Perspective on Bias in Biomedical AI: Preventing Downstream Healthcare Disparities

Michal Rosen-Zvi, Yoav Kan-Tor, Michael Danziger, Agata Ferretti, Javier Aula-Blasco, Júlia Falcão, Ron Shamir, Mira Marcus-Kalish, Mordechai Muszkat

IBM Research - IsraelFaculty of Medicine, The Hebrew UniversityIsrael Michal.Rosen-Zvi@mail.huji.ac.il · IBM Research - IsraelIsrael yoavkt@gmail.com · IBM Research - IsraelIsrael Michael.Danziger@ibm.com · IBM Research ZurichSwitzerland Agata.Ferretti@ibm.com · Barcelona Supercomputing CenterSpain javier.aulablasco@bsc.es · Barcelona Supercomputing CenterSpain julia.falcao@bsc.es · Tel Aviv UniversityIsrael rshamir@tauex.tau.ac.il · Tel Aviv UniversityIsrael miram@tauex.tau.ac.il · Hadassah University HospitalIsraelmordechai.muszkat@mail.huji.ac.il

ABSTRACT Healthcare disparities persist across socioeconomic boundaries, often attributed to unequal access to screening, diagnostics, and therapeutics. However, this perspective highlights that critical biases can emerge much earlier, during data collection and research prioritization, long before clinical implementation, particularly in studies focused on molecular and omics data. A vast number of studies focus on collecting omics data, but the demographic information associated with these datasets is often not reported, and when it is reported, it reveals substantial biases. An automated analysis of 4514 PubMed-indexed omics publications from 2015 to 2024, examining reporting across multiple demographic dimensions, reveals limited reporting overall; for example, only 2.7% of studies report ancestry or ethnicity information and geographic origin reporting is limited to 2.5%. Analysis of large-scale datasets commonly used for model training, such as CellxGene and GEO, reveals substantial population bias where European-ancestry data dominates. As biomedical foundation models become central to biomedical discovery with a paradigm in which base models are pretrained on large datasets and reusing them repeatedly for many different downstream tasks, they risk perpetuating or amplifying these early-stage biases, leading to cascading inequities that regulatory interventions cannot fully reverse. We propose a community-wide focus on three foundational principles: Provenance, Openness, and Reliability through Evaluation Transparency. Together, these principles can help make biases and limitations more visible to model developers and users, supporting more informed model development, evaluation, and deployment decisions in biomedical AI.

1 I Introduction

Healthcare disparities are especially visible among the aging population, with differences in lifespan and quality of life largely shaped by socioeconomic status and geographic location. Older adults in wealthier areas tend to live longer, healthier lives compared to those in lower-income or underserved regions [49]. A recent study examining the elderly population in the U.S. revealed biological differences underlying racial and ethnic disparities in health, showing accelerated aging among ethnic groups from lower socioeconomic backgrounds [18]. Socioeconomic status is closely linked to healthy aging. Common explanations for why greater wealth improves health outcomes in later life include reduced stress, lower trauma exposure, decreased allostatic load, and better access to timely, appropriate healthcare in higher socioeconomic groups [35].

While considerable attention has been given to addressing healthcare disparities at the point of care, including seminal findings that widely used clinical algorithms can encode significant racial bias [41], less focus has been directed toward upstream biases in the research enterprise that ultimately influence which therapeutic options become available and what guidance is provided when administering them. It is well established that, due to the high cost of drug development, the industry operates within an incentive system that prioritizes conditions affecting populations with greater ability to pay [50]. The primary force behind the industry is to create and deliver a continuous stream of novel, patented, and medically attractive drugs while capturing commercial value. Due to the enormous costs associated with drug development, companies often face challenges in remaining cost-effective. Only in recent years has the high-quality availability and advanced integration of human genetics and multi-omics technologies, enabling a bet-

ter understanding of disease mechanisms, begun to drive a new trend toward improved returns on investment [45].

We focus on the biases embedded in foundation models and their potential impact as these technologies begin to shape biomedical research and, increasingly, the broader medical domain. These models are driving new paradigms in medical AI, such as the generalist medical AI paradigm [37], which seeks to replace task-specific models with a unified system capable of handling diverse clinical tasks across multiple data modalities, and the virtual cell paradigm [6], which seeks to simulate cellular behavior using generative approaches. However, the foundational datasets used to train these models often exhibit significant geographic and socioeconomic biases. For instance, individuals of European ancestry account for over 78% of participants in genome-wide association studies, despite representing only 16% of the global population [47]. If such biases are not addressed at the source, they may propagate through the research pipeline, leading to inequities that downstream interventions, even if technically sound, cannot fully correct.

This perspective is structured as follows. We begin by introducing the role of biomedical foundation models in discovery. Next, we examine the datasets these models rely on and the biases they may carry. We then discuss how explainable AI (XAI) and benchmarking tools can help assess model robustness and uncover bias. Finally, we propose a practical framework to address these challenges.

2 II The Role of Biomedical Foundation Models in Biomedical Research

Biomedical research focuses on understanding health and disease and finding remedies to diseases. The rationale behind the drug discovery process often involves a series of critical research steps, beginning with the identification and validation of disease-relevant targets, followed by the development and optimization of therapeutic molecules.

During clinical development, the safety and efficacy of the drug candidates are rigorously tested. Biomarkers often play a crucial role in identifying responsive patient subgroups, tailoring dosing regimens based on genetic profiles, and enabling precise patient stratification—all of which increase the likelihood of successful clinical trials while minimizing the risk of adverse effects. Since the completion of the first human genome sequence over two decades ago through the Human Genome Project, the landscape of drug discovery has been fundamentally reshaped by genomics, with a 2021 study counting 7712 approved and experimental pharmaceuticals associated with the advances initiated by the project [20]. Genetic insights now play a pivotal role not only in identifying new drug targets, but also in predicting drug efficacy and safety across diverse populations. The growing recognition that genetic variation significantly influences medication response—both in terms of effectiveness and risk of adverse events—has fundamentally transformed pharmaceutical development strategies. The drug Warfarin provides a compelling illustration of this principle. As the most widely prescribed oral anticoagulant globally, Warfarin requires precise dosing due to its narrow therapeutic range and highly variable patient requirements. Although algorithms incorporating genetic information have been developed to optimize dosing, they are predominantly based on European populations. For example, genetic variants that effectively explain Warfarin dose variability in Europeans account for substantially less variability in patients of African descent. As a result, dosing algorithms derived from European datasets often fail to provide safer and more effective treatment across diverse ethnic groups [4]. Especially for complex diseases such as Alzheimer's, it has been shown that some treatments may affect certain ancestry groups more than others [58]. This underscores the urgent need to identify genetic variants influencing drug metabolism across global populations to support the development of universally effective therapies.

Recent advances in omic technologies—particularly the emergence of single-cell RNA sequencing (scRNA-seq)—are adding powerful new layers of resolution to biomedical research. These technologies are reshaping approaches ranging from disease mechanism discovery and target identification to lead optimization and biomarker discovery. By pairing cutting-edge computational tools with vast public datasets, scRNA-seq enables a deeper, more precise understanding of disease biology at the cellular level, transforming how therapeutic targets are identified and validated [52].

Gene expression levels are influenced by genetic mutations in enhancer and promoter regions, as studied in transcriptome-wide association studies (TWAS) and expression quantitative trait loci (eQTL) analyses [7]. Transcriptomic data has been shown to vary by ancestry group, both in healthy and cancerous cells [27, 3]. Integrating multiple ancestry groups in TWAS analyses can improve statistical robustness [10]. Epigenetic effects also impact cellular activity. Unlike genetic mutations, epigenetic modifications—such as DNA methylation, chromatin remodeling, nucleosome positioning, and changes in noncoding RNA profiles—are reversible and play a fundamental role in regulating gene expression.

This wealth of biological information, whether available as free text in books and journals or captured in structured forms (including genes, their variants and disease associations, proteins, protein–protein interactions, small molecules, transcriptomic data, and more), has been leveraged by numerous research groups to develop large-scale foundation models trained on uni- and multi-modal data. Examples include large language models (LLMs) trained on extensive biological corpora, such as BioGPT [32] and BioBERT [28]; models trained on massive protein datasets to enable structure and function prediction, such as ESM-3 [25] and AlphaFold 3 [1]; models trained on DNA sequences and genomic variation, such as DNABERT [26]; models trained on human transcriptomic profiles (e.g.,

scBERT [55] and BMFM-RNA [15]) or epigenomic profiles (e.g., MethylGPT [56]); and models spanning multiple domains, including multimodal and cross-species models (see, e.g., [46, 59, 40]).

These foundation models are increasingly used to address downstream tasks in biomedical research, including understanding disease mechanisms, discovering biomarkers for prognosis, and supporting the drug discovery process. They are becoming valuable resources in their own right. The community continues to progress toward broader data coverage and the integration of tools to create more powerful multi-omic foundation models [14, 13], alongside streamlined environments that enable agentic workflows, where intelligent agents autonomously deploy and refine these models [60, 29, 62].

Recent work suggests that agentic AI in bioinformatics can incorporate bias-awareness mechanisms, such as fairness-aware learning and multi-agent coordination, to mitigate known data biases [61]. However, the effectiveness of such approaches remains dependent on the underlying data and requires explicit design and evaluation to ensure equitable outcomes across populations. As foundation models become central to biomedical innovation, their growing influence highlights the need for careful scrutiny of both the data they rely on and how these systems are deployed, in order to mitigate bias and ensure equitable outcomes.

3 III Data Used in Foundation Models: A Potential Source for Bias

At the core of the success of foundation models lies the data itself—the essential fuel powering these systems. Large-scale pretraining data is key to biomedical model performance. A comprehensive list of datasets used in such models can be found in [59].

In this section, we examine various data types utilized by foundation models and their potential population biases. Here, we refer to bias as the systematic under-representation of biomedical information characterizing certain groups, such as ethnic populations, age groups, or genders. We exclude ”batch effects,” common in omics data, arising from differences in timing, personnel, reagents, and instrumentation. These have been extensively studied (see e.g. transcriptomic/proteomic [21] and microbiome data [39]). As noted in [21], batch effects often confound hidden biological heterogeneity. They are typically removed, as they lack biological meaning. This is done via correction algorithms or model architectures. For example, in single-cell RNA-seq, read depth, a key technical variable, is explicitly modeled [24].

In contrast, subpopulation effects are potentially biologically significant and should be preserved. Ignoring factors such as sex, gender, or ethnicity, and their contributions to health and disease, can lead to suboptimal results, analytical errors, and potentially discriminatory outcomes.

Demographic information is rarely reported in omics studies. To quantify this issue, we analyzed 4514 PubMed-indexed omics publications published between 2015 and 2024 and assessed whether abstracts reported basic characteristics of the study population, including ancestry or ethnicity, geographic origin, and sex or gender composition (see Appendix A for methodology). Across all five major omics domains, only 2.7% of publications report ancestry or ethnicity information, 4.5% report geographic origin, and just 3.1% report sex or gender composition. Reporting rates vary considerably by field, with genomics showing the highest ancestry reporting at 6.0%, while transcriptomics (0.6%) and proteomics (1.0%) rarely report it. These findings, summarized in Table I, suggest that concerns regarding population bias extend beyond the composition of individual datasets. Across omics domains, even the information needed to assess representativeness is frequently unavailable, limiting our ability to evaluate whether findings and downstream models are likely to generalize across populations. Beyond its implications for bias assessment, this lack of demographic transparency also conflicts with broader FAIR data stewardship principles, which emphasize rich metadata and provenance as prerequisites for the discoverability, interpretation, and reuse of scientific data [54].

Several limitations should be noted. First, the analysis was based on abstracts rather than full-text articles, and some demographic information may therefore have been reported elsewhere. Nevertheless, abstracts typically emphasize information that authors consider central to the study and thus provide insight into reporting priorities. Second, to assess annotation quality, we compared automated annotations against blinded human annotations for a randomly selected subset of 40 publications. While the automated approach showed a tendency toward under-detection of demographic reporting, resulting in conservative prevalence estimates, it rarely generated false positive demographic assignments and therefore is more likely to underestimate than overestimate reporting rates. Finally, our analysis focused on demographic reporting and did not assess other important determinants of health, including environmental exposures. A growing body of work demonstrates that gene–environment interactions involving lifestyle, pollutants, and socioeconomic factors substantially influence molecular phenotypes and disease risk. Consequently, the reporting gaps identified here likely represent only one component of a broader challenge in capturing the contextual information required for robust, reliable, and generalizable biomedical AI [2, 53].

Demographic reporting rates across omics fields based on automated text mining of 4719 PubMed-indexed abstracts (2015–2024). Values indicate the percentage of publications in each field reporting each demographic dimension.

Omics Field	Ancestry/ Ethnicity	Geographic Origin	Age Reporting	Sex/ Gender	Socio-economic
Genomics (748)	6.0%	4.7%	2.1%	0.9%	4.1%
Transcriptomics (944)	0.6%	3.6%	2.8%	2.6%	0.0%
Proteomics (912)	1.0%	4.8%	2.2%	1.1%	0.0%
Epigenomics (917)	4.0%	5.0%	6.2%	4.3%	1.7%
Microbiome (993)	2.3%	4.2%	7.6%	6.1%	0.9%
All fields	**2.7% (120)**	**4.5% (201)**	**4.3% (194)**	**3.1% (142)**	**1.2% (56)**

Genomics: As previously discussed, models designed to learn from genomic variation often rely on whole genome sequencing data obtained from individuals. However, the data available to date has been shown to have a skewed population distribution, over-representing certain groups while under-representing others [19]. This imbalance limits the model's ability to generalize across diverse ancestries and may result in biased genetic risk predictions or missed therapeutic opportunities for underrepresented populations. For example, polygenic risk scores estimate an individual's genetic risk for complex diseases by combining the effects of many genetic variants identified through GWAS. However, as noted above, most existing GWAS data comes from individuals of European ancestry, and studies have shown that the accuracy of risk scores derived from such data declines as the genetic distance from the GWAS population to the target population increases [34]. This situation may deepen health disparities between ancestral groups.

Transcriptomics: Bulk RNA-seq, and more recently, single-cell RNA-seq (as well as single-nucleus RNA-seq), have become key research tools in biomedical research. Large databases, such as CellxGene, have emerged to support this work [12, 44]. Treatment targets are often selected based on studies that use these data to uncover the mechanisms underlying various diseases. While CellxGene includes structured metadata fields at both the cell and dataset levels, such as sex and self-reported ethnicity, these annotations are inconsistently reported and remain incomplete across a substantial fraction of datasets. Consequently, the presence of metadata does not translate into representative coverage, and CellxGene, the largest known source of human single-cell RNA-seq data to date, remains skewed toward samples from adults of European or unknown ethnicity [44].

Epigenomics: Recent efforts have focused on learning from DNA methylation, ChIP-seq, and histone modification data, which offer insights into gene regulation and cellular identity [16]. These datasets, however, also suffer from population and tissue sampling biases. A related repository that collects transcriptomic and epigenomic data is the Gene Expression Omnibus (GEO) [11]. A recent study assessed whether GEO datasets reflect the diversity of the U.S. population using DEI criteria (gender, ethnicity, ancestry, race). The findings revealed that while many datasets displayed balanced gender ratios, significant bias in ethnicity representation was observed, with a predominance of White and African American participants [22].

Proteomics: Similarly, protein-focused models face their own limitations. Despite the large number of known proteins, only a small subset is actively studied as potential drug targets [9]. Unlike genomic or transcriptomic data, the selection of proteins for study is often dictated by the availability of biological tools—such as antibodies or promoters—that enable functional investigation in laboratory settings, rather than by considerations of population diversity.

Microbiome: Microbiome profiles from various human body niches (e.g., gut, vagina, airways) and environmental sources have become widely accessible, thanks to the efforts of many research teams. These datasets span diverse age groups, genders, disease conditions, and anatomical sites. While computational methods have long been applied to microbiome data, only recently have foundation models demonstrated the ability to enable precise and efficient microbiome-based classification and prediction [23]. A large-scale metagenomic assembly approach aimed at reconstructing bacterial and archaeal genomes across multiple populations, body sites, and host ages revealed substantial phylogenetic and functional diversity—much of which remains uncaptured, particularly from rare organisms, non-stool sample types, underrepresented global populations, and varied lifestyles [43].

Texts: Pre-trained language models are also subject to bias. Studies have shown that these models tend to over-represent dominant cultural and scientific perspectives, while encoding biases that may marginalize underrepresented groups [5]. In particular, a recent systematic review revealed pervasive demographic biases in LLMs trained on clinical notes, medical literature and guidelines, with gender and racial/ethnic biases being particularly common [42]. This concern extends to biomedical applications, where language models trained on biased literature may perpetuate an overemphasis on data from individuals of European descent. Consequently, such models risk reinforcing existing disparities in research and healthcare by under-representing findings relevant to diverse populations.

To summarize, it is well known that current datasets contain biases. Modern large models, with hundreds of millions or even billions of parameters trained on massive datasets, make identifying and tracking these biases particularly challenging. Their apparent success may be misleading and may be mistakenly perceived as delivering objective truth.

Because all datasets are collected under practical, scientific, and historical constraints, no dataset can fully capture the diversity of biological systems and patient populations. Consequently, transparency regarding data provenance, study populations, and model behavior becomes critically important. Making these factors more visible can help researchers better understand the scope and limitations of foundation models, assess the contexts in which they are likely to be reliable, and identify situations in which additional validation may be warranted. The next section discusses methods for uncovering and characterizing model bias.

4 IV Interrogating the Models: XAI and Benchmarks

Explainable AI (XAI) technologies can support developers in interrogating their models and uncovering potential biases. These technologies generally fall into three broad categories [33]: causal inference methods, inherently explainable models, and post hoc analysis techniques.

Causal inference methods use AI to infer causal relationships and perform root-cause analysis. These approaches often assume that causal relationships are known and apply weighting techniques to estimate causal effects. A recent paper [57] demonstrated how the attention mechanisms in foundation models can be leveraged to address causal questions at a larger scale and with greater generalizability than traditional statistical tools. This emerging direction toward causally aware foundation models is expected to gain increasing attention [30].

Inherently explainable models, such as decision trees, offer transparency in their decision-making processes, making them more interpretable to human users. Post hoc analysis techniques aim to explain the outputs of complex, black-box models after training. Unfortunately, foundation models often suffer from poor interpretability due to their intricate and opaque architectures. A recent study on explainability in LLMs reviews several established methods, such as SHAP and attention mechanisms, while also pointing out their inherent limitations [38]. SHAP (SHapley Additive exPlanations), a game-theoretic technique, assigns an importance value to each feature in a prediction, thereby quantifying the contribution of individual inputs to the model's output. Furthermore, research has demonstrated that the data generation process itself can influence the resulting explanations, underscoring the critical need to examine and understand this process [36].

Explainability methods may also contribute to the identification of population-related biases in biomedical foundation models. For example, if a model trained predominantly on data from a specific population systematically relies on molecular features, genetic variants, or clinical correlates that are unevenly represented across populations, feature attribution approaches such as SHAP can help reveal these dependencies. Similarly, attention-based analyses may identify biological signals that consistently drive model predictions, while causal approaches may help distinguish population-associated correlations from biologically meaningful mechanisms. Although explainability alone cannot establish the presence or absence of bias, it can provide an additional layer of transparency that, when combined with rich metadata and provenance information, helps researchers evaluate whether model behavior is likely to generalize across diverse populations.

There is a clear need for further development of post hoc methods specifically tailored to foundation models. Interestingly, LLMs themselves may be leveraged to help explain the reasoning behind the outputs of other black-box biomedical models. However, to do so effectively, these LLMs must possess sufficient domain expertise in biomedicine.

A widely used approach for evaluating algorithmic performance in data science is the common task framework. In this setup, participants develop models using publicly available training datasets to perform a shared task—such as class prediction or content generation. These models are then submitted to a scoring system (the referee), which evaluates them on a hidden test dataset [17].

To assess biases in biomedical models more effectively, we advocate for the development of targeted common tasks and benchmarks. A recent meta-assessment of single-cell analysis benchmarks found that, although many are reproducible and make their code publicly available, they are often difficult to extend—particularly as new methods and evaluation criteria emerge [48]. Moreover, to the best of our knowledge, none of the currently available benchmarks are specifically designed to rigorously evaluate aspects related to population-based biases.

There are already platforms specifically designed to simplify the creation and extension of benchmarks. For example, the **Open Problems in Single-Cell Analysis** platform is an open-source, extensible, and continuously evolving benchmarking framework that enables rigorous, quantitative evaluation of best practices in single-cell analysis [31]. Another example is **BenchEMark** [8], a community-driven infrastructure built to support continuous, automated benchmarking of bioinformatics methods, tools, and web services. These platforms provide a strong foundation for designing new benchmarks that explicitly assess population-level disparities. Encouragingly, the community is already moving in this direction—laying the groundwork for more inclusive, transparent, and bias-aware evaluation practices in biomedical AI.

5 V Recommendations

Data paucity remains a pervasive challenge in efforts to capture biological complexity through pretraining biomedical foundation models on raw laboratory data. While initiatives aimed at diversifying biomedical datasets, such as inclusive genome sequencing roadmaps [19] and proteome exploration efforts [9], are essential, they are often resource-intensive and slow to implement. In the meantime, this paper advocates for complementary strategies that focus on making existing biases, and the complex ways in which they influence early stages of foundation model development, more visible, thereby raising awareness and enabling more informed choices.

We recommend that the foundation model development community collaborate with key stakeholders across the biomedical ecosystem, including clinicians, patients, researchers, industry leaders, and regulators, to consistently work along the following three dimensions:

Provenance: Capturing and reporting the origin of data, including ethnicity, geography, and socioeconomic context, is essential for assessing model generalizability and identifying population-specific risks. Without this transparency, AI tools risk reinforcing existing disparities. The CONSORT

Guidelines [51], widely used in clinical trials, offer a model for how provenance reporting could be standardized in biomedical AI.

Openness: Open science practices and collaborative initiatives are vital for democratizing access to tools, datasets, and evaluation protocols. These efforts foster transparency, reproducibility, and accountability in AI development.

Reliability, Benchmarking and Evaluation Transparency: Ensuring the reliability of biomedical AI systems, including their robustness, consistency across populations, and stability under distribution shifts, is critical for their safe and effective use. Current benchmarks rarely assess population-level biases. We advocate for the creation of new, inclusive benchmarks that explicitly evaluate model performance across diverse subgroups. Transparent reporting of evaluation metrics, especially those related to fairness and equity, should become standard practice.

By embedding these principles into the AI development pipeline, we can improve the visibility of biases and limitations before models are widely deployed. This proactive approach complements regulatory oversight and helps promote more transparent, reliable, and appropriately evaluated biomedical AI systems.

Acknowledgments

The research leading to these results has received funding from the Horizon Europe Programme under the Marie Skłodowska-Curie grant agreement No 101183031. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, et al. (2024)Accurate structure prediction of biomolecular interactions with alphafold 3. Nature 630 (8016), pp. 493–500.

[2] R. Alemu, N. T. Sharew, Y. Y. Arsano, M. Ahmed, F. Tekola-Ayele, T. B. Mersha, and A. T. Amare (2025)Multi-omics approaches for understanding gene-environment interactions in noncommunicable diseases: techniques, translation, and equity issues. Human genomics 19 (1), pp. 8.

[3] K. Arora and M. F. Berger (2023-06)Inferring genetic ancestry from cancer sequencing data. Trends In Genetics 39 (6), pp. 431–432.

[4] I. G. Asiimwe and M. Pirmohamed (2022)Ethnic diversity and warfarin pharmacogenomics. Frontiers In Pharmacology 13, pp. 866058.

[5] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell (2021)On the dangers of stochastic parrots: can language models be too big?.In Proceedings Of The 2021 Acm Conference On Fairness, Accountability, And Transparency, pp. 610–623.

[6] C. Bunne, Y. Roohani, Y. Rosen, A. Gupta, X. Zhang, M. Roed, T. Alexandrov, M. AlQuraishi, P. Brennan, D. B. Burkhardt, et al. (2024)How to build the virtual cell with artificial intelligence: priorities and opportunities. Cell 187 (25), pp. 7045–7063.

[7] M. Bykova, Y. Hou, C. Eng, and F. Cheng (2022-08)Quantitative trait locus (xqtl) approaches identify risk genes and drug targets from human non-coding genomes. Human Molecular Genetics 31 (R1), pp. R105–R113.

[8] S. Capella-Gutierrez, D. d. l. Iglesia, J. Haas, A. Lourenco, J. M. Fernández, D. Repchevsky, C. Dessimoz, T. Schwede, C. Notredame, J. L. Gelpi, et al. (2017)Lessons learned: recommendations for establishing critical periodic scientific benchmarking. BioRxiv, pp. 181677.

[9] A. J. Carter, O. Kraemer, M. Zwick, A. Mueller-Fahrnow, C. H. Arrowsmith, and A. M. Edwards (2019)Target 2035: probing the human proteome. Drug Discovery Today 24 (11), pp. 2111–2115.

[10] F. Chen, X. Wang, S. Jang, B. C. Quach, J. D. Weissenkampen, C. Khunsriraksakul, L. Yang, R. Sauteraud, C. M. Albert, N. D. D. Allred, D. K. Arnett, A. E. Ashley-Koch, K. C. Barnes, R. G. Barr, D. M. Becker, L. F. Bielak, J. C. Bis, J. Blangero, M. P. Boorgula, D. I. Chasman, S. Chavan, Y. I. Chen, L. Chuang, A. Correa, J. E. Curran, S. P. David, L. d. l. Fuentes, R. Deka, R. Duggirala, J. D. Faul, M. E. Garrett, S. A. Gharib, X. Guo, M. E. Hall, N. L. Hawley, J. He, B. D. Hobbs, J. E. Hokanson, C. A. Hsiung, S. Hwang, T. M. Hyde, M. R. Irvin, A. E. Jaffe, E. O. Johnson, R. Kaplan, S. L. R. Kardia, J. D. Kaufman, T. N. Kelly, J. E. Kleinman, C. Kooperberg, I. Lee, D. Levy, S. M. Lutz, A. W. Manichaikul, L. W. Martin, O. Marx, S. T. McGarvey, R. L. Minster, M. Moll, K. A. Moussa, T. Naseri, K. E. North, E. C. Oelsner, J. M. Peralta, P. A. Peyser, B. M. Psaty, N. Rafaels, L. M. Raffield, M. S. Reupena, S. S. Rich, J. I. Rotter, D. A. Schwartz, A. H. Shadyab, W. H-H. Sheu, M. Sims, J. A. Smith, X. Sun, K. D. Taylor, M. J. Telen, H. Watson, D. E. Weeks, D. R. Weir, L. R. Yanek, K. A. Young, K. L. Young, W. Zhao, D. B. Hancock, B. Jiang, S. Vrieze, and D. J. Liu (2023-01)Multi-ancestry transcriptome-wide association analyses yield insights into tobacco use biology and drug repurposing. Nature Genetics 55 (2), pp. 291–300.

[11] E. Clough and T. Barrett (2016)The gene expression omnibus database. In Statistical Genomics: Methods and Protocols, pp. 93–110.

[12] T. T. S. Consortium, R. C. Jones, J. Karkanias, M. A. Krasnow, A. O. Pisco, S. R. Quake, J. Salzman, N. Yosef, B. Bulthaup, P. Brown, et al. (2022)The tabula sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. Science 376 (6594), pp. eabl4896.

[13] H. Cui, A. Tejada-Lapuerta, M. Brbić, J. Saez-Rodriguez, S. Cristea, H. Goodarzi, M. Lotfollahi, F. J. Theis, and B. Wang (2025)Towards multimodal foundation models in molecular cell biology. Nature 640 (8059), pp. 623–633.

[14] H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, N. Duan, and B. Wang (2024)Scgpt: toward building a foundation model for single-cell multi-omics using generative ai. Nature Methods, pp. 1–11.

[15] B. Danziger, V. Gurev, M. Madgwick, S. Ravid, T. Rumbell, A. Koseki, T. Kozlovski, C. Tsou, E. Barkan, T. Biswas, J. Xu, Y. Shimoni, J. Hu, and M. Rosen-Zvi (2025)Bmfm-rna: an open framework for building and evaluating transcriptomic foundation models.arXiv preprint arXiv:2506.14861.

[16] L. P. e. al. de Lima Camillo (2024)CpGPT: a foundation model for dna methylation.bioRxiv.

[17] D. Donoho (2017)50 years of data science. Journal of Computational and Graphical Statistics 26 (4), pp. 745–766.

[18] M. P. Farina, J. K. Kim, and E. M. Crimmins (2023)Racial/ethnic differences in biological aging and their life course socioeconomic determinants: the 2016 health and retirement study. Journal Of Aging And Health 35 (3-4), pp. 209–220.

[19] S. Fatumo, T. Chikowore, A. Choudhury, M. Ayub, A. R. Martin, and K. Kuchenbäcker (2022)Diversity in genomic studies: a roadmap to address the imbalance. Nature Medicine 28 (2), pp. 243.

[20] A. J. Gates, D. M. Gysi, M. Kellis, and A. Barabási (2021)A wealth of discovery built on the human genome project—by the numbers. Nature 590 (7845), pp. 212–215.

[21] W. W. B. Goh, W. Wang, and L. Wong (2017)Why batch effects matter in omics data, and how to avoid them. Trends In Biotechnology 35 (6), pp. 498–507.

[22] M. N. Gondal (2024)Assessing dei bias in gene expression omnibus (geo) datasets based on gender and ethnicity. BioRxiv, pp. 2024–11.

[23] J. Han, H. Zhang, and K. Ning (2025)Techniques for learning and transferring knowledge for microbiome-based classification and prediction: review and assessment. Briefings in Bioinformatics 26 (1), pp. bbaf015.

[24] M. Hao, J. Gong, X. Zeng, C. Liu, Y. Guo, X. Cheng, T. Wang, J. Ma, X. Zhang, and L. Song (2024)Large-scale foundation model on single-cell transcriptomics. Nature Methods 21 (8), pp. 1481–1491.

[25] T. Hayes, R. Rao, H. Akin, N. J. Sofroniew, D. Oktay, Z. Lin, R. Verkuil, V. Q. Tran, J. Deaton, M. Wiggert, et al. (2025)Simulating 500 million years of evolution with a language model. Science 387 (6736), pp. 850–858.

[26] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri (2021)Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. Bioinformatics 37 (15), pp. 2112–2120.

[27] L. Kachuri, A. C. Y. Mak, D. Hu, C. Eng, S. Huntsman, J. R. Elhawary, N. Gupta, S. Gabriel, S. Xiao, K. L. Keys, A. Oni-Orisan, J. R. Rodríguez-Santana, M. A. LeNoir, L. N. Borrell, N. A. Zaitlen, L. K. Williams, C. R. Gignoux, E. G. Burchard, and E. Ziv (2023-05)Gene expression in african americans, puerto ricans and mexican americans reveals ancestry-specific patterns of genetic architecture. Nature Genetics 55 (6), pp. 952–963.

[28] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang (2020)BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36 (4), pp. 1234–1240.

[29] B. Liu, X. Li, J. Zhang, J. Wang, T. He, S. Hong, H. Liu, S. Zhang, K. Song, K. Zhu, et al. (2025)Advances and challenges in foundation agents: from brain-inspired intelligence to evolutionary, collaborative, and safe systems. Arxiv Preprint Arxiv:2504.01990.

[30] X. Liu, P. Xu, J. Wu, J. Yuan, Y. Yang, Y. Zhou, F. Liu, T. Guan, H. Wang, T. Yu, J. McAuley, W. Ai, and F. Huang (2025-04)Large language models and causal inference in collaboration: a comprehensive survey. In Findings of the Association for Computational Linguistics: NAACL 2025, L. Chiruzzo, A. Ritter, and L. Wang (Eds.),Albuquerque, New Mexico, pp. 7668–7684.

[31] M. D. Luecken, S. Gigante, D. B. Burkhardt, R. Cannoodt, D. C. Strobl, N. S. Markov, L. Zappia, G. Palla, W. Lewis, D. Dimitrov, et al. (2025)Defining and benchmarking open problems in single-cell analysis. Nature Biotechnology, pp. 1–6.

[32] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T. Liu (2022)Biogpt: generative pre-trained transformer for biomedical text generation and mining. Briefings In Bioinformatics 23 (6), pp. bbac409.

[33] L. Malinverno, V. Barros, F. Ghisoni, G. Visonà, R. Kern, P. J. Nickel, B. E. Ventura, I. Šimić, S. Stryeck, F. Manni, et al. (2023)A historical perspective of biomedical explainable ai research. Patterns 4 (9).

[34] A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, and M. J. Daly (2019)Clinical use of current polygenic risk scores may exacerbate health disparities. Nature genetics 51 (4), pp. 584–591.

[35] D. J. McMaughan, O. Oluyomi, and S. L. SmithSocioeconomic status and access to healthcare: interrelated drivers for healthy aging. front public health. 2020; 8: 231.

[36] V. Mhasawade, S. Rahman, Z. Haskell-Craig, and R. Chunara (2024)Understanding disparities in post hoc machine learning explanation.

[37] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar (2023)Foundation models for generalist medical artificial intelligence. Nature 616 (7956), pp. 259–265.

[38] F. Mumuni and A. Mumuni (2025)Explainable artificial intelligence (xai): from inherent explainability to large language models. Arxiv Preprint Arxiv:2501.09967.

[39] J. T. Nearing, A. M. Comeau, and M. G. Langille (2021)Identifying biases and their potential solutions in human microbiome studies. Microbiome 9 (1), pp. 113.

[40] E. Nguyen, M. Poli, M. G. Durrant, B. Kang, D. Katrekar, D. B. Li, L. J. Bartie, A. W. Thomas, S. H. King, G. Brixi, et al. (2024)Sequence modeling and design from molecular to genome scale with evo. Science 386 (6723), pp. eado9336.

[41] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan (2019)Dissecting racial bias in an algorithm used to manage the health of populations. Science 366 (6464), pp. 447–453.

[42] M. Omar, V. Sorin, R. Agbareia, D. U. Apakama, A. Soroush, A. Sakhuja, R. Freeman, C. R. Horowitz, L. D. Richardson, G. N. Nadkarni, et al. (2025)Evaluating and addressing demographic disparities in medical large language models: a systematic review. International Journal for Equity in Health 24 (1), pp. 57.

[43] E. Pasolli, F. Asnicar, S. Manara, M. Zolfo, N. Karcher, F. Armanini, F. Beghini, P. Manghi, A. Tett, P. Ghensi, et al. (2019)Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell 176 (3), pp. 649–662.

[44] C. C. S. Program, S. Abdulla, B. Aevermann, P. Assis, S. Badajoz, S. M. Bell, E. Bezzi, B. Cakir, J. Chaffer, S. Chambers, et al. (2025)CZ cellxgene discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. Nucleic acids research 53 (D1), pp. D886–D900.

[45] A. Schuhmacher, M. Hinder, A. v. S. und Stein, D. Hartl, and O. Gassmann (2023)Analysis of pharma r&d productivity–a new perspective needed. Drug Discovery Today 28 (10), pp. 103726.

[46] Y. Shoshan, M. Raboh, M. Ozery-Flato, V. Ratner, A. Golts, J. K. Weber, E. Barkan, S. Rabinovici-Cohen, S. Polaczek, I. Amos, et al. (2024)Mammal–molecular aligned multi-modal architecture and language. Arxiv Preprint Arxiv:2410.22367.

[47] G. Sirugo, S. M. Williams, and S. A. Tishkoff (2019)The missing diversity in human genetic studies. Cell 177 (1), pp. 26–31.

[48] A. Sonrel, A. Luetge, C. Soneson, I. Mallona, P. Germain, S. Knyazev, J. Gilis, R. Gerber, R. Seurinck, D. Paul, et al. (2023)Meta-analysis of (single-cell method) benchmarks reveals the need for extensibility and interoperability. Genome Biology 24 (1), pp. 119.

[49] S. Stringhini, C. Carmeli, M. Jokela, M. Avendaño, P. Muennig, F. Guida, F. Ricceri, A. d'Errico, H. Barros, M. Bochud, et al. (2017)Socioeconomic status and the 25× 25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1· 7 million men and women. The Lancet 389 (10075), pp. 1229–1237.

[50] P. Trouiller, P. Olliaro, E. Torreele, J. Orbinski, R. Laing, and N. Ford (2002)Drug development for neglected diseases: a deficient market and a public-health policy failure. The Lancet 359 (9324), pp. 2188–2194.

[51] L. Turner, L. Shamseer, D. G. Altman, L. Weeks, J. Peters, T. Kober, S. Dias, K. F. Schulz, A. C. Plint, and D. Moher (2012)Consolidated standards of reporting trials (consort) and the completeness of reporting of randomised controlled trials (rcts) published in medical journals. Cochrane Database Of Systematic Reviews (11).

[52] B. Van de Sande, J. S. Lee, E. Mutasa-Gottgens, B. Naughton, W. Bacon, J. Manning, Y. Wang, J. Pollard, M. Mendez, J. Hill, et al. (2023)Applications of single-cell rna sequencing in drug discovery and development. Nature Reviews Drug Discovery 22 (6), pp. 496–520.

[53] P. Vineis, O. Robinson, M. Chadeau-Hyam, A. Dehghan, I. Mudway, and S. Dagnino (2020)What is new in the exposome?.Environment international 143, pp. 105887.

[54] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. (2016)The fair guiding principles for scientific data management and stewardship. Scientific data 3 (1), pp. 1–9.

[55] F. Yang, W. Wang, F. Wang, Y. Fang, D. Tang, J. Huang, H. Lu, and J. Yao (2022)Scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. Nature Machine Intelligence 4 (10), pp. 852–866.

[56] K. Ying, J. Song, H. Cui, Y. Zhang, S. Li, X. Chen, H. Liu, A. Eames, D. L. McCartney, R. E. Marioni, et al. (2024)MethylGPT: a foundation model for the dna methylome.bioRxiv.

[57] J. Zhang, J. Jennings, A. Hilmkil, N. Pawlowski, C. Zhang, and C. Ma (2024)Towards causal foundation model: on duality between optimal balancing and attention. In Forty-first International Conference On Machine Learning,

[58] P. Zhang, Y. Hou, W. Tu, N. Campbell, A. A. Pieper, J. B. Leverenz, S. Gao, J. Cummings, and F. Cheng (2022-11)Population-based discovery and mendelian randomization analysis identify telmisartan as a candidate medicine for alzheimer's disease in african americans. Alzheimer's and Dementia 19 (5), pp. 1876–1887.

[59] Q. Zhang, K. Ding, T. Lv, X. Wang, Q. Yin, Y. Zhang, J. Yu, Y. Wang, X. Li, Z. Xiang, et al. (2025)Scientific large language models: a survey on biological & chemical domains. Acm Computing Surveys 57 (6), pp. 1–38.

[60] H. Zheng, L. Shen, A. Tang, Y. Luo, H. Hu, B. Du, Y. Wen, and D. Tao (2025)Learning from models beyond fine-tuning. Nature Machine Intelligence, pp. 1–12.

[61] J. Zhou, J. Jiang, Z. Han, Z. Wang, and X. Gao (2025)Streamline automated biomedical discoveries with agentic bioinformatics. Briefings in Bioinformatics 26 (5), pp. bbaf505.

[62] J. Zou and E. J. Topol (2025)The rise of agentic ai teammates in medicine. The Lancet 405 (10477), pp. 457.

Appendix A Demographic Analysis Details

A-1 Study Design and Data Source

To quantify demographic reporting practices in omics research, we conducted a systematic analysis of PubMed abstracts published between January 2015 and December 2024. We queried the PubMed database using the NCBI Entrez Programming Utilities via the Biopython library (version 1.87), adhering to NCBI usage guidelines.

A-2 Search Strategy

We analyzed five major omics domains: genomics, transcriptomics, proteomics, epigenomics, and microbiome research. For each field, we searched for publications containing the respective term (e.g., "genomics") in the title or abstract and indexed with the "human" Medical Subject Heading (MeSH) term. Publications were stratified into three time periods: 2015–2019, 2020–2022, and 2023–2024. We retrieved up to 350 top-ranked abstracts (by relevance) per field per time period, yielding 4719 abstracts with complete data (89.9% retention rate from 5250 retrieved). Further analysis was performed to identify duplications. 205 duplications were found leading to a starting cohort of 4514 papers.

A-3 Text Mining and Classification

We assessed five demographic dimensions using keyword-based pattern matching with regular expressions: (1) ancestry/ethnicity (64 patterns), (2) geographic origin (9 patterns), (3) age reporting (11 patterns), (4) sex/gender (10 patterns), and (5) socioeconomic status (10 patterns). Keywords included terms such as "multi-ethnic," "European ancestry," "participants from [location]," "age range," "sex distribution," and "socioeconomic status." Each abstract was classified as reporting (present) or not reporting (absent) each dimension based on case-insensitive pattern matching in the combined title and abstract text.

A-4 AI-Assisted Analysis Development

The analytical framework and scientific questions were defined by the authors. Large language models were used as programming and research assistants to accelerate implementation, including code generation, debugging, and development of candidate extraction patterns (IBM's BoB (Build on Bedrock) and Anthropic's Claude Opus 4.6). All extraction rules, keyword dictionaries, validation proce-

dures, and final methodological choices were reviewed and approved by the authors. The resulting analysis was performed using a deterministic rule-based text-mining pipeline implemented in Python, ensuring that all reported results are fully reproducible and independent of the behavior of any specific language model.

A-5 Statistical Analysis

Reporting rates were calculated as the percentage of abstracts containing at least one keyword for each demographic dimension. We analyzed temporal trends across the three time periods and field-specific patterns across the five omics domains. Cross-tabulations examined the interaction between field and time period.

A-6 Validation of the Automated Text-Mining Pipeline

To assess the reliability of the automated extraction approach, we conducted a blinded validation study on a randomly selected subset of 40 publications from the analyzed corpus. Human expert annotations were independently generated and compared against the automated annotations for both omics field classification and demographic reporting detection.

For omics field classification, the automated pipeline demonstrated substantial agreement with human annotation Cohen's $\kappa = 0.78$. Only a single complete disagreement was observed. All discrepancies occurred in studies involving multiple omics modalities, reflecting the inherent ambiguity of assigning a single omics category to multimodal studies rather than systematic classification errors.

For demographic reporting detection, the automated pipeline achieved high specificity 97.8% and an overall agreement of 91% with human annotations. However, recall was limited (12.5%), indicating that the system was substantially more likely to miss demographic information than to report it incorrectly. Examination of the validation set suggests that some false negatives occurred when demographic characteristics were implied by the study population description rather than explicitly reported. The ratio of false negatives to false positives was approximately 3.5:1, demonstrating a systematic tendency toward underdetection rather than over-detection. Geographic origin and sex/gender reporting achieved fair agreement with human annotation (Cohen's $\kappa = 0.29$), whereas ancestry/ethnicity, age reporting, and socioeconomic indicators showed particularly low recall in the validation set.

A-7 Limitations

This analysis has several limitations. First, keyword-based pattern matching may miss implicit demographic information or alternative terminology. Second, we analyzed abstracts only; full-text articles may contain additional demographic details. Third, our binary classification (present/absent) does not capture the quality or completeness of reporting. Fourth, we used simplified single-term queries (e.g., "genomics") rather than comprehensive Boolean searches, potentially affecting the representativeness of our sample. Finally, our analysis was limited to English-language publications indexed in PubMed.