MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220047

Exploring Determinants of Longevity of Biomedical Databases

Joseph Finkelstein[a], Jennifer Guarino[a], Xingyue Huo [a], Kirill Borziak[a], Irena Parvanova[a]

[a]*Center for Biomedical and Population Health Informatics, Icahn School of Medicine at Mount Sinai, New York, NY, USA*

Abstract

The maintenance of biomedical databases requires ongoing and systematic efforts in keeping them up-to-date which may affect long-term sustainability. Since research has become more reliant on publicly available biomedical data collections, it is important to understand factors affecting their longevity. The aim of this article was to explore potential determinants of biomedical database longevity. To build an analytical dataset, we used Database journal that have been created as an open access platform for presenting biological databases. A stratified analysis of all the original databases published in Database journal between 2009 and 2016 was conducted depending on their accessibility status. Overall, 35% of 518 analyzed databases were not accessible in 2020. We showed that databases with higher citation counts from institutions with higher scientific output were significantly more likely to be currently accessible. Databases from researchers with higher h-index were more likely to be accessible. Further investigation is warranted to identify factors affecting longevity of high impact databases.

Keywords:

Biomedical databases, longevity, predictive analytics

Introduction

Biomedical data management has been receiving increasing attention as Open Science approaches gain popularity as major vehicle for research collaboration, transparency and reproducibility [1]. During the recent decades the scientific community witnessed an explosive proliferation of biomedical databases [2]. Many of these databases have proven to be indispensable resources that greatly catalyzed knowledge discovery and predictive analytics [3]. To further promote effective use of big data in biomedical research the guiding principles of data Findability, Accessibility, Interoperability, and Reusability (FAIR) have been introduced [4]. Following a considerable discussion on ways to enhance reliability and trustworthiness of shared research data, a common framework for best practices in digital preservation has been developed which includes Transparency, Responsibility, User focus, Sustainability and Technology (TRUST) [5]. Design, implementation and maintenance of biomedical databases in compliance with the FAIR and TRUST principles may be a time consuming and costly endeavor which requires continuous allocation of significant resources [6]. A recent longevity analysis of 326 biological databases demonstrated that after a period of 18 years 76.7% of the databases were abandoned [7]. Thus, it is important to prioritize allocation of resources and efforts into databases with potentially higher longevity and long-lasting impact. However, determinants of longevity of biomedical databases have not been studied systematically. The aim of this pilot study is to identify potential predictors of biomedical databases' longevity.

Methods

To analyze longevity of biomedical databases an analytical dataset has been constructed which included characteristics of all databases discussed in a scientific journal specifically aimed at presenting information on biomedical databases. Use of a single journal allowed us avoid bias resulting from varying journal impact factors. To curate the dataset for our analysis of biomedical database longevity, we collected all publications indexed in PubMed for the journal Database: The Journal of Biological Databases and Curation for the years spanning 2009 to 2016 were. For the purposes of this analysis, only articles which focused on the presentation of a database were retained for further analysis, excluding articles that focused on reviews, workflows, workshops and methodologies. For the resultant set of 518 articles, citation information was retrieved from PubMed, including citations from 2009 to 2020. The h-index for the first and last author of each article was obtained from Google Scholar. The last author's institution ranking was retrieved from two sources: Nature Index institution outputs covering December 1, 2019 to November 30, 2020, for both the article count and fractional authorship count, and from the Scrimago Institution Rankings for university rankings. Databases were classified based on their current accessibility through the links provided in their respective publications as of April 1, 2021. Databases which were not accessible or which were missing the data they were designed to contain were classified as offline. To ensure classification integrity, efforts were made to verify that database have not migrated to a different URL. If offline databases were located at a different address, they were then reclassified as accessible. Database publications were classified on whether they were presenting a novel database or a database update, based on the presence of previous articles or citations in the article or on the database website. Databases were also categorized based on the basis of their main form of data collection. Databases were classified as computational data collection if the data was primarily computer generated, and manual data collection if the data collection involved extensive manual curation. Database publications were classified by funding type (government or other) based on the reported funding in the relevant section in the publication. Other funding included nonprofit, profit and no funding categories.

The databases were divided into 'accessible' and 'offline' groups, and the related predictive variables were compared between the two groups. Dead databases which were not accessible online via an original or proxy link were labeled 'offline.' The continuous and categorical variables were presented as mean and frequency (percentage). They were compared by two-sample T-test and chi-square test, respectively, to identify significant differences between 'accessible' and 'offline' databases. A multivariate logistic regression model was used to identify potential factors affecting database longevity. The model robustness was assessed by receiver operating characteristic (ROC) curves and the Concordance (C) statistic to evaluate the accuracy of database "dead/alive" status prediction.

Results

To assess the ability of published biomedical databases to remain maintained and accessible following publication, we retrieved all biological databases published in *Database: The Journal of Biological Databases and Curation* for the years 2009 through 2016. In total, our dataset contained 518 published biological databases. Of these 518 databases, a surprisingly high percentage were offline. In total, 34.9% (181) databases were not accessible. We further note that the percentage of databases offline by year range from a high of 44.6% for the year of 2013 to a low of 26.5% for the year 2016 (Figure 1). To further understand why over a quarter of published biological databases failed to be maintained for even a minimum of five years, we examined the effects of publication citations, author h-index and the author's institutional ranking affected the maintenance and public accessibility of these databases.

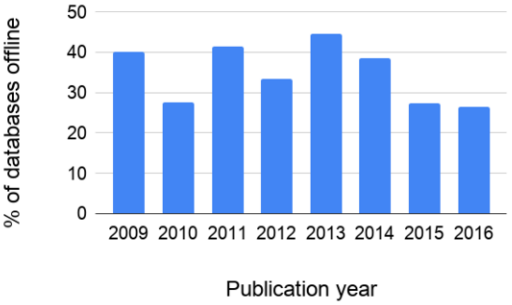

Figure 1– Distribution of published biological databases that are classified as offline by publication year.

We found that the number of citations a database publication receives has a significant predictive ability (p-value < 0.0001) on the chances that the database is currently accessible (Table 1). Compared to live databases, which have an average of 26.36 citations in PubMed, offline databases have on average only 11.77 citation. These results are further confirmed using Odds Ratio Estimates (Table 2), where the number of PubMed citations have a significant predictive ability (p-value = 0.004, point estimate = 1.06) for the probability that the database is currently accessible.

Table 1– Comparison of characteristics of accessible and offline databases.

	Database status						
	all		accessible		offline		
	count	Mean	count	Mean	count	Mean	p-value
PubMed citations	518	21.26	337	26.36	181	11.77	<.0001
First author h-index	279	16.48	175	15.85	104	17.55	0.19
Last author h-index	358	45.68	233	47.70	125	41.92	0.05
nature index publication output	455	440.75	291	487.85	164	357.18	0.03
nature index authorship share	455	128.45	291	145.59	164	98.02	0.01
Scrimago university rank	323	215.43	200	200.98	123	238.92	0.04
	count	**Percentage**	**count**	**Percentage**	**count**	**Percentage**	**p-value**
Database publication type							0.006
new	375	72.39	230	68.25	145	80.11	
update	143	27.61	107	31.75	36	19.89	
Total	518	100.00	337	100.00	181	100.00	
Data collection type							0.27
computational	284	54.83	179	53.12	105	58.01	
manual curation	234	42.66	158	46.88	76	41.99	
Total	518	100.00	337	100.00	181	100.00	
Funding type							0.04
government	401	77.41	251	74.48	150	82.97	
others	117	22.59	86	25.52	31	17.13	
Total	518	100.00	337	100.00	181	100.00	

We next examined whether the authors' publication records, as reflected in the h-index, can be a predictive factor for their database's current accessibility. We found no significant predictive value for the first author's citation index (p-value = 0.19). In fact, first authors of accessible databases have a slightly lower average h-index of 15.85, compared to 17.55 for first authors of offline databases. However, the odds ratio in logistic regression suggests a significant predictive probability that the databases of first authors with lower h-indexes are currently accessible (p-value = 0.006, point estimate = 0.919). The citation index of last authors also had a significant predictive value for accessibility of their databases (p-value = 0.05), with last authors of accessible databases having an average citation index of 47.7, compared to an average of only 41.92 for last authors of offline databases. Similarly, the odds ratio logistic regression shown some predictive ability that last authors with higher h-index are more likely to have their databases currently accessible (p-value = 0.089, point estimate = 1.017).

Table 2– Odds ratios from predicitve model for database accessibility.

		95% Wald		
Effect	**Point Estimate**	**Confidence Limits**		**p-value**
PubMed citations	1.060	1.020	1.103	**0.004**
First author h-index	0.919	0.866	0.976	**0.006**
Last author h-index	1.017	0.997	1.037	0.089
Nature Index publication output	0.992	0.986	0.998	**0.014**
Nature Index authorship share	1.017	1.000	1.033	**0.047**
Scrimago university rank	0.995	0.991	0.999	**0.028**
Database publication type	0.257	0.078	0.843	**0.025**
Data collection type	1.637	0.705	3.802	0.252
Funding type	1.013	0.363	2.830	0.980

Next, we examined the predictive ability of the ranking of the last authors' institutions on database accessibility. The institutional rankings were based on the Scrimago Institution Ranking and on the Nature Index publication output for the institutions as both total publications and as authorship share. Our statistical analysis suggested that databases that originate from higher ranked universities and those with higher publication output are more likely to be currently accessible for all three indexes. Higher Scrimago university ranking is a significant predictor (p-value = 0.04) of database accessibility, with an average ranking of 200.98 for accessible databases versus 238.92 for offline databases. Similarly, higher institutional publication output as measured by nature index is a significant predictor (p-value = 0.03) of database accessibility, with an average institutional output of 487.85 articles for accessible databases versus 357.18 for offline databases. Higher institutional publication output by authorship share is also a significant predictor (p-value = 0.01) of database accessibility, with an average institutional authorship share of 145.59 articles for accessible databases versus 98.02 for offline databases. The odds ratio from logistic regression showed similar significant predictive effect for Scrimago rankings (p-value = 0.028, point estimate = 0.995) and Nature Index authorship share (p-value = 0.047, point estimate = 1.017).

To further understand what characteristics affect database accessibility, we also examined characteristics specific to the databases themselves. We first examined if newly reported databases in their respective index publications versus being a database update had a predictive ability of database accessibility. New database publications accounted for 72.39% of articles in our data set. We saw that databases that have been published previous to their Database Journal articles are significantly more likely (p-value = 0.006) to be currently accessible, with 74.8% of previously published databases accessible versus only 61.3% of newly reported databases. This result is strongly supported by odds ratio from logistic regression (p-value = 0.025, point estimate = 0.257). We also examined if the main mode of data collection (computational vs manual curation) was predictive of database survival. However, the increased effort of manual curation did not affect database accessibility (p-value = 0.27), with manually curated databases being only slightly more likely to be currently accessible (67.5% vs 63% for computational data collection). Finally, we examined the effect of funding type on database accessibility. Surprisingly, we found that while government funding was used to develop the majority of databases (77.41%), databases that were developed with government funding (e.g. NIH or National Basic Research Program of China) were significantly less likely (p-value = 0.04) to be currently accessible compared to other funding types, such as non-profit, institutional, or no funding. Only 62.6% of government funded databases are currently accessible, compared to 73.5% funded through other methods.

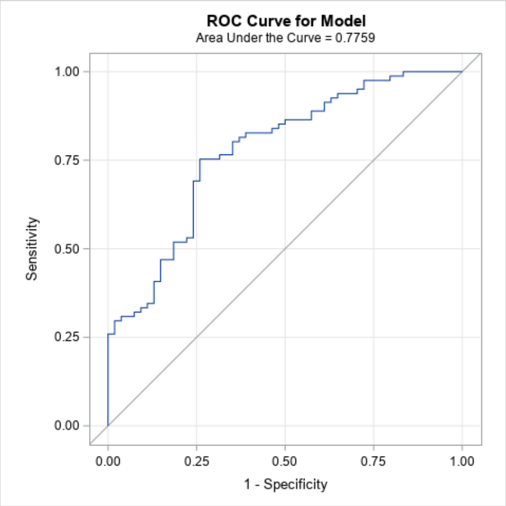

Figure 2– The ROC Curve of the predictive model assessing database accessibility.

In order to measure how well the combination of these variables is able to predict database accessibility, PubMed citations, first author h-index, last author h-index, Nature Index publication output and authorship share, Scrimago university rank, database publication type, data collection type, and funding type were used in a logistic regression with ‘accessible” status as a primary dependent variable. Each database was treated as a unit of analysis and assumed independent from other databases. The results showed that with these parameters we were able to effectively predict database accessibility with 77.59% accuracy (Figure 2).

Discussion

We analyzed characteristics of 518 publicly available online databases presented in peer-reviewed articles published between 2009 and 2016 in Database journal devoted to introduction of biomedical databases to scientific community. As of April 1, 2020, 35% of these databases were not accessible indicating significant attrition of biomedical databases over the time. Significant differences were identified between characteristics of surviving and dead databases. The surviving databases were published by authors with significantly higher h-index from institutions with higher publication output. The surviving databases were more likely to have prior publications and were more frequently cited in PubMed. Surprisingly, government funding was not identified as a significant predictor of database longevity. A multivariate logistic regression accounting for all potential covariates demonstrated sufficient accuracy of predicting database accessibility with C-statistic of 78%.

Our results are congruent with previous report which demonstrated that in the course of 18 years out of 326 biological databases only 16.3% remained alive and 7% were rebranded with the remaining databases not being accessible [7]. This article positioned weaker financial support as one of the primary factors affecting database longevity which was not fully supported by our initial analysis. However, we were in an agreement that databases originating from institutions with stronger academic environments or whose core mission was aligned with supporting that database development and maintenance were more likely to have prolonged longevity. As in our work, citation count was shown to be a significant predictor of biological database longevity in a recent analysis of 1.727 biological databases [8].

Recently introduced guidelines for improved data availability and reusability entitled FAIR [9] combined with a comprehensive set of approaches to enhance reliability and trustworthiness of shared research data entitled TRUST [10] are likely to facilitate meaningful data sharing and storage [11]. Intelligent integrative informatics approaches [12] utilizing cross-linked biomedical ontologies [13], common data models [14], and core outcomes sets [15] will promote data harmonization and longevity of evolving biomedical databases [16]. Application of appropriate data exchange standards with domain-relevant content standards combined with accessible rich metadata based on applicable terminologies will catalyze effective and sustainable data sharing in the future [17].

Our pilot study was restricted to databases published in a single peer-reviewed journal and included analysis of a limited number of database characteristics. Temporal trends, compliance with FAIR and TRUST policies, database size, functionality and subject area were not addressed in this initial analysis. Nevertheless, we were able to identify significant determinants of biomedical database longevity and build sufficiently accurate predictive model.

Conclusions

Significant number of publicly available biomedical databases became abandoned over the time. Development of biomedical databases with higher longevity and potential long-lasting impact may optimize resource and effort allocation. Future studies of biomedical database longevity should include broader spectrum of diverse databases from multiple sources and expand number of potential characteristics which may affect database survival.

Acknowledgements

This work was funded in part by NIH contract UL1TR001433.

References

[1] E. Foster, A. Deardorff. Open Science Framework (OSF), *Journal of the Medical Library Association* **105(2)** (2017), 203-206.

[2] U. Sivarajah, M. Kamal, Z. Irani, V. Weerakkody. Critical analysis of Big Data challenges and analytical methods, *Journal of Business Research* **70** (2017), 263-286.

[3] I. Masic, K. Milinovic. On-line Biomedical Databases – the Best Source for Quick Search of the Scientific Information in the Biomedicine, *Acta Inform Med* **20(2)** (2012), 72-84.

[4] M.D. Wilkinson, M. Dumontier, I. J. Aalbersberg, et al. The FAIR Guiding Principles for scientific data management and stewardship, *Sci Data* **6(1)** (2019), 6.

[5] D. Lin, J. Crabtree, I. Dillo, et al. The TRUST Principles for digital repositories, *Sci Data* **7(1)** (2020), 144.

[6] A. K. Jha, D. Doolan, D. Grandt, T. Scott, D. W. Bates. The use of health information technology in seven nations, *Int J Med Inform* **77** (2008), 848-854.

[7] T. K. Attwood, B. Agit, L. Ellis. Longevity of Biological Databases, *EMBnet J* **21** (2015), e803.

[8] H. J. Imker. 25 years of molecular biology databases: A study of proliferation, impact, and maintenance, *Frontiers in Research Metrics and Analytics* **3** (2018), 18.

[9] L. Reiser, L. Harper L, M. Freeling, B. Han, S. Luan. FAIR: A Call to Make Published Data More Findable, Accessible, Interoperable, and Reusable, *Mol Plant* **11(9)** (2018), 1105-1108.

[10] R. Horn, A. Kerasidou. Sharing whilst caring: solidarity and public trust in a data-driven healthcare system, *BMC Med Ethics* **21(1)** (2020), 110.

[11] P. Holub, F. Kohlmayer, F. Prasser, et al. Enhancing Reuse of Data and Biological Material in Medical Research: From FAIR to FAIR-Health, *Biopreserv Biobank* **16(2)** (2018), 97-105.

[12] J. Finkelstein, I. Parvanova, F. Zhang. Informatics Approaches for Harmonized Intelligent Integration of Stem Cell Research, *Stem Cells Cloning* **13** (2020), 1-20.

[13] F. Gutierrez. Semantic Technologies and Bio-Ontologies, *Methods Mol Biol* **1617** (2017), 83-91.

[14] I. Parvanova, J. Finkelstein. Data Integration Approaches for Representing Stem Cell Studies, Stud Health Technol Inform **270** (2020), 1235-1236.

[15] A. Elghafari, J. Finkelstein. Automated Identification of Common Disease-Specific Outcomes for Comparative Effectiveness Research Using ClinicalTrials.gov: Algorithm Development and Validation Study, *JMIR Med Inform* **9(2)** (2021), e18298.

[16] K. Borziak, I. Parvanova, J. Finkelstein. ReMeDy: a platform for integrating and sharing published stem cell research data with a focus on iPSC trials, *Database (Oxford)* **2021** (2021), baab038.

[17] R. D. Kush, D. Warzel, M. A. Kush, et al. FAIR data sharing: The roles of common data elements and harmonization, *J Biomed Inform* **107** (2020), 103421.

Address for correspondence

Joseph Finkelstein, MD PhD
Joseph.Finkelsein@mssm.edu