

ARTICLE

Navigating Privacy and Progress

Regulating Genetic Information Utilization in AI-Driven Drug Development Information*

Jessica Combs[†]

Abstract

There is an urgent need to develop a global regulatory framework for using and reusing genetic information in pharmaceutical development. As artificial intelligence (AI) and deep learning (DL) continue to advance, the pharmaceutical industry will increasingly utilize this technology to discover new therapeutics. This progress will create a significant demand for large quantities of genetic information, commodifying already available genetic data through pharmaceutical and biobanking companies.

If a proactive global regulatory framework is not established, reactive regulations will likely be implemented in response to inevitable global privacy violations. However, proactive regulation is far more effective than reactive regulation, as it protects individual rights and prevents significant harm before it occurs. In contrast, reactive measures often come too late, allowing significant damage to unfold, which erodes public trust and diminishes social capital in the process.

Keywords: genetic data, artificial intelligence, drug development, deep learning, machine learning.

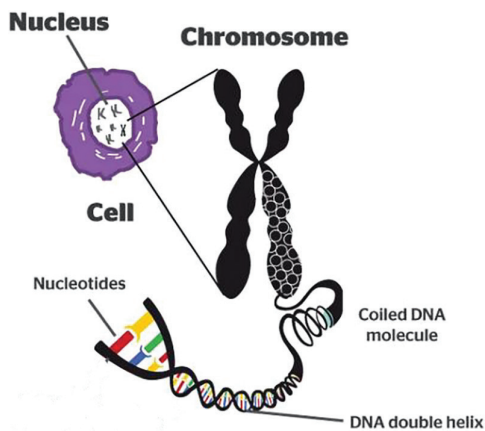
* The research for this article was aided by the artificial intelligence tools Perplexity and SciSummary. Grammarly was used for editing.

† Jessica Combs, J.D. 2026, McKinney School of Law, Indiana University. My Master of Science from Purdue University was focused on 3D structural characterization of large molecules, like DNA. I have over 20 years of experience in the pharmaceutical industry, holding technical and leadership positions in drug discovery, early development, clinical trial testing and drug manufacturing. I am indebted to Frank Emmert, John S. Grimes Professor of Law, for his guidance in this research.

A Introduction

Genetic data can be obtained for all living organisms, including plants and bacteria. The focus of this article is genetic information from human DNA. All the genetic information for one human is stored in a personal DNA library known as the genome. Most cells in the body contain a complete copy of an individual's DNA library (genome) within the cell nucleus. The DNA resides within chromosomes located in the nucleus of the cells.¹ To understand why genetic information differs from other personal information, one must first have a high-level understanding of the organization of the genome (library) (Figure 1).

Figure 1 *The correlation between the cell, chromosome and DNA strand²*



The organization of this library is important because researchers can navigate it and point to locations associated with specific traits. The DNA library (genome) consists of shelves that represent chromosomes. Humans typically have 23 pairs of chromosomes (shelves). The number of shelves in the library is specific to the species.³ Humans have 46 “shelves” in their library, whereas dogs have 78 shelves.⁴

Each shelf (chromosome) in the library (genome) holds two books. The two books are complementary to each other and tell a single story. These books represent each side of the DNA double helix. The first book contains how individuals should grow, look and function. The second book, the other side of the DNA helix, serves as a backup copy, maintaining your genetic information's accuracy. Each

- 1 Marian, 'Sequencing Your Genome: What Does It Mean?', 10 *Methodist Debakey Cardiovas J* (2014) 3.
- 2 Human Origin Project, 'What Is a Chromosome and Its Function?' (Medium, 2019) <https://medium.com/@humanoriginproject/what-is-a-chromosome-and-its-function-4635ce549748> (accessed 18 November 2024).
- 3 National Human Genome Research Institute, 'A Brief Guide to Genomics' (Genome, 2019) <https://www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics> (accessed 18 November 2024).
- 4 National Human Genome Research Institute, 'Dog Genome Project Information' (Research.nhgri.nih.gov, 2021) https://research.nhgri.nih.gov/dog_genome/canine_genome.shtml (accessed 18 November 2024).

chapter in a book is analogous to a gene. Sometimes, numerous chapters (genes) are required to compose a specific attribute, like eye colour. For instance, the genetic information determining an individual's eye colour could be encoded in as many as 16 chapters (genes).⁵

The language of these books is written in a four-letter alphabet: A (adenine), G (guanine), C (cytosine) and T (thymine).⁶ These letters are the fundamental components of DNA, known as nucleotides. The uniqueness of DNA in each human, apart from identical twins, lies in the arrangement of these letters (nucleotides) within the volumes on the 46 shelves (chromosomes) of a person's DNA library (genome).⁷

Genetic information comes in many sizes. It can refer to sequences of letters (nucleotides) in a specific book or the entire library (genome). This genetic information comes from biological samples such as blood, saliva, urine or tissue.⁸ The samples are sequenced using one of the many sequencing technologies currently available. This genetic information may be combined with age, sex, childbearing potential and ethnicity. This type of information is known as phenotype data. The dataset can be analysed by integrating the DNA sequence with the phenotype information. In the context of genetic information used for medical diagnosis, specific patient identifiers such as name, date of birth and address may also be present alongside the phenotype data.⁹

The human genome is approximately 3.1 billion letters (nucleotides). Of these 3.1 billion letters (nucleotides), the sequence between two humans is only approximately 0.6% different.¹⁰ To identify the uniqueness of an individual using only genetic information, researchers analyse Short Tandem Repeats (STRs), which are repeating sequences of nucleotides at a specific location in the genome.¹¹ Since all humans share the same basic genome organization (library), each sample will feature the same location, known as "loci", in their individual library (genome).¹² These loci are akin to bookmarked locations in a book on a certain shelf (chromosome). STRs are bold words at these bookmarked locations that create a unique

- 5 White and Rabago-Smith, 'Genotype-Phenotype Associations and Human Eye Color', 56 *J Hum Genet* (2011) 5.
- 6 National Human Genome Research Institute, 'Deoxyribonucleic Acid (DNA) Fact Sheet' (Genome.gov, 2020) <https://www.genome.gov/about-genomics/fact-sheets/Deoxyribonucleic-Acid-Fact-Sheet> (accessed 18 November 2024).
- 7 'Difference between DNA, Genes, Chromosomes- for Kids!' (Human Body Learning, n.d.) <https://humanbodylearning.com/difference-between-dna-genes-chromosomes-kids/> (accessed 18 November 2024).
- 8 'What Is DNA?' (National Institute of Justice, 2023) <https://nij.ojp.gov/library/publications/whatever-law-enforcement-officer-should-know-about-dna-evidence> (accessed 5 December 2024).
- 9 National Human Genome Research Institute, 'Human Genomic Variation Fact Sheet' (genome.gov, n.d.) <https://www.genome.gov/genetics-glossary/Phenotype> (accessed 18 November 2024).
- 10 National Human Genome Research Institute, 'Human Genomic Variation Fact Sheet' (Genome.gov, n.d.) <https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genomic-variation> (accessed 5 December 2024).
- 11 Fan and Chu, 'A Brief Review of Short Tandem Repeat Mutation', 5 *Genom Proteom Bioinform* (2007) 7.
- 12 'Variable Number Tandem Repeats' (CUNY Academic Commons, n.d.) <https://bioclimate.commonsgc.cuny.edu/analyzing-dna/variable-number-tandem-repeats/> (accessed 5 December 2024).

pattern for each person.¹³ This is how DNA is used to identify individuals from crime scenes. The FBI uses 20 bookmarked loci across 19 chromosomes to conclusively identify a person's DNA.^{14,15,16}

The uniqueness of a person's DNA is not the only reason each person's library (genome) is exceptional. Knowing the shelf (chromosomes) and chapters (genes) attributed to certain traits like eye colour can identify these traits in anonymous samples. For this reason, an individual's genome is deeply personal information. It is a collection of stories about a person's ancestry, ethnicity, personal traits, medical conditions, carrier status for inherited diseases, and predispositions for certain diseases.¹⁷ An individual's DNA tells not just their personal story but also the story of their relatives. It can identify sex and ethnicity. As science advances, a person's DNA library may reveal even more information, like predictive information for lifespan or intelligence. It is difficult to predict the scope of future use of genetic data. The nature of genetic information means it does not change over time for an individual. If a person chooses to sequence their DNA at 18, the information deduced about that person will be much larger in scope in a few decades.¹⁸

B Where Is Genetic Information Stored?

There are already various large repositories of genetic information. These repositories fall into three categories: open access, government controlled and privately controlled.¹⁹ The following are examples of genomic databases within these categories.

- a Open Access Genetic Databases:
 - i Large-Scale Genomic Projects: Whole genomic projects like the 1,000 Genomics project, a catalogue of common human genetic variation using a diverse population of 2,500 individuals.^{20,21}

- 13 Butler, 'Genetics and Genomics of Core Short Tandem Repeat Loci Used in Human Identity Testing', 51 *J Forensic Sci* (2006) 253.
- 14 Nwawuba et al., 'Forensic DNA Profiling: Autosomal Short Tandem Repeat as a Prominent Marker in Crime Investigation', 27 *Malays J Med Sci* (2020) 22.
- 15 FBI, 'CODIS and NDIS Fact Sheet' (FBI.gov, 2021) <https://www.fbi.gov/how-we-can-help-you/dna-fingerprint-act-of-2005-expungement-policy/codis-and-ndis-fact-sheet#DNA-Data%20Requirements> (accessed 5 December 2024).
- 16 FBI, 'How We Can Help You' (FBI.gov, 2021) <https://www.fbi.gov/how-we-can-help-you/dna-fingerprint-act-of-2005-expungement-policy/codis-and-ndis-fact-sheet#DNA-Data%20Requirements>.
- 17 İlkiliç, 'Coming to Grips with Genetic Exceptionalism: Roots and Reach of an Explanatory Model', 1 *Med Stud* (2009) 131.
- 18 Naveed et al., 'Privacy in the Genomic Era', 48 *ACM Comput Surv* (2015) 6.
- 19 Annaratone et al., *Basic Principles of Biobanking: From Biological Samples to Precision Medicine for Patients* (2021).
- 20 'The International Genome Sample Resource' (IGSR, 2024) <https://www.internationalgenome.org/> (accessed 5 December 2024).
- 21 'Principles and Best Practices for Protecting Participant Privacy' (NIH, n.d.) <https://sharing.nih.gov/data-management-and-sharing-policy/protecting-participant-privacy-when-sharing-scientific-data/principles-and-best-practices-for-protecting-participant-privacy> (accessed 5 December 2024).

- ii Major Genomic Databases: Genetic sequence databases can be accessed publicly and some require registration.²² An example is the Database of Genotypes and Phenotypes (dbGaP) which distributes data investigating the interaction of genotypes and phenotypes in humans.²³
- b Government-controlled Genomic Databases:
 - i Forensic profiles used in law enforcement:²⁴ over 50 countries have DNA databases, the most prominent being China, the US and the UK.²⁵ However, smaller countries like Saudi Arabia, Qatar and Israel also have genomic databases.²⁶
 - ii Government agencies collect data for public health purposes. Examples of this are the European Genome-phenome (EGA), the UK's Population Biomedical Collection,²⁷ and the Qatar Genome database, which incorporates data from 19 Arab countries²⁸ and Newborn databases.²⁹
- c Other Government DNA banks:
 - i In 2024, Iran began creating a DNA database of high-risk professions, including journalists, flight crews, armed forces and mine workers.³⁰ The collection of journalist's genetic information raises concerns about ulterior motives.
 - ii The Chinese government database of Uyghur Muslims in Xinjiang has been performing mass collection of samples for genetic testing along with other biometric data.³¹
- d Privately controlled Genomic Databases:
 - i Direct-to-consumer genetic testing companies like 23andMe. Greater than >39 million profiles collectively.³²

22 'Finding Datasets for Secondary Analysis' (Genomic Databases) <https://browse.welch.jhmi.edu/datasets/genomic-databases>.

23 'dbGaP Overview' (NIH. n.d.) <https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.html> (accessed 5 December 2024).

24 de Groot, van Beers and Meynen, 'Commercial DNA Tests and Police Investigations: A Broad Bioethical Perspective', 47 *J Med Ethics* (2021) 788.

25 Hazel and Clayton, 'Law Enforcement and Genetic Data' (The Hastings Center, 2021) <https://www.thehastingscenter.org/briefingbook/law-enforcement-and-genetic-data/> (accessed 7 December 2024).

26 Ma et al., 'Database Commons: A Catalog of Worldwide Biological Databases', 21 *Genom Proteom Bioinf* (2023) 1054.

27 'Celebrating 15 years of UK Biobank' (University of Oxford Nuffield Department of Population Health, 2022) <https://www.ndph.ox.ac.uk/longer-reads/celebrating-15-years-of-uk-biobank> (accessed 18 November 2024).

28 'Qatar Genome Program' (Global Alliance for Genomics and Health, n.d.) https://www.ga4gh.org/driver_project/qatar-genome-program/ (accessed 18 November 2024).

29 'Widespread Newborn Sequencing Will Worsen Risks to Genetic Privacy' (American Civil Liberties Union, 2023) <https://www.aclu.org/news/privacy-technology/widespread-newborn-dna-sequencing-will-worsen-risks-to-genetic-privacy> (accessed 18 November 2024).

30 'Iran' (Iran International, 2024) https://www.iranintl.com/en/202403036159?t&utm_source (accessed 18 November 2024).

31 'Chinese Authorities Collecting DNA from Residents in Xinjiang' (The Guardian, 13 December 2017) <https://www.theguardian.com/world/2017/dec/13/chinese-authorities-collecting-dna-residents-xinjiang> (accessed 5 December 2024).

32 Guerrini et al., 'Family Secrets: Experiences and Outcomes of Participating in Direct-to-Consumer Genetic Relative-Finder Services', 109 *Am J Hum Genet* (2022) P486, <https://www.sciencedirect.com/science/article/pii/S0002929722000131>.

- ii Medical records are maintained at hospitals and other health organizations.
- iii Pharmaceutical and other private research institutions.

In addition to the DNA sequence information, these repositories may have varying degrees of information identifying the individual. Much of the data have had identifiers, like name and date of birth information, removed. Some refer to this as “de-identified” data. Data are de-identified to the greatest extent that maintains scientific utility.³³

C Genetic Information, AI and Drug Development

The preceding defined what genetic data are and where they can be found. The remainder of this document will examine (a) the relationship between genetic information, AI and drug development; (b) the current regulatory framework governing the utilization of genetic information in conjunction with AI by private entities; (c) the risks associated with the integration of genetic information and AI; and (d) proactive measures to mitigate potential violations of privacy in the future.

I Why Genetic Information Is Important to the Drug Development Process

There is a strong positive correlation between genetic information and the success of drug candidates, which will drive the use of genetic data with AI as this technology develops. Significant evidence supports that genetic information positively affects the success rate of drug candidates. Between 1970 and 2010, the quantity of new drugs brought to market dramatically declined as research and development costs rose. In the 1970s, ten new drugs were approved for every billion dollars spent on research and development. By 2000, the costs of new drugs had increased globally 10-fold.³⁴ Only one new drug was approved for every billion spent on research and development. In 2010, this decline began to stabilize. By 2018, this trend had begun to reverse, slightly improving the cost per successful drug candidate.³⁵ This is partly due to a better understanding of disease causes, leading to more successful drug selection. Much of this insight came from using human genomic data in drug development.³⁶

Some evidentiary support for the positive effect of genetic information on the success rate of drug candidates comes from recent US Food and Drug Administration (FDA) approvals of drugs. Candidates with genetic support are twice as likely to be approved by the FDA.³⁷ In 2021, it was found that two out of three FDA-ap-

33 Sorani et al., ‘Genetic Data Sharing and Privacy’, 13 *Neuroinformatics* (2015) 1.

34 Ibid.

35 Scannell, Blanckley and Boldon, ‘Diagnosing the Decline in Pharmaceutical R&D Efficiency’, 11 *Nat Rev Drug Discov* (2012) 191, <https://doi.org/10.1038/nrd3681>.

36 Ringel et al., ‘Breaking Eroom’s Law’, 19 *Nat Rev Drug Discov* (2020) 833.

37 Nelson et al., ‘The Support of Human Genetic Evidence for Approved Drug Indications’, 47 *Nat Genet* (2015) 856.

proved drugs were supported by human genetic evidence.³⁸ In 2024, the journal *Nature* published an article detailing the clear advantage genetic information provided to the probability of success in drug development candidates.³⁹ Leveraging genetic insights can improve the efficiency, success and overall effectiveness of drug discovery and development by improving drug candidate selection, better patient selection for clinical trials, prediction of drug effects and safety and revealing new uses of existing drugs.⁴⁰

Genetic information can only be used effectively in drug development when leveraged through extensive datasets and algorithms. Large datasets are fundamental to AI, which makes utilizing AI with genetic data a natural pairing.⁴¹ This natural pairing, combined with the lucrative nature of the drug development business, creates strong drivers for pharmaceutical companies to utilize genetic information with AI. The sensitive nature of genetic information and the absence of consistent global regulations governing its use with AI raise significant risks to individual privacy rights. While these data can facilitate advancements in medical research and the development of new therapeutic drugs, understanding how to manage such information is crucial for safeguarding individual privacy.

II Artificial Intelligence Used in the Drug Development Process

As genetic information has increased in drug development, the tools used have evolved to process these larger datasets.⁴² This need for complex analytical tools has led to the rise of AI in drug development.⁴³ Supervised Traditional Machine Learning (ML) models have primarily been used, relying heavily on human-engineered features and have been used in conjunction with computer modelling for decades.⁴⁴ The availability of larger datasets has promoted using a different approach to ML called deep learning (DL). DL is a promising and relatively new technique comprised of simple but nonlinear processing units.⁴⁵ DL uses artificial multilayer neural networks to mimic the human brain's learning process.⁴⁶

38 Ghousaini, Nelson and Dunham, 'Future Prospects for Human Genetics and Genomics in Drug Discovery', 80 *Curr Opin Struct Biol* (2023) 102568, <https://www.sciencedirect.com/science/article/pii/S0959440X23000428>.

39 Minikel et al., 'Refining the Impact of Genetic Evidence on Clinical Success', 629 *Nature* (2024) 1.

40 Heilbron et al., 'Advancing Drug Discovery Using the Power of the Human Genome', 254 *J Pathol* (2021) 418.

41 Sundblad, 'Data Is the Foundation for Artificial Intelligence and Machine Learning' (*Forbes*, 18 October 2018) <https://www.forbes.com/sites/willemsundbladeurope/2018/10/18/data-is-the-foundation-for-artificial-intelligence-and-machine-learning/> (accessed 18 November 2024).

42 Niazi and Mariam, 'Computer-Aided Drug Design and Drug Discovery: A Prospective Analysis', 17 *Pharmaceuticals (Basel)* (2023) 22.

43 Geevarghese, 'Explainable Artificial Intelligence in Drug Discovery', *Computational Intelligence Methods and Applications* (2024), https://doi.org/10.1007/978-981-97-3705-5_6.

44 Lee et al., 'Big Data and Artificial Intelligence (AI) Methodologies for Computer-Aided Drug Design (CADD)', 50 *Biochem Soc Trans* (2022) 241.

45 Ching et al., 'Opportunities and Obstacles for Deep Learning in Biology and Medicine', 15 *J R Soc Interface* (2018) 20170387, <https://doi.org/10.1098/rsif.2017.0387>.

46 Zhang et al., 'From Machine Learning to Deep Learning: Progress in Machine Intelligence for Rational Drug Discovery', 22 *Drug Discov Today* (2017) 1680, <https://doi.org/10.1016/j.drudis.2017.08.010>.

DL requires less human intervention and can automatically learn and improve from data without human guidance. The full potential of these advanced forms of AI called DL models has not been completely realized.⁴⁷

There are already vast amounts of data in repositories in different systems and formats that have not been able to be merged, which would allow them to be used as one larger dataset. The continued advancements in deep neural networks will increase the ability to integrate these large genetic datasets into even larger datasets.⁴⁸ These advancements could potentially be transformative at every stage of drug development.⁴⁹ A 2024 article in the *World Journal of Experimental Medicine* highlights how AI-driven drug discovery and development processes will significantly impact each drug development phase by enhancing efficiency, accuracy and innovation.⁵⁰ The result will be a higher number of better-quality novel medicines in less time.⁵¹

A notable example is the use of DL to identify new antibiotics. Antibiotics have a low success rate of moving to the clinical phase of drug development, which results in a lack of economic incentives for private sector companies to pursue them.⁵² With the emergence of antibiotic-resistant strains of bacteria, there has come a rising need for new antibiotics. To address this challenge, MIT scientists trained a deep neural network capable of predicting molecules with antibacterial activity. Eight antibacterial compounds were found that had detectable levels of bacteria growth inhibition.⁵³ One of these, Halicin, is currently in clinical trials and shows promising results against antibiotic-resistant bacterial infections.⁵⁴

Another prominent example is the case of a small molecule, DSP-1181, which had impressive efficiency gains in timing from discovery to clinical testing.⁵⁵ In 2020, this compound became the first AI-designed drug to enter clinical trials. Incorporating AI in its development resulted in remarkable efficiency gains, allowing it to reach clinical testing in 12 months. This marks a significant improvement compared to traditional methods, which typically require five years.⁵⁶

As of 2024, the top 20 pharmaceutical companies are using AI to develop new drugs.⁵⁷ Genentech has made it widely known that AI is an integral part of its drug

47 Kokudeva et al., 'Artificial Intelligence as a Tool in Drug Discovery and Development', 14 *World J Exp Med* (2024) 96042.

48 Eraslan et al., 'Deep Learning: New Computational Modelling Techniques for Genomics', 20 *Nat Rev Genet* (2019) 389.

49 Blanco-González et al., 'The Role of AI in Drug Discovery: Challenges, Opportunities, and Strategies', 16 *Pharmaceuticals* (2023) 891, <https://doi.org/10.3390/ph16060891>.

50 Kokudeva et al., 'Artificial Intelligence as a Tool in Drug Discovery and Development', 14 *World J Exp Med* (2024) 96042.

51 Vora et al., 'Artificial Intelligence in Pharmaceutical Technology and Drug Delivery Design', 15 *Pharmaceutics* (2023) 1916.

52 Brown and Wright, 'Antibacterial Drug Discovery in the Resistance Era', 529 *Nature* (2016) 336.

53 Stokes et al., 'A Deep Learning Approach to Antibiotic Discovery', 180 *Cell* (2020) 688, <https://doi.org/10.1016/j.cell.2020.01.021>.

54 Booq et al., 'Assessment of the Antibacterial Efficacy of Halicin against Pathogenic Bacteria', 10 *Antibiotics (Basel)* (2021) 1480.

55 Burki, 'A New Paradigm for Drug Development', 2 *Lancet Digit Health* (2020) e226.

56 Arnold, 'Inside the Nascent Industry of AI-Designed Drugs', 29 *Nat Med* (2023) 1292

57 Jayatunga, 'How Successful Are AI-Discovered Drugs in Clinical Trials? A First Analysis and Emerging Lessons', 29 *Drug Discov Today* (2024) 104009, <https://www.sciencedirect.com/science/article/pii/S135964462400134X>.

development lifecycle. They have coined this foundation “lab in a loop”, where data from the lab and clinic are fed into AI models, which leads to predictions and more experiments based on the AI predictions.⁵⁸

In 2022, the AI pipeline of drug candidates from 20 companies contained 160 candidates in early development and 15 in clinical trials.⁵⁹ It is estimated that the global AI drug discovery market is US\$1.7 billion and is expected to reach US\$11.93 billion by 2033.⁶⁰ This expected increase in investment in AI is not surprising when considering that shortening drug development timelines would decrease the cost of developing drugs. Considering the cost of developing a new drug is approximately US\$2.3 billion, there is a substantial financial incentive for pharmaceutical companies to invest in any tool that would improve costs.⁶¹

III *Big Data Era of Genomics and Drug Development*

The volume of available genomic data has been rapidly increasing. One source stated it is the largest and fastest growing data source.⁶² One estimate suggests that each year, 40 times more genetic data are produced than content on YouTube data. For the first time, ML tools can extract more and better information from this large volume of genetic data.⁶³

‘Big Data’ is a term used to refer to extremely large amounts of complex data. These are high-volume datasets with a variety of data formats, making it challenging to extract meaningful information.⁶⁴ There is not just a single definition for Big Data. Still, most sources agree that it is characterized by enormous volume, a variety of data inputs and a high velocity of data generation.

Genomics is already considered a “Big Data” science.⁶⁵ On average, one person’s DNA library (genome) requires 100 to 150 GB of storage. The US National Human Genome Research Institute predicts that genomics research will generate between 2 and 40 exabytes of data in the next ten years.⁶⁶ This is 40 gigabytes of data,

58 ‘Redefining Drug Discovery with AI’ (Genentech, Inc. 2025) <https://www.gene.com/stories/redefining-drug-discovery-with-ai> (accessed 7 February 2025).

59 ‘From the Analysts Couch: AI in Small Molecule Drug Discovery: A Coming Wave?’, 21 *Nat Rev Drug Discov* (2022) 175, <https://doi.org/10.1038/d41573-022-00025-1>.

60 ‘AI in Drug Discovery Market Size to Expand US 11.93 Bn by 2033’ (Biospace, n.d.) <https://www.biospace.com/ai-in-drug-discovery-market-size-to-expand-us-11-93-bn-by-2033> (accessed 18 November 2024).

61 ‘Measuring the Return from Pharmaceutical Innovation’ (Deloitte United States, n.d.) <https://www2.deloitte.com/us/en/pages/life-sciences-and-health-care/articles/measuring-return-from-pharmaceutical-innovation.html> (accessed 5 December 2024).

62 ‘The Human OS Is at the Top of Big Data’ (IEEE Spectrum, 16 February 2012) <https://spectrum.ieee.org/the-human-os-is-at-the-top-of-big-data> (accessed 7 February 2025).

63 ‘Big Data among Big Data, Genome Data’ (3 billion, 6 January 2021) <https://3billion.io/blog/big-data-among-big-data-genome-data> (accessed 17 November 2024).

64 Cremin, Dash and Huang, ‘Big Data: Historic Advances and Emerging Trends in Biomedical Research’, 4 *Current Research in Biotechnology* (2022) 138, <https://www.sciencedirect.com/science/article/pii/S2590262822000090>.

65 Brown, *Genomes* (2nd edition, 2002), <https://www.ncbi.nlm.nih.gov/books/NBK21134/>.

66 National Human Genome Research Institute, ‘Genomic Data Science’ (Genome.gov, 2022) <https://www.genome.gov/about-genomics/fact-sheets/Genomic-Data-Science> (accessed 17 November 2024).

roughly equivalent to the data required to stream 2.86 billion movies in 4K high definition.⁶⁷

Although these DL models require less human expertise, they require significantly more extensive datasets. DL is a significant driver in consuming larger and more complex sets of biomedical information. This is why France's Department of Public Health called the 21st century an era of Big Data involving all aspects of human life, including biology and medicine.⁶⁸

The availability of a large amount of high-quality data will affect the performance and reliability of DL modelling. DL learning requires large-scale datasets, often containing millions of datasets. In contrast, the traditional models work on smaller, curated datasets on the magnitude of thousands of data points.⁶⁹ For this reason, genomic data will be one of the best datasets for training AI models. The completeness of the genetic dataset has been shown to impact the usefulness of the genetic data in drug development. If the data were a complete genome sequence, the chances of success in the subsequent development phase increased.⁷⁰ This may drive the retesting of stored samples to get complete genomes, which, in many cases, was not needed for the primary use of the sample.

Yet, neither data collection nor its usage with AI tools has a regulatory framework. This significant growth in the use of genetic data in drug development raises concerns about the regulations surrounding privacy and individuals' rights to control the future of their genetic makeup biospecimens. The growth rate of genetic data will continue to increase, particularly its use in drug development, because of the previously discussed correlation with the success of drug candidates getting to market. With the enormous costs of bringing medicines to market, any small improvements in cost and efficiency are valuable to the bottom line of pharmaceutical companies. AI will go beyond offering small improvements; it has the potential to revolutionize the lifecycle of drug development. When the revenues of the global pharmaceutical industry total over a trillion dollars, there is a significant financial incentive to invest in AI-based DL models trained with genetic data.⁷¹

In 2023, GlaxoSmithKline paid 23andMe Holding Co., a genetic testing company, US\$20 million upfront to access its genetic information and related phenotype data database. Ultimately, Glaxo Smith Kline (GSK) will pay US\$300 million to access this data over five years.⁷² As of 2024, 23andMe has over 12 million customers with genomic and phenotype data who have consented to participate in research.⁷³ According to a press release in October 2023, 23andMe reported the

67 Ibid.

68 Baro et al., 'Toward a Literature-Driven Definition of Big Data in Healthcare', 2015 *Biomed Res Int* (2015) 639021.

69 Lecun et al., 'Deep Learning', 521 *Nature* (2015) 436, hal-04206682.

70 Minikel, 'Refining the Impact of Genetic Evidence on Clinical Success', 629 *Nature* (2024) 436-44.

71 'Global Pharmaceutical Industry' (Statista, n.d.) <https://www.statista.com/topics/1764/global-pharmaceutical-industry/> (accessed 18 November 2024).

72 'Drugmaker Pays 23andMe \$20 Million for Access to User DNA Information' (Scripps News, 2023) <https://www.scrippsnews.com/business/company-news/drugmaker-pays-23andme-20-million-for-access-to-user-dna-information> (accessed 5 December 2024).

73 Wang et al., 'The Impact on Clinical Success from the 23andMe Cohort', *medRxiv* (2024), <https://doi.org/10.1101/2024.06.17.24309059>.

GSK partnership has yielded approximately 50 drug development programmes from this secondary use of genetic information.⁷⁴

D Current Regulations Governing the Use of Genetic Information and AI

Currently, no comprehensive set of regulations governs using genetic information with AI in drug development. To understand the current regulations governing the use of genetic data with AI, one must understand the legal framework that allows the collection and use of genetic data and any associated AI regulations. There are three applicable legal subjects: (a) informed consent requirements for primary and secondary use of biospecimens and associated data; (b) individual privacy and property rights to their sample and data associated data; and (c) current AI regulations governing drug development. Here within, these legal principles will be reviewed from a global perspective.

I *Informed Consent Requirements for Primary and Secondary Use of Biospecimens and Associated Data*

One must start with the legal basis for obtaining the sample to understand individual rights to genetic information and its use with AI. Informed consent is a critical ethical and legal principle that requires individuals to consent to medical or scientific experimentation. Informed consent must be obtained whenever a biospecimen is collected from a human. It ensures that participants voluntarily agree after being fully informed about the nature, purpose and potential implications when the sample is collected.⁷⁵ A biospecimen is obtained for multiple reasons, including research, medical diagnosis or learning about familial history. The type of informed consent acquired will determine the future use of that sample and the scope of secondary use. The necessity of consent may also change if the samples are de-identified.⁷⁶

There are two mechanisms where consent is communicated: implied or explicit.⁷⁷ The difference between the two is the quality and depth of information supplied to the individual before sample extraction. Explicit or express consent requires formal documentation and entails a thorough discussion of the associated risks and benefits. This approach provides clear evidence that the individual comprehends the information presented. In contrast, implied consent is inferred from the individual's actions or circumstances, stemming from reasonable assumptions regarding their agreement. This form of consent necessitates less formal documentation.⁷⁸

Different types of consent models allow samples to be used for their primary purpose and stored and subsequently used for secondary purposes. Broad consent

74 '23andMe Announces Collaboration Extension and New Data Licensing Deal with GSK' (23andMe Holding Co, 16 July 2018) <https://investors.23andme.com/news-releases/news-release-details/23andme-announces-collaboration-extension-new-data-licensing/> (accessed 5 December 2024).

75 Lartey, *Ethical Considerations in Obtaining Informed Consent for Research Participation: The Role of Review Boards* (2024).

76 Grady et al., 'Broad Consent for Research with Biological Samples: Workshop Conclusions', 15 *Am J Bioeth* (2015) 34.

77 Chen, Rosenstein and Muthappan, 'Research with Stored Biological Samples: What Do Research Participants Want?', 165 *Arch Intern Med* (2005) 652.

78 Council of Europe, 'Additional Protocol to the Convention on Human Rights and Biomedicine, Concerning Genetic Testing for Health Purposes', 15 *Eur J Health Law* (2008) 441.

provides the broadest storage and allowable secondary use of identifiable private information or biospecimens.⁷⁹ In the US, the common rule, officially known as the Federal Policy for the Protection of Human Subjects, is codified in 45 CFR 46, Subpart A. The rule provides guidelines for protecting human subjects. It allows researchers to obtain broad consent for future unspecified research uses of identifiable private information or biospecimens.⁸⁰ An analysis of 1,670 consent forms found that most (87.1%) research participants in clinical research give broad consent for unlimited future research use of their biological study.⁸¹ There are also tiered consent and dynamic consent categories. Figure 2 illustrates the type of informed consent questions individuals may answer when collecting biological samples.

Figure 2 *Sample of informed consent verbiage individuals receive when biological samples are collected*⁸²

Sample Language: Broad versus Specific Consent
(Broad consent example)

Your samples, genomic data and health information will be stored and shared with other researchers. The samples and information will be available for any research question, such as research to understand what causes certain diseases (for example heart disease, cancer, or psychiatric disorders), development of new scientific methods, or the study of where different groups of people may have come from.

(Tiered consent example)

If you wish to participate, please check your answers to the following questions.

May we collect your tissue samples, health information, and genomic information to study [state specific research project]?

Yes No

May we share your tissue samples, health information, and genomic information with other researchers to study [state specific disease or disorder]?

Yes No

May we share your tissue samples, genomic data, and health information with other researchers for future research projects related to other topics?

Yes No

(Permission to re-contact example)

Researchers might want to ask you to participate in additional studies. In some cases, you might be a particularly good candidate for a particular study because of your health history or genomic information.

May we contact you in the future to get your permission to use your samples, health information, and genomic information for additional studies?

Yes No

May we contact you in the future to ask your permission for additional samples or follow-up information about your health or medical care?

Yes No

Persons giving informed consent should be legally and cognitively capable of consenting, be free from external control and have adequate information and understanding.⁸³ It will determine how the sample and any resulting information

79 Maloy and Bass, 'Understanding Broad Consent', 20 *Ochsner J* (2020) 81.

80 Ibid.

81 Chen et al., 'Research with Stored Biological Samples: What Do Research Participants Want?', 165 *Arch Intern Med* (2005) 652.

82 'Special Considerations for Genome Research' (Genome.gov, n.d.) <https://www.genome.gov/about-genomics/policy-issues/Informed-Consent/Special-Considerations-for-Genome-Research#participants-no-consent> (accessed 5 February 2025).

83 Khan et al., 'Informed Consent for Human Genetic and Genomic Studies: A Systematic Review', 86 *Clin Genet* (2014) 199, <https://doi.org/10.1111/cge.12384>.

may be used and reused in the future. International guidelines are provided by the World Health Organization (WHO) along with other international, national and regional policies.⁸⁴ The most vigorous guidelines around informed consent are for the European Union (EU). The General Data Protection Regulation (GDPR) is the legal framework driving the informed consent requirements impacting the secondary use of biospecimens.⁸⁵

The data generated from genetic testing and retesting of stored samples will likely constitute the primary pool of genetic information used by AI for drug development. This highlights the importance of legal regulations concerning the secondary use of biospecimens. There is already significant societal and scientific support for the secondary use of biological samples, given the advancements in medicine development, population studies, the reduced need for future study participants and the prevention of waste of biological material that could contribute to medical progress knowledge.⁸⁶ Human biospecimens are vital to biomedical research and form the intellectual foundation of medicine because they represent human biology more accurately than animals do models.⁸⁷ There is tension in balancing the collective benefit of reusing biological samples for research and an individual's right to decide how these samples are used.⁸⁸

Once identifiers, like names, are removed from the biospecimens and associated, these are called 'de-identified'. This also refers to all medical and personal data associated with the biomaterials or even data from electronic health records. In most cases, the type of informed consent is no longer associated with these specimens and the resulting data.⁸⁹ When traceability is lost, it complicates an individual's right to privacy and right to control the future use of data and samples.⁹⁰

II Individual Rights to Genetic Data and the Originating Sample

When discussing genetic information, there are two types of individual rights: (i) privacy rights of genetic data and (ii) property rights of the physical sample and resulting data.

84 Mascalzoni et al., 'International Charter of Principles for Sharing Bio-Specimens and Data', 23 *Eur J Hum Genet* (2015) 721, <https://doi.org/10.1038/ejhg.2014.197>.

85 Peloquin et al., 'Disruptive and Avoidable: GDPR Challenges to Secondary Research Uses of Data', 28 *Eur J Hum Genet* (2020) 697.

86 Alkaraki et al., 'Sharing, Reuse, and Storage of Biosamples among Biomedical Researchers in Jordan: Practice and Concerns', 17 *PLoS One* (2022) e0267552.

87 Bledsoe and Grizzle, 'Use of Human Specimens in Research: The Evolving United States Regulatory, Policy, and Scientific Landscape', 19 *Diagn Histopathol (Oxf)* (2013) 322.

88 Kaye, 'The Tension between Data Sharing and the Protection of Privacy in Genomics Research', 13 *Annu Rev Genomics Hum Genet* (2012) 415.

89 Comandè and Schneider, 'Can the GDPR Make Data Flow for Research Easier? Yes It Can, by Differentiating! A Careful Reading of the GDPR Shows How EU Data Protection Law Leaves Open Some Significant Flexibilities for Data Protection – Sound Research Activities', 41 *Comput Law Security Rev* (2021) 105539.

90 Schneider and Comandè, 'Differential Data Protection Regimes in Data-Driven Research: Why the GDPR Is More Research-Friendly Than You Think', 23 *German Law Journal* (2021) 559-596.

1 Data Privacy Protections for Genetic Information

Regional privacy laws governing personal data protection regulate how most nations store, process and transfer genetic information. While specific regulations exist worldwide, some common principles regarding informed consent, non-discrimination, data minimization, de-identification and security measures apply. This document compares how genetic information and human samples are utilized and can be reused under the current frameworks in the EU, the US and China, as these countries hold the largest repositories of genetic information worldwide.

From a global perspective, the strongest privacy protections for personal information, including genetic information, are in the EU. In 2018, the EU GDPR became effective for the 27 member-states of the EU and the 3 additional countries in the European Economic Area (EEA).⁹¹ This omnibus regulation is the standard framework for processing, storing and transferring all personal data, including large-scale data collection for scientific purposes within the EU. Individual privacy rights include the right to access one's data, the right to correct one's data, the right to erasure and the right to object to processing, which are core rights under Articles 15 through 21 of the GDPR.⁹² Additionally, Articles 44-50 are overarching regulations prohibiting data transfer outside the EU unless complying with GDPR standards.⁹³ This has led to a compliant framework being adopted by 165 countries that provides for data transfer across borders.⁹⁴

Article 5(1)(b) of the GDPR introduces the concept of "purpose limitation", which states that personal data must be "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes".⁹⁵ This restricts the use of broad informed consent when collecting the sample as a means to allow all unspecified future research. However, the GDPR also states that there is a provision for archiving purposes that serve the public interest and scientific or historical research if these purposes align with the original intent purposes.⁹⁶ This is stricter than the broad consent for future unspecified research, which is allowable in the US. Under the GDPR, genetic information holds a privileged position in relation to other data categories. Member-states of the EU can establish stricter regulations. Article 9(2)(j) of GDPR permits member-states to establish a legal basis in national law that allows for processing personal data for scientific research purposes without consent.⁹⁷ The GDPR explicitly classifies

91 Solove and Schwartz, *Information Privacy Law* (8th edition, 2024), pp. 1058-1150.

92 'Global Data Protection Regulation – How GDPR Sets the International Standard' (GDPR Local, 28 October 2024) <https://gdprlocal.com/global-data-protection-regulation-how-gdpr-sets-the-international-standard/> (accessed 5 February 2025).

93 'Third Countries' (GDPR-info.eu, n.d.) <https://gdpr-info.eu/issues/third-countries/> (accessed 5 February 2025).

94 Marikyan, 'General Data Protection Regulation: A Study on Attitude and Emotional Empowerment', 43 *Behav Inf Technol* (2023) 3561, <https://doi.org/10.1080/0144929X.2023.2285341>.

95 Becker et al., 'Purpose Definition as a Crucial Step for Determining the Legal Basis under the GDPR: Implications for Scientific Research', 11 *J Law Biosci* (2024) lsae001.

96 Comandé and Schneider, 'It's Time: Leveraging the GDPR to Shift the Balance towards Research-Friendly EU Data Spaces', 59 *Common Mark Law Rev* (2022) 739.

97 'Article 9 GDPR' (GDPR-info.eu, n.d.) <https://gdpr-info.eu/art-9-gdpr/> (accessed 5 February 2025).

genetic data as a ‘special category of personal data’ or sensitive data. Article 4(13) defines genetic data as:

Personal data relating to the inherited or acquired genetic characteristics of a natural person which give unique information about the physiology or the health of that natural person and which result from an analysis of a biological sample from the natural person in question.⁹⁸

However, the GDPR distinguishes between anonymous data and personal data. Anonymous data do not have GDPR protections and can be used and reused with less scrutiny. Anonymous genetic data require considering ‘all the means reasonably likely to be used’ directly or indirectly. There are differing opinions on whether genetic information can ever be anonymous.⁹⁹ The GDPR does not explicitly state that genetic data cannot be anonymized. However, it establishes a high standard for what constitutes anonymous data. In many instances, de-identified genetic information is viewed as ‘pseudonymized’, which is still considered personal data under the GDPR.¹⁰⁰ Article 4(5) defines pseudonymization, which means that identifiers are assigned and the code linking the identifiers to the person is kept separately.¹⁰¹

Article 89 is the key provision that regulates scientific research under the GDPR.¹⁰² Article 9(2) provides specific exceptions where the processing of genetic information is allowed, including (a) explicit consent from the data subject, (b) protection of vital interests, (c) legitimate activities of certain non-profit bodies, (d) data made public by the data subject, (e) legal claims, (f) substantial public interest, (g) preventative or occupational medicine and (h) public health reasons and scientific research purposes. This research exception is applicable for AI utilization of genetic information in drug development.¹⁰³

In January 2025, a new regulation, the European Health Data Space (EHDS), was approved, establishing a cross-border digital infrastructure for secure health data exchange. This is a monumental endeavour that will take until 2029 to complete. The health data will be decentralized, but the domains will follow centralized policies that permit seamless secondary use with harmonized data formats.¹⁰⁴ This

98 ‘Article 4 GDPR’ (GDPR-info.eu, n.d.) [https://gdprhub.eu/Article_4_GDPR#\(13\)_Genetic_data](https://gdprhub.eu/Article_4_GDPR#(13)_Genetic_data) (accessed 5 February 2025).

99 Rahnasto, ‘Genetic Data Are Not Always Personal – Disaggregating the Identifiability and Sensitivity of Genetic Data’, 10 *J Law Biosci* (2023), <https://doi.org/10.1093/jlb/lsad029>.

100 Shabani and Marelli, ‘Re-Identifiability of Genomic Data and the GDPR: Assessing the Re-identifiability of Genomic Data in Light of the EU General Data Protection Regulation’, 20 *EMBO Rep* (2019) e48316.

101 Finck and Pallas, ‘They Who Must Not Be Identified – Distinguishing Personal from Non-Personal Data under the GDPR’, 10 *Int Data Priv Law* (2020) 11, <https://doi.org/10.1093/idpl/ipy026>.

102 Tzortzatou et al., ‘Biobanking Across Europe Post-GDPR: A Deliberately Fragmented Landscape’, 43 *GDPR and Biobanking* (2021), https://doi.org/10.1007/978-3-030-49388-2_22.

103 Shabani and Borry, ‘Rules for Processing Genetic Data for Research Purposes in View of the New EU General Data Protection Regulation’, 26 *Eur J Hum Genet* (2018) 149, <https://doi.org/10.1038/s41431-017-0045-7>.

104 ‘European Health Data Space’ (European Commission, 5 March 2025) https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en (accessed 17 January 2025).

will allow standardized genetic data to be fed into AI tools. The regulation supports explicitly using pseudonymized datasets under secure conditions. There will be intermediary government bodies, such as Health Data Access Bodies (HDABs), that review requests to use the data.¹⁰⁵ It has the potential to revolutionize the use and reuse of genetic information.¹⁰⁶

The GDPR has been criticized for creating unnecessary challenges to the secondary use of health data like genetic data.¹⁰⁷ The intent of EHDS is to build on GDPR Articles 6 and 9 for the lawful processing of health data and to address gaps in GDPR, such as the secondary use of data and cross-border exchange.¹⁰⁸ However, the current information about the initiative leaves some ambiguity over its ability to harmonize with the GDPR.

One example of inconsistency is the legal basis for the secondary use of health data between the EHDS and GDPR. The EHDS allows secondary use of health data for broad purposes such as “public health” and “innovation”, while Article 5 of the GDPR mandates explicit consent.¹⁰⁹ Another example is Article 9 of the GDPR, which allows member-states to enact national laws tied to their legal traditions or public policy priorities. These national laws could conflict with the EHDS framework. For instance, Spain and Italy require additional safeguards for genetic data. If the national laws create inconsistencies across member-states regarding national data protections, it is unknown how it will be addressed for harmonization with the EHDS.¹¹⁰ As implementation is done, proactive actions will be needed to align the frameworks and regulatory goals of GDPR and EHDS.

This new regulation offers many opportunities for drug development companies to access genetic data that will facilitate new therapeutics. However, there is a lack of clarity on how intellectual property (IP) would be protected when using the EHDS. IP is the foundation of the drug development business. IP protections are the foundation of the pharmaceutical industry’s innovation.¹¹¹ Without IP provisions, companies would hesitate to invest in the research and development that ensures social access to lifesaving resources and drugs.¹¹² Little information exists

105 ‘What Is the European Health Data Space (EHDS)?’ (european-health-data-space.com, n.d.) <https://www.european-health-data-space.com/> (accessed 17 January 2025).

106 Van Quathem, ‘European Health Data Space: Revolutionizing Health Care, Scientific Research in the EU’ (IAPP, n.d.) <https://iapp.org/news/a/european-health-data-space-revolutionizing-health-care-scientific-research-in-the-eu> (accessed 17 January 2025).

107 Peloquin, DiMaio and Bierer, ‘Disruptive and Avoidable: GDPR Challenges to Secondary Research Uses of Data’, 28 *Eur J Hum Genet* (2020) 697-705, <https://doi.org/10.1038/s41431-020-0596-x>.

108 Li and Quinn, ‘The European Health Data Space: An Expanded Right to Data Portability?’, 52 *Comput Law Secur Rev* (2024) 105913, <https://doi.org/10.1016/j.clsr.2023.105913>.

109 Kertesz, ‘Collaboration in Healthcare: Implications of Data Sharing for Secondary Use in the European Union’, 31(5) *Eur J Health Law* (2024) 497-517, <https://doi.org/10.1163/15718093-bja10133>.

110 Tschider, Compagnucci and Minssen, ‘The New EU-US Data Protection Framework’s Implications for Healthcare’, 11(2) *J Law Biosci* (2024) lsae022, <https://doi.org/10.1093/jlb/lsae022>.

111 ‘Intellectual Property’ (European Federation of Pharmaceutical Industries and Associations, n.d.) <https://www.efpia.eu/about-medicines/development-of-medicines/intellectual-property> (accessed 5 December 2024).

112 Consulting, ‘The Role of Intellectual Property in the Biopharmaceutical Sector’ (2022). FTI Consulting can be accessed at https://www.ifpma.org/wp-content/uploads/2023/01/i2023_2022-The-role-of-IP-in-the-biopharmaceutical-sector.pdf

on preserving IP rights and trade secrets that may exist in datasets.¹¹³ The lack of clarity regarding data protection generated by pharma has led to significant uncertainty. It is a concern that must be addressed to ensure the EHDS does not hinder innovation, which would discourage investment in research and development activities.¹¹⁴

With the EHDS, the EU will set the global standard for regulatory framework of genetic data use in the same way the GDPR has set the standard for privacy for global business. Although the US has not adopted a federal data privacy law, some sector-specific laws within the US are worth noting. In 2009, the Genetic Information Nondiscrimination Act (GINA) became effective. The regulations implementing this law are in 29 C.F.R. Part 1635.¹¹⁵ This protects individuals from genetic discrimination by health insurers and employers. This law does not explicitly affect the use of this data for drug development; however, it emphasizes the US recognition of the sensitive nature of this data.

Additionally, in February 2024, President Biden signed Executive Order 14117, “Preventing Access to American’s Bulk Sensitive Data and United States Government-Related Data by Countries of Concern”.¹¹⁶ This order lists genomic data as one of the six categories of sensitive personal data that countries of concern could exploit.

These risks may be exacerbated when countries of concern use bulk sensitive personal data to develop AI capabilities and algorithms that, in turn, enable the use of large datasets in increasingly sophisticated and effective ways to the detriment of United States national security. Countries of concern can use AI to target United States persons for espionage or blackmail by, for example, recognizing patterns across multiple unrelated datasets to identify potential individuals whose links to the Federal Government would be otherwise obscured in a single dataset. (EO 14117)¹¹⁷

Together with EO 14117, the US Department of Justice issued an Advance Notice of Proposed Rulemaking, which suggested that foreign researchers’ access to hu-

113 ‘European Health Data: Pharmaceutical Companies’ (Pinsent Masons LLP, n.d.) <https://www.pinsentmasons.com/out-law/analysis/european-health-data-pharmaceutical-companies> (accessed 17 January 2025).

114 Barton, ‘Finalizing the European Health Data Space’, 48(1) *Pharm Technol* (2024) 7-9.

115 ‘The Genetic Information Nondiscrimination Act of 2008 (GINA)’ (U.S. Department of Labor, n.d.) [https://www.dol.gov/agencies/oasam/centers-offices/civil-rights-center/statutes/genetic-information-nondiscrimination-act-of-2008/guidance#:~:text=The%20Genetic%20Information%20Nondiscrimination%20Act%20of%202008%20\(GINA\)%2C%2042,effect%20on%20November%2021%2C%202009](https://www.dol.gov/agencies/oasam/centers-offices/civil-rights-center/statutes/genetic-information-nondiscrimination-act-of-2008/guidance#:~:text=The%20Genetic%20Information%20Nondiscrimination%20Act%20of%202008%20(GINA)%2C%2042,effect%20on%20November%2021%2C%202009) (accessed 17 January 2025).

116 ‘Executive Order on Preventing Access to Americans’ Bulk Sensitive Personal Data...’ (White House, 2024) <https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2024/02/28/message-to-the-congress-on-preventing-access-to-americans-bulk-sensitive-personal-data-and-united-states-government-related-data-by-countries-of-concern/> (accessed 5 December 2024).

117 ‘Preventing Access to Americans’ Bulk Sensitive Personal Data...’ (Federal Register, 1 March 2024) <https://www.federalregister.gov/documents/2024/03/01/2024-04573/preventing-access-to-americans-bulk-sensitive-personal-data-and-united-states-government-related> (accessed 5 December 2024).

man specimens should be considered the same as genomic bulk data and prohibited data transactions for the countries concerned.¹¹⁸ The listed countries are China, Russia, Iran, North Korea, Cuba and Venezuela. Also, in 2024, there was a proposed bipartisan bill that passed the Senate Homeland Security and Governmental Affairs Committee entitled the Prohibiting Foreign Access to American Genetic Information Act of 2024.¹¹⁹ If this bill becomes law, it will ban biotechnology companies found to be owned or controlled by foreign adversaries from accessing US citizens' genetic data. At the time of writing this article in November 2024, the fate of this executive order is uncertain with the changing political landscape in the US considering the upcoming Trump presidency.

Lastly, the regulations of China should be explored because drug development companies will certainly want to market new drug products here. The Chinese pharmaceutical market is the second largest in the world, just after the US.¹²⁰ China has established strict control over genetic data through multiple regulatory frameworks that treat genetic information as a national strategic resource. The relevant regulations are (a) the 2019 Administrative Provisions on Human Genetic Resources, (b) the 2021 Biosecurity Law and (c) the 2023 Detailed Implementation Rules. These regulations primarily focus on collecting and storing samples and data within Chinese control.¹²¹ Foreign entities are prohibited from collecting Chinese genetic material or providing it abroad.¹²² This has led to clinical trial testing being done within China's borders to comply with rules around genetic data.¹²³ Large pharmaceutical companies have expressed continued commitment to doing business in China despite geopolitical challenges.¹²⁴ Although drug development companies will likely use the genetic data of Chinese citizens in conjunction with AI, the significant difference in regional security will mean drug companies will need to find a specific approach for these regions and will not be part of any harmonized global strategy.

In summary, the common principles around informed consent, non-discrimination, data minimization, de-identification and security measures exist in the EU

118 'Justice Department Issues Final Rule Addressing Threat Posed by Foreign Adversaries' Access...' (U.S. Department of Justice, 11 March 2024) <https://www.justice.gov/opa/pr/justice-department-issues-final-rule-addressing-threat-posed-foreign-adversaries-access> (accessed 5 December 2024).

119 'Hagerty, Peters Applaud Committee Approval of Bipartisan Legislation...' (Senator Bill Hagerty, 6 March 2024) <https://www.hagerty.senate.gov/press-releases/2024/03/06/hagerty-peters-applaud-committee-approval-of-bipartisan-legislation-to-protect-american-genetic-data-from-china-controlled-companies/> (accessed 5 December 2024).

120 'China Pharmaceutical Market Estimated to Reach a CAGR of 7.50% during 2024-2032' (BioSpace, n.d.) <https://www.biospace.com/china-pharmaceutical-market-estimated-to-reach-a-cagr-of-7-50-during-2024-2032-impelled-by-the-rising-geriatric-population> (accessed 5 December 2024).

121 'China's Human Genetic Resources Regulation: Implementation Rules' (China Briefing, 8 July 2022) <https://www.china-briefing.com/news/chinas-human-genetic-resources-regulation-implementation-rules-key-points-for-foreign-stakeholders/> (accessed 5 December 2024).

122 Wang, Wang and Zhang, 'Bioethics in China's Biosecurity Law: Forms, Effects, and Unsettled Issues', 8(1) *J Law Biosci* (2021), <https://doi.org/10.1093/jlb/lsab019>.

123 <https://www.cnn.com/2023/08/11/china/china-human-genetic-resources-regulations-intl-hnk-dst/index.html> (accessed 5 December 2024).

124 <https://www.sidley.com/en/insights/newsupdates/2024/06/china-provides-policy-clarifications-to-foreign-life-sciences-companies> (accessed 5 December 2024).

and US. These principles will need to be incorporated into any harmonized guidelines for utilizing genetic information with AI.

2 Property Rights to the Physical Sample and the Corresponding Genetic Data

The property rights of genetic information and associated human samples represent a complex and evolving legal landscape. In the EU and the US, the property rights of human samples have been distinguished from human reproductive materials. Here the focus will be on the property common law and regulations around non-reproductive samples like blood and saliva. Theoretically, property rights to genetic information and samples could represent different property rights. Here, a global perspective on the property rights of the physical sample and the information generated from that sample will be summarized.

The US law on this subject is based on common law, and each state within the US can have different property laws. A California Supreme Court case is the landmark case establishing the foundational precedent regarding the property rights of medical and research samples. In *Moore v. Regents of the University of California*, the court established that patients cannot claim ownership of their discarded tissues or cells used in research; patient interests should be protected through informed consent rather than property rights.¹²⁵ The court reasoned that any excised tissue was akin to donated organs, which do not carry personal property interest. The court also commented that medical research could be hindered if researchers had to verify willing donors for every cell line. The courts rely on informed consent contracts that patients sign before any procedure, establishing clear guidelines for future ownership of material. This was also upheld when Native American Havasupai Tribe sued Arizona State University for misuse of biological property.¹²⁶ The university conducted a study of the tribe, researching chronic diabetes. Their original samples were used in various other studies beyond the scope of diabetes research. These examples show the court's importance of research using human samples by not wanting to create too much hindrance to medical research.

Although the US regulation generally supports the public good created by medical research over individual's property rights over biospecimens, a few states have enacted legislation to increase the property rights of the sample donors. These states have begun legislating new genetic privacy protections through property rights legislation. For example, Florida and Alaska have enacted legislation protecting genetic privacy, declaring that DNA samples and analysis results are the property of the person sampled.^{127,128}

European courts are also reluctant to recognize a property right in human biological materials as do the US courts. The fundamental principle governing human biological materials in the EU is the "no property rule".¹²⁹ The EU lacks a unified,

125 *Moore v. Regents of Univ. of California*, 51 Cal. 3d 120, 793 P.2d 479 (1990).

126 *Tilousi v. Ariz. State Bd. of Regents*, Order no. 04-CV-1290-PCT-FJM (2004).

127 Alaska Stat. Ann. § 18.13.010 (West).

128 Fla. Stat. Ann. § 760.40 (West).

129 Demir and Stamhuis, 'Patenting Human Biological Materials and Data: Balancing the Reward of Innovation with the Order Public and Morality Exception', 18(7) *J Intellect Prop Law Pract* (2023) 546-553, <https://doi.org/10.1093/jiplp/jpad052>.

single property regulation for biomedical samples in the way the EU's GDPR protects the privacy of genetic information. The property laws over biospecimens would be at the regional level. For example, in Germany, once extracted from the body, samples are considered inanimate matter covered by German Civil Code §§854-1296, and donors are primary owners. Explicit written contracts are required to transfer property rights to biobanks or other institutions for secondary use. Biobanks cannot claim ownership through "implied consent".¹³⁰

In summary, there is no legal basis for property rights over biological specimens in most of the EU member-states and US states. However, there are a few regional exceptions. As genetic information is being used with AI, it will be challenging for drug development companies to differentiate these few outliers in large data pools to ensure the owner's legal rights are respected.

III Current Regulations of Genetic Information with AI in Drug Development

To examine the current regulations of genetic information, specifically with AI use in drug development, the overall AI policies need to be reviewed that would apply to genetic information, and, secondly, any drug development-specific policies should be reviewed. As the largest repositories of data and samples along with largest pharmaceutical companies are in EU, US and China, these are the regions where this regulation framework will be examined. The regulatory framework is dynamic at the time of writing, November 2024, and would expect to change quickly. Here the current framework will be divided into two parts: (i) AI framework in EU, US and China and (ii) current AI policies specific to drug development.

1 AI Framework in EU, US and China

EU, US and China, all have different approaches to AI risk management. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (EU AI Act) became effective in August 2024 as the world's first comprehensive regulatory framework for AI.¹³¹ This regulatory framework applies uniformly across all industries and applications rather than developing separate regulations for individual sectors. The Act employs a risk-based classification system that categorizes the AI system into four tiers based on the potential impact on fundamental rights, public safety and health.¹³² Systems posing unacceptable risks, like untargeted facial recognition, are outright

130 Simon et al., 'A Legal Framework for Biobanking: The German Experience', 15 *Eur J Hum Genet* (2007) 528-532, <https://doi.org/10.1038/sj.ejhg.5201810>.

131 'Artificial Intelligence Act' (artificial-intelligence-act.com, n.d.) <https://www.artificial-intelligence-act.com/> (accessed 17 November 2024).

132 'EU AI Act' (European Parliament Topics, 2023) <https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence> (accessed 17 November 2024).

prohibited.¹³³ The framework applies to all organization operations within the EU, serving EU users and companies utilizing AI outputs in the EU Compliance, with this legislation enforced through steep financial penalties dependent on the violation's severity.¹³⁴

Genetic information used by AI would be classified as 'high-risk', requiring strict compliance with requirements for data quality, documentation, human oversight and risk management systems. Article 6 of the AI Act describes the thresholds that lead to an AI system being 'high risk'. AI systems processing genetic data for healthcare purposes are automatically classified as high risk due to the information's highly sensitive and personal nature.¹³⁵ Its processing would also need to comply with GDPR requirements while working with the AI Act's provisions. However, the AI Act provides for specific exemptions for research purposes. AI systems and models developed and serviced solely for scientific research are excluded from the Act's scope. Recital 25, Articles 2.6 and 2.8 specify that AI systems and models specifically developed and put into service for the sole purpose of scientific research and development are excluded from the scope.¹³⁶ The European Federation of Pharmaceutical Industries and Associations (EFPIA) interprets this exemption as applying to AI-based drug development tools used solely in the research and development of medicine. The rationale behind this is the development of medicines in Europe is already subject to a sufficient regulatory framework that can address the development of medicines.¹³⁷

In contrast to the EU, the US approach to AI risk management can be characterized as more decentralized, with federal regulation, guidance and state-level initiatives. The National Artificial Intelligence Initiative Act (NAIIA) became effective in 2021. The Act establishes the National Artificial Intelligence Initiative under 15 U.S.C. § 941169.¹³⁸ The Initiative aims to ensure continued US leadership in AI research and development, promote the development of trustworthy AI systems, prepare the workforce for AI integration and coordinate AI activities across government agencies.¹³⁹

In October 2023, President Biden expanded the foundational work from the NAIIA with an executive order to establish a framework for "Safe, Secure and Trustworthy Development and Use of AI", which delegated implementation to fed-

133 'Artificial Intelligence Act' (artificial-intelligence-act.com, n.d.) <https://artificialintelligenceact.eu/high-level-summary/> (accessed 17 November 2024).

134 Aboy, Minssen and Vayena, 'Navigating the EU AI Act: Implications for Regulated Digital Medical Products', 7 *NPJ Digit Med* (2024) 237, <https://doi.org/10.1038/s41746-024-01232-3>.

135 Braun, Vallery and Benizri, 'What Are the High-Risk AI Systems within the Meaning of the EU's AI Act, and What Requirements Apply to Them?' 1(6) *Global Regulatory Dev J* (2024) 429-435.

136 'Recital 25' (artificialintelligenceact.eu, n.d.) <https://artificialintelligenceact.eu/recital/25/> (accessed 17 November 2024).

137 'EFPIA Statement on the Use of AI in the Medicinal Product Lifecycle in the Context of the AI Act' (The European Federation of Pharmaceutical Industries and Associations, 29 January 2024) <https://efpia.eu/news-events/the-efpia-view/statements-press-releases/efpia-statement-on-the-use-of-ai-in-the-medicinal-product-lifecycle-in-the-context-of-the-ai-act/> (accessed 17 November 2024).

138 15 U.S. Code § 9411.

139 H.R. 6216 (116th Congress).

eral agencies like the FTC, EEOC and DOJ.¹⁴⁰ Additionally, multiple states have enacted differing bills involving AI systems. It is a fragmented approach that seems to lack a cohesive strategy.¹⁴¹

National Institute of Standards and Technology's (NIST) AI Risk Management Framework (RMF) was first released as a draft on 17 March 2022, with a final release on 26 January 2023.¹⁴² The NIST AI RMF is a voluntary framework that builds off the Organisation for Economic Co-operation and Development's (OECD) Framework for the Classification of AI Systems by offering comprehensive suggestions on when and how risk can be managed throughout the AI lifecycle. A recent study provides an in-depth analysis of the NIST AI RMF's application to facial recognition systems. The study demonstrates the framework's utility in identifying and mitigating risks in high-stakes AI applications. The study proposes a six-step process tailored to the specific challenges of surveillance technology, aiming to produce a more systematic and effective risk management practice based on the NIST AI RMF. This study could be used as a blueprint for companies that are looking to implement a robust RMF.¹⁴³

China does not have an overarching AI law, but several AI regulations govern AI development and deployment.¹⁴⁴ The current legal framework is focused on regulating algorithms, application development and use. The laws apply to Chinese citizens as well as foreign entities that would use AI products to provide services to Chinese citizens. A draft ethical review measure would be applicable if it becomes effective.¹⁴⁵ This measure applies broadly to science and technology activities having 'ethical risks'. Examples of these activities are the R&D of algorithm models, applications and systems with public opinion attributes or social mobilization capabilities.¹⁴⁶ China also has significant AI regulation in the works.¹⁴⁷ The government oversight from

140 'Fact Sheet: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence' (The White House, 30 October 2023) <https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/> (accessed 17 November 2024).

141 'The State of State AI Laws 2023' (Electronic Privacy Information Center, 2023) <https://epic.org/the-state-of-state-ai-laws-2023/> (accessed 17 November 2024).

142 'AI RMF Knowledge Base' (National Institute of Standards and Technology, n.d.) https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF (accessed 17 November 2024).

143 'Application of the NIST AI Risk Management Framework to Surveillance Technology' (ARXIV, n.d.) <https://arxiv.org/pdf/2403.15646> (accessed 17 November 2024).

144 'China's New AI Regulations' (Latham Watkins, 2023) <https://www.lw.com/admin/upload/SiteAttachments/Chinas-New-AI-Regulations.pdf> (accessed 17 November 2024).

145 'AI Watch Global Regulatory Tracker: China' (White & Case LLP, n.d.) <https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-china> (accessed 17 November 2024).

146 Chun, de Witt and Elkins, 'Comparative Global AI Regulation: Policy Perspectives from the EU, China, and the US' (2024), <https://doi.org/10.48550/arXiv.2410.21279>.

147 'Navigating AI's Uncharted Waters: Insights from China's Model AI Law, EU AI Act' (IAPP, 2023) <https://iapp.org/news/a/navigating-ai-s-uncharted-waters-insights-from-china-s-model-ai-law-eu-ai-act> (accessed 17 November 2024).

China makes it an unlikely place for global drug companies to develop AI technology but much more likely to choose a more regulatory-friendly environment like the US.

Table 1 High-level comparison of approaches to AI regulation¹⁴⁸

	EU	US	China
Regulatory Approach	Centralized approach with comprehensive AI Act	Decentralized, sector-specific approach with voluntary guidelines	Regulations for algorithm registering, addressing deep fakes, security assessments
Risk Classification	Four-tier system: unacceptable, high, limited and minimal risk	No formal classification	Focus on specific high-risk applications
Enforcement	European AI Board and national authorities	Distributed across federal agencies	Government oversight with a focus on the private sector
Key Focus	Protection of fundamental rights and consumer safety	Innovation and self-regulation	State control and economic development
Prohibited Uses	Specific bans (social scoring, facial recognition)	No blank prohibitions	Strict control
Implementation	Comprehensive AI Act	State by State	Three administrative regulations for specific AI types
Compliance Requirements	Documentation and reporting requirements	Voluntary guidelines and industry self-regulation	Service provider liability and transparency requirements

2 Current AI Policies Specific to Drug Development

As pharmaceutical drug development is an international endeavour, there is a global systematic approach to discovering, developing and bringing medicines to market worldwide. The International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH) currently plays a pivotal role in this effort. ICH has successfully developed globally harmonized guidelines through ICH Expert Working Groups.¹⁴⁹ These work with national regulatory bodies to create laws with basic scientific, legal and ethical standards that promote good laboratory, clinical and manufacturing practices. Though there may be some differences

148 Chun, de Witt and Elkins, ‘Comparative Global AI Regulation: Policy Perspectives from the EU, China, and the US’ (2024), <https://doi.org/10.48550/arXiv.2410.21279>.

149 ‘ICH Offers Updates on Guidelines Working Groups’ (Regulatory Affairs Professionals Society, 10 January 2019) <https://www.raps.org/News-and-Articles/News-Articles/2019/1/ICH-Offers-Updates-on-Guidelines-Working-Groups> (accessed 17 November 2024).

in the laws, the guiding principles are the same globally to promote safe and effective drugs worldwide.¹⁵⁰

As genetic information has become widely available, there has been some work to align its use and reuse from a scientific, legal and ethical perspective. Legal and ethical concerns are rooted in the right to privacy for sensitive information, such as genetic information. Particularly in the EU, data protection is a central element in the rights-oriented society. Secondary use of genetic information is subject to some regulations that balance privacy protection and enable beneficial research. On a global scale, multiple regulatory bodies have some form of guidelines around secondary use of data. However, there has not been a concerted effort for global harmonization on guidelines for the secondary use of genetic information.¹⁵¹ With the rapid advancements of AI, the reuse of this data will become more prevalent. ICH has yet to create specific guidelines dedicated to AI. The European Medicines Agency (EMA) has published a draft reflection paper on the use of AI in the drug product lifecycles.¹⁵²

E Risks of Not Proactively Establishing a Global Framework for Genetic Data Use with AI in Drug Development

Genetic information's uniqueness and its potential for the public good in developing lifesaving medicines create tensions with privacy rights, security and consent interests. In the 1990s, when genetic sequencing became a reality, bioethical considerations were raised that genetic information, which is uniquely personal, deserves special handling. This school of thought, where genetic information is handled separately from other medical information, is called "Genetic Exceptionalism".¹⁵³ The uniqueness is attributable to many personally identifiable facts in a single and portable biological sample that can be turned into data at present and in the future.¹⁵⁴ Genetic information's proven advances and success in identifying new drug products create a strong financial incentive for its increasing use in drug development. The global pharmaceuti-

150 'International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH)' (European Medicines Agency, n.d.) <https://www.ema.europa.eu/en/partners-networks/international-activities/multilateral-coalitions-initiatives/international-council-harmonisation-technical-requirements-registration-pharmaceuticals-human-use-ich> (accessed 17 November 2024).

151 Massella, Dri and Gramaglia, 'Regulatory Considerations on the Use of Machine Learning Based Tools in Clinical Trials' 12(6) *Health Technol (Berl)* (2022) 1085-1096, <https://doi.org/10.1007/s12553-022-00708-0>.

152 'Reflection Paper on use of Artificial Intelligence in the Medicinal Product Lifecycle' (European Medicines Agency, 21 June 2024) <https://www.ema.europa.eu/en/news/reflection-paper-use-artificial-intelligence-lifecycle-medicines> (accessed 17 November 2024).

153 Ilkilic, 'Coming to Grips with Genetic Exceptionalism: Roots and Reach of an Explanatory Model', 1 *Med Stud* (2009) 131-142, <https://doi.org/10.1007/s12376-009-0015-7>.

154 Slokenberga, 'Special Issue: You Can't Put the Genie Back in the Bottle: On the Legal and Conceptual Understanding of Genetic Privacy in the Era of Personal Data Protection in Europe', 15 *BioLaw J* (2021) 223-250, <https://doi.org/10.15168/2284-4503-783>.

cal industry revenue reached US\$1.6 trillion in 2023.¹⁵⁵ This lucrative nature of the industry creates a compelling economic incentive to incorporate genetic information in the drug development process.

Advancements in DL technology in AI have greatly enhanced the capacity to analyse large amounts of complex data, including genetic information. However, not all industries have the resources to be early adopters of this new technology, as developing and training DL models require substantial investment.

Currently, 35% of global revenue in the drug development sector is concentrated among the top 10 pharmaceutical companies.¹⁵⁶ These key players possess the capital to invest in technologies that facilitate the creation of more innovative medicines. Consequently, the integration of genetic information with AI tools is already underway, exemplified by the partnership between GSK and 23andMe. The economic incentive will only accelerate this trend alongside the rapid progress of DL technology.

Nevertheless, increasing reliance on genetic information raises significant concerns regarding individual privacy, security and control over managing this sensitive data. These concerns create tensions with the potential for public health in developing lifesaving medicines. Understanding the risks can lead to proactive measures to avoid any infringement on an individual's rights. The following summarizes some of the risks associated with privacy rights, security and issues of consent.

I Privacy Risks Associated with the Use of Genetic Information in Drug Development

Genetic data ARE uniquely identifiable and cannot be truly anonymized since each person's DNA sequence is unique (except for identical twins).¹⁵⁷ There are limited regulations in both the US and EU around the use of de-identified genetic information. The regulation of de-identified genetic information from being used for research purposes and shared with other research organizations is limited. Even when data are de-identified, individuals can potentially be re-identified by combining genomic data with public records and genealogical databases. Traditional privacy models designed for health data provide limited protection for genomic data. An attacker may learn sensitive information about a target individual by exploiting the dependency between genomic data and other publicly available information, such as family name, demographic data and observable features (e.g. eyes and hair colour).¹⁵⁸ This not only impacts the individual but also blood relatives and future generations. No other single source of personal information can detail so much about an individual and their family.

155 'Global Pharmaceutical Industry' (Statista, n.d.) <https://www.statista.com/topics/1764/global-pharmaceutical-industry/> (accessed 17 November 2024).

156 'World's Top 10 Pharma Companies by 2023 Revenue' (Xtalks, 12 April 2024) <https://xtalks.com/worlds-top-10-pharma-companies-by-2023-revenue-3850/> (accessed 17 November 2024).

157 'Privacy' (National Human Genome Research Institute, 13 March 2023) <https://www.genome.gov/about-genomics/policy-issues/Privacy> (accessed 17 November 2024).

158 Bonomi, Huang and Ohno-Machado, 'Privacy Challenges and Research Opportunities for Genomic Data Sharing', 52(7) *Nat Genet* (2020) 646-654, <https://doi.org/10.1038/s41588-020-0651-0>.

Re-identification has already been shown to be possible from raw anonymous genetic code. A 2013 article highlighted the high privacy risk that accompanies genetic information.¹⁵⁹ A profile from the database GenBank was used in conjunction with a computer algorithm, which cross-referenced with genealogical information to identify the family name. Even if the raw genetic information is unavailable, multiple examples of processed genetic information have been retrieved from shared ML models.¹⁶⁰

Another way genomic data could be exposed is through shared ML models.¹⁶¹ One example is the researcher's use of yeast genomic data to demonstrate the privacy leaking of a ML model. They also proposed a solution for keeping this from happening, along with one way to defend against it called Differential Privacy (DP). DP helps to prevent de-identification by adding just enough "noise" to the data to protect individuals while still getting useful overall information.¹⁶²

Although genetic data analysis has the potential to unlock lifesaving medicines and is being used in academia and research for the good of society, it also has the potential to be used by bad actors. This information could be exploited in many ways, so the exposure of this individual's genetic information raises concerns. For example, genetic information could be used for insurance discrimination to deny coverage. It could be used to identify women of childbearing potential, which has implications in any society that means to control reproductive rights. It also can be used to target specific ethnic or demographic groups. This has already been done with genetic data from the 23andME data breach targeting ancestry data.¹⁶³ The breach specifically targets users with Ashkenazi Jewish heritage, exposing approximately 1 million data profiles of people with Jewish heritage. The hacker referred to the leaked information as "Ashkenazi DNA data" and said it was in retribution for the Israel-Hamas war.¹⁶⁴ These data were sold and shared on the dark web.¹⁶⁵ It contained the name, sex, date of birth, geographical location, genetic ancestry results, family surname, grandparents' birthplace and profile photos.¹⁶⁶

Just as genetic information could be used to develop therapeutic treatments by targeting specific genetic code variations, it can also be used to create harmful

159 Gymrek et al., 'Identifying Personal Genomes by Surname Inference', 339(6117) *Science* (2013) 321-324, <https://doi.org/10.1126/science.1229566>.

160 Ibid.

161 Chen, Wang and Shi, 'Differential Privacy Protection against Membership Inference Attack on Machine Learning for Genomic Data', 26 *Pac Symp Biocomput* (2021) 26-37.

162 Bonomi, Huang and Ohno-Machado, 'Privacy Challenges and Research Opportunities for Genomic Data Sharing', 52(7) *Nat Genet* (2020) 646-654, <https://doi.org/10.1038/s41588-020-0651-0>.

163 'What to Do If You're Concerned about the 23andMe Breach' (Electronic Frontier Foundation, 9 October 2023) <https://www.eff.org/deeplinks/2023/10/what-do-if-youre-concerned-about-23andme-breach> (5 December 2024).

164 Collier, '23andMe User Data Targeting Ashkenazi Jews Leaked Online' (*NBC News*, 6 October 2023) <https://www.nbcnews.com/news/us-news/23andme-user-data-targeting-ashkenazi-jews-leaked-online-rcna119324> (accessed 5 December 2024).

165 'Attorney General Tong Issues Inquiry Letter to 23andMe Following Data Breach' (State of Connecticut Department of Attorney General, 6 October 2023) <https://portal.ct.gov/ag/press-releases/2023-press-releases/attorney-general-tong-issues-inquiry-letter-to-23andme-following-data-breach> (accessed 5 December 2024).

166 'Attorneys Weigh the Fallout of 23andMe Data Breach' (*The Daily Journal*, 17 October 2023) https://www.dailyjournal.com/articles/376899?t&utm_source (accessed 5 December 2024).

biological agents targeting particular genetic codes. The concern here is that genetic information has the potential to be used to develop bioweapons targeting particular ethnic groups. This was even mentioned as the US House of Representatives passed the BIOSECURE Act in September 2024.¹⁶⁷ This act would protect Americans' genetic information from being accessed by foreign governments. Republican Congressman and Chairman of the Select Committee on Strategic Competition with China Mike Gallaher specifically raised concerns about the Chinese Communist Party developing bioweapons with data to target Americans.¹⁶⁸ At the time of writing, this had not passed the US Senate.

In addition to concerns about bad actors using genetic information, there could be a concern that the market demand for more genetic data will lead to the reuse of genetic information. This could lead to autonomy and control issues. Sample donors may risk losing control over how their personal information is used and shared.

II Reuse of Genetic Information and Samples in Storage

As pharmaceutical companies train DL models for drug development, there will be an increased demand for genetic information. This demand for genetic information will surpass what is currently available in large academic biobanks. The convenient way for pharmaceutical companies to fill this need is to reuse genetic information from previous studies or create new genetic information from samples stored in earlier studies that were not used initially for genetic testing. As previously mentioned, this has already been observed in the four-year collaboration between GSK and 23andMe.¹⁶⁹

A huge biobanking industry has already developed storing samples that have been or are currently used in research and personal DNA testing. Biobanks are a large collection of biospecimens linked to relevant personal and health information for health and medical research. Many of these biobank's samples are pseudonymized to assure donor privacy. Pseudonymization is a data management and de-identification procedure that replaces personally identifiable information with artificial identifiers called pseudonyms. The link between the identifiers and the personally identifiable information is separately stored. This is distinct from anonymization, which has no link to personally identifiable data once removed.

This secondary use of DNA samples and data has increased biobanking efforts worldwide.¹⁷⁰ As clinical and research institutions generate vast amounts of genetic and genomic data through large-scale biobanking efforts, sharing these data be-

167 'The US Government Takes Actions to Protect Vital Biological Datasets' (Council on Strategic Risks, 15 January 2025) <https://councilonstrategicrisks.org/2025/01/15/the-us-government-takes-actions-to-protect-vital-biological-datasets/> (accessed 7 February 2025).

168 'Bill to Ban Foreign Adversary Biotech Companies, including BGI Group' (Select Committee on the Chinese Communist Party, 15 May 2024) <https://selectcommitteeontheccp.house.gov/media/bills/bill-ban-foreign-adversary-biotech-companies-including-bgi-group> (accessed 7 February 2025).

169 '23andMe Announces Collaboration Extension & New Data Licensing Deal with GSK' (23andMe Holding Co., 16 July 2018) <https://investors.23andme.com/news-releases/news-release-details/23andme-announces-collaboration-extension-new-data-licensing> (accessed 7 February 2025).

170 Brothers et al., 'Eliciting Preferences on Secondary Findings: The Preferences Instrument for Genomic Secondary Results', 19(3) *Genet Med* (2017) 337-344, <https://doi.org/10.1038/gim.2016.110>.

tween institutions has become an important topic. The global biobanking industry was valued in 2023 between US\$75 and US\$80 billion, with anticipated growth trajectories of US\$200 billion by 2033.¹⁷¹ Regionally, the largest players in the industry are in Europe and the US, with the Asia Pacific region showing a fast growth rate. Biobanks can be independent companies that research organizations compensate for storing these samples, or biobanks can be within the research organizations themselves. Large biobanks exist within hospitals, research centres, pharmaceutical companies and patient advocacy organizations.¹⁷²

No estimates publicly available on the number of samples stored worldwide in these biobanking facilities, but one large-scale biorepository reports maintaining 60 million samples. An indication of the size of the business may be inferred by the size of the automated biospecimen storage systems commercially available and marketed to biobanks. One of those automated systems accommodates up to 14,400,000 samples.¹⁷³ The biobanking industry supports a market for biospecimen storage systems, and managing millions of samples with one system may indicate enormous quantities of biospecimens being stored globally. It is not unreasonable to speculate the number of samples in these biobanks globally is at least in the hundreds of millions.

The volumes of samples in storage make biobanks a great source of genetic information for drug development companies. The biobanking industry is already primed to provide this large amount of data that have already been collected or to analyse samples in storage. This leads to multiple concerns: the voluntariness of the testing becomes a grey area as an individual is essentially volunteering for testing for their family, for which they share significant portions of DNA.¹⁷⁴ Lastly, the issue of future reanalysis of genomic data raises questions about the obligation to discuss the possibility as part of the initial informed consent.

The secondary use of genetic information becomes more relevant as AI capabilities increase. The market for genetic information reused for drug development in AI has already started. Commercialization of health data has become increasingly prevalent. Private biobanks are ultimately a business. New revenue streams may likely be developed as the market grows for genetic information to train and be used in AI-driven DL models. As the value of an individual's data increases, it is likely to be sold and resold, and an individual will not know in what type of research their data are being used. The nation's largest private equity firm, Blackstone, acquired Ancestry.com for US\$4.7 billion in 2020. Blackstone was prepared to pay US\$261 for each person's genetic profile at the time of sale. It is worth noting that Blackstone also holds stakes in insurers, drugmakers and research labs. This raises concerns about the future use of genetic information, considering that

171 'Biobanking Market' (Straits Research, n.d.) <https://straitsresearch.com/report/biobanking-market> (accessed 7 February 2025).

172 Annaratone et al. 'Basic Principles of Biobanking: From Biological Samples to Precision Medicine for Patients', 479(2) *Virchows Arch* (2021) 233-246, <https://doi.org/10.1007/s00428-021-03151-0>.

173 'Automated Sample Storage System' (Azenta Life Sciences, n.d.) <https://www.azenta.com/products/bioarc-flex-80-c-automated-sample-storage-system> (accessed 7 February 2025).

174 Rego et al., 'Informed Consent in the Genomics Era', 10(8) *Cold Spring Harb Perspect Med* (2020) a036582, <https://doi.org/10.1101/cshperspect.a036582>.

Ancestry's privacy policy states that the company may use individuals' genetic data for scientific research and developing new products and services.¹⁷⁵

As genetic information increases, there may also be an increase in the cost of individuals paying for samples to generate genetic information.¹⁷⁶ Some smaller "bio-broker" companies have already developed business models around paying individuals for genetic data.¹⁷⁷ Luna DNA allows users to monetize existing genetic data from services like 23and ME.¹⁷⁸ EncryptoGen has launched the first blockchain-based genomic data marketplace.¹⁷⁹ Nebula Genomics offers whole-genome sequencing with the option to sell or rent data to pharmaceutical companies. This company has also deployed blockchain technology to secure data management. It enables direct sales between individuals and researchers.¹⁸⁰

As in any industry, some biobanking and genetic testing companies will not be successful; what happens with all these samples when companies fail? Genetic information can also be sold when a biobank business fails, as these samples would most likely be sold as assets. The fate of data and samples may be uncertain. For example, Atlas Biomed, a London-based DNA testing company, offered insight into people's genetic predispositions. The website for this company went offline, and the firm seems to have disappeared, with no trace of it since 2023.¹⁸¹ The people who gave DNA samples and associated data do not know the fate of their data or samples. An investigation by the BBC reported in October 2024 that this company has links to Russia, with the two officers listed as having the same address in Moscow. There was no evidence of misuse of the genetic data, but the links to Russia could be concerning for those who had given genetic data to the company.¹⁸²

Once someone's genetic information is in the commercial market, there is no more individual control. The critical question that needs answering is whether this lack of control is outweighed by the collective good that can arise from researching

175 Lazarus, 'Why Spend Billions for Ancestry's DNA Data if You Don't Plan to Use It?' (*Los Angeles Times*, 13 April 2021) <https://www.latimes.com/business/story/2021-04-13/column-blackstone-ancestry-genetic-privacy> (accessed 7 February 2025).

176 Koplin, Skeggs and Gyngell, 'Ethics of Buying DNA', 19(3) *J Bioeth Inq* (2022) 395-406, <https://doi.org/10.1007/s11673-022-10192-w>.

177 'Companies to Help People Sell or Rent Out Their Health Data' (*The Scientist*, 31 July 2017) <https://www.the-scientist.com/companies-to-help-people-sell-or-rent-out-their-health-data-41890> (accessed 7 February 2025).

178 'Luna DNA Review' (*Nebula Genomics*, 17 February 2023) <https://nebula.org/blog/lunadna-review/?srsltid=AfmBOOpUtApDrs5zl9McWgxuEQkoxXGxqi6yCIPxPhnLTxTLruwBJFts> (7 February 2025).

179 'First Blockchain-Based Genomic Data Marketplace Launches' (*The Scientist*, 4 February 2019) <https://www.the-scientist.com/first-blockchain-based-genomic-data-marketplace-launches-65041> (7 February 2025).

180 'Nebula Genomics Launches \$299 Direct-to-Consumer Whole Genome Sequencing' (*MobiHealthNews*, 20 November 2018) <https://www.mobihealthnews.com/news/nebula-genomics-launches-299-direct-consumer-whole-genome-sequencing> (7 February 2025).

181 'Atlas Biomed Ceases Operations, Leaving Client Genetic Data Inaccessible' (Today's General Counsel, 15 February 2023) <https://todaysgeneralcounsel.com/atlas-biomed-ceases-operations-leaving-client-genetic-data-inaccessible/> (7 February 2025).

182 'Atlas Biomed Ceases Operations, Leaving Client Genetic Data Inaccessible' (Today's General Counsel, 15 February 2023) <https://www.bbc.com/news/articles/cz7wl7rpdj0> (7 February 2025).

this genetic information.¹⁸³ This discussion needs to be proactive before capitalism has commodified genetic information to a scale that would hinder attempts to regulate.

As previously shown, re-identification is sometimes possible once someone's genetic information is available. Even if an individual consented with full knowledge of the risks of re-identification, the relatives of the same individuals likely did not consent to having their genetic information on the marketplace. This legal issue and the ethical issue of commodification of DNA itself need to be proactively assessed. As the value of genetic information increases, the risks to individuals' privacy and control over their genetic information also will increase.

F Approach to Global Regulatory Framework for Genetic Data Use

A balanced approach is essential for safeguarding privacy and preventing the misuse of genetic information while promoting beneficial research and development. Implementing a proactive international strategy for AI governance in drug development could help alleviate potential future challenges. Without a unified framework, agencies such as the FDA and EMA may respond independently by establishing region-specific guidelines. The EHDS framework aims to serve as a solution to these challenges; however, the complexities of implementation mean it will take years to realize its full potential. During the EHDS implementation phase, an organization with extensive expertise in global drug development standards could bridge the existing gaps and provide a structured framework for the use of genetic information in drug development with AI.

An organization already exists whose function is to harmonize global drug development regulations. The International Council for Harmonization of Technical Requirements for Human Use (ICH) aims to develop standard guidelines and collaborate with regulatory agencies.¹⁸⁴ The ICH has not directly addressed the issue of AI in drug development. They will need to address this topic, specifically the use of genetic data, before different agencies across the globe create independent regulations. A multidisciplinary group from regulatory agencies, drug development businesses and academia should be brought together to develop the guidelines. These guidelines should take into consideration the long-term goals of EHDS. These guidelines should balance the risks to individuals' rights to control their genetic information and the public good derived from advanced medical research. Bringing together representatives from drug development companies, regulatory agencies and academic leaders will provide that balanced perspective. Based on the risks identified through this research, the guidelines should be developed from assessments in four key areas focusing on informed consent requirements for primary

183 Koplin, Skeggs and Gyngell, 'Ethics of Buying DNA', 19(3) *J Bioeth Inq* (2022) 395-406, <https://doi.org/10.1007/s11673-022-10192-w>.

184 'Demystifying the Definition of ICH (International Council for Harmonisation)' (BioAccess, 28 March 2023) <https://blog.bioaccessla.com/demystifying-the-definition-of-ich-international-council-for-harmonisation/> (7 February 2025).

and secondary use of biospecimens and associated data and control over future use of one's sample.

- a Genetic data collection and use of human biospecimens stored globally in the biobanking industry.
- b Individual property and privacy rights to their sample and data associated data considering differing global regulations.
- c Data traceability.
- d Data storage governance and cybersecurity.

In addition to global standards, drug development companies should implement proactive internal policies to minimize the risks associated with using genetic information alongside AI. These internal policies should encourage the responsible utilization of AI and genetic data, enabling drug companies to harness this valuable tool. Addressing potential risks proactively would demonstrate a good-faith effort to protect individuals' rights related to genetic information data.

Since the pharmaceutical industry is already highly regulated, the existing ICH framework for drug development can be used as a guide for the identification of internal policies that drug development companies could implement to address potential risks associated with the use of AI on genetic information in drug development. The ICH guidelines for drug development are in four categories: Quality, Safety, Efficacy and Multidisciplinary.¹⁸⁵ The ICH guidelines serve as a roadmap for identifying policies that companies utilizing genetic data and AI can implement, ensuring alignment with the existing frameworks common in many drug development organizations. Below are some recommended policy considerations, streamlined into concise recommendations for drug development companies to adopt voluntarily, promoting the responsible use of genetic information in conjunction with AI. This proactive and systematic assessment of the key components of the ICH guidelines related to the internal use of genetic data and AI would strategically position pharmaceutical companies for future submissions to regulatory agencies. Furthermore, leveraging the ICH guidelines to develop a framework facilitates easier integration into the existing quality infrastructure of most drug development businesses.

The ICH Quality Category has 14 topics, with guidance on quality infrastructure for drug development.¹⁸⁶ Three of these topics are Pharmaceutical Development (ICH Q8), Quality Risk Management (ICH Q9) and Pharmaceutical Quality System (ICH Q10). These three guidelines work synergistically throughout the drug development lifecycle to create a comprehensive approach to pharmaceutical quality management, emphasizing preemptive, risk-based decision-making.¹⁸⁷ These topics can be used to direct the policy considerations for AI use in drug

185 'ICH Guidelines' (International Council for Harmonisation, n.d.) <https://www.ich.org/page/ich-guidelines> (accessed 7 February 2025).

186 'Quality Guidelines' (International Council for Harmonisation, n.d.) <https://www.ich.org/page/quality-guidelines> (accessed 7 February 2025).

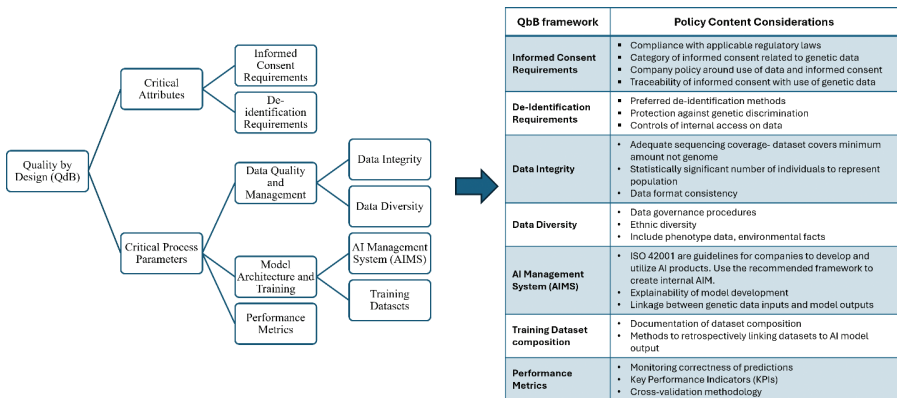
187 'Integrating Artificial Intelligence and Big Data in Pharmaceutical Development: Ethical Considerations and Legislative Frameworks' (eternal project, n.d.) https://www.eternalproject.eu/downloads/publications/ai_and_big_data_in_pharma_dev.pdf (accessed 7 February 2025).

development. Elements from these three ICH Quality Topics can serve as a roadmap for pharmaceutical companies to manage the risks associated with using genetic information with AI.

ICH Q8, Q9 and Q10 are not legally binding regulations by either the FDA or EMA. However, both regulatory bodies have issued recommended adoption guidelines. This has led to an internal infrastructure in many drug development companies based on the regulatory agency's recommendation. This quality infrastructure can be adapted for drug development companies to identify key internal policies needed when utilizing genetic information with AI. The following is an example of how guidelines established following the ICH guidelines in Q8, Q9 and Q10 could be used to proactively identify policy content to mitigate their use of genetic information in drug development.

ICH Q8 focuses on pharmaceutical development and introduces the concept of quality by design, which is relevant to AI applications because, in many cases, these are being developed within these organizations.¹⁸⁸ Quality by Design (QbD) is a proactive, science-based approach to determining critical process parameters and quality attributes. The guidelines in this area give direction on content requirements for technical documents and enhanced process understanding. Using a QbD approach can lead to inherent quality policies that mitigate some risks associated with using genetic data in drug development.¹⁸⁹ Figure 3 demonstrates how elements from the ICH guidelines can create a quality infrastructure for AI use within drug development organizations. A robust AI management system will be one of the most important parts of effectively using genetic information with AI in drug development.

Figure 3 *ICH Q8 guidelines can direct policy content considerations on the use of genetic information with AI during drug development that develops AI mitigating the critical risks*



188 'ICH Guidelines' (International Council for Harmonisation, n.d.) https://database.ich.org/sites/default/files/Q8_R2_Guideline.pdf (accessed 7 February 2025).

189 Yu et al. 'Understanding Pharmaceutical Quality by Design', 16(4) AAPS J (2014) 771-783, <https://doi.org/10.1208/s12248-014-9598-3>.

Another global organization, the International Organization for Standardization (ISO), has already released a standard on AI management that can be referred to by drug development companies to implement a robust AI management system. The ISO develops and publishes international standards for various industries, including the pharmaceutical sector. Many quality groups within drug development companies already utilize ISO standards to establish, implement and maintain robust quality management systems. ISO 42001 is a framework for implementing robust AI management systems within organizations. These standards offer comprehensive guidelines for corporations to establish, implement, maintain and continually improve their AI governance system.¹⁹⁰ ISO 42001 is the world's first international standard focused on the governance of AI management systems (AIMS). It provides a structured approach to implementing and maintaining AI systems within organizations. The key features of ISO 42001 include ethical AI development and risk and impact assessment. To achieve this, corporations will need to establish policies and controls.¹⁹¹

ISO 42001 establishes a comprehensive framework for transparency and accountability in AI decision-making processes. This standard emphasizes several key principles and practices to achieve these goals. ISO 42001 acknowledges the complexity of DL models and provides several approaches to address this challenge. First, it requires detailed documentation of AI processes, including data sources, algorithms and decision rationale. Second, it uses a modular approach of breaking the AI system into manageable modules. Lastly, it emphasizes the execution of AI risk assessment processes.

The quality management system within ISO 42001 requires organizations to maintain detailed documentation of their AI systems, including the data sources, algorithms applied and rationale behind system decisions. This documentation helps identify and rectify bias and errors, bridging process gaps in AI development. The standard also emphasizes the need for explainability and interpretability of AI models, addressing technical gaps created by opaque systems, like DL networks. These opaque systems create technical challenges for AI transparency. Advanced AI systems, particularly those based on DL, are often called “black boxes” due to their intricate computations, making it challenging to explain their decisions clearly.¹⁹² This complexity poses a significant obstacle to achieving the transparency and explainability requirements of ISO 4200 if pharmaceutical companies begin to build the tools needed for auditing AI systems.¹⁹³ A strong AI management system is an

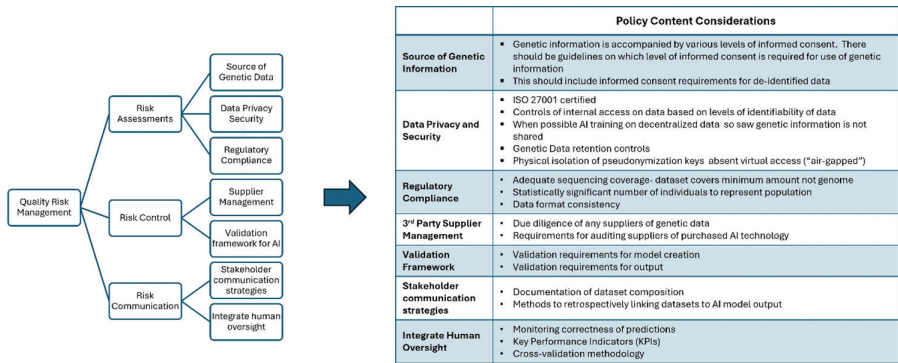
190 ‘ISO/IEC 42001:2023 Information Technology -Artificial Intelligence- Management System’ (ISO, 2023) <https://www.iso.org/standard/81230.html> (accessed 7 February 2025).

191 Ranjbar et al., ‘Managing Risk and Quality of AI in Healthcare: Are Hospitals Ready for Implementation?’ 17 *Risk Manag Health Policy* (2024) 877-882, <https://doi.org/10.2147/RMHP.S452337>.

192 ‘The Black Box Problem: Opaque Inner Workings of Large Language Models’ (Prompt Engineering Guide, 29 April 2024) <https://promptengineering.org/the-black-box-problem-opaque-inner-workings-of-large-language-models/> (accessed 10 January 2025).

193 De Palma et al., ‘Identification and Assessment of Risks in Biobanking: The Case of the Cancer Institute of Bari’, 14(14) *Cancers (Basel)* (2022) 3460, <https://doi.org/10.3390/cancers14143460>.

Figure 4 *Policy considerations for quality risk management of using genetic information with AI as part of the drug development process*



essential piece of designing quality into the processes and systems that use sensitive information like genetic information.

ICH Q9 focuses on a framework for risk management, ensuring processes around risk assessments, risk control and risk communication are part of drug development. Figure 4 highlights how a quality organization could identify risks and policy content to mitigate these risks. One key policy consideration drug development should address is data privacy and security. Most global drug companies have already implemented policies to adhere to the GDPR. However, there should still be an assessment to ensure adequate data privacy controls when using genetic information with AI. Again, drug development companies can turn to ISO for a framework for managing data privacy and security. ISO has developed a framework for organizations to protect sensitive information by managing security risks. If drug companies using genetic information become ISO 27001-certified, they have already implemented a robust framework for protecting sensitive genetic data.¹⁹⁴

ISO 27001 is a widely recognized standard for establishing, implementing, maintaining and continually improving an Information Security Management System (ISMS).¹⁹⁵ It provides the framework for identifying, assessing and mitigating information security risks. It also offers organizational, people, physical and technological controls that can be implemented. Its primary purpose is to protect an organization's information assets' confidentiality, integrity and availability.¹⁹⁶

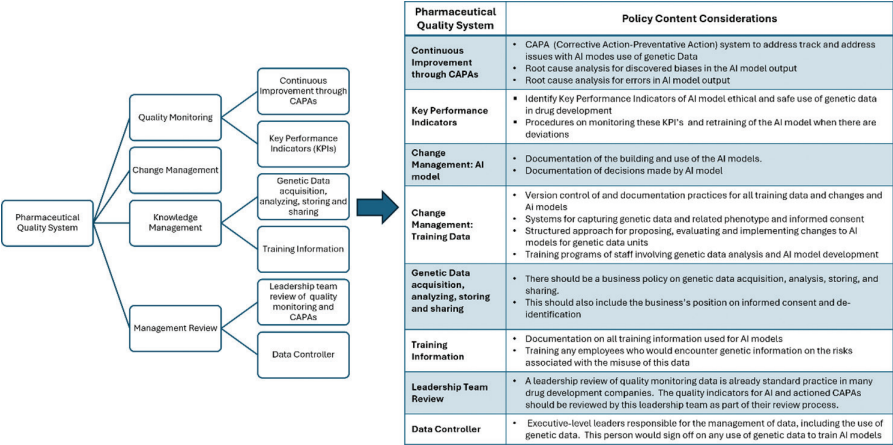
Risks associated with genetic data will require internal company policies to ensure the ethical and responsible use of this sensitive information, for example, a

194 McIntosh et al. 'From COBIT to ISO 42001: Evaluating Cybersecurity Frameworks for Opportunities, Risks, and Regulatory Compliance in Commercializing Large Language Models', 144 *Comput Sec* (2024) 103964. <https://doi.org/10.1016/j.cose.2024.103964>.

195 'Information Security, Cybersecurity and Privacy Protection Requirements' (ISO, 2022) <https://www.iso.org/standard/27001> (accessed 5 January 2025).

196 'The Ultimate Guide to ISO 27001' (isms.online, 2024) <https://www.isms.online/iso-27001/> (accessed 5 January 2025).

Figure 5 Shows the Types of policy content that would result from the pharmaceutical quality system guidelines (ICH Q10)



policy on the source of genetic information and the level of informed consent required. Figure 4 identifies policy content to create a quality RMF for genetic data use with AI.

ICH Q10 outlines a model for an effective pharmaceutical quality system (PQS).¹⁹⁷ The guidance applies to all stages of the drug development lifecycle. The key elements are management responsibilities, process performance, product monitoring, CAPAs (corrective and preventive actions), change management and management review (Figure 5). It outlines an effective PQS that integrates quality monitoring, change management and knowledge management.¹⁹⁸

As part of knowledge management, one of the key policies that should be in every pharmaceutical company's policy is acquiring, storing and sharing genetic information. The Global Alliance for Genomics and Health (GA4GH) has established foundational genomic data lifecycle principles.¹⁹⁹ This is an international, non-profit collaborative alliance of over 500 organizations across various sectors interested in the responsible use of genomic information to promote health.²⁰⁰ GA4GH has developed policy guidance documents around informed consent, data privacy and security, ethical review and cloud security.²⁰¹ The policies share core

197 'ICH Guidelines' <https://database.ich.org/sites/default/files/Q10%20Guideline.pdf> (accessed 7 February 2025).

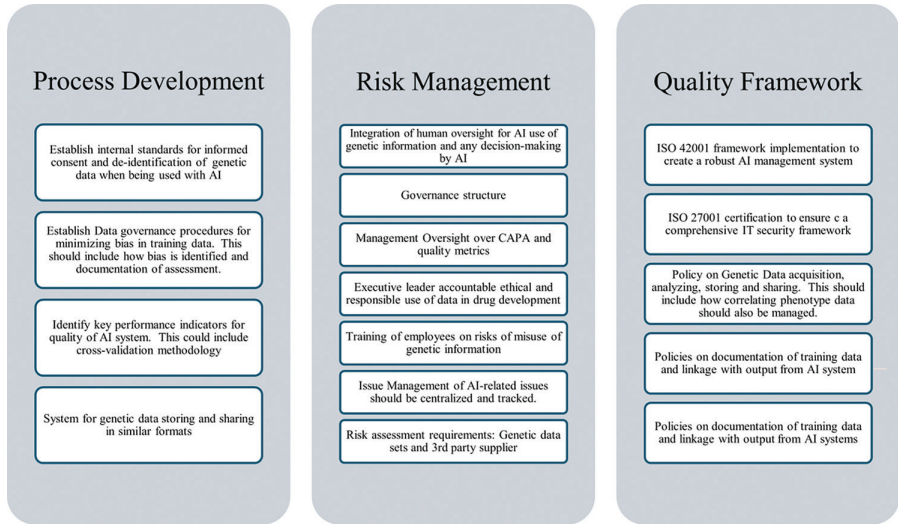
198 International Conference on Harmonisation. ICH Q10 PHARMACEUTICAL QUALITY SYSTEM. 2008.

199 Page et al., 'A Federated Ecosystem for Sharing Genomic, Clinical Data', 352 *Science* (2016) 1278-1280, <https://doi.org/10.1126/science.aaf6162>.

200 'GA4GH Constitution' (Global Alliance for Genomics and Health, 1 March 2023) <https://www.ga4gh.org/constitution/> (accessed 15 December 2024).

201 Rehm et al., 'GA4GH: International Policies and Standards for Data Sharing across Genomic Research and Healthcare', 1(2) *Cell Genom* (2021) 100029. <https://doi.org/10.1016/j.xgen.2021.100029>.

Figure 6 *High-level summary of recommendations for drug development companies using AI and genetic information*



elements of transparency, accountability and sustainability. There are tools and standards to support the implementation of the framework.²⁰²

The policy considerations identified by following the ICH Q8, Q9 and Q10 guidelines as a compass can be organized into three categories: Process Development, Risk Management and the Overarching Quality framework. These categories provide high-level guidance for drug development companies using genetic information with AI (Figure 6).

As AI deep learning improves, the pharmaceutical industry will increasingly leverage this technology to discover new therapeutics. The global nature of drug development means a strong need for a harmonized, international approach to AI governance in drug development needs to exist. As this will take time, pharmaceutical companies should proactively establish their internal quality control framework to mitigate the risks associated with using genetic information. Using the already established guidelines from ICH can be a compass for creating an internal regulatory framework. This study illustrated this approach and identified 16 topics within process development, risk management and a quality framework (Figure 6).

In conclusion, if proactive measures are not taken to balance the privacy of individual genetic information, then reactive regulations are likely to be implemented in response to inevitable privacy violations on a worldwide scale. However, proactive regulation is far more effective, as it safeguards individual rights and

202 'International Partners Describe Global Alliance to Enable Secure Sharing of Genomic and Clinical Data' (Broad Institute, 15 June 2016) <https://www.broadinstitute.org/news/international-partners-describe-global-alliance-enable-secure-sharing-genomic-and-clinical-data> (accessed 15 December 2024).

prevents significant harm. In contrast, reactive measures tend to come too late, allowing extensive damage to unfold, which erodes public trust and diminishes social capital in the process. Lifesaving drugs will be developed using AI in conjunction with genetic data. A proactive regulatory framework can minimize the social cost of these medicines.