcontract 75N92020D00018).

References

[1]. The world's most valuable resource is no longer oil, but data. The Economist. London, UK2017.

[2]. Initiative MG. Big Data: The next frontier for innovation, competition and producitivty. 2011.

[3]. Hilbert M, López P. The world's technological capacity to store, communicate, and compute information. Science (New York, NY). 2011;332:60–5.

[4]. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. Nat Rev Clin Oncol. 2019;16:703–15. [PubMed: 31399699]

[5]. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. CA: a cancer journal for clinicians. 2019;69:127–57. [PubMed: 30720861]

[6]. Nensa F, Demircioglu A, Rischpler C. Artificial Intelligence in Nuclear Medicine. Journal of nuclear medicine : official publication, Society of Nuclear Medicine. 2019;60:29s–37s.

[7]. Kulikowski CA. Beginnings of Artificial Intelligence in Medicine (AIM): Computational Artifice Assisting Scientific Inquiry and Clinical Art - with Reflections on Present AIM Challenges. Yearbook of medical informatics. 2019;28:249–56. [PubMed: 31022744]

[8]. El Naqa I, Haider MA, Giger ML, Ten Haken RK. Artificial Intelligence: reshaping the practice of radiological sciences in the 21st century. The British journal of radiology. 2020;93:20190855. [PubMed: 31965813]

[9]. Munakata T Thoughts on Deep Blue Vs. Kasarov. Communications of the ACM: Automation of Computer Machinery; 1996.

[10]. IBM. IBM computer watson wins jeopardy clash. The Guardian: Guardian Media Group; 2011.

[11]. Gibney E What Google's winning Go algorithm will do next. Nature. 2016;531:284–5. [PubMed: 26983517]

[12]. Mulfari D, Palla A, Fanucci L. Embedded Systems and TensorFlow Frameworks as Assistive Technology Solutions. Studies in health technology and informatics. 2017;242:396–400. [PubMed: 28873830]

[13]. Sun Y, Zhu S, Ma K, Liu W, Yue Y, Hu G, et al. Identification of 12 cancer types through genome deep learning. Scientific reports. 2019;9:17256. [PubMed: 31754222]

[14]. Shah H The DeepMind debacle demands dialogue on data. Nature. 2017;547:259. [PubMed: 28726841]

[15]. Krizhevsky Alex; Sutskever IaH, Geoffrey E. ImageNet Classification with Deep Convolutional Neural Networks: Curran Associates, Inc.; 2012.

[16]. Dreyfus SE. Artificial neural networks, back propagation, and the Kelley-Bryson gradient procedure. Journal of Guidance, Control, and Dynamics. 1990;13:926–8.

[17]. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44. [PubMed: 26017442]

[18]. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of Clinical Epidemiology. 1996;49:1225–31. [PubMed: 8892489]

[19]. Strickland E IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. IEEE Spectrum. 2019;56:24–31.

[20]. Sayers EW, Agarwala R, Bolton EE, Brister JR, Canese K, Clark K, et al. Database resources of the National Center for Biotechnology Information. Nucleic acids research. 2019;47:D23–d8. [PubMed: 30395293]

[21]. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. GenBank. Nucleic acids research. 2019;47:D94–D9. [PubMed: 30365038]

[22]. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. Journal of digital imaging. 2013;26:1045–57. [PubMed: 23884657]

[23]. Prior FW, Clark K, Commean P, Freymann J, Jaffe C, Kirby J, et al. TCIA: An information resource to enable open science. Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual International Conference. 2013;2013:1282–5.

[24]. Kirby J, Prior F, Petrick N, Hadjiski L, Farahani K, Drukker K, et al. Introduction to Special Issue on Datasets hosted in The Cancer Imaging Archive (TCIA). Medical physics. 2020.

[25]. Duda RO, Hart PE, Stork DG. Pattern Classification (2nd Edition): Wiley-Interscience; 2000.

[26]. Cherkassky V, Mulier F. Learning from Data: Concepts, Theory, and Methods: John Wiley & Sons, Inc.; 2006.

[27]. Yankeelov TE, Mankoff DA, Schwartz LH, Lieberman FS, Buatti JM, Mountz JM, et al. Quantitative Imaging in Cancer Clinical Trials. Clinical cancer research : an official journal of the American Association for Cancer Research. 2016;22:284–90. [PubMed: 26773162]

[28]. Winters IP, Murray CW, Winslow MM. Towards quantitative and multiplexed in vivo functional cancer genomics. Nature Reviews Genetics. 2018;19:741–55.

[29]. O'Loughlin TA, Gilbert LA. Functional Genomics for Cancer Research: Applications In Vivo and In Vitro. Annual Review of Cancer Biology. 2019;3:345–63.

[30]. Shortliffe EH, Davis R, Axline SG, Buchanan BG, Green CC, Cohen SN. Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system. Computers and Biomedical Research. 1975;8:303–20. [PubMed: 1157471]

[31]. Giger ML, Chan HP, Boone J. Anniversary paper: History and status of CAD and quantitative image analysis: the role of Medical Physics and AAPM. Medical physics. 2008;35:5799–820. [PubMed: 19175137]

[32]. Schwartz WB, Patil RS, Szolovits P. Artificial Intelligence in Medicine. New England Journal of Medicine. 1987;316:685–8.

[33]. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. New England Journal of Medicine. 2019;380:1347–58.

[34]. El Naqa I, Ruan D, Valdes G, Dekker A, McNutt T, Ge Y, et al. Machine learning and modeling: Data, validation, communication challenges. Medical physics. 2018;45:e834–e40. [PubMed: 30144098]

[35]. Goldhahn J, Rampton V, Spinas GA. Could artificial intelligence make doctors obsolete? BMJ (Clinical research ed). 2018;363:k4563.

[36]. Geis JR, Brady AP, Wu CC, Spencer J, Ranschaert E, Jaremko JL, et al. Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement. Radiology. 2019;293:436–40. [PubMed: 31573399]

[37]. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. Artificial Intelligence in Healthcare. 2020:295–336.

[38]. Luo Y, Tseng H-H, Cui S, Wei L, Haken RKT, Naqa IE. Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling. BJR|Open. 2019;1:20190021. [PubMed: 33178948]

[39]. Naqa IE, Kosorok MR, Jin J, Mierzwa M, Haken RKT. Prospects and Challenges for Clinical Decision Support in the Era of Big Data. JCO Clinical Cancer Informatics. 2018:1–12.

[40]. Khan UAH, Stürenberg C, Gencoglu O, Sandeman K, Heikkinen T, Rannikko A, et al. Improving Prostate Cancer Detection with Breast Histopathology Images. In: Reyes-Aldasoro CC, Janowczyk A, Veta M, Bankhead P, Sirinukunwattana K, editors. Digital Pathology. Cham: Springer International Publishing; 2019. p. 91–9.

[41]. Rai T, Morisi A, Bacci B, Bacon N, Thomas S, La Ragione R, et al. Can ImageNet feature maps be applied to small histopathological datasets for the classification of breast cancer metastatic tissue in whole slide images?: SPIE; 2019.

[42]. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115–8. [PubMed: 28117445]

[43]. Cerwall PR EM https://www.ericsson.com/assets/local/mobilityreport/documents/2016/Ericsson-mobility-report-june-2016.pdf. 2016.

[44]. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. Jama. 2016;316:2402–10. [PubMed: 27898976]

[45]. Varadarajan AV, Bavishi P, Ruamviboonsuk P, Chotcomwongse P, Venugopalan S, Narayanaswamy A, et al. Predicting optical coherence tomography-derived diabetic macular edema grades from fundus photographs using deep learning. Nature communications. 2020;11:130.

[46]. Roth HR, Lu L, Liu J, Yao J, Seff A, Cherry K, et al. Improving Computer-Aided Detection Using Convolutional Neural Networks and Random View Aggregation. IEEE transactions on medical imaging. 2016;35:1170–81. [PubMed: 26441412]

[47]. Albayrak A, Bilgin G. Mitosis detection using convolutional neural network based features. 2016 IEEE 17th International Symposium on Computational Intelligence and Informatics (CINTI)2016. p. 000335–40.

[48]. Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, et al. Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. Cell reports. 2018;23:181–93.e7. [PubMed: 29617659]

[49]. Janowczyk A, Zuo R, Gilmore H, Feldman M, Madabhushi A. HistoQC: An Open-Source Quality Control Tool for Digital Pathology Slides. JCO Clin Cancer Inform. 2019;3:1–7.

[50]. Senaras C, Niazi MKK, Lozanski G, Gurcan MN. DeepFocus: Detection of out-of-focus regions in whole slide digital images using deep learning. PloS one. 2018;13:e0205387. [PubMed: 30359393]

[51]. Ueda D, Shimazaki A, Miki Y. Technical and clinical overview of deep learning in radiology. Japanese journal of radiology. 2019;37:15–33. [PubMed: 30506448]

[52]. Shen D, Wu G, Suk H-I. Deep Learning in Medical Image Analysis. Annu Rev Biomed Eng. 2017;19:221–48. [PubMed: 28301734]

[53]. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med. 2019;25:954–61. [PubMed: 31110349]

[54]. Heaven WD. Google's medical AI was super accurate in a lab. Real life was a differnt story. MIT technology review. USA: MIT; 2020.

[55]. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ (Clinical research ed). 2020;368:m689.

[56]. Chen P-HC, Liu Y, Peng L. How to develop machine learning models for healthcare. Nature Materials. 2019;18:410–4. [PubMed: 31000806]

[57]. Nagy M, Radakovich N, Nazha A. Machine Learning in Oncology: What Should Clinicians Know? JCO Clinical Cancer Informatics. 2020:799–810. [PubMed: 32926637]

[58]. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ (Clinical research ed). 2015;350:g7594.

[59]. Mongan J, Moy L, Charles E. Kahn J. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. Radiology: Artificial Intelligence. 2020;2:e200029. [PubMed: 33937821]

[60]. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nature Medicine. 2020;26:1320–4.

[61]. Huynh E, Hosny A, Guthier C, Bitterman DS, Petit SF, Haas-Kogan DA, et al. Artificial intelligence in radiation oncology. Nat Rev Clin Oncol. 2020;17:771–81. [PubMed: 32843739]

[62]. Thompson RF, Valdes G, Fuller CD, Carpenter CM, Morin O, Aneja S, et al. Artificial intelligence in radiation oncology: A specialty-wide disruptive transformation? Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology. 2018;129:421–6. [PubMed: 29907338]

[63]. Wolff J, Pauling J, Keck A, Baumbach J. The Economic Impact of Artificial Intelligence in Health Care: Systematic Review. J Med Internet Res. 2020;22:e16866–e. [PubMed: 32130134]

[64]. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. Journal of the American Medical Informatics Association : JAMIA. 2020;27:491–7. [PubMed: 31682262]

[65]. O'Sullivan S, Nevejans N, Allen C, Blyth A, Leonard S, Pagallo U, et al. Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. The international journal of medical robotics + computer assisted surgery : MRCAS. 2019;15:e1968. [PubMed: 30397993]

[66]. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. npj Digital Medicine. 2020;3:118. [PubMed: 32984550]

[67]. Administration FaD. Artificial Intelligence and Machine Learning in Software as a Medical Device. 2020. p. FDA Regulation on AI.

[68]. Bychkov D, Linder N, Turkki R, Nordling S, Kovanen P, Verrill C, et al. Deep learning based tissue analysis predicts outcome in colorectal cancer. Scientific reports. 2018;8. [PubMed: 29311689]

[69]. Eddy DM, Hollingworth W, Caro JJ, Tsevat J, McDonald KM, Wong JB. Model Transparency and Validation: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. Value in Health. 2012;15:843–50. [PubMed: 22999134]

[70]. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: Association for Computing Machinery; 2016. p. 1135–44.

[71]. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020;58:82–115.

[72]. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: Curran Associates Inc.; 2017. p. 4768–77.

[73]. Boulbes DR, Costello T, Baggerly K, Fan F, Wang R, Bhattacharya R, et al. A Survey on Data Reproducibility and the Effect of Publication Process on the Ethical Reporting of Laboratory Research. Clinical cancer research : an official journal of the American Association for Cancer Research. 2018;24:3447–55. [PubMed: 29643062]

[74]. Zhao B, James LP, Moskowitz CS, Guo P, Ginsberg MS, Lefkowitz RA, et al. Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non-small cell lung cancer. Radiology. 2009;252:263–72. [PubMed: 19561260]

[75]. Balagurunathan Y, Kumar V, Gu Y, Kim J, Wang H, Liu Y, et al. Test-retest reproducibility analysis of lung CT image features. Journal of digital imaging. 2014;27:805–23. [PubMed: 24990346]

[76]. Zhao B, Tan Y, Tsai WY, Qi J, Xie C, Lu L, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. Scientific reports. 2016;6:23428. [PubMed: 27009765]

[77]. Marshall CR, Chowdhury S, Taft RJ, Lebo MS, Buchan JG, Harrison SM, et al. Best practices for the analytical validation of clinical whole-genome sequencing intended for the diagnosis of germline disease. npj Genomic Medicine. 2020;5:47. [PubMed: 33110627]

[78]. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008;18:1509–17. [PubMed: 18550803]

[79]. Kanwal S, Khan FZ, Lonie A, Sinnott RO. Investigating reproducibility and tracking provenance - A genomic workflow case study. BMC bioinformatics. 2017;18:337-. [PubMed: 28701218]

[80]. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. npj Digital Medicine. 2020;3:119. [PubMed: 33015372]

[81]. Jochems A, Deist TM, van Soest J, Eble M, Bulens P, Coucke P, et al. Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital - A real life proof of concept. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology. 2016;121:459–67. [PubMed: 28029405]

[82]. Jochems A, Deist TM, El Naqa I, Kessler M, Mayo C, Reeves J, et al. Developing and Validating a Survival Prediction Model for NSCLC Patients Through Distributed Learning Across 3 Countries. International journal of radiation oncology, biology, physics. 2017;99:344–52.

[83]. Ryan M In AI We Trust: Ethics, Artificial Intelligence, and Reliability. Science and Engineering Ethics. 2020;26:2749–67. [PubMed: 32524425]

[84]. Asaro PM. AI Ethics in Predictive Policing: From Models of Threat to an Ethics of Care. IEEE Technology and Society Magazine. 2019;38:40–53.

[85]. Leslie D Understanding artificial intelligence ethics and safety. London: The Alan turing institute; 2019.

[86]. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine. 2019;17:195. [PubMed: 31665002]

[87]. Van Den Bos J, Rustagi K, Gray T, Halford M, Ziemkiewicz E, Shreve J. The $17.1 billion problem: the annual cost of measurable medical errors. Health affairs (Project Hope). 2011;30:596–603. [PubMed: 21471478]

[88]. Obermeyer Z, Lee TH. Lost in Thought - The Limits of the Human Mind and the Future of Medicine. The New England journal of medicine. 2017;377:1209–11. [PubMed: 28953443]

[89]. Dong J, Geng Y, Lu D, Li B, Tian L, Lin D, et al. Clinical Trials for Artificial Intelligence in Cancer Diagnosis: A Cross-Sectional Study of Registered Trials in ClinicalTrials.gov. Front Oncol. 2020;10:1629-. [PubMed: 33042806]

[90]. COMEST. Preliminary study on the ethics of Artificial Intelligence. USA: UNESCO 2019.

[91]. Light G. Race, Policing, and Detroit's Project Green Light. 2019.

[92]. Harmon A As Cameras Track Detroit's Residents, a Debate Ensues Over Racial Bias. NY Times2019.

[93]. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic acids research. 2018;46:10546–62. [PubMed: 30295871]

[94]. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC medical informatics and decision making. 2019;19:281. [PubMed: 31864346]

[95]. Levine AB, Schlosser C, Grewal J, Coope R, Jones SJM, Yip S. Rise of the Machines: Advances in Deep Learning for Cancer Diagnosis. Trends in cancer. 2019;5:157–69. [PubMed: 30898263]

[96]. Bernard O, Lalande A, Zotti C, Cervenansky F, Yang X, Heng P, et al. Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? IEEE transactions on medical imaging. 2018;37:2514–25. [PubMed: 29994302]

[97]. Hosny A, Parmar C, Coroller TP, Grossmann P, Zeleznik R, Kumar A, et al. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. PLoS medicine. 2018;15:e1002711–e. [PubMed: 30500819]

[98]. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nature genetics. 2019;51:12–8. [PubMed: 30478442]

[99]. Li Y, Shi W, Wasserman WW. Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. BMC bioinformatics. 2018;19:202. [PubMed: 29855387]

[100]. Lu J, Jin S, Liang J, Zhang C. Robust Few-Shot Learning for User-Provided Data. IEEE transactions on neural networks and learning systems. 2020.

[101]. Lentz S How to meet the challenges of auto recalls. SME Society of Mechanical Engineers; 2020.

[102]. Haibe-Kains B, Adam GA, Hosny A, Khodakarami F, Shraddha T, Kusko R, et al. Transparency and reproducibility in artificial intelligence. Nature. 2020;586:E14–E6. [PubMed: 33057217]

[103]. Miller DD, Brown EW. Artificial Intelligence in Medical Practice: The Question to the Answer? The American journal of medicine. 2018;131:129–33. [PubMed: 29126825]

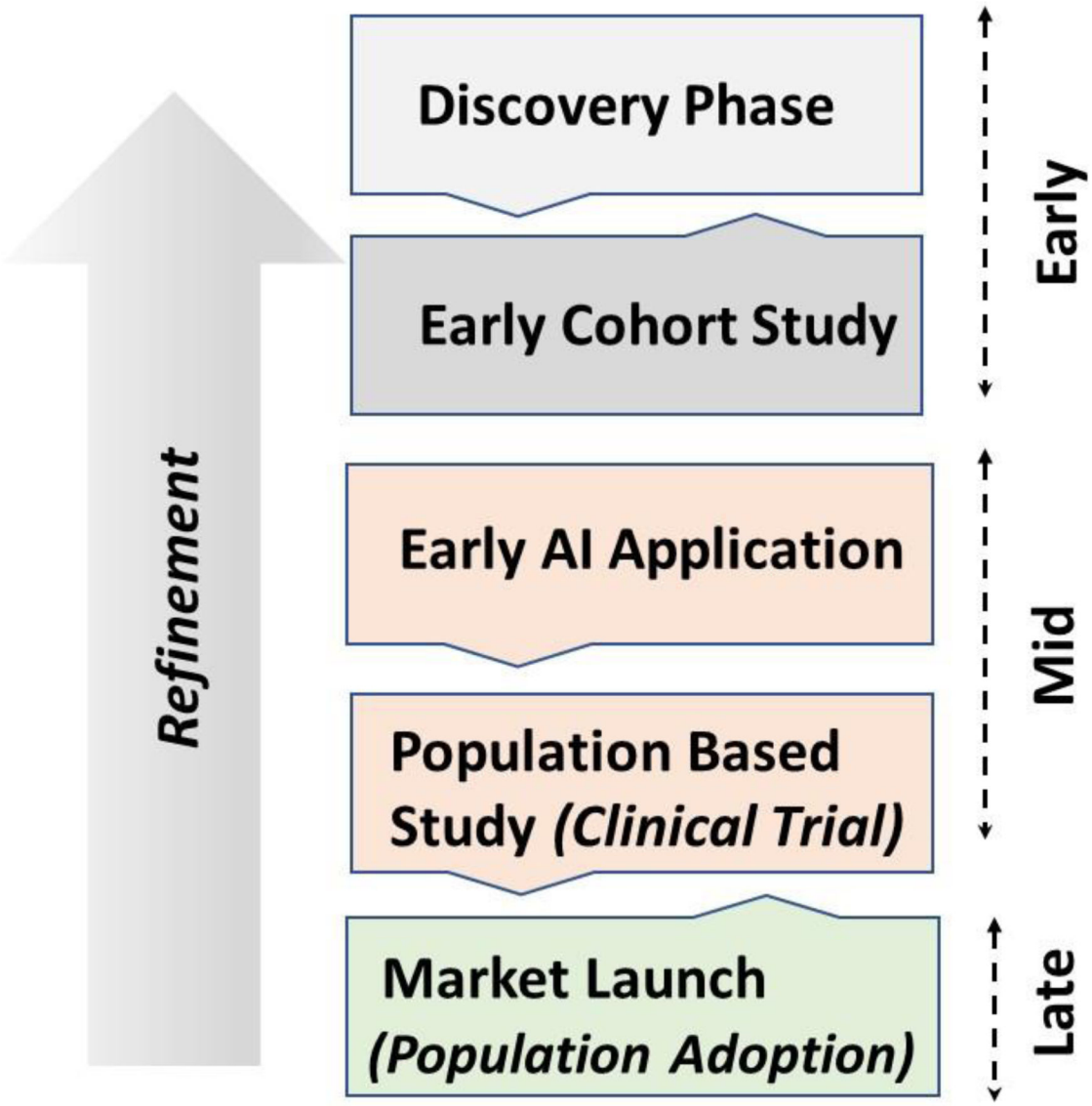

Figure 1.
Overview of different phases (early, mid, late) in developing a reliable AI application are outlined from discovery to market launch.