POLICY

dbGaP Data Access Requests: A Call for Greater Transparency

Lorelei Walker,[1,2] Helene Starks,[1,2,3] Kathleen M. West,[1,2,3] Stephanie M. Fullerton,[1,2,3*]

The scientific and public health benefits of mandatory data-sharing mechanisms must be actively demonstrated. To this end, we manually reviewed 2724 data access requests approved between June 2007 and August 2010 through the U.S. National Center for Biotechnology Information database of genotypes and phenotypes (dbGaP). Our analysis demonstrates that dbGaP enables a wide range of secondary research by investigators from academic, governmental, and nonprofit and for-profit institutions in the United States and abroad. However, limitations in public reporting preclude the tracing of outcomes from secondary research to longer-term translational benefit.

Researchers and funders promote broad sharing of data and specimens from human study participants as a key prerequisite for large-scale discovery science and, ultimately, translational advances (*1*). Data-sharing recommendations have been based on an expectation that centralized access to consistently cleaned and well-annotated data can promote research efficiency and maximize resources. Accordingly, the U.S. National Institutes of Health (NIH) has adopted explicit data-sharing policies (*2*), and resources have been devoted to the development of research infrastructure to facilitate widespread sharing (*1*). Yet, the net impact of these policies and initiatives has been underinvestigated. Although data sharing places new demands on informed consent, institutional oversight, and repository governance (*3*–*5*), these demands may be amply compensated by the nature and extent of the science enabled by such sharing. We cannot undertake this calculus, however, without knowing more about the types of research and associated benefits that are generated through specific data-sharing initiatives.

A major infrastructure investment aimed at promoting the sharing of participant data in genetic epidemiology and the genome sciences is the U.S. National Center for Biotechnology Information (NCBI) database for genotypes and phenotypes (dbGaP) (*6*). The NIH genome-wide association study (GWAS) data-sharing policy (*2*) strongly encourages deposition of GWAS data into dbGaP and, as of August 2010, researchers had deposited genotype and linked phenotype data from more than 100 primary studies. Deidentified participant data are made available to secondary users in two forms: unrestricted public access to summaries of aggregate study data and restricted access to individual-level data (*7*). Scientists who are interested in accessing individual-level genotypic and phenotypic data must submit a Data Access Request (DAR) to 1 of 14 NIH Data Access Committees (DACs) and gain approval. Approved DARs are noted on the dbGaP Web site under specific studies (*8*).

The public accessibility of approved dbGaP DARs provides an opportunity to evaluate use of this federal database. We manually reviewed the approved DARs posted on the dbGaP Web site between June 2007 and August 2010 to identify how and to what extent the sharing of these data has contributed to population-based genomic investigation and translational science.

DATA REQUESTS

At the time we conducted our analysis (August 2010), dbGaP listed 48 parent projects, which encompassed a total of 103 primary studies that were each classified as 1 or more of 26 different study types or disease groups (*9*) (see Supporting Online Material for further details). Of the 103 primary studies, 33 had no approved DAR records listed on the Web site. Six studies had not been assigned to a DAC because they had no individual-level data to share. The remaining 27 studies had been assigned to a DAC but had no approved DARs; a year later (August 2011), 26 had approved requests. Each of the other 70 primary studies had at least one DAR record, with an average of 33 DARs (range 1 to 241) per primary study. On simple inspection, we could not discern any systematic patterns between the primary studies and the number of DARs, either with respect to the amount of time the primary study had been available in dbGaP, sample size, or relevant consent restrictions (for example, secondary use limitations linked to specific classes of health conditions or types of investigators).

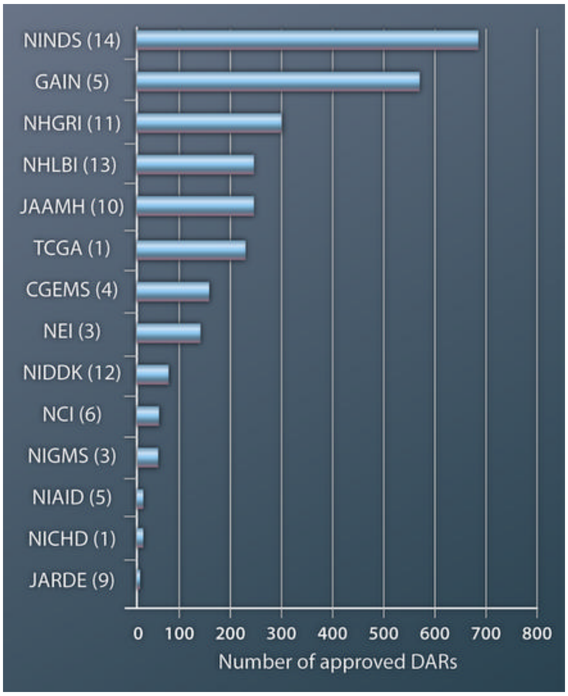

Fig. 1. Requested and approved. Total number of DARs approved by each DAC. Shown in parentheses is the number of primary studies for which each DAC is responsible. [NHGRI, National Human Genome Research Institute; NHLBI, National Heart, Lung, and Blood Institute; JAAMH, Joint Addiction, Aging, and Mental Health; CGEMS, Cancer Genetic Markers of Susceptibility; NEI, National Eye Institute; NIDDK, National Institute of Diabetes and Digestive and Kidney Disease; NCI, National Cancer Institute; NIGMS, National Institute of General Medical Sciences; NIAID, National Institute of Allergies and Infectious Diseases; JARDE, Joint NIAMS (National Institute of Arthritis and Musculoskeletal and Skin Diseases)–NIDCR (National Institute of Dental and Craniofacial Research) DAC]

[1] Institute for Public Health Genetics, University of Washington, Seattle, WA 98195, USA. [2] Center for Genomics and Healthcare Equality, University of Washington, Seattle, WA 98195, USA. [3] Department of Bioethics and Humanities, University of Washington, Seattle, WA 98195, USA.

*Corresponding author. E-mail: smfllrtn@u.washington.edu

CREDIT: B. STRAUCH/SCIENCE TRANSLATIONAL MEDICINE

Table 1. Distribution of proposed secondary research use by investigator affiliation (*n* = 2724 DARs).

Affiliation	Discovery	Methods	Replication	Controls	Not listed	Population structure	Quality control
U.S. academic	37%	30%	14%	13%	4%	1%	0%
U.S. nonprofit	38%	12%	25%	11%	8%	0%	5%
U.S. for-profit	41%	16%	29%	7%	3%	4%	0%
Non-U.S. academic	43%	29%	18%	2%	5%	4%	1%
Non-U.S. nonprofit	33%	10%	25%	29%	4%	0%	0%
Non-U.S. for-profit	47%	0%	35%	6%	12%	0%	0%

The 2724 DARs identified were approved by 1 of 14 different DACs, which are each sponsored by one or more of the NIH institutes for the primary studies under their purview; special DACs were formed for a few very large parent studies [such as the Genetic Association Information Network (GAIN) and The Cancer Genome Atlas (TCGA) studies]. Each DAC is responsible for ensuring that requests for secondary use comply with the original participant’s informed consent obtained by the primary study investigators (*7*). The DAC that manages requests for studies sponsored by the National Institute of Neurological Disorders and Stroke (NINDS) approved the largest number of DARs (693 for 14 primary studies), whereas the National Institute of Child Health and Human Development (NICHD) DAC approved the fewest DARs (2 for 1 primary study) (Fig. 1). Because only approved DARs are indicated on the dbGaP Web site, we cannot assess whether the different numbers reflect rates of request for data or whether DACs employ distinct review and approval criteria. From the information made available to the public, there is no obvious relationship between the number of studies managed by a DAC and the number of approved DARs.

After the 2008 publication of an analysis demonstrating the feasibility of individual-level reidentification from aggregate participant genotype data (*10*), NIH placed all genotypic data in dbGaP behind the NIH firewall, effectively requiring DARs for previously open-access data (*11*). Before this policy change, approved DARs averaged about one per day. After this shift, approved DARs averaged just over three per day. It is unclear if the observed increase in DARs is a result of the firewall change, the increased awareness of the database among new users, an increased number of data sets available over time, or a combination of these.

WHO’S ASKING?

A total of 851 investigators from 330 institutions have requested data from dbGaP: 490 (57%) made a single request for a single primary dataset, while 361 made between 2 and 73 requests for different primary data sets. The majority of requests came from investigators at U.S.-based organizations (224, 68%), and 73, 18, and 9% of total requests (within and outside the United States) were from academic, nonprofit/government, and for-profit institutions, respectively. The 46 for-profit institutions included 16 pharmaceutical, 12 biotechnology, 9 bioinformatics and software, 4 direct-to-consumer genetic testing, and 3 health care services companies; 1 clinical research organization; and 1 accredited online learning institute. The top 10 institutions with the most DARs by institutional type (U.S. and non-U.S. combined) are shown in table S1. For-profit, academic, nonprofit, and non-U.S. institutions averaged 8.6, 5.5, 3.7, and 3.6 DARs per investigator, respectively.

Of the 330 institutions with approved DARs, 224 U.S. institutions were granted a total of 2210 (81%) DARs; the remaining 514 DARs (19%) were granted to institutions in 27 countries outside of the United States, and the majority of these requests (89% of the 514) were from 78 institutions in 10 countries (table S1). Investigators from Canada, the U.K., and the Netherlands accounted for 57% of non-U.S. approved DARs; investigators at six institutions in China made up 9% of the non-U.S. approved DARs.

WHAT FOR?

More than half of the DARs (*n* = 1593, 58%) involved linked requests, meaning that individual investigators received approvals for the analysis of data from multiple primary studies for the same (or similarly described) proposed secondary research use. The remaining 1131 DARs (42%) were single requests for secondary use of data from one primary study. Requests from different investigators with a similar research objective for the same primary data likely represent multi-institutional collaborations, because dbGaP requires that each institution make its own request (rather than have a lead in-

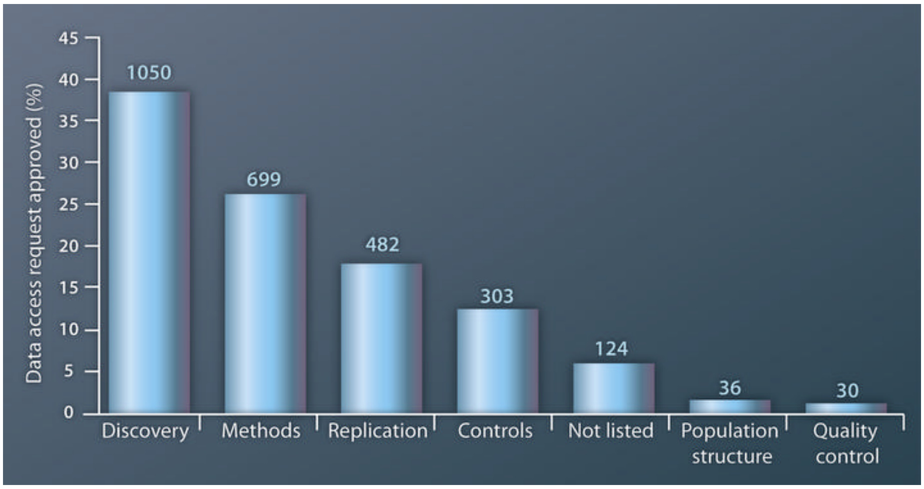

Fig. 2. Fueling new research. Total numbers of DARs approved per type of proposed secondary research use (with categories as defined in table S2).

CREDIT: B. STRAUCH/SCIENCE TRANSLATIONAL MEDICINE

stitution request the data and then share with collaborators). However, the information in the DAR records was not specific enough to verify whether multiple requests for the same data represent collaboration or simply similar or overlapping research interests.

Figure 2 depicts the distribution of proposed secondary research uses described in the approved DARs, based on our classification scheme (table S2). The most common secondary uses were the discovery of new genotype-phenotype associations (39%), identification of new methods (26%), replication of previous findings (18%), and identification of control populations (11%). Additional uses (at much lower rates of request) included analysis of population structures and internal (NINDS) quality control; of the approved DARs, 4% did not specify a proposed research use. The right-most column of table S1 indicates the most frequent category of proposed secondary use in the DARs of the top requesting institutions.

The most common proposed use, irrespective of affiliation, was the discovery of new associations (Table 1). Investigators from non-U.S. for-profit institutions more often failed to propose a specific research use (see “Not listed,” Fig. 2) than investigators with other affiliations. Such investigators also made no requests to advance methods development, a notable difference relative to other affiliations. Investigators from non-U.S. nonprofit institutions requested data more often for use as control samples than did investigators with other affiliations.

BENEFITS OF SHARING

Our analysis of these DARs demonstrates that most of the data held in dbGaP have been requested for secondary research use. In the first three and a half years since its inception, 14 DACs have approved 2724 DARs for 70 primary studies. Although it is impossible to know for sure, it seems unlikely that so much secondary research, by so many independent researchers, could have been attempted in the same time period in the absence of a dedicated data repository.

Although our analysis provides new insights into the level and diversity of secondary research uses enabled by dbGaP-mediated data sharing, our efforts could not demonstrate the scientific and public health benefits of this new data-sharing mechanism, because of limitations in public reporting. dbGaP DARs provide information only about the *proposed,* and not the *actual,* secondary uses of data shared via the repository. Currently, the NCBI provides no direct links to any published research that resulted from specific approved DARs. In addition, despite recommendations that “published analyses of data from dbGaP should explicitly reference unique and stable accession numbers in method descriptions and acknowledge each study used” (*6*), we were largely unable to identify articles that reported secondary analyses of dbGaP data in PubMed or related online databases. It is thus impossible for the public to evaluate whether the ambitions of the NIH GWAS data-sharing policy (*2*) have been realized either through creation of the repository or through the sharing of information in dbGaP that was generated with the use of NIH funding.

The long-term viability of a publicly funded data-sharing mechanism such as dbGaP rests in the comprehensive and ongoing assessment of outcomes enabled by the resource. Although public disclosure of approved DARs is an important first step, the NCBI could and should do more to provide information about DARs, data requestors, and attendant scientific advances summarized in the peer-reviewed literature. Specifically, the considerable time that it took us to extract, summarize, and analyze the aggregate dbGaP DARs points to the need for an automatically updated and readily queryable database of DARs and linked publications. Key variables of interest, including many we have reported in our analysis here, could then be summarized for rapid inspection, or the full database could be interrogated by stakeholders with particular questions. If the outcomes of wide data sharing were made easily accessible in this way, more investigators would likely avail themselves of this valuable public resource.

Greater transparency with regard to the nature and extent of dbGaP data sharing will increase the value of the repository for potential users and enhance the trustworthiness of the resource for data submitters and their parent institutions. Currently, submitting investigators and responsible institutional officials are required to certify that they have “considered the risks to individuals, their families, and groups or populations associated with data submitted to” dbGaP (*2*). However, in the absence of detailed information about how data are shared, with whom, and for what purposes, full consideration of potential risks is impossible. Increasing the transparency and accessibility of information about the nature of secondary uses and associated outcomes will go a long way toward maximizing accountability and reassuring investigators that their hard-won data are being well controlled and effectively disseminated.

Finally, and perhaps most importantly, increased transparency will help ensure that research participants’ interests are recognized and protected. Although participant data in dbGaP are deidentified at submission, and hence not governed by human subjects regulations, DACs take care to ensure that approved DARs are consistent with the original informed consent. In addition, prior research suggests that some participants have concerns about placing their data in a federally controlled repository or about their data being used by for-profit organizations (*12*–*14*). However, it is not easy for the average research participant to assess how their cohort’s data have been shared or with whom, and it is also not clear what proportion of participants are even aware that their data have been submitted to dbGaP (*4*). Recently proposed revisions to the Common Rule (*15*) include a requirement for written informed consent for primary and secondary research use of biospecimens using a (proposed) simplified uniform consent form, which may go much of the way toward addressing this concern.

Our analysis suggests that the dbGaP repository is enabling a wide range of secondary research uses with probable near- and longer-term translational science and public health benefit. However, more can—and should—be done to make the scientific advantages of data sharing transparently accessible to interested stakeholders.

SUPPLEMENTARY MATERIAL

www.sciencetranslationalmedicine.org/cgi/content/full/113/113cm34/DC1

Materials and Methods

References

Table S1. Top ten institutions by type and non-U.S. countries with granted DARs.

Table S2. Classification scheme for proposed secondary uses of dbGaP data.

REFERENCES AND NOTES

1. NIH Data Sharing Policy and Implementation Guidance. http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm.
2. Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html.
3. S. M. Fullerton, N. R. Anderson, G. Guzauskas, D. Freeman, K. Fryer-Edwards, Meeting the governance challenges of next-generation biorepository research. *Sci. Transl. Med.* **2**, 15cm3 (2010).

4. S. B. Trinidad, S. M. Fullerton, E. J. Ludman, G. P. Jarvik, E. B. Larson, W. Burke, Research ethics. Research practice and participant preferences: The growing gulf. *Science* **331**, 287–288 (2011).
5. A. L. McGuire, M. Basford, L. G. Dressler, S. M. Fullerton, B. A. Koenig, R. Li, C. A. McCarty, E. Ramos, M. E. Smith, C. P. Somkin, C. Waudby, W. A. Wolf, E. W. Clayton, Ethical and practical challenges of sharing data from genome-wide association studies: The eMERGE Consortium experience. *Genome Res.* **21**, 1001–1007 (2011).
6. M. D. Mailman, M. Feolo, Y. Jin, M. Kimura, K. Tryka, R. Bagoutdinov, L. Hao, A. Kiang, J. Paschall, L. Phan, N. Popova, S. Pretel, L. Ziyabari, M. Lee, Y. Shao, Z. Y. Wang, K. Sirotkin, M. Ward, M. Kholodov, K. Zbicz, J. Beck, M. Kimelman, S. Shevelev, D. Preuss, E. Yaschenko, A. Graeff, J. Ostell, S. T. Sherry, The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* **39**, 1181–1186 (2007).
7. dbGaP overview. http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.html.
8. dbGaP authorized access. https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?login=&page=login.
9. dbGaP projects. http://www.ncbi.nlm.nih.gov/gap.
10. N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, D. W. Craig, Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* **4**, e1000167 (2008).
11. E. A. Zerhouni, E. G. Nabel, Protecting aggregate genomic data. *Science* **322**, 44 (2008).
12. E. J. Ludman, S. M. Fullerton, L. Spangler, S. B. Trinidad, M. M. Fujii, G. P. Jarvik, E. B. Larson, W. Burke, Glad you asked: Participants' opinions of re-consent for dbGap data submission. *J. Empir. Res. Hum. Res. Ethics* **5**, 9–16 (2010).
13. A. A. Lemke, S. B. Trinidad, K. L. Edwards, H. Starks, G. L. Wiesner, GRRIP Consortium, Attitudes toward genetic research review: Results from a national survey of professionals involved in human subjects protection. *J. Empir. Res. Hum. Res. Ethics* **5**, 83–91 (2010).
14. D. J. Kaufman, J. Murphy-Bollinger, J. Scott, K. L. Hudson, Public opinion about the importance of privacy in biobank research. *Am. J. Hum. Genet.* **85**, 643–654 (2009).
15. http://www.hhs.gov/ohrp/humansubjects/anprm-2011page.html.

Acknowledgments: This work was supported in part by NHGRI grant number P50-HG-3374 (to W. Burke). The authors wish to thank L. L. Rodriguez, Director of the Office of Policy, Communications, and Education at the National Human Genome Research Institute, for her assistance with obtaining collated lay and technical descriptions of dbGaP DARs.

10.1126/scitranslmed.3002788

Citation: L. Walker, H. Starks, K. M. West, S. M. Fullerton, dbGaP data access requests: A call for greater transparency. *Sci. Transl. Med.* **3**, 113cm34 (2011).

Science Translational Medicine

dbGaP Data Access Requests: A Call for Greater Transparency

Lorelei Walker, Helene Starks, Kathleen M. West and Stephanie M. Fullerton

Sci Transl Med **3**, 113cm34113cm34.
DOI: 10.1126/scitranslmed.3002788

Science Translational Medicine (ISSN 1946-6242) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science Translational Medicine* is a registered trademark of AAAS.