

Statement for the Record for the U.S.-China Economic and Security Review Commission
“Taking a Bigger Byte: China’s Expanding Strategy for Data Dominance”

April 30, 2026
Prof. Yves Moreau
Full professor, University of Leuven, Belgium

China’s strategic vision of genomic data

Chinese authorities have developed a strategic vision across multiple industrial sectors. This has led to absolute dominance not only in cheap industrial manufacturing, but also solar energy, electric vehicles, lithium-ion batteries, or 5G telecom infrastructure. By contrast, China still lags in the pharmaceutical and biotechnology sectors, but over the past twenty years Chinese authorities have systematically put the biopharmaceutical sector as a strategic priority in each of its five-year plans. Just as China is trying to rapidly catch up in semiconductor design, it is also pushing ahead in biotech and pharma. At the level of biological research and health research, as a country, China is now second only to the United States (Nature Index, 2026). This rapidly increasing research performance is driven by a major commitment to research funding that is now overtaking the United States (Nature, 2026).

Data (rather than equipment, as for example for semiconductor technology) is now the main fuel of biotech research. While governments from all developed countries recognize the value of data, Chinese authorities consider genetic data as a strategic national asset and sets its policies accordingly. This has been framed through the 2019 P.R.C. Regulation on the Management of Human Genetic Resources and its 2022 Implementation Guidelines (Allen-Ebrahimian, 2022). This regulation formalized and strengthened earlier guidance. Key elements of this regulation are (1) human genetic data as a strategic national asset, (2) requirement for licensing from the Ministry of Science and Technology (MOST) to export human genetic data and samples, (3) the mapping of human genetic resources available across the country through surveys, and (4) a focus on “important gene families” and “human genetic resources of specified regions”. As a result collaboration between Western and Chinese scientists on the analysis of biological data from Chinese participants has become increasingly difficult. From a limited number of personal discussions, my impression is that Western scientists are simply disengaging from research on Chinese research participants, see the discussion on the Chinese Kadoorie Biobank below as a case in point. Chinese authorities do enforce this regulation. In 2018, AstraZeneca (Sweden/Britain) and BGI (China, aka Huada Gene) were sanctioned separately for violations of the Human Genetic Resources Regulation (TKDB, 2018). In 2024, AstraZeneca was again investigated by Chinese authorities for “data privacy and import breaches” (Makortoff, 2024). This limited enforcement seems to have been sufficient to make researchers and companies aware that breaches of this regulation carry serious consequences. It seems that essentially zero genetic data trickles out of China without a license from MOST and only few licenses are awarded.

Genomic vs. biometric data

It is important to distinguish several categories of data that raise different issues. Human biomedical and genomic data primarily raises privacy and economic issues. Biometric data raises privacy and human rights issues.

Genomic and medical data

Genome sequencing (full genome or partial) is the first type of data that comes to mind. Its specificity is that the genome of an individual is essentially unique to that person and stable across his or her lifetime. If leaked, it cannot be replaced by a new identifier. Techniques exist to reidentify a person from their genome sequence alone (Gymrek, 2013), so that a genome should not be considered “anonymized”. Some medical and physical characteristics can to a variable extent be predicted from the genome sequence. Which characteristics will become highly predictable in the future is still unclear. From that perspective, it is difficult to assess future risks associated with the disclosure of genomic data.

Other types of genomic data, such as tumor genome sequencing, transcriptomic, and epigenomic data, often contain a substantial fraction of the genome information of the person from which it was collected, even if many researchers overlook this fact.

Non-genomic clinical data can contain highly sensitive medical information (about mental health, infectious disease, etc.). While the leakage of genome data tends to evoke strong concerns, the leakage of sensitive medical data, if it can be traced back to a specific person, can be a much more concrete risk. Even when a data set does not appear to directly identify individuals (e.g., when names, addresses, and other HIPAA identifiers have been removed), the appropriate privacy standard for data to be considered personal data is whether concrete strategies can be outlined, possibly using other data sets to which an attacker could reasonably gain access, to reidentify participants (El Emam, 2011; de Montjoye, 2013).

The right to privacy for patients and research participants does not require the demonstration of an actual harm. Privacy is a right based in personal autonomy, the right of people to carry out their lives without undue interference and without being subjected to the unchecked power of the state or other individuals. As put by Justice Brandeis, it is “the right to be let alone” (Warren and Brandeis, 1890).

Next to privacy issues, human genomic and clinical data raise economic issues. Countries want their economy to prosper and thus gain competitive advantages. Globalized unfettered trade was for several decades a prevailing goal. More protectionist policies, sovereignty, and resilience have in the past few years gained traction. Given the economic weight of the healthcare and pharmaceutical sectors, geopolitical concerns cannot be fully set aside. It is reasonable to ask for economic competition to be “fair”, although what fair means might be perceived very differently by different countries. If research and its industrial translation is to be fair, it should respect privacy and be transparent.

Biometric data

Biometric data is data that allows the reidentification of individuals for security, law enforcement, surveillance purposes, etc. This includes fingerprints, facial recognition, video tracking, gait recognition, iris scans, DNA, etc. Forensic DNA data comprises a variety of technologies aiming at unambiguously assigning a DNA trace to an individual, or predicting ancestry or physical characteristics (eye color, hair color, etc.). The national Chinese DNA database (the Chinese equivalent of the US CODIS database) is the largest in the world. Chinese authorities are secretive about its current size with the last reliable figure of 100 million DNA profiles dating back to 2020 (Dirks and Leibold, 2020). Reasonable estimates would put the size of this database to over 150 million DNA profiles as compared to about 20 million profiles in the US CODIS database.

A surprising instance of an international database containing extensive genetic data from Chinese citizens is the male forensic DNA database YHRD (www.yhrd.org) based in Germany, which contains 350,000 male Y-Chromosome forensic DNA profiles, 132,000 (38%) of which are from men across China. This database has been developed in close collaboration with Chinese public security forces, yet it is based outside China (Nothnagel, 2022, Retracted).

In my work, I have not seen substantial evidence that Chinese authorities have been collecting DNA profiles from foreign citizens for use by law enforcement or for mass surveillance purposes. Chinese forensic genetics researchers have contributed to research on foreign population and may have retained some foreign forensic genetic data, but this does not appear to amount to a structured effort by authorities.

BGI Genomics

BGI Genomics is the leading provider of genome sequencing services worldwide. In terms of supplying sequencing instrument, the US company Illumina is the leading supplier. BGI also sells instruments, primarily through its MGI subsidiary.

BGI has been linked to the People's Liberation Army and forensic DNA profiling in Xinjiang. Through my research, I have confirmed that BGI has been involved in forensic DNA profiling in Xinjiang through its subsidiary FGI. Nevertheless, BGI did not appear to be among the top suppliers of such services in Xinjiang. US companies, Thermo Fisher Scientific and Promega, as well as several Chinese companies (AGCU ScienTech, PeopleSpot, etc.) seemed at least as important actors as FGI in Xinjiang (Wee, 2021). Through a review of the scientific literature via Web of Science, I identified a few research articles involving both researchers affiliated with BGI and the PLA. This was a small fraction (25 out of 6,935) of all research articles involving BGI. Most of these articles did not appear problematic and involved typical medical research with researchers from PLA affiliated hospitals. The only research topic of concern were two articles related to high-altitude adaptation, a major research topic for the PLA as Han soldiers and settlers controlling the Tibetan plateau face serious health concerns due to chronic mountain sickness (as opposed to the genetically and physiologically adapted Tibetan population). Given the scale of the research in which BGI is involved (with at least seven thousand research publications), this does not provide evidence of a large-scale structured collaboration between BGI and the PLA (Needham and Jacobsen, 2021). BGI was also linked to the PLA through the use of the Tianhe supercomputers (Needham and Baldwin, 2021). While this observation is correct, it does not provide evidence of a strong link between BGI and the PLA. In the US, academic and commercial researchers can access the world's Top 3 supercomputers at Department of Energy national labs (top500.org). This does not imply that all research carried out at these supercomputers is linked to the DoE's nuclear research programs.

BGI has also been linked to broad collection of data for prenatal genetic tests (which detect chromosomal anomalies, such as Trisomy 21, before birth) offered by its subsidiary NIFTY (Needham and Baldwin, 2021). While a few publications seem to involve studies of the NIFTY test by clinicians associated with PLA affiliated hospitals, the claim that "BGI developed the prenatal tests with the PLA" seems a stretch. NIFTY's activities do nevertheless raise substantial privacy concerns. It has carried out millions of tests worldwide. Where this data is stored, when it is deleted, who has access to it, and for which purposes remain unclear. BGI's privacy policy (which applies to NIFTY) is woefully inadequate. It states that "BGI will retain the Personal Data for no longer than it is necessary for the purposes as long as needed to provide the Services requested from BGI or requested by applicable laws and regulations. Beyond the above retention period, we will delete or anonymize your Personal Data." The possibility that "anonymized" genomic data could be retained after the service has been provided is highly problematic. It also states that "we will never share or disclose your Personal Data without obtaining your consent unless when:

- It is directly relevant to public health or significant public interest;
- It is directly relevant to investigation, prosecution, trial and execution of judgment of crimes;
- It is for the purpose of protecting life, property or other significant legal rights and interests of Personal Data subjects or others, and it is difficult to obtain consents from such person;
- The Personal Data collected has been disclosed by the Personal Data Subject to the public actively; or
- It is otherwise provided by applicable laws and regulations."

Since it is unclear under which jurisdiction the data falls, this standard might be interpreted under Chinese law.

Also, pictures of an electronic a display monitoring the NIFTY database, taken at BGI China National GeneBank in Shenzhen in May 2017 (see Appendix 1), indicated that the Global NIFTY database contained 1.99 million samples at the time, while the China NIFTY database contained 1.77 million samples. This strongly suggests that researchers in mainland China had then access to NIFTY DNA profiles from foreign customers.

BGI also offers genomic sequencing services where customers send their DNA samples to one of eleven BGI labs across the world and sequencing data is then returned electronically.

Because of these substantial privacy concerns, I consider that the use of NIFTY services and more generally BGI genome sequencing services present an irreducible and unacceptable privacy risk. While BGI contracts with clinical and research institutions have requirements that the data shall be deleted after the delivery of the service has been finalized (KCE, 2018), for these clauses to be enough of a deterrent, it would require that (1) breaches of contract can be detected and (2) local legal enforcement acts as a sufficient deterrent. If diversion of copies of the data were to be organized internally, this might be extremely difficult to detect, even if a lab is externally audited (e.g., ISO/IEC 27001 Information Security Management Systems). Moreover, while breaches of contract and of privacy laws might bring a company like BGI to a local court, the Chinese state reaches powerfully abroad and could be much more threatening to BGI management than foreign courts. The disappearance of several CEOs include Jack Ma (Alibaba), Bao Fan (China Renaissance), Xiao Jianhua (Tomorrow Group) for extended periods certainly signals that the management of major Chinese companies is not beyond the reach of the Chinese state.

Case study: UK Biobank vs. China Kadoorie Biobank vs. All of Us Research Program

To illustrate the risk of asymmetry in access to clinical and genomic data internationally, we consider three major health and genomic data initiatives: the UK Biobank, the China Kadoorie Biobank, and the All of Us Research Program.

The UK Biobank is a long-term prospective biobank study in the United Kingdom (UK) that houses the de-identified biological samples and health-related data of half a million people. The volunteer participants aged 40–69 were recruited between 2006 and 2010 from across Great Britain and consented both to share their health data and to be followed for at least 30 years thereafter, with the aim to enable scientific discoveries into the prevention, diagnosis, and treatment of disease. (From https://en.wikipedia.org/wiki/UK_Biobank)

The China Kadoorie Biobank (CKB) is one of the world's largest prospective studies and is a long-term collaboration between the University of Oxford, Peking University and the Chinese Academy of Medical Sciences. More than 512,000 adults were recruited from ten geographically defined and diverse areas of China between 2004 and 2008 and extensive data was collected at baseline and subsequent periodic resurveys. The health of participants has been monitored over about two decades. (From <https://www.ckbiobank.org/about-us/aims-and-rationale>)

Note that the CEO of UK Biobank is Prof. Rory Collins from the University of Oxford and its Chief Scientific Officer is Prof. Naomi Campbell, also from the University of Oxford. Note that for researchers outside China and Hong Kong, the China Kadoorie Biobank is operated from Oxford University.

The All of Us Research Program is a large-scale precision medicine initiative launched by the U.S. National Institutes of Health (NIH) that aims to gather health data from at least one million diverse participants across the United States to accelerate biomedical research and improve health outcomes. The aggregated, de-identified dataset is made available to approved researchers to study the genetic, environmental, and lifestyle factors underlying a broad range of diseases and conditions, with the long-term goal of enabling more personalized approaches to prevention, diagnosis, and treatment. The evolution of these initiatives provides a clear view of data sharing asymmetry. The UK Biobank has been a resounding academic success (albeit with notable concerns regarding privacy risks) with over 20,000 researchers from over 60 countries using its data in over 18,000 peer-reviewed publications. The impact of the All of Us program is now increasing rapidly. By contrast, the China Kadoorie Biobank has not achieved the same level of success. I carried out a rough analysis of publications labeled with the topic “UK Biobank” or “Kadoorie” in Web of Science, see Appendix 2. While over 16,000 publications are linked to “UK Biobank” in Web of Science, over 500 publications are linked to “Kadoorie”, and over 3,500 publications are linked to “All of Us”.

UK Biobank data seems to be studied by researchers across many countries, although the UK, the US, and China appear dominant in terms of countries and institutions. All of US data seems to be studied by researchers across many countries, although the US appear dominant as a country and in terms of institutions. By contrast, it appears that studies on the China Kadoorie Biobank almost always involve a Chinese collaborator. From outside China, contributors seem to mainly come from the University of Oxford. Growth of references to the China Kadoorie Biobank appear to have stalled since 2021. It is unclear whether in practice access to this data corresponds to its declared open science commitment. One key potential explanation is the stringent restrictions on the sharing of primary genetic data. One of the key articles about the genetic analysis of the China Kadoorie Biobank” (Walters, 2023) explicitly states that “sharing of genotyping data is currently constrained by the Administrative Regulations on Human Genetic Resources of the People’s Republic of China. Access to these and certain other data is available through collaboration with CKB researchers.”

Other biometric data: video tracking and person identification and misuse of the DukeMTMC data set

Going beyond genomic data, I want to briefly mention another surveillance technology: video tracking and person reidentification. I am currently carrying out a comprehensive study of the misuse of a data set collected on the campus of Duke University. The Duke Multi-Target Multi-Camera (DukeMTMC) tracking data set was initially collected in 2014 and published in 2016 (Ristani, 2016). In 2019, it appeared that the data set had been collected in breach of its ethical approval (https://exposing.ai/duke_mtmc/) and Duke quickly removed the data. This data set is still widely used. We have collected and are currently analyzing over 3,000 scientific articles referencing the DukeMTMC data in 2020 and later, after its official removal. We currently estimate that three quarters of these articles actually use the data set in spite of its removal and in clear breach of basic ethical principles of academic research. We currently estimate that over 80% of these articles involve researchers from China. As such, this data set has served, through ethically questionable research, as a major tool for the development of surveillance technology by Chinese researchers. A detailed report will be available after the summer.

Privacy-preserving methods as a privacy-enhancing layer

Several approaches have been proposed to protect the privacy of research participants, such as research analysis platforms, federated learning, and synthetic data.

The most mature approach are Research Analysis Platforms, which are cloud-based environments where researchers log in to process the parts of the database to which they given access. This means that researchers are vetted and their research proposal is evaluated against the restrictions placed on the data (for example, in terms of allowed research purposes). Moreover, the use of the data on the platform is logged by the data controller. Nevertheless, the UK Biobank recently suffered a notable breach (Devlin, 2026) where part of the UK Biobank data was listed for sale on Alibaba. Note that Alibaba and Chinese authorities were fully cooperative in removing the offending files from the internet. (This incident followed earlier ones not involving China (Burgis, 2025)). In response, the UK Biobank promised to “implement an automated ‘airlock’ that checks files and data” that are sent out of the platform. Moreover, the UK Biobank seems to have an active process in place to scan internet sources for UK Biobank files. Even in this tightly controlled environment, it would be practically impossible to detect statistical analysis code that produce results that “encode” individual-level participant data in an obfuscated manner.

Another approach is federated learning, where several partners keep their data under their control at all times and the analysis protocol is sent to all of the partners and carried out locally, and only statistically aggregated results are sent back to an aggregating node. For example, if a consortium wanted to compute the average weight of all participants in a research study, each center would count the number of research participants it holds and compute the sum of their weights. Then all participating centers would send back to the aggregating node their research participant counts and the sum of their weights. The aggregating node then divides the sums of the sums of the weights by the

sum of the participant counts. In this way, the average weight is computed but no individual weights were shared with the aggregating node. Thus, the average weight is computed in a fully privacy-preserving way. Unfortunately, when this approach is repeated on many variables at the same time or similar approaches are applied to learning complex models with many parameters, a number of statistical attacks emerge that can retrieve some information about the data or the model and can potentially breach privacy. It is difficult to fully preclude such attacks because they leverage core statistical properties of complex, high-dimensional data.

A third family of approaches is synthetic data generation where the privacy-sensitive data is used to generate data that does not correspond to any participant, but retains the essential statistical properties of the underlying data for further analysis. This can be done by for example injecting noise that “blurs” the data to make the original records unrecognizable. The key challenge is that, to guarantee privacy, synthetic data may necessitate an amount of perturbation that severely limits its usefulness for further analysis. Obtaining reliable synthetic data is notoriously tricky and may essentially require carrying out the analysis for which it is intended to check whether it is truly useful – which defeats the purpose of using synthetic data.

In short, no approach provides an airtight mathematical guarantee that no attack against the data is possible. This should not be interpreted as meaning that these approaches are without merit. As a comparison, no real-world IT security platform can provide airtight mathematical guarantees against intrusion, denial of service, phishing, etc. Real-world cybersecurity depends on a layered approach against defined threat models and commensurate cybersecurity efforts. For example, private devices cannot be fully guarded against a motivated state actor, but can be protected against routine attacks. Approaches such as federated learning, synthetic data, or data analysis platforms can be embedded into a global approach to data security that blends such approaches with contractual commitments, legal enforcement, logging, user education, data surveillance, etc. to bring risks against well-defined threat models down to an acceptable level. The evolution of UK Biobank practices in reaction to breaches and in response to criticism illustrates how such frameworks need to be continuously developed and deployed.

Balancing privacy risks vs. harms to the research enterprise

While we wish that research participants face no privacy risk, such zero risk seems unachievable with current technology without catastrophically impeding vital research. The central role of genomic research as the foundation of much modern clinical and pharmaceutical research cannot be overemphasized. Advances such as targeted cancer therapies, CAR-T immunotherapy in cancer, advanced prenatal testing and newborn screening, AlphaFold for protein structure prediction (a key enabler of future AI-driven drug discovery), etc. would not have been possible without this foundation of genomic research.

Under the double effect principle, the goal should not be to eliminate all risks if doing so also eliminates all benefits, but rather to search for ways to obtain most benefits while bringing risks and harms down to a level that is acceptable and as low as possible practically (i.e., searching for a low-risk, high-benefit tradeoff). By way of analogy, participants to clinical trials often face a small managed risk of foreseen and unforeseen harms from a new pharmaceutical product or therapeutic intervention. Foreseen risks (for example of an allergic reaction) are reduced to the lowest level practically possible (for example, through an appropriate protocol to handle any allergic reaction). Unforeseen risks are by essence difficult to manage, but strenuous efforts are made to anticipate risks as best as possible. As a result of this proper management, severe harms in clinical trials are now relatively rare.

The academic research also differs substantially from commercial environments. While for example in a pharmaceutical research environment, the cost of the regulatory burden (e.g., cost of clinical trials, including regulatory compliance) will be inputted to the cost of the final product, the dynamics of academic research differs substantially. Academic research is financially driven by research grants and socially by research publications. While the regulatory burden translates into an accounting cost

in a commercial environment, it largely translates into an opportunity cost in an academic environment. If red tape slows down research resulting in substantially fewer or lower profile academic publications, scientists might anticipate that they will not be able to get their next grant or that they or their researchers might not be able to move their career forward. Given the immense pressure to “publish or perish” and towards grant acquisition, biomedical researchers are in a constant cycle of allocating scarce resources (funding, human resources) to ensure scientific output to ensure future funding. It will not take much for many of them to decide that research on human data is a career dead-end and that they should switch to some other field of biological research, for example animal research, or for data scientist to move to some other application area of artificial intelligence. This would lead to a critical gap in ensuring that biomedical research translates into actual clinical and pharmaceutical advances.

The research enterprise is fragile, yet incredibly valuable as a source of welfare and well-being. We cannot afford to break it while trying to fix it.

References

- (Allen-Ebrahimian, 2022) Allen-Ebrahimian B. China makes genetic data a national resource. Axios. 2022 Mar 29. <https://www.axios.com/2022/03/29/china-makes-genetics-data-national-resource>
- (Burgis, 2025) Burgis T. Revealed: Chinese researchers can access half a million UK GP records. The Guardian. 2025 Apr 15. <https://www.theguardian.com/technology/2025/apr/15/revealed-chinese-researchers-access-half-a-million-uk-gp-records>
- (de Montjoye, 2013) de Montjoye YA et al. Unique in the Crowd: The privacy bounds of human mobility. Sci Rep. 2013;3:1376.
- (Devlin, 2026) Devlin H. Private health records of half a million Britons offered for sale on Chinese website. The Guardian. 2026 Apr 23. <https://www.theguardian.com/technology/2026/apr/23/private-health-records-uk-biobank-chinese-website-alibaba>
- (Devlin and Burgis, 2026) Devlin H. and Burgis T. Confidential health records from UK BioBank project exposed online. The Guardian. 2026 Mar 14. <https://www.theguardian.com/science/2026/mar/14/confidential-health-records-exposed-online-uk-biobank>
- (Deighton, 2026) Deighton B. China could be the world’s biggest public funder of science within two years. Nature. 2026 Mar 19. <https://www.nature.com/articles/d41586-026-00618-5>
- (Dirks and Leibold, 2020) Dirk E and Leibold J. Genomic surveillance: Inside China’s DNA dragnet. Policy Brief Report No. 34/2020. Australian Strategic Policy Institute. <https://www.aspi.org.au/report/genomic-surveillance/>
- (El Emam, 2011) El Emam K et al. A systematic review of re-identification attacks on health data. PLoS One. 2011;6(12):e28071.
- (Gymrek, 2013) Gymrek M et al. Identifying personal genomes by surname inference. Science. 2013 Jan 18;339(6117):321-4.
- (KCE, 2018) Belgian Health Care Knowledge Center. The use of whole genome sequencing in clinical practice: challenges and organisational considerations for Belgium. 2018 Feb 19. https://kce.fgov.be/sites/default/files/2021-11/KCE_300_Whole_genome_Sequencing_Report.pdf
- (Makortoff, 2024) Makortoff K. China detains five AstraZeneca staff over ‘data privacy and import breaches’. The Guardian. 2024 Sep 5. <https://www.theguardian.com/business/article/2024/sep/05/china-detains-five-astrazeneca-staff-in-investigation-over-data-privacy-and-import-breaches>
- (Murgia, 2019) Murgia M. Who’s using your face? The ugly truth about facial recognition. The Financial Times. 2019 Apr 19. <https://www.ft.com/content/cf19b956-60a2-11e9-b285-3acd5d43599e?syn-25a6b1a6=1>
- (Nature Index, 2026) Nature Index. Country/territory outputs. Nature. 2026. <https://www.nature.com/nature-index/country-outputs/China>
- (Needham and Baldwin, 2021) Needham K and Baldwin C. China’s gene giant harvests data from millions of women. Reuters. 2021 Jul 7. <https://www.reuters.com/investigates/special-report/health-china-bgi-dna/>

(Needham and Jacobsen, 2021) Needham K and Jacobsen S. Monkey-brain study with link to China's military roils top European university. Reuters. 2021 Nov 18. <https://www.reuters.com/world/exclusive-monkey-brain-study-with-link-chinas-military-roils-top-european-2021-11-18/>

(Nothnagel, 2022, Retracted) Nothnagel et al. Retraction Note to: Revisiting the male genetic landscape of China: a multi-center study of almost 38,000 Y-STR haplotypes. Hum Genet. 2022 Jan;141(1):175-176.

(Ristani et al., 2016) Ristani E et al. Performance measures and a data set for multi-target, multi-camera tracking. European conference on computer vision. 2016 Oct 8.

(Smyth and Peel, 2026) Smyth C. and Peel M. UK Biobank says security checks not imposed because of 'harms' to research. The Financial Times. 2026 Apr 24. <https://www.ft.com/content/99dcafff-f4e9-4361-8d59-e765ca9fc6f0?syn-25a6b1a6=1>

(TKDB, 2018) Administrative penalty decisions. Traditional Knowledge Database. 2018 Nov 23. <http://www.cntkdb.com/index.php/Index/News/index/ClassA/1/ClassB/3/id/735/use/1>

(Walters, 2023) Walters RG et al. Genotyping and population characteristics of the China Kadoorie Biobank. Cell Genom. 2023 Jul 20;3(8):100361.

(Warren and Brandeis, 1890) Warren SD and Brandeis L. The Right to Privacy. Harvard Law Review. 1890 Dec 15;193.

(Wee, 2021) Wee S-L. China Still Buys American DNA Equipment for Xinjiang Despite Blocks. The New York Times. 2021 Jun 11. <https://www.nytimes.com/2021/06/11/business/china-dna-xinjiang-american.html>

Appendix 1 – BGI NIFTY Shenzhen database

Pictures taken at BGI China National GeneBank in Shenzhen in May 2017. The Global NIFTY database contains 1.99 million samples, while the China NIFTY database contains 1.77 million samples. Credit: Dr. Ausma Bernot, Griffith University, Australia.



Appendix 2 – Bibliometric analysis of UK Biobank, China Kadoorie Biobank, and All of Us

The data is obtained from Web of Science (webofknowledge.com) using the advanced query on topics linked to a publication:

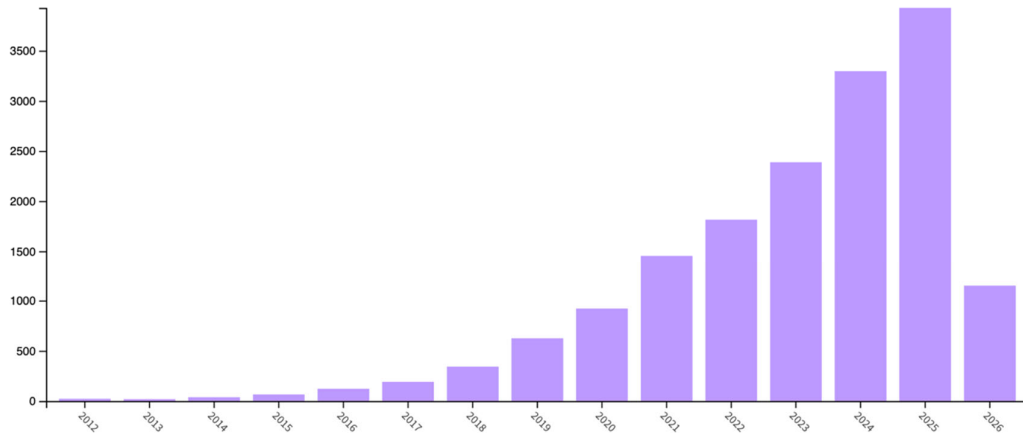
- ts = “UK Biobank” (N=16,446)
- ts = “Kadoorie” (N=551)
- ts = “All of Us” (N=3,672).

Ideally, the analysis should be further validated (as an example, mentions of the Palestine Technical University Kadoorie led to false positive matches).

Note that the scale of the histograms is substantially different. Note that the size of the boxes in the color diagrams is unfortunately not directly proportional to their value.

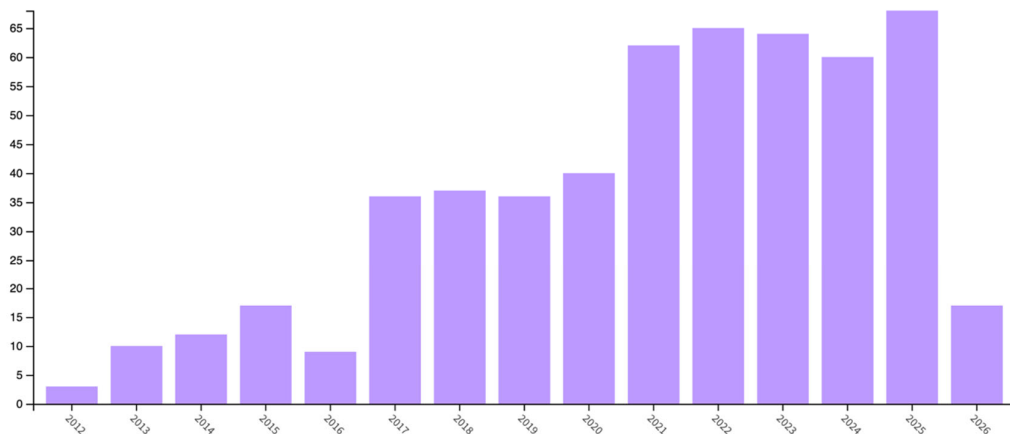
UK Biobank – Mentions per year

Number of publications with topic “UK Biobank” per year of publication. (N=16,446)



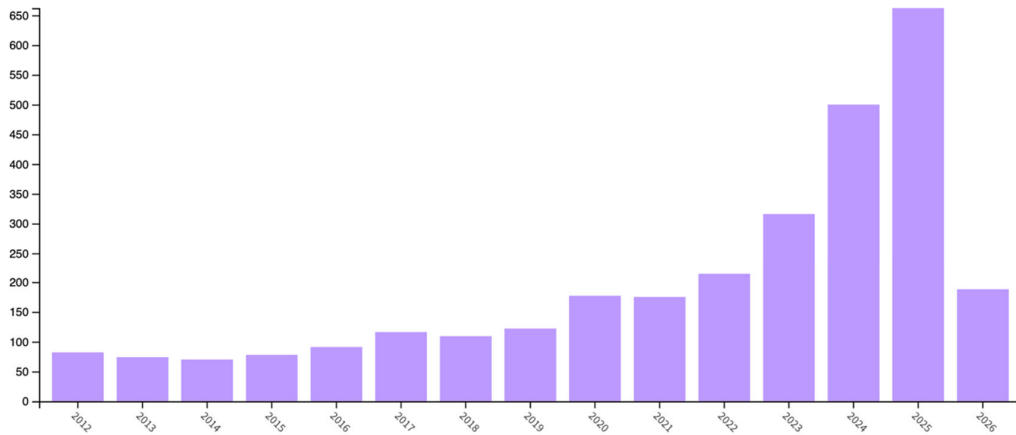
Kadoorie Biobank – Mentions per year

Number of publications with topic “Kadoorie” per year of publication. (N=551)



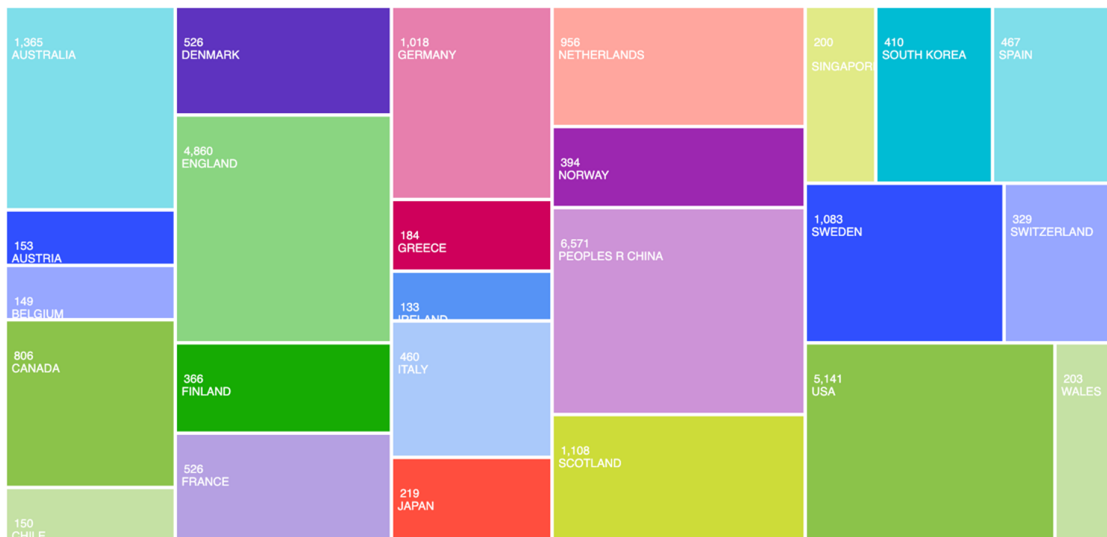
All of Us – Mentions per year

Number of publications with topic “All of Us” per year of publication. (N=3,672)



UK Biobank – Top 25 countries

Number of publications with topic “Kadoorie” having at least one co-author from a given country (Top 25). (N=16,446)



All of Us – Top 25 affiliations

Number of publications with topic “All of Us” having at least one co-author affiliated with a given institution (Top 25). (N=3,672)



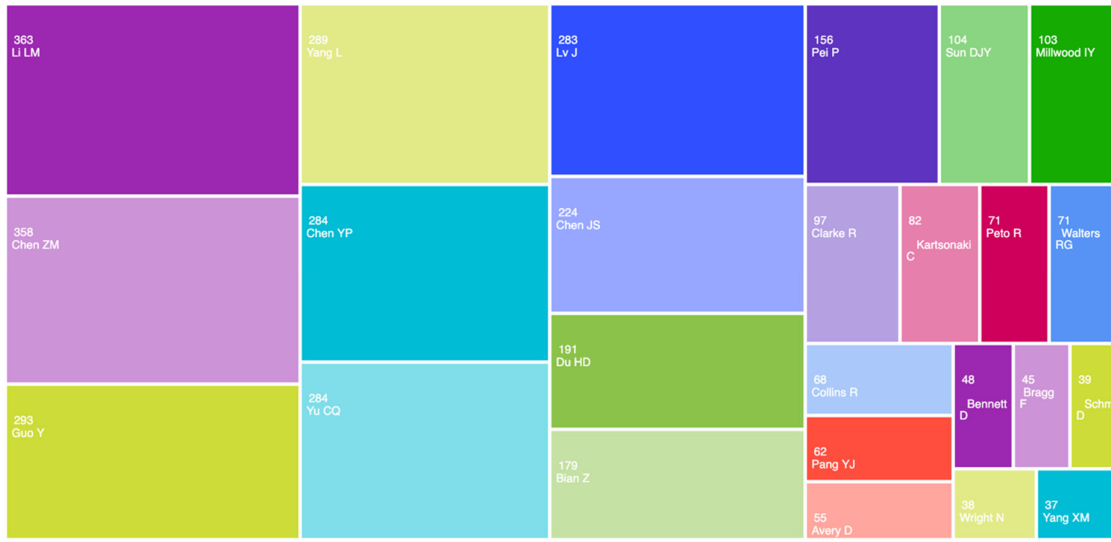
UK Biobank – Top 25 authors

Number of publications with topic “Kadoorie” per given author (Top 25). (N=16,446)



China Kadoorie Biobank – Top 25 authors

Number of publications with topic “Kadoorie” per given author (Top 25). (N=551)



All of Us – Top 25 authors

Number of publications with topic “All of Us” per given author (Top 25). (N=3,672)

