Journal of the American Medical Informatics Association, 25(1), 2018, 25–31
doi: 10.1093/jamia/ocx101
Advance Access Publication Date: 26 September 2017
Research and Applications

Research and Applications

It's all in the timing: calibrating temporal penalties for biomedical data sharing

Weiyi Xia,[1,7] Zhiyu Wan,[2,7] Zhijun Yin,[2,3,7] James Gaupp,[1,7] Yongtai Liu,[2,7] Ellen Wright Clayton,[4,5,6,7] Murat Kantarcioglu,[8] Yevgeniy Vorobeychik,[1,2,7] and Bradley A Malin[1,2,3,7]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA, [2]Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN, USA, [3]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA, [4]Center for Biomedical Ethics and Society, Vanderbilt University Medical Center, Nashville, TN, USA, [5]Law School, Vanderbilt University, Nashville, TN, USA, [6]Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN, USA, [7]Center for Genetic Privacy and Identity in Community Settings, Vanderbilt University Medical Center, Nashville, TN, USA and [8]Department of Computer Science, University of Texas at Dallas, Richardson, TX, USA

Corresponding Author: Weiyi Xia, Department of Biomedical Informatics, 2525 West End Avenue, Suite 1475, Nashville, TN 37203 USA. E-mail: weiyi.xia@vanderbilt.edu. Phone: +1 615 887 4798

Received 30 May 2017; Revised 5 August 2017; Editorial Decision 16 August 2017; Accepted 23 August 2017

ABSTRACT

Objective: Biomedical science is driven by datasets that are being accumulated at an unprecedented rate, with ever-growing volume and richness. There are various initiatives to make these datasets more widely available to recipients who sign Data Use Certificate agreements, whereby penalties are levied for violations. A particularly popular penalty is the temporary revocation, often for several months, of the recipient's data usage rights. This policy is based on the assumption that the value of biomedical research data depreciates significantly over time; however, no studies have been performed to substantiate this belief. This study investigates whether this assumption holds true and the data science policy implications.

Methods: This study tests the hypothesis that the value of data for scientific investigators, in terms of the impact of the publications based on the data, decreases over time. The hypothesis is tested formally through a mixed linear effects model using approximately 1200 publications between 2007 and 2013 that used datasets from the Database of Genotypes and Phenotypes, a data-sharing initiative of the National Institutes of Health.

Results: The analysis shows that the impact factors for publications based on Database of Genotypes and Phenotypes datasets depreciate in a statistically significant manner. However, we further discover that the depreciation rate is slow, only ~10% per year, on average.

Conclusion: The enduring value of data for subsequent studies implies that revoking usage for short periods of time may not sufficiently deter those who would violate Data Use Certificate agreements and that alternative penalty mechanisms may need to be invoked.

Key words: biomedical data science, data sharing, policy, economics of data, genomics

INTRODUCTION

Biomedical research is increasingly data-driven.[1] This phenomenon is facilitated by advances in technologies that enable the collection, storage, and processing of information in a finely detailed and high-throughput manner.[2] These technologies are being adopted in traditional clinical domains (eg, electronic medical record systems[3]) and research settings (eg, whole-genome sequencing[4]), but are also expanding into nontraditional environments (eg, sensors in mobile phones[5] and social media[6]). We are creating new datasets at an unprecedented rate, with growing volume and richness.

While these datasets are valuable for the primary scientific investigations for which they are created, many believe they should be shared and reused for multiple purposes, including verification of research findings, new scientific inquiries, and formation of larger data compendiums to enable more powerful statistical claims. Given the public resources devoted to their creation, funding agencies, including the National Institutes of Health (NIH),[7] the National Science Foundation,[8] and the Patient-Centered Outcomes Research Institute,[9] have adopted policies to promote data-sharing activities. For instance, the NIH issued its Genomic Data Sharing Policy in 2014, which requires all genome-based studies that receive NIH funding to have a data-sharing plan.[10] To support sharing, the NIH constructed various repositories that are accessible to investigators across the globe, such as the Database of Genotypes and Phenotypes (dbGaP).[11] This is a publicly accessible central repository created to host individual-level phenotype, exposure, genotype, and sequence datasets, as well as the associations among such factors. dbGaP has provided data to a number of investigators with highly varied portfolios.[12] The overarching goal of this study is to promote the development of effective policies to accelerate biomedical research data sharing while mitigating the concerns and protecting the public's trust by analyzing the efficacy of current enforcement.

There are 2 distinct concerns raised from data sharing: (1) that those who initially collect the data will not have adequate time to publish their findings, and (2) that uses that are discordant with the conditions placed on the data may undermine public acceptance of data sharing.[13] To mitigate concerns, laws and policies of funding agencies often require data recipients to sign a contract, such as a Data Use Certificate (DUC) agreement, that indicates responsibilities and liabilities. While there are concerns over blatant abuse of data (eg, reidentification of deidentified records[14]), the majority of violations to date have arisen from a failure to comply with policy (eg, reuse of data without approval and insufficient documentation of security procedures). Thus, policymakers, as well as the authorities managing access to data, need to design and enforce penalties that induce sufficient, but reasonable, losses to the data recipients to deter them from undesirable actions.

One penalty adopted by various funding agencies, including the NIH's dbGaP and the Wellcome Trust Case Control Consortium,[15] is temporal exclusion from access. In this situation, the violator is barred from accessing the database, as well as conducting research, publishing papers, or writing grant proposals using the database for a period of time. The length of time is influenced by the type and severity of the violation, as well as whether the individual is a repeat offender. The temporal penalty is based, in part, on the rationale that data depreciate in value over time for their users. This is rooted in the assumption that, when data are shared, multiple researchers may be interested in reusing the data for similar investigations and will compete to publish their findings first, thereby decreasing the value for other users. However, it is unknown whether this competition actually occurs.

In this study, we examine whether or not data decreases in value over time and provide guidance on how to set temporal penalties in the event that data value does indeed depreciate. We specifically invoke the anticipated impact of a publication that uses a dataset, represented by the journal impact factor (JIF) and the journal Eigenfactor score (JES) as a proxy for the value an investigator gains from the dataset. We test the hypothesis that the value of dbGaP datasets decreases over time through use of a corpus of >1200 publications based on dbGaP datasets. The results indicate that the value depreciates in a statistically significant manner, but the rate of change is small (~10% per year). The remainder of this paper discusses how this analysis was performed and the implications for biomedical data science policymaking.

BACKGROUND

Temporal penalty

One of the major disincentives for investigators to sharing data is the concern over the loss of a competitive advantage.[16] Thus, temporal penalties have been adopted by several data-sharing initiatives. For example, the DUC agreement of the Genetic Association Information Network International Multi-Center ADHD Genetics Project (available through dbGaP: https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?view_pdf&stacc=phs000016.v2.p2) states that a recipient who violates the terms may forfeit access to all NIH genomic datasets. The period of time for which violators are barred from the repository is documented in the dbGaP DUC agreement compliance violation report.[17] At the time of this study, there were only 27 documented incidents. A brief summary of each incident, the policy expectations violated, and the action taken and/or preventive measures implemented are detailed in the report. Our informal review of this report suggests that users who violated the DUC agreement are typically suspended from accessing dbGaP data for a period of 3–6 months. An example of such an incident transpired in 2009, when a user of the Genome-Wide Association Study of Schizophrenia data (dbGaP accession number phs000021) conducted research that was not documented in the data access request. This was a violation, because the DUC agreement requires that users use the data only for the purpose described in the approved data access request. Once this incident was detected and reviewed, access to all NIH genomic datasets was revoked for 3 months.

The value of sharing and reusing research data

Sharing research data can affect different stakeholders in a variety of ways. A notable survey of the impact of sharing research data[18] reviewed the different rationales and the corresponding beneficiaries behind the promotion of data sharing. These rationales include (1) making results available to the public, (2) stewarding the resources applied to collect and curate data, (3) reproducing and verifying results, (4) enabling new investigations using one or more data sources, and (5) advancing research and innovation. The rationales most relevant to our investigation are (3), (4), and (5) because they are research-driven and concern the benefits for scientific investigators.

Studies have also assessed the amount of value primary investigators can gain when they share their data and secondary investigators can gain when they have access. A representative example is a study of the association between the increased citation rate of a publication and whether or not the detailed research data used in the publication are shared.[19] The method used in[19] is similar to ours in that it relied on a cohort of publications and tested a hypothesis about the value resulting from data sharing. However, it relied on a different notion of value. Specifically, the analysis focused on the value of a publication that described a primary investigation that produced a dataset, while our study focuses on the value of the secondary publications that reuse a dataset produced by primary investigators. Another major difference to note is that in[19] the value of the publication was represented as the number of times a paper was cited. While citation count is a proxy for the value gained through sharing, other proxies could be invoked, such as JIF[20] and JES.[21] We focus on an investigator's perceived valuation at the time of publication rather than the value realized afterward, as the former is more in line with our overarching hypothesis.

Our focus on impact factor, as opposed to actual citations, is further motivated by the fact that the citation count can be affected by the date of the publication to a large extent. An older publication, for instance, is more likely to receive a larger citation count than one that is more recent. And normalizing for the number of years a paper has existed will not account for the different citation rate trajectories that papers generate. By contrast, JIF and JES can be relied on to compare the values of 2 publications, regardless of when the publication was accepted, based on the commonly held belief that a paper with more scientific contributions is more likely to be accepted by a high-impact journal. Thus, in our study, we rely on JIF and JES as proxies of the value that both the primary and secondary investigators obtain by analyzing a dataset and publishing the results.

Another related examination of the value of sharing data is the analysis conducted by Paltoo and colleagues[12] on the value that secondary investigators gain from initiatives of the 2007 NIH Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies[22] and dbGaP.[12] This work constructed a corpus of the publications resulting from investigations using the data deposited in dbGaP from 2007, the year when it came online, to 2013, the year when the study was completed. Using this publication corpus, this study assessed the impact of the secondary use of the dbGaP data. In particular, this study reported on the annual increase in secondary publications based on dbGaP data. It further reported case studies of novel scientific discoveries and the increased strength of association made from individual or combined dbGaP datasets. Our study is based on this publication corpus, but our focus is on the value of each publication for a scientific investigator instead of the total scientific return the funding agency receives.

Finally, Piwowar and colleagues[23] constructed a corpus of secondary publications that use data from the Gene Expression Omnibus database (also maintained by the National Center for Biotechnology Information). The focus of this study was on the overall scientific return of initiative, which was substantiated by showing that the total number of secondary publications yielded from all the datasets was 1150 by the end of 2010. This study is notable in that it shows that valuation based on publications over time will not be limited to the dbGaP initiative.

MATERIALS AND METHODS

Materials

The data for this investigation were derived from a corpus of 1205 publications involving the analysis of dbGaP datasets that were authored by approved recipients as of 2013.[12] Each instance in our data corresponds to a ⟨publication, dbGaP dataset⟩ pair. Multiple instances were derived from a publication if it used more than one dbGaP dataset. For each instance, we collected information on the following variables:

1. $\text{date}_{\text{embargo\_expire}}$: dbGaP dataset embargo expiration date
2. PHSID: dbGaP dataset phs id (ie, the accession number)
3. $\text{date}_{\text{received}}$: Received date of the manuscript
4. $\text{date}_{\text{publish}}$: Published print date of the manuscript
5. $date_{e\_publish}$: Published online date of the manuscript
6. JIF: Journal impact factor for the publication
7. JES: Journal Eigenfactor score for the publication

The dbGaP embargo expiration dates were provided by the team who published.[12] Notably, there is a release date and an embargo expiration date for each version (.v#) and participant set (.p#) of a study (phs#). The release date is when the data are made accessible by dbGaP. The embargo expiration date corresponds to when the secondary investigators are permitted to publish. We use the embargo expiration date of the first version and first participant set (phs#.v1.p1).

We downloaded the Journal Citation Reports of the involved journals from 2006 and 2015. For each paper in the dataset, we assigned JIF and JES in the year prior to when the paper was published.

The received date, the published online date, and the published print date were obtained from the PubMed and PMC databases via the National Center for Biotechnology Information Entrez system.[24] Received date is not reported for a large proportion of publications, including high-impact journals such as *Science*, *Nature*, and *Nature Genetics*. Additionally, a subset of journals are published in print form only, such that their papers lack a published online date. Thus, we designed an imputation process (details in Supplementary Appendix) to avoid having to triage a large number of instances. For publications missing a received date, if the publication is online, we impute this value using a linear regression model of the received date versus the published online date. Otherwise, we impute the received date using a linear regression model of the received date versus the published print date. We consider both models to be valid, because they indicate a strong linear relationship between the 2 variables involved. Specifically, the R^2 for the models of received date as a function of published online date and published print date were 0.971 and 0.953, respectively.

All data have been deposited in the online Dryad repository.[25]

Data triage pipeline

We relied on instances that satisfied the following conditions: (1) devoid of missing values, (2) the publication was received after the embargo expiration date of the dbGaP dataset, and (3) the dbGaP dataset was used in more than one publication. We use the second condition because an instance in our dataset could correspond to a publication received before the embargo release date of its dbGaP dataset. This happens when a primary investigator is among the group of authors. The third condition is applied because a dbGaP dataset needs multiple observation points to show a change in value.

Figure 1 depicts the triage pipeline based on the aforementioned conditions. We began with 1451 instances. In the first step, we removed 166 instances with missing values for the dbGaP dataset embargo expiration date, JIF or JES. In the second step, we removed 281 instances where the publication received date transpired before the dbGaP dataset embargo expiration date. In the third step, we removed 44 instances where the dbGaP study dataset appears in only one sample. The final dataset consists of 960 observations.

Data analysis

We test the hypothesis that the value a scientific investigator can gain from a dbGaP dataset decreases over time. The dataset value is represented by the JIF and JES of the publication. To reduce the skew in the distributions, we log transformed the response variables to $\log_2(\text{JIF})$ and $\log_2(\text{JES})$. We use $period = \text{date}_{\text{received}} - \text{date}_{\text{embargo\_expire}}$ as the independent variable.

To test our hypothesis, we adopt a linear mixed-effects (LME) model to capture the relationship between the response variable and the independent variable *period*. We use LME instead of a more conventional linear regression, because each dbGaP dataset is affiliated with a varying number of publications that are not

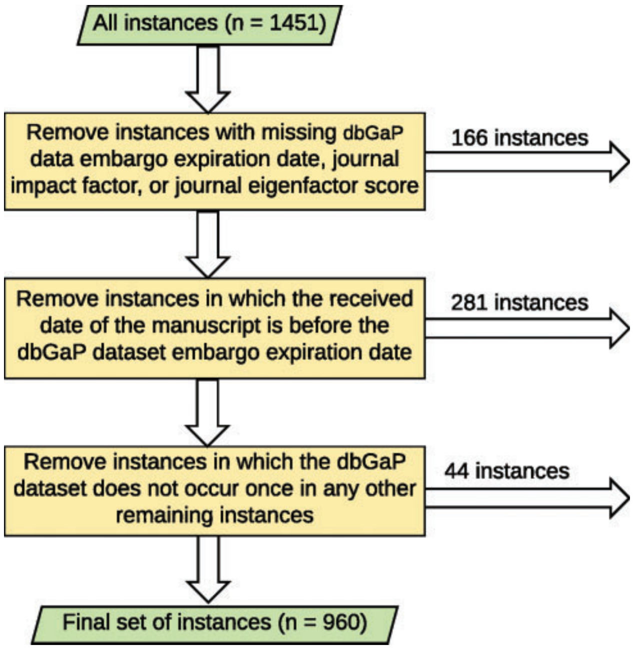
All instances (n = 1451)
Remove instances with missing dbGaP data embargo expiration date, journal impact factor, or journal eigenfactor score
166 instances
Remove instances in which the received date of the manuscript is before the dbGaP dataset embargo expiration date
281 instances
Remove instances in which the dbGaP dataset does not occur once in any other remaining instances
44 instances
Final set of instances (n = 960)

Figure 1. The triage process for the data used in this study.

independent. As such, we rely on PHSID as a grouping variable. The LME model uses a random intercept and slope with one continuous independent variable *period*, where both the intercept and slope vary by PHSID.

Formally, we adopt the LME model implementation of the R nlme package (version 3.1.126). The model was defined as follows:

$$E[y_{ij}|\ \text{period}_{i,j}\ ,\gamma_i] = \beta_0 + \beta_1\ \times \text{period}_{i,j}\ + \gamma_{i0} + \gamma_{i1}\ \times \text{period}_{i,j} + \ \varepsilon_{i,j}$$

$$(\gamma_{i0},\ \gamma_{i1})\ \sim\ N(0, D)$$

$$\varepsilon_i\ \sim\ N\left(0, \sigma^2\right)$$

$$i = 1, \ldots, m$$

$$j = 1, \ldots, n_j$$

where the response variable in the model (which corresponds to $\log_2$ (JIF) and $\log_2$(JES)) is y, while y_{ij} corresponds to the jth publication in group i. The other variables in the model are defined as follows:

m: Number of groups
n_i: Number of instances in group i
$\text{period}_{\text{i,j}}$: Period of the jth publication in group i
ε_{ij}: Random error
β_0: Fixed intercept
β_1: Fixed slope of the period
γ_{0i}: Random intercept
γ_{1i} : Random slope of period
D: Covariance matrix of random effect γ_i

Given the LME model, we formulate the null hypothesis as $\boldsymbol{H}_0 : \boldsymbol{\beta}_1 = 0$ and the alternative hypothesis as $\boldsymbol{H}_a : \boldsymbol{\beta}_1 \neq 0$. We reject the null hypothesis at a significance level of 0.05. If the null hypothesis is rejected, we further analyze the value of $\boldsymbol{\beta}_1$ to determine the rate at which the value of the dataset is changing.

RESULTS

This section begins with the results from model selection and validation, including the probability density of the JIF and JES of the publications and the residual Q-Q plot of the fitted models. We then present a visualization of $\log_2$(JIF) and $\log_2$(JES) as a function of *period* to illustrate the relationships between the corresponding variables and the resulting LME models and the hypothesis test result. It should be noted that our focus is on the fixed-effect coefficient β_1, which is used in the hypothesis test. Based on the fitted models, we computed the amount of value depreciation of the dbGaP dataset for scientific investigators.

The publications involved in this analysis are from 163 different journals. The journal names, impact factors, and Eigenfactor scores are listed in the Supplementary Appendix. The impact factors of these journals range between 51.658 (*New England Journal of Medicine* in 2012) and 0.864 (*Chinese Medical Journal* in 2011). The Eigenfactor scores of these journals range between 1.76345 (*Nature* in 2008) and 0.0005 (*Biodemography and Social Biology* in 2012). The probability densities of JIF and JES are depicted in Figure 2. Both JIF and JES are highly skewed to the right, such that a log transformation is applied to JIF and JES before being fitted to the LME model. The residual normal Q-Q plot of $\log_2$(JIF) and $\log_2$(JES) are shown in Figures 3 and 4, respectively. The JIF density plot also shows that there are 2 coarse groups of journals: those with values above 20 and those below 20. We recognized that the 2 groups could affect the validity of our LME model. Thus, instead of fitting a separate model to each group of journals, which runs counter to the hypothesis we were testing, we assessed the hypothesis that the probability of publishing in a high-impact journal using a dataset decreases over time. The results of this analysis are omitted due to the length limitation; instead, they are reported in the Supplementary Appendix. In short, the results suggest that this hypothesis holds true in a statistically significant manner as well.

The scatterplots with a LOESS curve of period vs $\log_2$(JIF) and period vs $\log_2$(JES) are shown in Figures 5 and 6, respectively. These plots show that $\log_2$(JES) exhibits a decreasing pattern along with an increase in period, while $\log_2$(JIF) exhibits the same pattern during the first 4 years.

The fixed-effects parameter estimates of the models are shown in Tables 1 and 2. The slope β_1 of the 2 fitted models suggests that both $\log_2$(JIF) and $\log_2$(JES) decrease as the period increases. In particular, the β_1 of the $\log_2$(JIF) model is −0.1348 with a *P*-value of .0028, and β_1 of the $\log_2$(JES) model is −0.1422 with a *P*-value of .0479. Therefore, in both models the null hypothesis, that the value of data does not decrease over time, can be rejected. Since the response variable of each model is the $\log_2$ transformation of the original JIF and JES, the average actual JIF and JES of the current year is 2^{β_1} times the average value of the previous year. In particular, the average JIF and JES of the current year is $2^{-0.1348} = 0.91$ and $2^{-0.1422} = 0.91$ times, respectively, that of the previous year. The results indicate that there is consistency in the JIF and JES valuations of the dataset, both of which drop at a relatively small rate, ∼10% annually.

In Table 3, we show the predicted value depreciation of a dbGaP dataset for scientific investigators subjected to a temporal penalty of a period ranging from 3 months to 3 years using our model. The depreciation quantities in Table 3 do not rely on the particular measure used, because the rates of JIF and JES are consistent. The results show that the amount of value depreciation caused by a temporal penalty of 3 or 6 months, which is typical for existing incidences where the temporal penalty is imposed, is small at 2–4.6%.

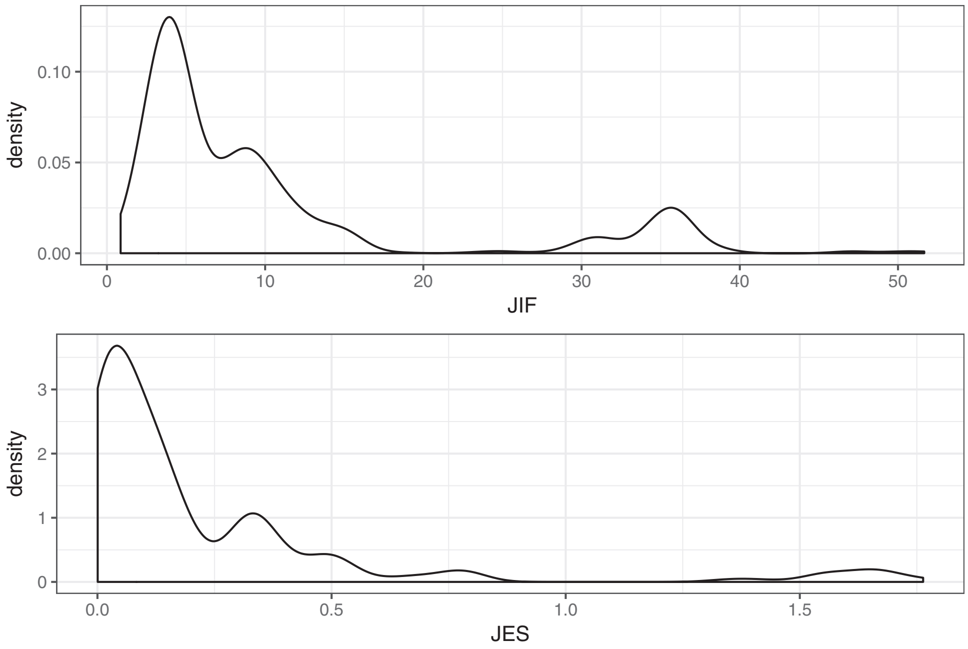

Figure 2. The probability densities of journal impact factor (JIF) and journal Eigenfactor score (JES).

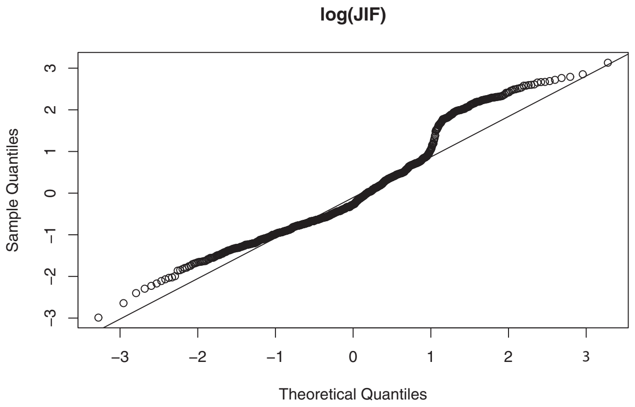

Figure 3. The residual normal Q-Q plot of the log transformation of the journal impact factor ($\log_2$ (JIF)).

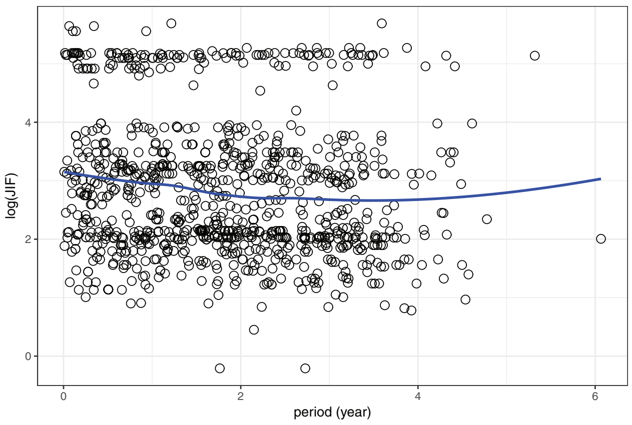

Figure 5. A scatterplot, with a LOESS smoothing curve, of the log transformation of the journal impact factor ($\log_2$ (JIF)) vs the period.

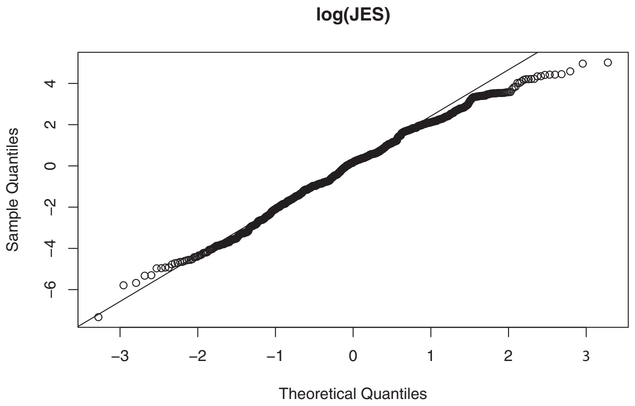

Figure 4. The residual normal Q-Q plot of the log transformation of the journal Eigenfactor score ($\log_2$ (JES)).

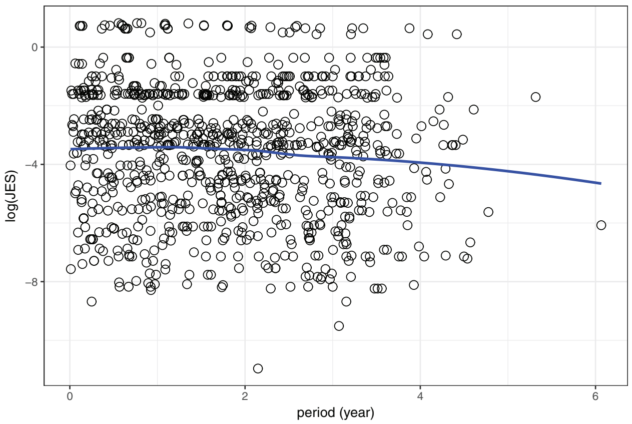

Figure 6. A scatterplot, with a LOESS smoothing curve, of log transformation of the journal Eigenfactor score ($\log_2$ (JES)) vs the period.

Table 1. Fixed-effects parameter estimates of the JIF model

Coefficient	Value	Std. Error	DF	*t*-value	*P*-value
β_0	3.0245	0.0993	879	30.4725	.0000
β_1	−0.1348	0.0449	879	−2.9985	.0028

Table 2. Fixed-effects parameter estimates of the JES model

Coefficient	Value	Std. Error	DF	*t*-value	*P*-value
β_0	−3.3484	0.1653	879	−20.2573	.0000
β_1	−0.1422	0.0718	879	−1.9812	.0479

Table 3. The predicted value depreciation of a dbGaP dataset as a function of the temporal penalty period length

Period (months)	3	6	12	18	24	30	36
Value depreciation (%)	2.0	4.6	8.9	13.0	17.0	21.0	24.0

DISCUSSION

In summary, the primary finding of this study is that the value of dbGaP datasets, as assessed by the journal impact and Eigenfactor scores of the publications in which they are included, depreciates at an annual rate of ∼10%. Here, we take a moment to reflect on the benefits and drawbacks of this discovery for biomedical data science policymaking, and to consider several limitations regarding this result.

Implications of our findings

There are several notable aspects of our findings that should be highlighted. First, the slow depreciation rate is good news, in that our findings support the hypothesis that storing and sharing data are valuable for the scientific enterprise, contributing to new discoveries, and that value endures over time.

At the same time, the enduring value of these data does suggest that excluding embargo violators for 3–6 months may not be an effective deterrent, as the value of the data will still be significant after the penalty ends. It may be that other strategies need to be implemented as alternatives, or addendums, to sufficiently deter violations of DUC agreements. One such approach would be to prevent investigators from accessing other resources, such as being unable to compete for new grant funding over a certain number of review cycles. Another approach to consider would be to shift from temporal holdouts, which depend on a particular view of data value, to explicit financial penalties, whereby violators are fined for violations. Of course, we recognize that federal agencies like the NIH may not be able to impose such fines without additional authority. Still, this could be a solution worth pursuing if the concern for violation grows, which is likely to occur as the quantity and quality of such data-sharing resources escalate. The National Academies of Sciences, Engineering, and Medicine recently urged in its report "Fostering Integrity in Research" that research institutions must assume greater responsibility to ensure the appropriate behavior of researchers.[26] Failing to exercise appropriate oversight to ensure compliance with data use agreements, then, could open an institution to criticism and perhaps additional financial penalties. Moreover, financial penalties could be a readily useable solution for organizations or consortia that can draw up use agreements that embed liquid damages for violation of the agreed-upon terms.

As an alternative to direct monetary penalties, it might be worth identifying violators publicly, a policy that has been adopted by the Office for Civil Rights at the US Department of Health and Human Services in support of the Security Rule of the Health Insurance Portability and Accountability Act (HIPAA) of 1996. Specifically, the Office for Civil Rights publicly publishes both the name of the organization and the amount of fine levied on a public "Wall of Shame" website for those who have had more than 500 patients' records breached.[27]

Limitations of the study

This study has limitations that suggest several future research directions. First, our analysis involves datasets only from dbGaP, which may introduce a bias, as it neglects other resources available in the United States and abroad. While we anticipate that our findings will generalize, we do not have evidence at present to confirm this claim.

Second, the value of each dataset was treated independently. This may inflate the value of a specific dataset, because multiple datasets are, at times, brought together for more powerful analyses. This may help to explain such a slow depreciation rate. Still, this would not detract from our conclusions, as it simply shows that value can be gained through data aggregation.

Third, our LME models consider the random effects caused only by the dbGaP dataset used in a publication. Other variables could induce effects, such as the type of study (eg, methodological development vs biomedical discovery) and the specific biological phenomenon investigated.

Fourth, we used the embargo expiration date of the first version and participant set of a dbGaP study. This neglects the particular version that was actually used in a publication. This decision was motivated by 2 reasons: (1) the version and participant set are not available for all publications, and (2) the difference between versions was anticipated to be small. Nonetheless, the latter assumption may not hold for certain dbGaP datasets.

Finally, we recognize that there may be other undesirable consequences for researchers who are subjected to a temporal penalty, including the inconvenience caused by the forced delay in work or the loss of opportunities to compete for grant funding due to a forced delay in applications. In this paper, however, we believe that the value of data over time is a good starting point to study the effects of the temporal penalty, because (1) the notion that value of data decreases over time is an important motivation for temporal penalties, and (2) to the best of our knowledge, there are no data readily available to study the effects of the delay in research work and funding applications.

CONCLUSION

This is the first study demonstrating that datasets in a large biomedical data research repository, dbGaP, retain substantial value over time when used in additional research. These findings provide powerful support for data sharing, but they also suggest that short-term exclusion from data access, the most common penalty currently imposed for violations of DUCs, is a weak deterrent in its own right. Thus, other mechanisms will need to be developed to ensure the appropriate use of these datasets in order to recognize the efforts of data collectors and, more importantly, ensure the public's trust in biomedical data science.

ACKNOWLEDGMENTS

This research was sponsored, in part, by the following grants from the National Institutes of Health: R01HG006844, RM1HG009034, U01HG006385, U01HG006378, and the the National Library of Medicine Training Grant T15 LM007450. The authors also wish to thank the following employees of the NIH for providing some of the core of the dataset used in this study, as well as helpful discussions on modeling and initial findings: Mike Feolo, Masato Kimura, Kristofor Langlais, Dina Paltoo, and Laura Rodriguez.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

REFERENCES

1. Margolis R, Derr L, Dunn M, *et al*. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *J Am Med Inform Assoc*. 2014;21:957–58.
2. Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA*. 2013;309:1351–52.
3. Adler-Milstein J, DesRoches C, Kralovec P, *et al*. Electronic health record adoption in US hospitals: progress continues, but challenges persist. *Heal Aff*. 2015;34:2174–80.
4. Hayden EC. Technology: the $1,000 genome. *Nature*. 2014;507:294–95.
5. Kumar S, Abowd GD, Abraham WT, *et al*. Center of excellence for mobile sensor data-to-knowledge (MD2K). *J Am Med Inform Assoc*. 2015;22:1137–42.
6. Grajales III FJ, Sheps S, Ho K, *et al*. Social media: a review and tutorial of applications in medicine and health care. *J Med Internet Res*. 2014;16:e13.
7. National Institutes of Health. *Final NIH Statement on Sharing Research Data*. Notice Number NOT-OD-03-032. February 26, 2003.
8. National Science Foundation. *Dissemination and Sharing of Research Results*. www.nsf.gov/bfa/dias/policy/dmp.jsp. Accessed August 3, 2017.
9. Patient Centered Outcomes Research Institute. *Data Access and Data Sharing Policy: Draft For Public Comment*. 2016. www.pcori.org/sites/default/files/PCORI-Data-Access-Data-Sharing-DRAFT-for-Public-Comment-October-2016.pdf . Accessed August 3, 2017.
10. National Institutes of Health. *NIH Genomic Data Sharing Policy*. Notice Number NOT-OD-14-124. August 27, 2014.
11. Mailman MD, Feolo M, Jin Y, *et al*. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet*. 2007;39:1181–86.
12. Paltoo DN, Rodriguez LL, Feolo M, *et al*. Data use under the NIH GWAS Data Sharing Policy and future directions. *Nat Genet*. 2014;46:934–38.
13. Pacheco CM, Daley SM, Brown T, *et al*. Moving forward: breaking the cycle of mistrust between American Indians and researchers. *Am J Public Health*. 2013;103:2152–59.
14. Rodriguez LL, Brooks LD, Greenberg JH, *et al*. The complexities of genomic identifiability. *Science (80-)*. 2013;339:275 LP–276.
15. The Wellcome Trust Case Control Consortium, Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447:661–78.
16. Borgman CL. Research data: who will share what, with whom, when, and why? China-North America Library Conference, Beijing. 2010.
17. National Institutes of Health. *Compliance Statistics for Policies that Govern Data Submission, Access, and Use of Genomic Data*. https://gds.nih.gov/20ComplianceStatistics_dbGap.html. Accessed August 3, 2017.
18. Borgman CL. The conundrum of sharing research data. *J Am Soc Inf Sci Technol*. 2012;63:1059–78.
19. Piwowar HA, Day RS, Fridsma DB. Sharing detailed research data is associated with increased citation rate. *PLoS One*. 2007;2:e308.
20. Garfield E. The history and meaning of the journal impact factor. *JAMA*. 2006;295:90–93.
21. West JD, Bergstrom TC, Bergstrom CT. The Eigenfactor metrics^{TM}: a network approach to assessing scholarly journals. *Coll Res Libr*. 2010;71:236–44.
22. National Institutes of Health. *Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies*. Notice Number NOT-OD-08-088. August 28, 2007.
23. Piwowar HA, Vision TJ, Whitlock MC. Data archiving is a good investment. *Nature*. 2011;473:285.
24. Acland A, Agarwala R, Barrett T, *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2014;42:D7–17.
25. Xia W, Wan Z, Yin Z, *et al*. Data from: *It's All in the Timing: Calibrating Temporal Penalties for Biomedical Data Sharing*. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.nr607. Accessed September 11, 2017.
26. National Academies of Sciences, Engineering, and Medicine. *Fostering Integrity in Research*. Washington, DC: The National Academies Press; 2017.
27. *Office for Civil Rights, U.S. Department of Health and Human Services, Breach Portal: Notice to the Secretary of HHS Breach of Unsecured Protected Health Information*. https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf. Accessed August 3, 2017.