Supplemental Information

Quantifying the Impact of Biobanks and Cohort Studies

The code presented here can be found at https://github.com/Barabasi-Lab/quantifying_biobanks. The code and the accompanying data is also stored in a Zenodo repository at 10.5281/zenodo.11671293. A dashboard is available at http://biobanks.pythonanywhere.com/.

1 Biobank dataset

The process to build a corpus of biobank names was based on two steps. First, we compiled a set of 16 biobank catalogs (SI 1.1), and second, we extracted biobank names from the text of 3,310,320 biobank-related articles (SI 1.2). After each step, we processed the names identified to find and remove duplicates based on word similarity and co-appearance in articles of each pair of names (SI 2.1). The resulting corpus of biobanks contains 2,663 unique biobank names that were then used to gather biobank mentions from textual documents (SI 8). Finally, the dataset was validated using an external database of GWAS studies (SI 3).

1.1 Biobank Catalogs

Although there is yet no central biobank repository including all biobanks, there have been several efforts to build catalogs for the scientific community. The largest to date is the BBMRI-ERIC Directory of Biobanks and their collections. The resource was established to "improve accessibility and interoperability between academic and industrial parties to benefit personalized medicine" [48]. As of 10/2023, the directory included a list of 617 European biobanks together with their 3,176 collections and related metadata, providing a large set of biobank names.

To complement this corpus, we integrated the biobank names found in 15 prominent directories of population bioresources (the resulting corpus of biobanks can be found in raw and processed formats in the data folders 'cohorts' and 'raw_cohorts', respectively). Specifically, this dataset includes biobanks from BIOLINCC [49], EPND [50], IADRP [51], Molgenis [52], P^3G (Public Population Project in Genomics and Society) [53], DCEG [54], DPUK [55], UKRI cohort directory, Birthcohorts [56], CEDCD [57], The Pooling Project of Prospective Studies of Diet and Cancer [58], Wikipedia's list of biobanks, SciCrunch [59], Maelstrom [60], dbGAP [61], and JPND [62]. The re-

sulting list contained 1,576 unique names. Most of these resources were available for download with the exception of Birthcohorts, IADRP, Pooling Project of Prospective Studies of Diet and Cancer, and UKRI for which we crawled the websites (code found at `python/crawl`).

Furthermore, we extracted the biobank names found in the titles of papers published in the 'Cohort Profile series' of the International Journal of Epidemiology, and BMJ Open Journal, both having strict guidelines to start the title with 'Cohort Profile:', followed by the biobank name. We identified these publications using regular expression on the full publication dataset of Dimensions (files `SQL/cohort_profile.sql` and `SQL/cohort_profile2.sql`), resulting in 488 publications containing 368 biobank names.

Finally, we used ChatGPT to generate a list of biobanks using the prompt 'produce a list of biobank names along with their country of origin', resulting in 738 names. After processing to remove special characters, trailing spaces, and parenthesis, the resulting list contained 2,207 unique biobank names (file `data/cohorts/final_cohorts.csv`). The code to produce the list is found at `python/cohort_names/raw_cohorts.py`.

1.2 Biobank names expansion

In order to expand our biobank list, we did a bibliography search based on a set of 3,310,320 biobank-related publications. These publications were obtained by first identifying 162,988 articles mentioning one of the 2,207 biobanks previously obtained, and then extracting their citing articles. The resulting set of publications is then either mentioning directly a biobank or citing a paper mentioning one. The corpus of mentioning publications was obtained by searching in the title, abstract, and acknowledgments sections of articles in Dimensions (`SQL/mentions/publications.sql`). The code to obtain the set of biobank-related publications is `SQL/expansion/papers_citations.sql`.

We used regular expressions based on common keywords found in the names of biobanks and similar repositories, including Biobank, Tissue Bank, Registry, Biorepository, Project, and 12 other keywords. The complete list of regular expressions as well as the code to obtain the potential biobank names can be found at `SQL/expansion/clean_expanded_cohorts.sql`. The resulting list of names yielded 23,435 potential biobank names. We filtered potential biobanks in this corpus of names based on each biobank's number of mentions in two different corpora: the set of biobank-related papers (N=3,310,320) and the set of papers mentioning a biobank from the catalogs (N=162,988). A potential biobank was removed from consideration if less than 5% of the articles mentioning it contained a mention to a biobank from the catalogs and less than 50% of its mentions came from biobank-related publications. The code to

obtain papers mentioning the potential biobanks is in SQL/expansion/expanded_papers.sql (set of papers mentioning a potential biobank) and SQL/expansion/expanded_papers_internal.sql (set of mentions from biobank-related papers). The code to obtain their number of co-mentions with biobanks from the catalogs is at SQL/expansion/clean_expanded_cohorts.sql and the code to calculate the percentages of mentions on each corpus is SQL/expansion/cohorts_counts.sql.

This expansion process resulted in an additional 1,924 biobank names (file data/expansion/selected_cohorts.csv), resulting in a total of 4,131 biobanks. The code to clean, preprocess, and obtain this list of biobanks is found at python/expansions/total_inner_ratio.py.

2 Biobank mentions

In order to assess the impact of biobanks we searched their mentions across multiple textual documents, from the Dimensions academic database we include scientific publications, clinical trials, grants, and public policy documents. From Google Patents Public Data we searched for patents mentioning biobanks, as they provide the full-text of each patent. The search for mentions was case insensitive and included the name of the biobank preceded by the word 'the' to reduce false positives. The code to search for mentions can be found in multiple files named after the type of document they were searched on in the folder SQL/mentions/ for 2,207 catalog biobanks, and SQL/expansion/mentions/ for the expanded list of 1,924 biobanks.

To focus on biobanks with traceable presence in biomedical science, we remove those without at least one mention in either the title, abstract, or acknowledgments section of a scientific publication. This included 1,163 biobanks from the catalogs and 1,882 biobanks from the expanded list, for a total of 3,045 biobanks.

2.1 Removing Biobank Duplicates

A common issue with biobanks is the multiplicity of names a single biobank can accumulate across publications. The reasons behind this issue are multiple: from name alterations in follow-up studies (e.g., Framingham Heart Study, Framingham Children's Study, and Framingham Offspring Study), use of abbreviations (e.g., ARIC instead of Atherosclerosis Risk in Communities), to inconsistencies in their naming across publications (e.g., Northern Sweden Medical Biobank and Medical Biobank of Northern Sweden or Leeds Biobank and Leeds Multidisciplinary Research Tissue Bank). Additionally, biobanks and their corresponding cohort studies can be used interchangeably, like in Lifelines Biobank and Lifelines Cohort Study, similarly to biobanks that are part of a consortium (e.g., African Neurobiobank for Precision

Stroke Medicine and H3Africa).

In order to minimize the number of name pairs referring to the same biobank, we performed a deduplication process based on two factors: (i) string similarity and (ii) co-mentions. The former allows to identify similar names of a single biobank and the latter to identify duplicates lacking string similarity such as abbreviations from their full name. String similarity of each biobank pair a and b was calculated using the Indel distance $I(a, b)$, measuring the minimal number of insertions and deletions needed to transform one string into the other, normalized by the maximum number of possible deletions and insertions ($I(a, b) = 1$ implying that $a = b$). We used two variants of these distance: the partial ratio I_{pr}, measuring the Indel distance needed to match the smallest string into a sub-sequence of the larger string (here $I_{pr}(a, b) = 1$ does not necessarily imply that $a = b$); and the token set ratio I_{sr}, based on the minimal number of word deletions and insertions needed to match a and b (the word order is irrelevant). Using Spacy, a natural language model, we 'cleaned' the names prior to their analysis to remove coordinating conjunctions, auxiliaries, punctuation, symbols, and numerals.

Next, we measured the co-mention similarity of each biobank pair across the 228,761 publications and 19,299 patents mentioning biobanks. The co-mention similarity $S(a, b)$ between two biobanks a and b is equal to the proportion of documents mentioning a that also mention b in their text. In other words, $S(a, b) = 1$ implies that a is exclusively mentioned in publications mentioning b, and $S(a, b) = 0$ that no publication co-mentions a and b. Note that S is not symmetric: for cases where mention counts of biobanks a and b differ, we have that $S(a, b) \neq S(b, a)$.

Finally, we identified a pair of biobanks a and b as duplicates if at least one of the following statements occurred: (i) $\min\{S(a, b), S(b, a)\} > 0.2$ and $I_{pr}(a, b) > 0.9$, (ii) $\min\{S(a, b), S(b, a)\} > 0.05$ and $I_{pr}(a, b) = 1$, (iii) $\max\{S(a, b), S(b, a)\} > 0.3$ and $I_{pr}(a, b) > 0.9$, (iv) $\max\{S(a, b), S(b, a)\} > 0.1$ and $I_{sr}(a, b) = 1$, where co-mentions where extracted from publications. For patent-based similarity, we only considered a and b to be duplicates if all three inequalities were true: $\max\{S(a, b), S(b, a)\} > 0.4$, $\min\{S(a, b), S(b, a)\} > 0.3$, and $I_{pr}(a, b) = 100$, where stricter similarity conditions accounted for a higher probability of co-mentions in full-text documents.

The deduplication process resulted in 337 duplicate pairs, often forming duplicated groups where a biobank had 3 or more alternative names (An example being the Framingham Heart Study or FHS, with 10 variations). To obtain all duplicate groups, we constructed the duplicate network in which nodes represent biobank names and edges duplicated pairs, and each connected component represents name variations of the same biobank. We completed the deduplication process by selecting the biobank name with the greatest number of article mentions in each duplicate group to be the name variation to replace the others.

After deduplication, the final biobank corpus contained 2,263 names mentioned

across 228,761 articles, 16,210 grants, 15,469 patents, 1,769 clinical trials, and 9,468 public policy documents. The code containing the deduplication process is found at `python/cohort_names/deduplicate.py`.

2.2 Scientific articles

In order to look for mentions across Dimensions' publications (from October 2023), we first processed the text from the title, abstract, and acknowledgments sections to remove Unicode characters, double spaces, commas, and periods. The code is found at `SQL/clean_text/publications.sql`. Once the articles' text is pre-processed, we look for textual mentions of biobanks within the title, abstract, and acknowledgment sections of 141,219,539 articles resulting in 250,857 biobank-article mentions on 228,761 articles (`data/expansion/cohort_patents.csv`).

2.3 Patents

We use the publicly available Google Patents Open Data to search for biobank mentions on the full text of 153,696,878 patents. The search was done using Google BigQuery on the table `patents-public-data.patents.publications`, resulting in 18,183 biobank-patent mentions on 15,469 unique patents (`data/expansion/cohort_patents.csv`. The patents found were then merged to the patent data from Dimensions using the grant application number id (`SQL/join_dim_google_patents.sql`).

2.4 Grants

We used grant data from Dimensions, data composed of 5,040,039 grants. We found 18,251 biobank-grant mentions on 6,210 unique grants (`data/expansion/cohort_grants.csv`).

2.5 Clinical trials

We used clinical trials from Dimensions, data composed of 801,708 clinical trials. We retrieved 1,881 biobank-clinical-trial mentions from 1,769 unique clinical trials (`data/expansions/cohort_clinical_trials.csv`).

2.6 Public Policy Documents

We used the public policy documents from Dimensions, a dataset composed of 1,783,533 public policy documents. We identified 10,285 biobank-public-policy mentions on

9,468 unique public policy documents (data/expansions/cohort_public_policy.csv).

3 Data Validation

To assess the comprehensiveness of our database of biobanks and related documents, we used the NHGRI-EBI Catalog of Genome-Wide Association Studies (GWAS) as a reference to measure the scope of our data (the catalog was accessed June 2024), compiling more than 100,000 genetic associations to disease from 6,899 research publications [63]. Indeed, biobanks and population studies play a major role in the discovery of biomarkers by providing samples and genetic data used in GWAS. They are not the only contributors, however, cross-sectional and case-control studies—largely found in clinical trials and not really considered in our dataset—remain the most used resources for biomarker discovery [64]. Another potential caveat to consider is our lack of full-text publications for our search, a limitation not present in the reference GWAS catalog, directly receiving its data from the authors of the articles.

To be able to compare the overlap between the two datasets, we used Dimensions to search the PubMed Ids of our 228,761 biobank-related publications, successfully doing so for 76.6%, but dropping 1/4 of the biobank-related papers in the process. We start the overlap assessment with a naive analysis by directly considering all 6,899 publications from the NHGRI-EBI catalog, finding that 2,602 GWAS publications, or 38% of the total, are also included in our publication data. In our dataset, these publications collectively mention 486 biobanks, the most mentioned by far being the UK Biobank with 748 publications, followed by FHS (169), Wellcome Trust Case Control Consortium (166), and the Rotterdam Study (156).

To measure our dataset's breadth in the reported mentions by individual biobank, we compared them to the reported numbers found in the NHGRI-EBI Catalog. Specifically, we extracted the number of publications that each biobank in the catalog is referenced as a data source and compared it to the number of mentions we find for that biobank across the catalog's publications. We calculate these numbers for 89 biobanks having the same name in both datasets, finding that our dataset identifies, on average, 10.3 more publication-biobank associations than NHGRI-EBI. The UK Biobank is found in 748 publications in our dataset, but only in 485 publications of the reference catalog, a difference showing a much higher breadth in our dataset for this Biobank. The 263 publications mentioning the UK Biobank that do not list this biobank in the reference catalog have 'NR' (230) missing values for the column 'cohort' (33), suggesting that authors of the GWAS studies did not include the biobank source in the NHGRI-EBI Catalog. On average, the 72 biobanks better or equally covered in our dataset have 14.4 missing publications in the reference catalog. On

the other hand, the 17 biobanks that are better covered in the NHGRI-EBI Catalog have 7.1 missing publications in our dataset.

Removing publications without a biobank assigned (cohort values 'NR', 'other', or 'NA'), our dataset covers 60% of the 727 remaining publications in the GWAS catalog. These results show a good coverage of our dataset not without its limitations: while not all of the GWAS publications not covered (40%) may have a population biobank as a data source, we miss mentions in the full-text of the paper. To have an approximate of what the percentage of missing mentions due to our lack of access to the full-text of a publication, we consider the 485 publications in the GWAS catalog identifying the UK biobank as their source. From these, 129 are not covered in our dataset, suggesting that 26% of mentions of the UK Biobank may only be found in the body of the manuscript (excluding title, abstract, and acknowledgements). The data including the number of covered and missing publications is `data/database/meta/coverage.csv` and the code for the coverage analysis at `python/cohort_metadata/gwas_papers.py`.

4 Co-citation network

The biobank co-citation network of a biobank is built on top of the publications mentioning biobanks. The nodes of the network correspond to biobanks, and two biobanks are connected by an edge if their corresponding publications are often cited together. We operationalize the citations of a biobank as the number of articles citing the biobank's articles where it is mentioned. Mathematically, let $C_i = i_1, i_2, \dots i_n$ be the citing articles of biobank i, and $C_j = j_1, j_2, \dots j_m$ the citing articles of biobank j, then the union $C_i \cap C_j$ is composed by the articles citing both biobanks i and j. For each biobank i, we only consider its most similar biobank by adding a single edge to the biobank j such that $|C_i \cap C_j| = \max_x |C_i \cap C_x|$, for each biobank $x \neq i$. In other words, each biobank i is connected by an edge to the biobank j that is more often cited together with i.

5 Hidden Citations

Hidden citations are a measure of the hidden impact of a biobank based on its reference papers [65], and refer to unambiguous allusions made to a biobank without a corresponding citation. The idea is rooted in eponyms [**cabanac˙2014˙eponym**], informal citations [**marx2009informal**], and "obliteration by incorperation" [**mccain2012obliteration**]. In the context of biobanks, these allusions appear as implicit citations based on contextual mentions of a biobank in the text of a scientific paper. A hidden citation

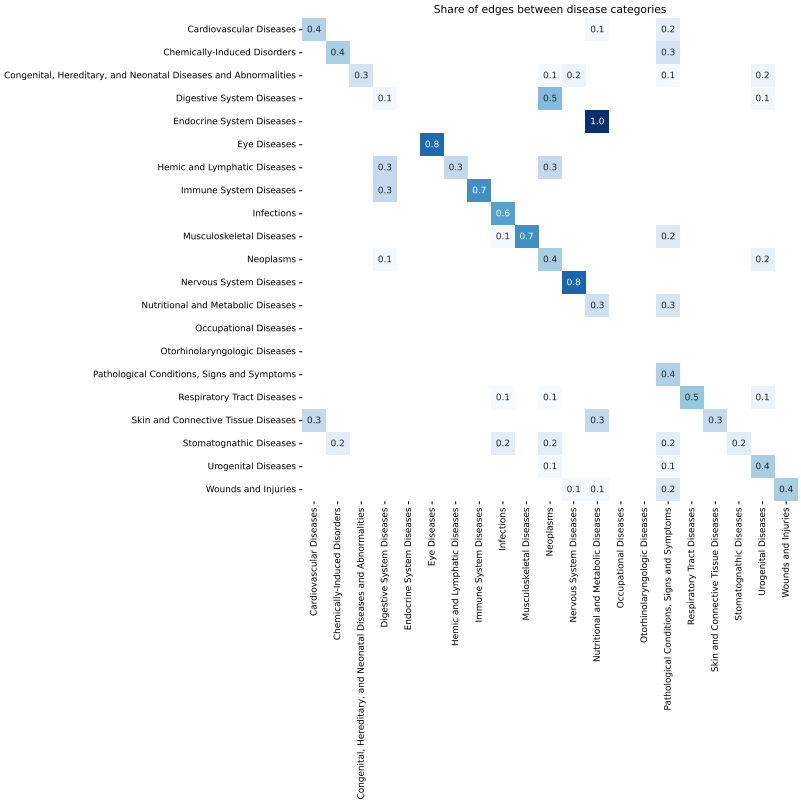

Figure 6: Share of edges between node categories.

to a biobank is defined as the difference between the explicit citations from papers mentioning the biobank and its total number of mentions.

5.1 Reference papers

In order to compute hidden citations, we first identify the foundational or reference papers of each biobank from the pool of papers that mention it. We identify a paper x a reference paper of a biobank if x mentions the biobank on its title or abstract and either (1) x is the first publication mentioning the biobank; (2) a paper y references x with probability $P(x|y = m) > 0.2 + P(x)$, where $P(x|y = m)$ is the probability that y references x given that y mentions the biobank and $P(x)$ is the probability that x is cited by any paper; or (3) the title of x contains the words 'design', 'baseline characteristics', or 'rationale'. We identify at least one reference paper for 500 biobanks. The code to compute the missing citations is found at `python/iumpact/recognition/foundational.py`.

5.2 Biobank Teams

The team of a biobank is composed of the union of all authors listed in one of its reference papers (Section 5.1). The data can be found at `data/expansion/database/cohort_impact/reference_papers.csv`.

6 Null model

In order to test the significance of the results showing that the impact of biobanks is local in terms of affiliation, country, and co-authorship, we create a null model in which we randomize the citing papers of each biobank such that the biobank remains with the same number of citations but with a random set of articles citing it. Once we have a randomized version of the citation network, where citing papers are randomized, we compute the mean share of in-house citations (coming from the same affiliation), the share of same-country citations, and the share of citations containing a collaborator of the biobank's team, and the share of citing papers containing its lead-author. We repeat this process 100 times and then use a z-test comparing the real value of each measure with its respective sequence of 100 randomized values to obtain a p-value.

7 Disease Impact

Based on MeSH terms, we link papers with diseases they have studied. For each disease, we count the total number of related articles per year for the period between 2000 and 2013. The number of articles in the sample is 10,486,605, studying 4,982 diseases. To have an idea of the disease impact of each biobank, we compute the ratio of publications relative to the total number of articles published for that disease in each year when the biobank was active. We limit the analysis to diseases with at least 10 publications per year in the global dataset. The resulting dataset then contains the proportion of articles related to each disease where the biobank is mentioned.

The last impact area of our proposed BIF is disease. This metric has three impact components of disease based on biobank publications: disease scope, disease depth, and rare disease (SI 7). The disease scope of the biobank is based on the number of different conditions studied from impacted papers relative to the total studies published for those diseases since the inception of the biobank, the UK Biobank leads this metric with 731 conditions studied, followed by NHANES (617), WHI, and the Telethon Network of Genetic Biobanks (TNGB, 199). The disease depth is based on the number of studies on a condition mentioning the biobank, relative to all publications on that condition, here the Avon Longitudinal Study of Parents and Children (ALSPAC) leads for environmental illness (14% of all publications in 2020, followed by the Human Microbiome Project (HMP) for neoplastic processes (9% of all publications in 2017), and the China National Genebank database (CNGBdb) for agricultural workers' diseases (7% of publications in 2021). Finally, rare disease impact is a measure of the number of biobank publications on rare diseases, relative to all biobank publications, and is led by the St. Jude Lifetime Cohort Study (St. Jude LIFE) with 101 papers out of 104 on rare diseases, followed by the Swiss Childhood Cancer Survivor Study (SCCSS, 96%), the NCI Childhood Cancer Survivor Study (CSS, 94%), and the Ovarian Cancer Association Consortium database (OCAC, 88%).

7.1 MeSH diseases

We use the definition of diseases from MeSH and operationalize a MeSH term as a disease if it is under the 'C' category of the MeSH tree. To compute the disease impact, we only consider articles published after the year 2000 mentioning a biobank, resulting in a total of 156,330 papers, we identify 75,269 as papers related to a condition (48% of the total of papers mentioning 985 biobanks in the given time period).

7.2 UK Clinical Research Collaboration: UKCRC Diseases

The HRCS system is used by a large number of health research funders in the UK and is subdivided into the Research Activity Classifications (RAC) and Health Categories (HC).

7.3 Research, Condition, and Disease Categorization (RCDC)

This categorization system is used by the NIH to report to Congress and is a biomedical system consisting of 237 categories, some of which are very specific in the topic (e.g. "ataxia-telangiectasia"), and others more general (e.g. "neuroscience").

7.4 Broad disease impact

For each biobank, we define its broad disease impact as the sum across diseases studied with the biobank data of the proportion of papers mentioning one of the 985 biobanks each year.

7.5 Specific disease impact

To account for the impact of biobanks focusing on a single disease, we define the specific disease impact as the highest proportion of articles published mentioning the biobank for a given disease relative to the total number of articles published for that disease during the active years of the biobank. Here, we limit our study to diseases in the second classification level of MeSH with at least 100 publications to account only for diseases with an established interest from the scientific community, resulting in 180 specific diseases related to 632 biobanks.

8 Mentions data

For each document mentioning a biobank, we collect the relevant metadata to implement each of the impact measures developed.

8.1 Grants data

For each grant, we extract its NIH activity code, the funder, funder amount, and disease categories (based on MeSH, RCDC, and HRCS). The NIH activity codes are then used to identify training, collaboration, and prestigious grants.

8.2 Patents data

For each patent, we identify its Cooperative Patent Classification (CPC) and its disease categories.

8.3 Clinical trials data

From each clinical trial, we extract its disease categories, as well as its assignee, and its institutional type.

8.4 Public policies data

For each public policy document, we extract its country of publication based on the affiliation of the institution implementing the policy, and the type of its implementing institution.

9 Biobank data

Biobanks differ in the scope and kind of data they offer. A Biobank may focus on a particular physical sample, e.g., brain tissue, while another may offer multiple sample types and questionnaire information from its participants. Genetic biomarkers, those gene variants related to higher risk or diagnosis of a disease, are among the most coveted goals of cohort studies and require DNA data from biobanks. The kind of genetic data, however, can define the methodology used by the study, however, and is therefore highly relevant. Similarly, medical records can be used to trace risk factors to multiple diseases and conditions suffered by the biobank's participants, complementing its data. Here, we assess the type of data available from the biobank across several factors including, genetic data type, environmental data, follow-up data, medical health records, and disease-specific registry data. In order to identify the type of data available from the biobank, we looked at the MeSH classification terms of the articles mentioning the biobank. If multiple biobank papers are Genome-Wide Association Studies (GWAS), it is highly likely the biobank contains GWAS data available. Here we present the different MeSH terms we used to associate each data type. The complete set of MeSH terms and tree numbers used can be found at data/mesh_and_meta/mesh_terms.md.

9.1 Genetic data

In order to assess whether a biobank has genetic data available, we looked for the following MeSH terms in the articles mentioning it: Genome-wide Association Study,

Whole Genome Sequencing, Genetic Association Studies, Genetic Techniques, as well as any MeSH term under the tree classification of Genetic Predisposition to Disease (MeSH tree number C23.550.291.906). We then sub-classified the DNA data on GWAS data and Whole Genome Sequencing. The code can be found at python/expansion/data/dna.py.

9.2 Environment data

We used the sub-branches of each of the following MeSH terms to identify papers with environmental data: Gene-environment Interaction, Environmental Exposure, Environmental Biomarkers, Environmental Indicators, Environment, and Environment Design. The code can be found at python/expansion/data/environment.py.

9.3 Follow-up data

Often, a biobank's cohort is followed through time, and samples and data are collected on each follow-up. To identify follow-up studies of a biobank, we used the MeSH term 'Follow-Up Studies' (D005500). The code can be found at python/expansion /data/follow_up.py.

9.4 Medical records

We used the MeSH terms under the tree classification of 'Medical Records' (MeSH tree ID E05.318.308.940.968). The code can be found at python/expansion /data/medical_records.py.

9.5 Registries and disease-specific data

In order to look for registries and other types of disease-specific data such as the one coming from panels, we used the following MeSH terms: Registries (E05.318.308.970), National Program of Cancer Registries (I01.409.418.750.600.650.200.760), and Mass Screening (E01.370.500). The code can be found at python/expansion /data/registries _disease_specific.py.

9.6 Surveys and Questionnaires

Some studies rely on data from surveys and questionnaires to build a social profile of the participants of the biobank. To identify those biobanks with survey and questionnaire data we identify the papers under the mesh term Surveys and Questionnaires (E05.318.308.980). The code can be found at surveys_and_questionnaires.py.

9.7 Age groups

Most biobank cohorts are designed with an age group in mind, usually defined by the first wave of data collected from its participants. In order to identify the age group of a biobank, we identified all the age groups under the MeSH classification of Age Groups (M01.060) and classified it into 7 categories depending on the tree number of the age group, namely: Middle Aged (M01.060.116.630), aged (M01.060.116.100), young adult (M01.060.116.815), adolescent (M01.060.057), infant (M01.060.703), child (M01.060.406), and birth (M01.060.261). The code can be found at persons.py.

9.8 Country of origin

We collected the country of origin of the cohort from each biobank by identifying the MeSH terms related to countries in their publications. This includes every term under the MeSH tree classification of Geographic Locations (Z01), except for cities (Z01.433). We also consider the countries where the authors of the papers are affiliated based on their institution on each paper. The code can be found at countries.py.

9.9 Open data index

In order to assess the level of *data openness* of a biobank, we consider three factors related to the papers mentioning the biobank: the percentage of papers led by the top corresponding author, the percentage of papers co-authored by the top 10 authors, and the percentage of papers with at least one author affiliated to the top affiliation. We then define the 'open data index' of a biobank as 100 minus the mean of these percentages.

9.10 Cohort size

The cohort size of a biobank, i.e., the number of participants or individuals included for sampling, surveys, and/or follow-up, may be of interest to researchers looking to access its data. Although Biobanks usually publish the size of their cohort on their design or foundational papers, it may be subject to change in subsequent updates of the biobank, or it may be a target number. Oftentimes, moreover, researchers may be interested in a sub-sample of the cohort depending on multiple factors such as the disease of interest, the analytical method of their study, or the type of data they need. In order to have an idea of the cohort size that is used across the research papers that make use of a biobank's data, we look for the self-disclosed sample sizes appearing in the abstracts of such studies. This approach makes use of papers mentioning a

biobank, resulting in a distribution of sample sizes used across time and topics. In order to identify sample sizes we use a regular expression matching numbers followed by a keyword (code available at SQL/expansion/cohort_size.sql), and to filter out those papers that may be using more than one biobank, we only consider articles mentioning one biobank in our sample. For each biobank, we then consider the 90th percentile as its cohort size to consider the larger sub-samples of the biobank and to avoid outliers that may represent noisy values. The code to create the table with the sample size of each biobank is found at `python/data/cohort_size.py`.

9.11 Population vs. health based biobanks

Biobank data is retrieved from a cohort of participants that may be recruited from a population defined by a geographical location, a line of work, or any other common factor, or it can be based on patients (dead or alive) coming from a hospital or clinic and usually sharing a disease [66]. The border between population and patients, however, is not always clear. This is the case when biobanks recruit individuals from a general population that are patients are local clinics, like in the UK Biobank. Indeed, often participants are invited to participate in the Biobank's cohort from a clinic or a hospital, not because they are patients there, but because it is strategic or convenient (e.g., to link their medical records to their samples).

To assess whether a biobank's design had a population or a group of patients in mind we identified several keywords from the abstracts of the papers mentioning the biobank, including 'population-based', 'patients', 'hospital', 'health-based', and 'clinic'. To classify the cohort type of the biobank, we then computed the ratio of papers with each keyword and assigned the type with the highest proportion of papers. Under this approach, the UK Biobank, even if around 10% of its papers contain the word 'patients', is classified as a population-based biobank as 29% of its articles contain the keyword 'population'. The code can be found at python/expansion/database/cohort_type.py.

10 Biobank Impact Factor

We considered two factors to calculate the Biobank Impact Factor (BIF): mentions and disease impact. To calculate the mention impact of a biobank, we standardize its number of mentions for each document type separately (i.e., publications, grants, patents, clinical trials, and public policy documents). For each document type, standardization is done by taking the number of mentions m of a biobank and returning $(m-\mu)/\sigma$, where μ and σ is the mean and standard deviation over all biobanks. Once every value is standardized, we sum the values of each biobank across all document

types, this is the mention impact. To consider the advantage in mentions of older biobanks, where the number of mentions may be related to the time period they have been available, we normalize mention impact by the number of years the biobank have been mentioned in scientific publications.

This mention impact takes into account the overall presence of biobanks across science, innovation, and policy, as well as their lack thereof. Because mentions are standardized by document type, a biobank can have a negative value that results in a 'penalty' for its impact. The BIF scores higher balanced biobanks, those having an impact, even if modest, across most sectors. The code to compute mention impact is found in `python/impact/cohort_impact_factor/impact_factor.py`.

Another factor that we consider for BIF is the scope and depth of research produced by each biobank (SI 7). This metric considers the number of publications each biobank helps produce for a disease, compared to the total number of publications about the disease (depth); as well as the total number of diseases studied on publications mentioning the biobank (scope). Additionally, disease impact considers the impact the biobank has on rare diseases, measured by the number of mentioning publications on a rare disease, relative to the all other biobanks. This ensures that biobanks contributing to a single disease, and those contributing to rare diseases, are considered in the BIF. The code to compute disease impact is found in `python/impact/cohort_impact_factor/disease_impact.py`.

The code to compute the BIF is found in `python/impact/cohort_impact_factor/target.py`.

11 Generalized Linear Model

We used a Generalized Linear Model (GLM) to measure the relation between biobank data characteristics and BIF. The dependent variable is then BIF and the 14 independent variables were all binary values depending on the data characteristics of each biobank (Table 11). Specifically, we used a Gaussian family GLM with identity function linked to the data, using the statsmodels library in Python. The model was applied to a set of 468 biobanks for which we could identify each feature (SI 9). For each dependent feature we obtained its coefficient and associated p-value, labeling it as significant if p-value $0.05/n$, where n is the number of features in the model (Bonferroni correction for each independent feature). The model presents a an R-square metric of 0.4 and log-likelihood of 862.65. The code to fit the data to the GLM and to obtain the significant variables is at `python/stat_models/glm.py`.

Dep. Variable:	target	**No. Observations:**	468
Model:	GLM	**Df Residuals:**	453
Model Family:	Gaussian	**Df Model:**	14
Link Function:	Identity	**Scale:**	0.0015157
Method:	IRLS	**Log-Likelihood:**	862.65
Date:	Mon, 15 Apr 2024	**Deviance:**	0.68663
Time:	10:02:51	**Pearson chi2:**	0.687
No. Iterations:	3	**Pseudo R-squ. (CS):**	0.4097
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-0.0186	0.009	-2.102	0.036	-0.036	-0.001
dna	0.0002	0.005	0.039	0.969	-0.010	0.011
gwas	0.0024	0.005	0.489	0.625	-0.007	0.012
whole_genome	0.0286	0.006	4.752	0.000	0.017	0.040
gene_env	0.0267	0.005	5.346	0.000	0.017	0.037
follow_up	0.0144	0.005	2.900	0.004	0.005	0.024
medical_records	0.0187	0.004	4.210	0.000	0.010	0.027
surveys	-0.0049	0.008	-0.608	0.543	-0.021	0.011
registries	0.0077	0.004	1.857	0.063	-0.000	0.016
population	-0.0010	0.005	-0.221	0.825	-0.010	0.008
large_cohort	0.0048	0.006	0.796	0.426	-0.007	0.017
open_data_high	0.0345	0.007	5.196	0.000	0.021	0.048
open_data_low	-0.0051	0.008	-0.618	0.537	-0.021	0.011
PI_high_impact	0.0198	0.006	3.079	0.002	0.007	0.032
PI_low_impact	-0.0092	0.007	-1.336	0.182	-0.023	0.004