American Heart Association Precision Medicine Platform

Taha A. Kass-Hout, MD, MS, **Laura M. Stevens, BS**, **Jennifer L. Hall, PhD**

Institute for Precision Cardiovascular Medicine, American Heart Association, Dallas, TX (T.A.K.-H., L.M.S., J.L.H.). Kass-Hout Consulting, LTD, Washington, DC (T.A.K.-H.); Department of Cardiology and Computational Bioscience Program, University of Colorado, Aurora (L.M.S.). Lillehei Heart Institute, University of Minnesota, Minneapolis (J.L.H.).

Abstract

Integrating the open science movement with impactful discoveries in science, velocity of technology, and raw power of cloud computing has led to an unprecedented opportunity for scientific discovery. The American Heart Association recently established the Precision Medicine Platform[1] through the efforts of multiple American Heart Association volunteers and a collaboration with Amazon Web Services. The cloud-based platform, powered by Amazon Web Services and available at https://precision.heart.org, was founded on the FAIR principles (findable, accessible, interoperable, and reusable)[2] and includes secure collaboration areas (workspaces) and an open sharing area. The goals of the platform are to democratize data, to make it easy to search across orthogonal data sets, to provide a secure workspace to leverage the power of cloud computing, and to provide a forum for users to share insights. Multiple learning tools are available, including video tutorials, templates using open interactive programming framework, and a forum for interaction among community members.[3]

Keywords

cloud computing; machine learning; risk factors; statistics [publication type]

DATA HARMONIZATION AND SEARCHING

When accessing large public data sets today, researchers have to find, access, download, and interpret each data set individually. Researchers must untangle and interpret the data and then expend resources to house the data. In addition, the lack of harmonization across multiple data sets obviates the ability of researchers to combine data sources and to confirm or generate cogent findings.

The platform aims to address these challenges through a transparent and explicit harmonization approach: identifying common parameters across all data sets and thus allowing users to interactively find or merge data of interest. Data harmonization on the platform is transparent, providing insight into how each data variable is defined. Filters and

Correspondence: Jennifer L. Hall, PhD, 7272 Greenville Avenue, Dallas, TX 75231. jennifer.hall@heart.org.

multiple graphics display the distributions of important clinical outcomes and risk factors and provide summaries of available data sets, allowing users to quickly identify a subcohort that meets their research needs. Users have access to both harmonized and raw data.

EDUCATIONAL OPPORTUNITIES

The biomedical field has evolved to become data intensive, but only a small percentage of scientists or health professionals have learned the computational methods to solve biomedical problems. Even for the technologically inclined, it is time-consuming to continually learn and apply new software. As a foundation for leading new discoveries and reproducible analyses, the platform provides tutorials that guide researchers through data analyses such as genome-wide association, population demographics, descriptive statistics, and deep learning.[4] These tutorials explain the structure of the data involved and necessary data preprocessing and provide references and explanations from biomedical, epidemiological, and data science perspectives, along with instructions in natural language for the computer code at each step.

LEVERAGING THE ANALYTICAL CAPABILITIES OF THE PLATFORM

A FAIR model for sharing data and analytics is required for high-quality digital publications that facilitate and simplify the process of original discovery, evaluation, and reproducibility.[5] To support this model, we elected to deploy our platform in a cloud-based environment for its economic advantages and to support a growing ecosystem of collaborators and artifacts. This is also important for our user community when dealing with data sets that require large storage and heavy computation such as is seen in genomic data management or deep learning predictive analytics. The technology stack uses mostly open source components that take advantage of the large storage and scalable computational environment of Amazon Web Services. These components include Apache Spark 2.0 for rapid in-memory processing of results, Elasticsearch and Kibana for indexing and visualizing information, and an open and interactive programming framework, Jupyter Notebooks. Researchers and software developers can take advantage of Jupyter Notebooks, which supports common and open programming languages we made available on the platform such as R, Python, and Scala for creating and sharing interactive documentation.

POWERING THE SEARCH, LEARN, DISCOVER CYCLE

To ensure transparency, reproducibility, and the continued evolution of scientific findings, the platform provides a forum for users to ask questions about the data available or analysis plans and interpretation. Users can learn from each other, publish data they bring to the platform, and share analyses they have completed. Our hope is for the platform to ultimately support community peer review of tutorials and published analyses, similar to bioRxiv.

A FEW CHALLENGES TO OVERCOME

A number of challenges had to be overcome to ensure platform scalability (ie, allowing more people to use the application), security, privacy, and ease of use. For example, challenges of using cloud computing for genomic data include lengthy data transfers for uploading data to

the cloud server. However, a portion of genomics data today exist in cloud-based environments, making this challenge feasible. A second challenge has been the community's perceived lack of information safety in cloud computing. We configured the platform's cloud computing infrastructure, computation, and software with diverse security, confidentiality, and authentication settings to adhere to widely adopted national and international standards and regulations. In addition, a third-party assessor deemed the environment to be compatible with the US Health Insurance Portability and Accountability Act–amenable services for analyzing biomedical data in a cloud-based environment. Furthermore, we encrypt data while they are being transferred to or stored on the platform. The road map for the platform also includes targeting a more rigorous authorization for exchanging information with US federal organizations. Another challenge was balancing the protection of intellectual property with enabling collaboration. To overcome this, the community search portal allows only summary-level views of results; detailed views are available only to those who own or have been granted access to data in private workspaces. Ultimately, the platform will be only as good as the researchers make it. Users can demonstrate the effectiveness of their methods, test the capabilities of the platform, learn from other experts, share data and lessons learned, participate in competitions, and generate new insights that will serve as a comprehensive source of information about cardiovascular diseases and stroke. The platform provides an opportunity to learn, search, and discover in new and efficient ways, and we will keep working with the community to weave in new diverse data to help us drill more deeply and enrich our understanding.

CONCLUSIONS

The American Heart Association Precision Medicine Platform provides a secure and collaborative cloud-based environment for FAIR data sharing, analysis, and tutorials accessible to a wide range of researchers and clinicians. The goal of the platform is to realize precision cardiovascular and stroke medicine through active community collaboration to accelerate discovery, training, evaluation, reusability, and scalability and to foster global innovation.

Acknowledgments

REFERENCES

1. Houser SR. The American Heart Association's new Institute for Precision Cardiovascular Medicine. Circulation. 2016;134:1913–1914. doi: 10.1161/CIRCULATIONAHA.116.022138. [PubMed: 27881558]

2. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, t Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater

T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. doi: 10.1038/sdata.2016.18. [PubMed: 26978244]

3. Stevens SM. AHA precision medicine platform tutorials, demographics and descriptives. 2017 https://s3.amazonaws.com/aha-pmp-researcher-publishbucket-172825994212/Demographics_Descriptives_Tutorial.html. Accessed November 6, 2017.

4. Deo RC. Machine learning in medicine. Circulation. 2015;132:1920– 1930. doi: 10.1161/CIRCULATIONAHA.115.001593. [PubMed: 26572668]

5. Boulton G Reproducibility: international accord on open data. Nature. 2016;530:281. doi: 10.1038/530281c.