Review Article

Open Access   CrossMark

Challenges to implementation of global translational collaboration platforms

Abstract

Translational Collaboration Platforms connect clinical, genomics, and patient-reported data for the advancement of biomedical research, providing an opportunity to speed up the translating of basic science findings into clinical applications and new medicines. These platforms bring together data from both clinical and research databases and provide opportunities for multi-disciplinary research. Recent years have seen a significant growth of these platforms and some global collaborations research networks have been established using these platforms. In this brief summary of these platforms, we examine the challenges in implementation for global international research collaborations and challenges for the sustainability of research networks.

Keywords: translational platforms, collaboration, global health, big data, genomics, clinical informatics, international research

Yuri Quintana

Global Health Informatics, Beth Israel Deaconess Medical Center, USA

Correspondence: Yuri Quintana, Ph.D., Director, Global Health Informatics, Division of Clinical Informatics, Beth Israel Deaconess Medical Center, Assistant Professor of Medicine, Harvard Medical School, Harvard University, Boston, Massachusetts, USA, Tel (617) 2788162, Email yquintan@bidmc.harvard.edu

Introduction

Translational Collaboration Platforms[1–3] provide opportunities to connect clinical, genomics, and patient-reported data that can be analyzed for the advancement of biomedical research. These platforms provide the opportunity for clinical researchers, basic science researchers, and data scientists to combine data sets to facilitate hypothesis generation and advanced multidisciplinary research studies. The rapid growth of data generated by electronic medical records, advanced diagnostics, and genomic sequencing has created a big data revolution in life sciences. New data research platforms provide an opportunity to speed up the translating of basic science findings into clinical applications, drug discovery, and new treatment protocols such as personalized medicine. In recent years, there has been a significant growth of platforms for translational research including caBig, caGrid i2b2, TranSMART, cBioPortal, BRISK, iDASH, iCOD, and G-DOC. In this brief summary of these platforms, we examine the challenges in their implementation for global international research collaborations.

Major platforms

Launched in 2004, caBig[4–6] was an infrastructure developed the US National Institutes of Health to integrate information technology and cancer data for multi-institutional data sharing and biomedical research. The original mission of caBIG® was to develop a collaborative information network that accelerates the discovery of new approaches for the detection, diagnosis, treatment, and prevention of cancer. The goals of caBIG® were to:

1. Connect scientists and practitioners through a shareable and interoperable infrastructure,

2. Develop standard rules and a common language to share information more easily, and

3. Build or adapt tools for collecting, analyzing, integrating, and disseminating information associated with cancer research and care.

In 2011, an NIH study[7] reported some of the problems with the caBig program. In May 2012, the program ended[8] and the National Cancer Informatics Program (NCIP) created caGrid as its successor.[9] Launched in 2007, the Informatics for Integrating Biology and the Bedside (i2b2)[10,11] infrastructure is based at Partners HealthCare System in Boston, Massachusetts, and is funded by United States National Institute of Health (NIH). The project is open source and has been adopted by numerous academic hospitals around the world for biomedical research. The system can store patient medications and laboratory values, and these can be combined with clinical research data, such as information from a case report form or genomic data, into a single cohesive unit that can be queried in an integrated manner. The i2b2 system differs from caBIG in that the core data in i2b2 is instantiated according to a single relational model, not a compendium of object models.[12] The i2b2 system has been used to set up the Shared Research Informatics Network (SHRINE) that can distribute i2b2 queries to data from several Harvard hospitals, particularly the Beth Israel Deaconess Medical Center, the Dana-Farber Cancer Institute, and Children's Hospital Boston.[13] Based on i2b2 architecture, the tranSMART platform[14–16] is a set of data models, shared data sets, data transformation utilities, and analytical web applications that accelerate discoveries within complex biological systems by creating a standardized and semantically integrated database of research results linked to reusable and scalable self-service analytics. TranSMART was initially funded by Johnson & Johnson Corporation and is now funded by the TranSMART Foundation as public-private cooperation.[17] Similarly, several European stakeholders have sponsored eTRICKS[18] for European life sciences research collaborations.

The cBioPortal for Cancer Genomics is an open-source platform[19] based at Memorial Sloan-Kettering Cancer Center, New York, funded by NIH grants and industry support. The goal is to provide translational researchers access to data sets generated by large-scale cancer genomics projects, such as the Cancer Genome Atlas (http://cancergenome.nih.gov) and the International Cancer Genome Consortium (http://icgc.org). The system has visualization and analysis tools and export functionalities. The public version contains large cancer genomics data sets. The system can also be privately installed and allows researchers to upload their data sets. The Biology-Related Information Storage Kit (BRISK)[20] is based at the University of British Columbia, Vancouver, Canada, and is funded by a partnership between private and private sources. It is a web-based

platform initially developed for researchers in the AllerGen (The Allergy, Genes and Environment Network) consortium (http://www. allergen-nce.ca). The Integrating Data for Analysis, Anonymization, and sharing (iDASH) platform[21] is based in San Diego, California, and is funded by NIH grants. The platform is a powerful high performance-computing platform for data integration for biomedical and behavioral researchers. It is focused on sharing data with privacy-preserving methods.

The integrated clinical omics database (iCOD)[22] is based at the Tokyo Medical Dental University, Japan, and is publically funded. The system can combine comprehensive clinical, pathological, and molecular information about patients. The system can show the interrelation of clinical and omics data for the discovery of plausible disease pathways. Georgetown Database of Cancer (G-DOC)[23] is based at Georgetown University, Washington, DC, and is funded by the US government's Health and Human Services agency. The system integrates patient demographics, structured clinical research data, and clinical outcomes data with high-throughput omics data (DNA, mRNA, microRNA, and metabolites). Launched in 2003, The Pediatric Oncology Network Database, (www.pond4kids.org)[24] is a secure, web-based, multilingual pediatric hematology/oncology database created for use in countries with limited resources to meet various clinical data management needs including cancer registration, delivery of protocol-based care, outcome evaluation, and assessment of psychosocial support programs. Established as a part of the International Outreach Program at St. Jude Children's Research Hospital in Memphis, Tennessee, USA. POND4Kids serves as a tool for oncology units to store patient data for easy retrieval and analysis and to achieve uniform data collection to facilitate meaningful comparison of information among international centers.

Discussion

There are several challenges to establishing and sustainably operating collaborative translational research platforms, particularly for centers that do not have extensive resources for data collection and management.

I. Technical Data Integration-The growing volume and complexity of data in biological data sets require more complex architectures to integrate data from diverse data sets. Data from different generations of lab and sequencing hardware make integration difficult because of different data formats and granularity. The process of uploading data is complicated and requires sustainable resources.

II. Data Quality-Data quality assurance remains a large problem for data that are collected from diverse institutions. Each institution may have different levels of capacity to review their data quality. The ability to track the level of review of data remains a problem. In some systems there are no detailed mechanisms to tag data (down to the individual data item) as to the level of certainty.

III. Data Sharing-Data sharing agreements must continue to evolve to manage the impact of ongoing changes in government regulations and evolving corporate compliance needs. This requires substantial dedicated efforts from various institutional departments (technical, legal, clinical, research, management) to review changes to agreements.

IV. Liability-Data breaches continue to be a growing problem for any online platform. This issue requires dedicated expert technical staff to manage access and legal agreements to

delineate the liability among collaborating partners. The problem becomes more complicated with the addition of international countries that have different laws and penalties for breaches.

V. Privacy-The increasing complexity of privacy laws requires changes to software to accommodate the tracking of consents for data and compliance with local, national, and international privacy laws pertaining to the data sources.

VI. Discovery-Novel discoveries from shared data are among the key objectives of these networks. Intellectual property agreements need to be established in advance to handle these opportunities, and the agreements are subject to change as institutions are merged, sold, or reorganized.

VII. Funding Sustainability-Sustainable funding models are unclear from the current emerging collaboration networks. Government research grants and/or industry funding initially fund most networks. Funding from governments continues to be strained. Government funding for any project will usually end once the proof of concept has been published. For industry-sponsored projects, industry will want to see a return on the investment. For industry, it is difficult to measure the return for a shared data network because of the length of time it takes to see outcomes that can be monetized in a commercial application.

Conclusion

Translational collaboration platforms have been successfully developed to support life science research with diverse types of data and from multiple centers. Among the challenges include data integration, quality, sharing models and policies and procedures to manage privacy, liability, and intellectual property. Despite the many challenges to the implementation of these platforms, there are some emerging networks for multi-national collaborations. Models for sustainability of these networks will need to be developed for these platforms and research networks to continue past the initial implementation phase. Careful planning with multiple stakeholders will be needed to create platforms that meet the needs of both clinical and life sciences researchers, and create sustainable research networks and funding models.

Acknowledgements

Conflict of interest

The author declares no conflict of interest.

References

1. Groves P, Basel K, Knott D, et al. *The 'big data revolution in healthcare: Accelerating value and innovation*. McKinsey & Company; 2013.

2. Biesecker LG, Burke W, Kohane I, et al. Next–generation sequencing in the clinic: are we ready? *Nat Rev Genet*. 2012;13(11):818–824.

3. Canuel V, Rance B, Avillach P, et al. Translational research platforms integrating clinical and omics data: a review of publicly available solutions. *Brief Bioinform*. 2014;16(2):280–290.

4. National Cancer Institute. *caBIG – Cancer Biomedical Informatics Grid*. 2015.

5. Saltz J, Oster S, Hastings S, et al. CaGrid: design and implementation of the core architecture of the cancer biomedical informatics grid. *Bioinformatics*. 2006;22(15):1910–1916.

6. McConnell P, Dash RC, Chilukuri R, et al. The cancer translational research informatics platform. *BMC Med Inform Decis Mak*. 2008;8:60.

7. National Institutes of Health. *An Assessment of the impact of the NCI Cancer Biomedical Informatics Grid (caBIG®)*. Report of the Board of Scientific Advisors Ad Hoc Working Group, Cancer Biomedical Informatics Grid (caBIG®). Program National Cancer Institute; 2011.

8. George A K. *"Program Announcement"*. National Cancer Institute; 2015.

9. National Cancer Institute. Center for Biomedical Informatics and Information Technology (CBIIT). *Informatics for Integrating Biology and the Bedside*. 2015.

10. i2b2: Informatics for Integrating Biology and the Bedside.

11. Murphy SN, Mendis M, Hackett K, et al. Architecture of the opensource clinical research chart from Informatics for Integrating Biology and the Bedside. *AMIA Annu Symp Proc*. 2007:548–552.

12. Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2. *J Am Med Inform Assoc*. 2010;17(2):124–130.

13. Weber GM, Murphy SN, McMurry AJ, et al. The shared health research information network (SHRINE):a prototype federated query tool for clinical data repositories. *J Am Med Inform Assoc*. 2009;16(5):624–630.

14. Perakslis ED, Van Dam J, Szalma S. How informatics can potentiate precompetitive open–source collaboration to jump–start drug discovery and development. *Clin Pharmacol Ther*. 2010;87(5):614–616.

15. Szalma S, Koka V, Khasanova T, et al. Effective knowledge management in translational medicine. *J Transl Med*. 2010;8:68.

16. Athey BD, Braxenthaler M, Haas M, et al. TranSMART: An Open Source and Community–Driven Informatics and Data Sharing Platform for Clinical and Translational Research. *AMIA Jt Summits Transl Sci Proc*. 2013:6–8.

17. TranSMART Foundation.

18. eTRICKS Consortium.

19. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. 2012;2(5):401–404.

20. Tan A, Tripp B, Daley D. BRISK–research–oriented storage kit for biology–related data. *Bioinformatics*. 2011;27(17):2422–2425.

21. Ohno–Machado L, Bafna V, Boxwala AA, et al. iDASH: integrating data for analysis, anonymization, and sharing. *J Am Med Inform Assoc*. 2012;219(2):196–201.

22. Shimokawa K, Mogushi K, Shoji S, et al. iCOD: an integrated clinical omics database based on the systemspathology view of disease. *BMC Genomics*. 2010;11(Suppl 4):S19.

23. Madhavan S, Gusev Y, Harris M, et al. G–DOC: a systems medicine platform for personalized oncology. *Neoplasia*. 2011;13(9):771–783.

24. Quintana Y, Patel AN, Arreola M, et al. POND4Kids: a global web–based database for pediatric hematology and oncology outcome evaluation and collaboration. *Stud Health Technol Inform*. 2013;183:251–256.