Quantifying Biobank Impact

Rodrigo Dorantes-Gilardi[1] and John Michael Gaziano[4,5] Albert-László Barabási[1,2,3]

[1] Network Science Institute, Northeastern University, Boston, USA
barabasi@gmail.com
[2] Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, USA
[3] Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, USA
[4] Department of Internal Medicine, Harvard Medical School, Boston, MA, USA
[5] Massachusetts Area Veterans Epidemiology Research and Information Center (MAVERIC), VA Cooperative Studies Program, VA Boston, MA 02130, USA

1 Introduction

Biobanks are repositories conceived for epidemiological research of biological samples and the associated data from a large group of individuals. During the last few decades, the use of biobanks has increased to the extent of being a fundamental component of biomedical research and has the potential to significantly improve future healthcare [1]. In 2009, Time magazine added biobanks as one of the ten ideas that were changing the world, and their creation has been promoted by research centers and gained international support from funding agencies and the research community as a whole [2]. In the last decade, several grants have been obtained for the purpose of creating, maintaining, or relying on biobank resources [3].

Biobanks vary widely in terms of purpose, scope, governance, and type of data. The team composition of a biobank can also differ in terms of gender composition, popularity of its lead, and size of the team. It is not surprising that their impact in science is dissimilar, and only a handful of well-known biobanks receive most of the necessary attention to be highly impactful in terms of publications.

2 Results

The characteristics of high-impact biological cohorts and the mechanisms of recognition given to their creators remain elusive, however, as quantitative measures applied to the universe of biobanks are hard to find. Here, we use a data mining approach to identify a list of more than one thousand biobanks together with their introductory paper to the academic community (Figure 1A). We use the corpus of articles to track their footprint on research under different angles, making it to our knowledge, a first case of a large-scale quantitative study of biobank academic impact. We show that biobank impact is unbalanced across the world by first describing the origin and usage of biobanks. Most of them are originated and used in the global north, and regions like south-east Asia and Africa have a usage rate smaller than their production rate.

In order to study the different methods biobanks' teams receive recognition from researchers, we measure the extent that biobank data-access results in collaboration with external researchers. On average, 56% of the citations to biobanks come through collaborations with third party authors (Figure 1B). In the case where data access is more restricted, we observe that a larger share of publications is co-authored with external researchers, and then discuss that stringent data sharing policies may be set in place to avoid lack of recognition. We use the UK Biobank case to observe that roughly half of the papers identifying the biobank as a data source in the article or abstract do not cite any of the resource articles written by the biobank team leading to reduced measurable impact (Figure 1C).

Finally, we implement two logistic regression models to predict the academic impact of a biobank based first on its inherent characteristics (Figure 2A, 2C), like cohort size and data openness, and another one based on the academic rank and gender composition of the team of the biobank (Figure 2B, 2D). We find that biobanks with genetic data for general purpose research (as opposed to disease specific) tend to be more impactful; unexpectedly, we observe that restrictive biobanks tend to have more citations as well. Biobank impact increases as the share of female team members increases, proving that highly diverse teams are good for biobank formation. The popularity of the lead author at the time the biobank was created is a major influence of its success.

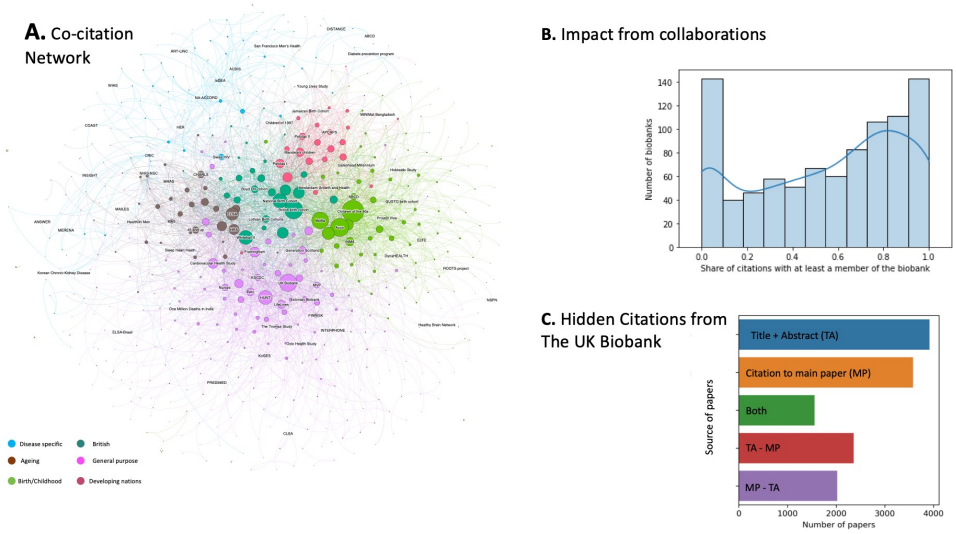
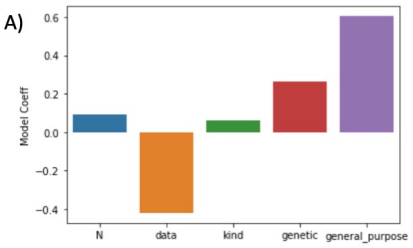
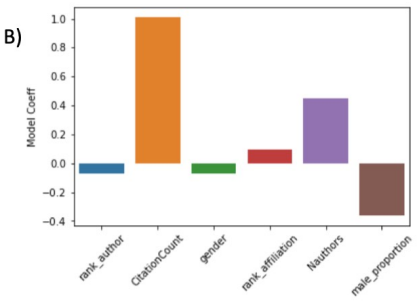


Fig. 1. Biobank Impact A. Biobanks represented by their co-citation network where nodes are biobanks that are connected if they are cited by the same paper. Different communities are given a color and represent a cluster of biobanks. B. Distribution of the share of citations from collaborations (at least one member of the biobank team is co-author of the citing article). C. Hidden citations of the UK Biobank are defined as the number of articles using the data of the biobank without citing any of the biobank's articles





C) Inherent properties model

	A	B	C
N	6	3	5
Data	0	1	0
Kind	0	1	1
Genetic	1	0	0
General purpose	1	0	1
P(Success)	0.72	0.36	0.66

D) Lead author model

	A	B	C
Rank	1k	10k	10k
cites	100k	10k	100k
gender	male	female	male
Rank affiliation	1k	10k	10k
Nauthors	8	16	8
Male proportion	50%	20%	75%
Pr(Success)	0.69	0.58	0.65

Fig. 2. Coefficients of the logistic regression models to predict biobank impact. A. Model is based on the characteristics of the biobank, namely, sample size (N), data openness, kind, genetic data available (genetic), and whether it is general purpose or disease specific. B. The model is based on the characteristics of the biobank's team, these include the rank of the lead scientist of the team, its popularity (based on citations received at the time the biobank was created), gender, and the rank of their affiliation, the team size of the biobank, and the male proportion of the team. Three examples of biobanks are shown in C) ad D) to represent different biobanks and their probabilities of being successful (highly cited).

References

1. Shilo, Smadar, Hagai Rossman, and Eran Segal. ”Axes of a revolution: challenges and promises of big data in healthcare.” Nature medicine 26.1 (2020): 29-38
2. Caulfield, Timothy, et al. ”A review of the key issues associated with the commercialization of biobanks.” Journal of Law and the Biosciences 1.1 (2014): 94-110
3. Kinkorová, Judita, and Ondřej Topolčan. ”Biobanks in Horizon 2020: sustainability and attractive perspectives.” Epma Journal 9.4 (2018): 345-353