Perspective

Enabling the clinical application of artificial intelligence in genomics: a perspective of the AMIA Genomics and Translational Bioinformatics Workgroup

Nephi A. Walton, MD, MS, FACMG[1,*], Radha Nagarajan, PhD[2,3], Chen Wang (iD), PhD, FAMIA[4],
Murat Sincan, MD[5,6], Robert R. Freimuth, PhD[7], David B. Everman, MD, FACMG[8], Derek C. Walton, JD[9],
Scott P. McGrath, PhD, FAMIA[10], Dominick J. Lemas, PhD[11], Panayiotis V. Benos (iD), PhD[12],
Alexander V. Alekseyenko (iD), PhD[13], Qianqian Song (iD), PhD[11], Ece Gamsiz Uzun, PhD[14,15],
Casey Overby Taylor, PhD[16], Alper Uzun, PhD[15,17], Thomas Nate Person, MS[18],
Nadav Rappoport (iD), PhD[19], Zhongming Zhao (iD), PhD, MS, FACMI, FAMIA[20],
Marc S. Williams, MD, FACMG[21]

[1]Division of Medical Genetics, University of Utah School of Medicine, Salt Lake City, UT 84112 ,United States, [2]Enterprise Information Services, Cedars-Sinai Medical Center, Los Angeles, CA 90025, United States, [3]Information Services Department, Children's Hospital of Orange County, Orange, CA 92868, United States, [4]Division of Computational Biology, Department of Quantitative Health Sciences, Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, United States, [5]Flatiron Health, New York, NY 10013, United States, [6]Department of Internal Medicine, Sanford School of Medicine, University of South Dakota, Sioux Falls, SD 57107, United States, [7]Department of Artificial Intelligence and Informatics, Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, United States, [8]EverMed Genetics and Genomics Consulting LLC, Greenville, SC 29607, United States, [9]Walton Legal PLLC, Murray, UT 84123, United States, [10]CITRIS Health, CITRIS and Banatao Institute, University of California Berkeley, Berkeley, CA 94720, United States, [11]Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL 32610, United States, [12]Department of Epidemiology, University of Florida, Gainesville, FL 32610, United States, [13]Department of Public Health Sciences, Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC 29403, United States, [14]Department of Pathology and Laboratory Medicine, Rhode Island Hospital and Lifespan Medical Center, Providence, RI 02915, United States, [15]Department of Pathology and Laboratory Medicine, Warren Alpert Medical School of Brown University, Providence, RI 02915, United States, [16]Departments of Medicine and Biomedical Engineering, Johns Hopkins School of Medicine, Baltimore, MD 21205, United States, [17]Legorreta Cancer Center, Brown University, Providence, RI 02915, United States, [18]Department of Bioinformatics and Genomics, Huck Institutes of the Life Sciences, Penn State University, Bloomsburg, PA 16802, United States, [19]Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Beer Sheva 8410501, Israel, [20]Center for Precision Health, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, United States, [21]Department of Genomic Health, Geisinger, Danville, PA 17822, United States

*Corresponding author: Nephi A. Walton, MD, MS, FACMG, Division of Medical Genetics, University of Utah School of Medicine, 30 South 2000 East, Salt Lake City, UT 84112 (nephi.walton@utah.edu)

Abstract

Objective: Given the importance AI in genomics and its potential impact on human health, the American Medical Informatics Association—Genomics and Translational Biomedical Informatics (GenTBI) Workgroup developed this assessment of factors that can further enable the clinical application of AI in this space.

Process: A list of relevant factors was developed through GenTBI workgroup discussions in multiple in-person and online meetings, along with review of pertinent publications. This list was then summarized and reviewed to achieve consensus among the group members.

Conclusions: Substantial informatics research and development are needed to fully realize the clinical potential of such technologies. The development of larger datasets is crucial to emulating the success AI is achieving in other domains. It is important that AI methods do not exacerbate existing socio-economic, racial, and ethnic disparities. Genomic data standards are critical to effectively scale such technologies across institutions. With so much uncertainty, complexity and novelty in genomics and medicine, and with an evolving regulatory environment, the current focus should be on using these technologies in an interface with clinicians that emphasizes the value each brings to clinical decision-making.

Key words: clinical genomics; artificial intelligence; machine learning; bioinformatics; genomics; translational bioinformatics.

Introduction

Artificial Intelligence (AI) is revolutionizing several domains, from art where diffusion models like MidJourney are driving transformative changes,[1] to law,[2] education,[3] and medicine where large language models such as ChatGPT are reshaping the landscape. The pace of AI advancements in these areas suggests what might be possible in genomics if done properly. While the potential of AI for genomic discovery is

tremendous, there is also significant potential to build AI models that utilize genomic data to make complex genetic diagnoses,[4,5] personalize treatment plans,[6,7] and identify and prevent health risks.[8]

The inherent complexities of genomic data including large dimensionality and lack of prior knowledge regarding the functional aspects of most genomic variants present major challenges for AI application. The human haploid genome consists of over 3 billion bases, and susceptibility to disease is usually a complex interplay of these variants.[9] There is uncertainty in the exact number of protein-coding genes,[9] the functions of most of these genes, and the impact of individual coding variants on these genes and their products. This uncertainty is dwarfed by variants in noncoding DNA, the understanding of which is in its infancy. While a complete human reference genome was recently released,[9] many populations remain underrepresented,[10] necessitating the further development of reference genomes such as the recently released pangenome.[11] DNA is transcribed in different ways depending on the stage of development, the type of tissue, its physical location in the body, and its function. Therefore a single genetic variant can have a myriad of effects on RNA transcription, protein production, and the ultimate cellular and phenotypic consequences, as is evidenced in many Mendelian disorders.[12,13] Finally, redundancy of compensatory pathways may play a role in penetrance of any specific variant; whereby a variant in one individual will have a deleterious effect but may have low or no effect in another individual based on the cumulative set of genetic and environmental factors acting on these individuals. In contrast to Mendelian diseases, most common diseases are orchestrated through variation in multiple genes with complex interactions.[14,15] This complexity underlies the concept of polygenic risk scores (PRS).[16]

Despite these challenges, the complexity and magnitude of genomic data make it an attractive target for AI, although clinical application adds to the challenges already identified. To address these, the American Medical Informatics Association Genomics and Translational Bioinformatics Workgroup (AMIA GEN-TBI), has identified a comprehensive list of factors that are crucial for advancing the adoption and clinical use of AI in genomics.

Factors facilitating clinical use of AI in genomics

These factors, grouped into 5 categories—Data Availability and Accessibility, Technical Aspects of Genomic Data, Ethical and Analytical Considerations, Stakeholder Engagement and Education, and Regulatory Frameworks—provide a roadmap for the successful clinical deployment of AI in genomics. Each category represents a critical aspect of the broader ecosystem that must be addressed to fully leverage the potential of AI in genomics.

Data availability and accessibility
Growth of genomic data
The cost of genome sequencing (GS) continues to decrease, from ~$300 million for the first genome[17] to a recently announced $100 genome.[18] The decrease in cost, along with the large number of population sequencing initiatives, has dramatically increased the amount of sequencing data available, making it easier to investigate genomic data with AI approaches. There are also increased efforts in the United States

to collect genomic data that better reflect the racial and ethnic diversity of the population such as the All of Us project[19] and TOPMed.[20] These efforts are expected to assist in developing AI models with enhanced generalizability, minimizing health disparities related to the current data gaps. We are also seeing an acceleration in the clinical adoption of exome sequencing and GS for diagnostic purposes, as these are now recommended as first-line tests for several groups of disorders[21,22] and are being investigated for use in newborn screening.[23]

Even with all these developments we do not have the number of sequences that would allow us to see the kind of progress being made with AI applied to text or visual images. While we are approaching the availability of millions of genomes, ChatGPT was trained on billions of text files scraped from the internet,[24] and MidJourney (a diffusion model for art) was trained on over 100 million images.[25] These domains have much more available data and are arguably less complex than genomic data particularly when one includes the associated healthcare data. Traditional research funding mechanisms have modest caps on costs which limits how much sequencing can be done in an investigator-initiated project. We argue that with lower sequencing costs there should be larger government funded sequencing initiatives that also include pertinent health information and require a degree of data sharing.

Democratization of data
Despite the existence of several large repositories of genomic data, the usability of the data is limited as much of it is siloed and accessible only by a few who have the resources to mine and extract its value.[26,27] Sharing genomic data are critical for improved statistical power. Moreover, it is expected to accelerate AI research, adoption, and crowd-sourcing efforts. Initiatives such as "All of Us" facilitate genomic data sharing.[28] This initiative currently has more than 250 000 genomes with a significant amount of racial and ethnic diversity.[19] The UK Biobank is another large genomic data sharing initiative.[29] These efforts have fueled data-driven approaches with potential AI applications that would have otherwise been impossible.[30,31] While we advocate for the continuation and expansion of these efforts, we also posit that there might be a more pressing need to develop government supported federated architectures for genomic data sharing. These structures could facilitate more efficient sharing of existing genomic datasets, all while safeguarding patient and organizational privacy. We believe such federated networks should be deployed in a way that allows for global data sharing as the data needed to drive this field are tremendous as is the need for diversity.

Clinical availability of genomic data
Clinical next generation sequencing (NGS) data are generated by CLIA-certified laboratories and only a summary of "actionable variants" is available at the point of care, usually in the form of a PDF.[32] A comprehensive adoption of NGS data would demand changes to clinical and laboratory processes surrounding genomics data, requiring a seamless interface to the larger dataset generated by the lab. Genomic data have longitudinal value that can be utilized at different points throughout an individual's healthcare journey. Therefore, it would make economic sense to separate the processes of sequencing and interpretation.[33] This would require an infrastructure that supports storage, query, and secondary analysis of the genomic data by both the clinic and competing

laboratories. Such infrastructure would facilitate exploratory analysis and empower patients and clinicians to interpret datasets with the aid of informatics tools. This approach must fall within the guidelines of established regulations to avoid misinterpretation. Currently, only genetic information delivered in a summary report from a CLIA laboratory can be acted on clinically. A framework is required to certify all information from the genome for clinical use. This would allow the clinical use of multiple PRS and encourage development of context-specific AI models for clinical decision support (CDS) while validating those reported by vendors.[33]

Technical aspects of genomic data
Functional genomics

Having prior knowledge about gene function can reduce the search space by focusing on variants in a context-specific manner. Resources such as ClinVar, which compiles community-contributed information about pathogenicity of variants and employs variant curation expert panels, has demonstrated tremendous benefit in deciphering clinical value from genomic data.[34] Large-scale functional studies, and a recent announcement by the NIH to launch the Molecular Phenotypes of Null Alleles in Cells (MorPhiC) program to establish the function of every human gene,[35] will generate significant improvements in functional knowledge and open venues for the application of AI in a context-specific manner. Continued efforts to uncover clinical impacts of genes and variants will improve our ability to build AI models with smaller datasets. While we are supportive of these efforts, we would argue that having larger datasets including genomic, transcriptomic, and clinical data, would enable the use of AI to more rapidly elucidate gene function.

Genomic data standards

Efforts continue to standardize the application and reporting of discrete genomic data and knowledge,[36–39] including its use in algorithms such as PRS.[40] Standards development should also support the integration of genomic data and derived knowledge (eg, risk scores) into the EHR. From an AI context, standardization is critical to ensuring the development of accurate and portable models. Standards should be established to both structure the input data to ensure consistency across health systems and set quality metrics to validate the data utilized in the models. Many standards currently exist to support genomic data. Global Alliance for Genomics and Health (GA4GH) is developing standards for the computable representation of genomic variations and knowledge,[36–39] which could enable the development of AI models. HL7 FHIR is widely used in EHRs and is primarily used for clinical reporting of genomic test results; the FHIR genomics specification is currently expanding to support more nuanced data types, improved harmonization with GA4GH standards, and improved computability of data. The ClinGen Allele Registry is actively creating global canonical IDs for genomic variants with mapping to other nomenclatures to reconcile competing variant nomenclatures.[41]

EHR integration

Efforts continue to integrate genomic data in EHRs in a structured format[32,42] but are primarily focused on specific variants and summaries that drive phenotypes as opposed to genome-level representation. Genome Archiving and Communication Systems (GACS)[43] interfaced to the EHR through standards-based interfaces, such as SMART on FHIR, would allow for use of large genetics datasets external to the EHR that can be ingested by AI approaches. Integrated AI platforms that can facilitate clinical use of genomic data in a real-time environment are critical to moving this field forward.

Ethical and analytical considerations
Privacy

While traditionally, large training datasets for AI algorithms have been based on real patient data, the landscape is evolving. Researchers are exploring the use of artificially generated data and leveraging pretrained algorithms to enhance the training process and overcome limitations associated with real patient data.[44–46] There are particularly sensitive aspects of genomics data that warrant special protections as addressed by recent guidance from the American College of Medical Genetics and Genomics; these include but are not limited to consanguinity, misattributed parentage, and presymptomatic test results.[47] Despite significant advancements in data security in the public cloud computing settings, there is still concern about uploading genomic data into the cloud environments. Even without traditional identifiers, genomic data can potentially be reidentified.[48] Genomic data must be protected, and algorithms deployed in a way that preserves patient privacy using techniques such as federated learning[49] and swarm learning.[50]

Accuracy

The accuracy of the output of ML algorithms is directly related to the accuracy of the input data. The genomic data used in such algorithms should be generated in a CLIA-certified lab to ensure data integrity and be subject to an acceptable threshold of data quality standards prior to clinical use. Equally important, the data must be expressed in semantically unambiguous and computable forms. Such standards need to be established as we venture into using genomic data outside of the clinical report.

Bias

The issue of bias in AI is of critical importance, particularly in healthcare where it can significantly impact patient outcomes. Biases can occur at various stages, from data collection to algorithm development and interpretation. Bias and variance can occur during the training phase of AI, often a result of underfitting or overfitting data. Regularization of these approaches can assist in minimizing bias-variance. This is especially critical in genomics data, where the number of features or dimensions is orders of magnitude higher than the sample size. As with all observational data, genomic data are prone to selection bias attributed to nonrandom distribution of the subjects between cohorts of interest (eg, case–control studies). Low representation of diverse populations in the data can also contribute to bias and lead to AI models with the potential to exacerbate inequities. Similarly, nonuniform representation can also lead to class-imbalance that may encourage prudent choice of performance measures in evaluating AI approaches. We encourage continued efforts to create large diverse genomic datasets with corresponding clinical outcome information on which AI models can train that more fully represent the population being served. Such datasets should be shared across federated networks to allow for building of complex AI models while safeguarding privacy, particularly for vulnerable populations. In addition to a more diverse and representative data collection, strategies to

mitigate these biases include the use of bias correction algorithms, and increased transparency in AI development.

Stakeholder engagement and education
Clinician education

AI/ML is increasingly prevalent in common consumer devices. While the fundamental mechanisms of AI/ML are well-understood, the challenge lies in effectively addressing the complexities and problems these mechanisms present. Ignorance can lead to mistrust and unrealistic expectations. It is critical that providers understand AI limitations and do not use it to override sound clinical judgment. Educating clinicians in this field is crucial as an understanding of the benefits and limitations of AI will increase trust and adoption.[51]

Explainability, transparency, and interpretability

Clinicians are less likely to adopt or utilize systems they do not understand.[52] Explainability in AI[53,54] is not just about addressing the "black box" problem. It is also about ensuring that AI models are interpretable and understandable to clinicians, patients, and regulators. Approaches to improving explainability include the use of interpretable models, explanation interfaces that provide insights into the model's decision-making process, and *post hoc* explanation techniques that explain the model's decisions after they are made. A clear rationale for their choice, along with understanding of underlying assumptions in a context-specific manner, can improve their adoption. AI models developed for clinical use should strive for transparency and interpretability that facilitate caregiver trust and allow caregivers to utilize the models' output in conjunction with their clinical judgment.

Regulatory frameworks

The FDA's existing framework does not account for adaptive AI approaches. Under current regulations, a manufacturer must gain additional clearance each time there is a significant change to a device.[55] In January 2021, the FDA issued the "Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan" from the Center for Devices and Radiological Health's Digital Health Center of Excellence.[56] In the SaMD Action Plan, the FDA proposed a new regulatory framework tailored to these new technologies. However, to date, no new regulatory framework has been developed. Recent FDA guidance regarding CDS may exempt such models if they are used as support to augment physician judgment but are not the primary factor on which the clinical decisions are based.[57] As new regulatory frameworks evolve, balance is needed between regulation and patient safety without stifling innovation through overregulation of AI approaches.[58,59]

Discussion

The factors presented inform challenges and opportunities that shape the application of AI to genomics. As we begin to implement AI applications in genomics, the health informatics community should develop best-practice guidelines, ideally based on an expanded research agenda[60] and a rigorous evidence-review process. Engagement should promote development of AI applications that are: (1) evaluated through best practices in a clinical environment, (2) transparent in contrast to black-box offerings, and (3) reproducible and generalizable across diverse settings. Algorithms should not introduce, reinforce, or exacerbate socio-economic or ethnic/racial biases that could lead to health disparities and inequity of care. It is also important to ensure that AI tools will protect patient privacy and information security. AI's legal and regulatory implications must be studied to ensure safe and adequate deployment if its tools are used in the treatment of patients and should be subject to appropriate oversight.

Active involvement of the clinical genomics and medical informatics communities is critical for successful validation and deployment of AI in genomics. AI algorithms implicitly subscribe to optimizing a chosen objective function. Thus, it is common for these algorithms to converge to locally optimal solutions that might not necessarily be the best solution overall. Therefore, domain experts in genetics will be integral to the implementation team and assisting patients in interpreting results, especially when involving clinical recommendations.

While it is widely accepted that AI deployments for patient care should outline known limitations of the algorithms, establish generalizability, and incorporate human expertise, we believe that these principles need to be reinforced with specific actions.[61] AI developers should provide detailed documentation of their algorithms' limitations and performance across different settings. Healthcare organizations should establish protocols for human oversight of AI decisions, ensuring that expert clinicians are involved in critical decision-making processes. Moreover, we advocate for rigorous, independent validation of AI algorithms in diverse real-world settings before they are deployed in patient care. As with any data source used for clinical decision making, AI results should be used with caution and always in the context of the patient's current health. Additionally, much genetic data can be augmented with other patient information to increase utility and adoption of these algorithms.

Given the limited knowledge surrounding the clinical impact of genetic variation, validation of the results generated by AI models requires critical assessment by domain experts. AI implementations should have enough transparency for validation across diverse settings and solicit feedback from clinicians which would support iterative improvement through creation of a virtuous cycle, a concept highlighted in the Institute of Medicine (now National Academy of Medicine) publication on the Genomics-enabled Learning Health Care System.[62] Although AI has potential to significantly impact genomics, dimensionality of genomic data necessitates larger sample sizes for optimal performance of these algorithms.[63] For AI models to discern complex patterns, detect uncommon genetic variations, and extrapolate results across varied populations, substantial datasets are essential. This becomes even more significant given the vast number of clinical outcomes associated with the genomic feature set. Without sufficient data, there is a heightened risk of models overfitting to specific examples, potentially overlooking crucial genetic nuances and compromising their real-world predictive precision and applicability. The use of large datasets requires advanced computing infrastructures for effective data processing. Ongoing national initiatives, in conjunction with standards in sharing genomic data, can assist in generating large sample sizes optimized for diversity while promoting enhanced transparency and external validation of the AI algorithms and minimizing bias. Commercial AI implementations in genomics data, with the stated purpose of recommending medications, supplements, treatment plans, and life changes to improve health outcomes, should clearly acknowledge known limitations and demonstrate evidence of utility.

As observed in other fields, the progress achievable in AI and genomics is proportional to the volume of data available for training our models. Even if existing initiatives like the "All of Us" program, which aims to sequence a million individuals,[19] achieve their targets, they would still not offer the volume of data required for comparable advancements in this field. Government-backed or private initiatives, and public–private partnerships that fund both sequencing and the development of infrastructures for responsible data sharing are pivotal for advancements in this domain. Maybe it is time for a second human genome project, emphasizing the volume of both genomic and clinical data, and engaging institutions nationwide with data sharing enabled through a federated network. While the scope of such a project would be vast, the outcomes could exceed that of the original endeavor.

Conclusion

Genomics based AI models have tremendous potential to impact patient care. Additional informatics research and development are needed to fully realize the clinical impact of such technologies. To achieve a similar impact to that seen with AI in other domains we must create larger and more diverse training datasets, establish responsible frameworks for data sharing, and standardized approaches for analysis of AI tools. With so much uncertainty, complexity, and novelty in genomics and medicine, and with an evolving regulatory space, the current focus should be on using these technologies to create an interface with clinicians that emphasizes the value each brings to clinical decision-making in healthcare.

Author contributions

N.A.W. and R.N. conceived of the presented idea. All authors contributed to discussions leading to the outline and development of the manuscript. All authors contributed to the writing of the final manuscript.

Funding

Conflicts of interest

None declared.

Data availability

No new data were generated or analyzed in support of this research.

References

1. Borji A. Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2. 2023. arXiv, arXiv:2210.00586v2, preprint: not peer reviewed.

2. UK Parliament. *How Will AI Large Language Models Shape the Future and What Is the Right Regulatory Approach?* Accessed July 7, 2023. https://www.parliament.uk/business/lords/media-centre/house-of-lords-media-notices/2023/july-2023/how-will-ai-large-language-models-shape-the-future-and-what-is-the-right-regulatory-approach/.

3. Kasneci E, Sessler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn Individ Differ*. 2023;103:102274.

4. Johnson KB, Wei WQ, Weeraratne D, et al. Precision medicine, AI, and the future of personalized health care. *Clin Transl Sci*. 2021;14(1):86-93.

5. Elemento O, Leslie C, Lundin J, Tourassi G. Artificial intelligence in cancer research, diagnosis and therapy. *Nat Rev Cancer*. 2021;21(12):747-752.

6. Sotoudeh H, Shafaat O, Bernstock JD, et al. Artificial intelligence in the management of glioma: era of personalized medicine. *Front Oncol*. 2019;9:768.

7. Badic B, Hatt M, Durand S, et al. Radiogenomics-based cancer prognosis in colorectal cancer. *Sci Rep*. 2019;9(1):9743.

8. Subramanian M, Wojtusciszyn A, Favre L, et al. Precision medicine in the era of artificial intelligence: implications in chronic disease management. *J Transl Med*. 2020;18(1):1-12.

9. Nurk S, Koren S, Rhie A, et al. The complete sequence of a human genome. *Science* 2022;376(6588):44-53.

10. Sirugo G, Williams SM, Tishkoff SA. The missing diversity in human genetic studies. *Cell* 2019;177(1):26-31.

11. Liao W-W, Asri M, Ebler J, et al. A draft human pangenome reference. *Nature* 2023;617(7960):312-324.

12. Reiners J, Nagel-Wolfrum K, Jürgens K, Märker T, Wolfrum U. Molecular basis of human Usher syndrome: deciphering the meshes of the Usher protein network provides insights into the pathomechanisms of the Usher disease. *Exp Eye Res*. 2006;83(1):97-119.

13. Ritelli M, Colombi M. Molecular genetics and pathogenesis of Ehlers–Danlos syndrome and related connective tissue disorders. *Genes (Basel)* 2020;11:547.

14. Hickey SL, McKim A, Mancuso CA, Krishnan A. A network-based approach for isolating the chronic inflammation gene signatures underlying complex diseases towards finding new treatment opportunities. *Front Pharmacol*. 2022;13:995459.

15. Jackson M, Marks L, May GH, Wilson JB. The genetic basis of disease. *Essays Biochem*. 2018;62(5):643-723.

16. Kumuthini J, Zick B, Balasopoulou A, et al.; G2MC Evidence Investigators. The clinical utility of polygenic risk scores in genomic medicine practices: a systematic review. *Hum Genet*. 2022;141(11):1697-1704.

17. International Human Genome Sequencing Consortium. *International Human Genome Sequencing Consortium Publishes Sequence and Analysis of the Human Genome*. Updated Feburary 12. Accessed July 7, 2023. https://www.genome.gov/10002192/2001-release-first-analysis-of-human-genome.

18. Pennisi E. A $100 genome? New DNA sequencers could be a 'game changer' for biology, medicine. *Scienceorg*. 2022;376(6599):1257-1258. https://doi.org/10.1126/science.add5060

19. Ramirez AH, Sulieman L, Schlueter DJ, et al. The All of Us Research Program: data quality, utility, and diversity. *Patterns* 2022;3(8):100570.

20. Burgess DJ. The TOPMed genomic resource for human health. *Nat Rev Genet*. 2021;22(4):200-200.

21. Smith L, Malinowski J, Ceulemans S, et al. Genetic testing and counseling for the unexplained epilepsies: an evidence-based practice guideline of the National Society of Genetic Counselors. *J Genet Couns*. 2023;32(2):266-280.

22. Manickam K, McClain MR, Demmer LA, et al.; ACMG Board of Directors. Exome and genome sequencing for pediatric patients with congenital anomalies or intellectual disability: an evidence-based clinical guideline of the American College of Medical Genetics and Genomics (ACMG). *Genet Med*. 2021;23(11):2029-2037.

23. Buxton J. Whole genome sequencing at birth: genomic data, a resource from cradle to grave. *BioNews* 2022;21:758.

24. Heaven W. *ChatGPT is Everywhere. Here's Where It Came From*. Accessed August 23, 2023. https://www.technologyreview.com/2023/02/08/1068068/chatgpt-is-everywhere-heres-where-it-came-from/.

25. Salkowitz R. *Midjourney Founder David Holz on the Impact of AI on Art, Imagination and the Creative Economy*. Forbes. Accessed August 23, 2023. https://www.forbes.com/sites/robsalkowitz/2022/09/16/midjourney-founder-david-holz-on-the-impact-of-ai-on-art-imagination-and-the-creative-economy.

26. Carey DJ, Fetterolf SN, Davis FD, et al. The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. *Genet Med*. Sep 2016;18(9):906-913. https://doi.org/10.1038/gim.2015.187

27. Walton NA, Hafen B, Graceffo S, et al. The development of an infrastructure to facilitate the use of whole genome sequencing for population health. *J Pers Med*. 2022;12(11):1867. https://doi.org/10.3390/jpm12111867

28. All of Us Research Program Investigators. The "All of Us" research program. *N Engl J Med*. 2019;381(7):668-676.

29. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12(3):e1001779.

30. All of US. *All of Us Publications*. Updated June 10, 2023. Accessed July 13, 2023. https://www.researchallofus.org/publications/.

31. UK Biobank. *Publications*. Updated November 29, 2022. Accessed July 13, 2023. https://www.ukbiobank.ac.uk/enable-your-research/publications.

32. Walton N, Johnson D, Heale B, Person T, Williams M. *Creating a Home for Genomic Data in the Electronic Health Record*. American Medical Informatics Association; 2021:1196.

33. Walton NA, Christensen GB. Paving a pathway for large-scale utilization of genomics in precision medicine and population health. *Front Sociol*. 2023;8:1122488.

34. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*. 2014;42(Database issue):D980-D985.

35. Soo S. *NIH Initiative to Systematically Investigate and Establish Function of Every Human Gene*. Genome.gov. Updated September 27, 2022. Accessed February 5, 2023. https://www.genome.gov/news/news-release/nih-initiative-to-systematically-investigate-and-establish-function-of-every-human-gene.

36. Goar W, Babb L, Chamala S, et al. *Development and Application of a Computable Genotype Model in the GA4GH Variation Representation Specification*. World Scientific; 2022:383-394.

37. Wagner AH, Babb L, Alterovitz G, et al. The GA4GH variation representation specification: a computational framework for variation representation and federated identification. *Cell Genom*. 2021;1(2):100027.

38. Murugan M, Babb LJ, Taylor CO, et al. Genomic considerations for FHIR®; eMERGE implementation lessons. *J Biomed Inform*. 2021;118:103795.

39. Garcia SJ, Zayas-Cabán T, Freimuth RR. Sync for genes: making clinical genomics available for precision medicine at the point-of-care. *Appl Clin Inform*. 2020;11(2):295-302.

40. Wand H, Lambert SA, Tamburro C, et al. Improving reporting standards for polygenic scores in risk prediction studies. *Nature* 2021;591(7849):211-219.

41. Pawliczek P, Patel RY, Ashmore LR, et al.; Clinical Genome (ClinGen) Resource. ClinGen Allele Registry links information about genetic variants. *Hum Mutat*. 2018;39(11):1690-1701.

42. Bielinski SJ, Olson JE, Pathak J, et al. *Preemptive Genotyping for Personalized Medicine: Design of the Right Drug, Right Dose, Right Time—Using Genomic Data to Individualize Treatment Protocol*. Elsevier; 2014:25-33.

43. Brunner M, Jauk F, Candenas N, Cancio A, Luna D, Benitez S. Building a genome archiving and communication system integrated into a health information systems. *Stud Health Technol Inform*. 2022;290:1036-1037.

44. Erol T, Mendi AF, Doğan D. *The Digital Twin Revolution in Healthcare*. IEEE; 2020:1-7.

45. Gao C, Killeen BD, Hu Y, et al. Synthetic data accelerates the development of generalizable learning-based algorithms for X-ray image analysis. *Nat Mach Intell*. 2023;5(3):294-308.

46. Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. *BMC Med Res Methodol*. 2020;20(1):108-140.

47. Grebe TA, Khushf G, Chen M, et al.; ACMG Social, Ethical and Legal Issues Committee. The interface of genomic information with the electronic health record: a points to consider statement of the American College of Medical Genetics and Genomics (ACMG). *Genet Med*. 2020;22(9):1431-1436.

48. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. *Science* 2013;339(6117):321-324.

49. Lu MY, Chen RJ, Kong D, et al. Federated learning for computational pathology on gigapixel whole slide images. *Med Image Anal*. 2022;76:102298.

50. Warnat-Herresthal S, Schultze H, Shastry KL, et al.; Deutsche COVID-19 Omics Initiative (DeCOI). Swarm learning for decentralized and confidential clinical machine learning. *Nature* 2021;594(7862):265-270.

51. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. *J Med Internet Res*. 2020;22(6):e15154.

52. Beglaryan M, Petrosyan V, Bunker E. Development of a tripolar model of technology acceptance: hospital-based physicians' perspective on EHR. *Int J Med Inform*. 2017;102:50-61.

53. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisc Rev Data Min Knowl Discov*. 2019;9(4):e1312.

54. Amann J, Blasimme A, Vayena E, Frey D, Madai VI; Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak*. 2020;20(1):310-319.

55. FDA. *Guidance for Industry and FDA Staff Modifications to Devices Subject to Premarket Approval (PMA) – The PMA Supplement Decision-Making Process*. Updated December 11, 2008. Accessed February 7, 2023. https://www.fda.gov/media/73328/download.

56. FDA. *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan*. Updated January 1, 2021. Accessed February 7, 2023. https://www.fda.gov/media/145022/download.

57. FDA. *Clinical Decision Support Software—Guidance for Industry and Food and Drug Administration Staff*. Updated September 28, 2022. Accessed February 7, 2023. https://www.fda.gov/media/109618/download.

58. Price W, Nicholson I. Regulating black-box medicine. *Mich Law Rev*. 2017;116(116.3):421.

59. Petersen C, Smith J, Freimuth RR, et al. Recommendations for the safe, effective use of adaptive CDS in the US healthcare system: an AMIA position paper. *J Am Med Inform Assoc*. 2021;28(4):677-684.

60. Wiley K, Findley L, Goldrich M, et al. A research agenda to support the development and implementation of genomics-based clinical informatics tools and resources. *J Am Med Inform Assoc*. 2022;29(8):1342-1349.

61. Wiley KK, Mendonca E, Blackburn J, Menachemi N, De Groot M, Vest JR. Quantifying electronic health record data quality in telehealth and office-based diabetes care. *Appl Clin Inform*. 2022;13(5):1172-1180.

62. Roundtable on Translating Genomic-Based Research for Health; Board on Health Sciences Policy; Health and Medicine Division; National Academies of Sciences, Engineering, and Medicine. *Applying an Implementation Science Approach to Genomic Medicine: Workshop Summary*. National Academies Press (US); 2016.

63. Berisha V, Krantsevich C, Hahn PR, et al. Digital medicine and the curse of dimensionality. *NPJ Digit Med*. 2021;4(1):153.