Cell

Data Harmonization for a Molecularly Driven Health System

Jerry Ssu-Hsien Lee,[1] Warren Alden Kibbe,[2,*] and Robert Lee Grossman[3]
[1]Department of Medicine/Oncology, Keck School of Medicine; Department of Chemical Engineering and Material Science, Viterbi School of Engineering; and Lawrence J. Ellison Institute for Transformative Medicine, University of Southern California, Los Angeles, CA, USA
[2]Duke Cancer Institute and Duke School of Medicine, Duke University, Durham, NC, USA
[3]Center for Translational Data Science, University of Chicago, Chicago, IL, USA
*Correspondence: warren.kibbe@duke.edu

https://doi.org/10.1016/j.cell.2018.08.012

Data commons have emerged as the best current method for enabling data aggregation across multiple projects and multiple data sources. Good data harmonization techniques are critical to maintain quality of data within a data commons, as well as to allow future meta-analysis across different data commons. We present some of the current best practices for data harmonization.

INTRODUCTION

The power of a learning health system (see the National Academies 2011 report at http://www.nationalacademies.org/hmd/Activities/Quality/~/media/Files/Activity%20Files/Quality/VSRT/Core%20Documents/ForEDistrib.pdf) grows as more organizations are willing to share their data. Humans have always been fascinated by natural events and attempt to advance our collective understanding through recording and sharing of observations and measurements. The value of aggregating data to develop knowledge, utilizing the knowledge to generate insight, and applying the insight to create impact is a common theme that repeats itself from documenting movements of stars in the sky to changes in body temperature caused by illness. Meaningful data sharing, data reuse, and knowledge extraction from experimental data have always been critical components of science.

In the past decade, improvements in our ability to generate data, particularly digital data, are having a profound impact on the way we collaborate and share knowledge. However, when biomedical data are produced by different groups, with different data models, different standards, and processed in different ways, it can be extremely challenging to draw meaningful conclusions from the aggregated data. While it is satisfying to use aggregated data to substantiate an existing hypothesis, disruption and new insights often emerge when carefully collected data contradict or expose limitations of existing models and dogma.

Our involvement in the NCI Genomic Data Commons (GDC), Blood Profiling Atlas Data Commons (BloodPAC), and the NCI Proteomics Data Commons has informed our perspective on data harmonization and its importance in enabling information and knowledge extraction from data generated by multiple groups and across projects. In this Commentary, we discuss the critical role of data harmonization and how data commons can support data harmonization at the scale required to build a national LHS. Data harmonization will play an even greater role as multiple data commons are brought together as components of a national data ecosystem.

What Is Data Harmonization?

For the purposes here, we view ‘‘data harmonization’’ as the process of bringing together data from multiples sources and ensuring uniform and consistent processes, including, but not limited to: (1) determining what data to accept; (2) cleaning and applying quality control metrics to the accepted data; (3) mapping the data to a common data model; (4) processing the data with common bioinformatics pipelines; and (5) applying common quality control metrics to the processed data (Figure 1).

At the beginning of the human cancer genome project (HCGP, later referred to as The Cancer Genome Atlas Project, or simply TCGA), the importance of data harmonization was not as widely appreciated as the largest pilot study for TCGA at the time characterized just 22 samples from colorectal and breast cancer patients (Sjöblom et. al., 2006). Some of the early analysis revealed candidate bioinformatic signatures that were due to batch effects from differences in sample handling, sample preparation, and other parts of the production pipelines across TCGA, not the underlying biology of the patient samples. Incorrect inferences and improper conclusions can result if these types of batch effects are not systematically identified and addressed as was done in TCGA. This is especially important in a learning health system, where many of the effects of interest may be quite small, while the batch effects (systematic differences in practice, billing procedures, documentation, patient population to name a few) are potentially much larger.

Reducing sources of heterogeneity that lead to batch effects when scaling up processes is well understood in engineering but was a relatively new concept to cancer research. A lot of time in TCGA was spent in standardizing data elements, data definitions, sample handling, and data generation protocols so that the results of assays from different groups and disease sites would be directly comparable.

Below, we review some of the techniques and technologies that have been developed to make data harmonization *repeatable*, so that the same organization can re-process the same data at some later time and get the same harmonized data, and *reproducible*, so that different organizations using different software

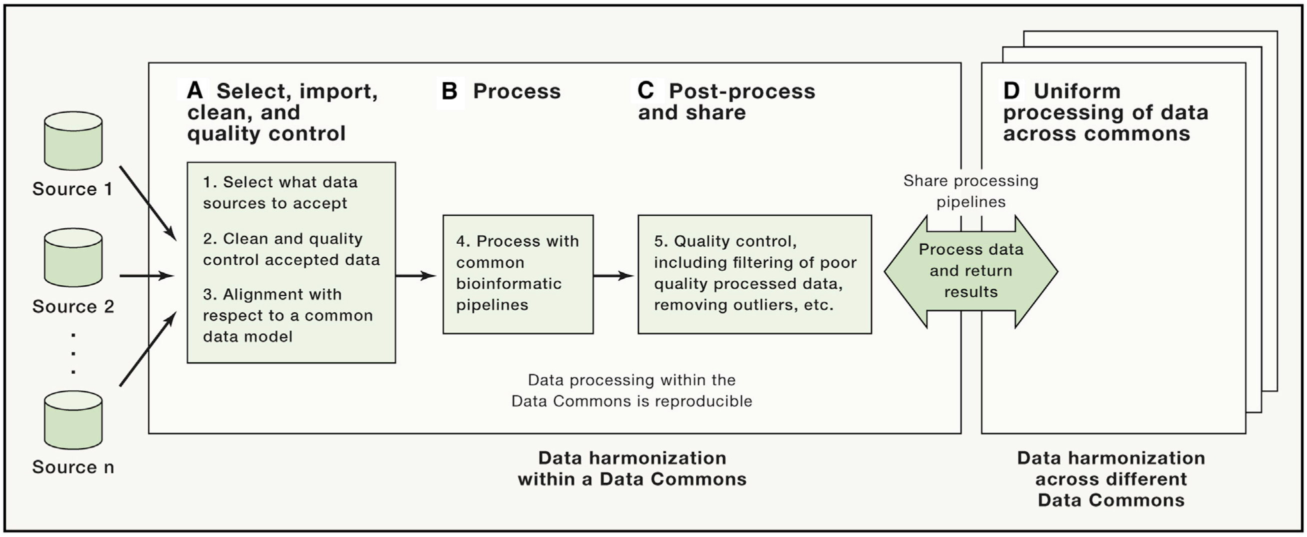
A Select, import, clean, and quality control
B Process
C Post-process and share
D Uniform processing of data across commons
Source 1
Source 2
Source n
1. Select what data sources to accept
2. Clean and quality control accepted data
3. Alignment with respect to a common data model
4. Process with common bioinformatic pipelines
5. Quality control, including filtering of poor quality processed data, removing outliers, etc.
Share processing pipelines
Process data and return results
Data processing within the Data Commons is reproducible
Data harmonization within a Data Commons
Data harmonization across different Data Commons

Figure 1. An Overview of the Data Harmonization Process

The main steps involved with data harmonization within a data commons and between data commons are depicted. It is helpful to divide data harmonization within a data commons into three phases. Phase A includes data cleaning, quality control, and structuring and curating the data with respect to a data model. Phase B uses bioinformatics pipelines to process the data. In phase C, it is common to apply additional quality control to a project's data as whole prior to release. The pipelines in phases A–C can be packaged using emerging standards, such as the Common Workflow Language (CWL), and applied to data in other data commons. The data processed uniformly in this way across data commons can then be combined in what can be thought of as cross-commons data harmonization.

platforms can reprocess the same data and get the same harmonized data.

Components of Data Harmonization

Metadata Standards

It all starts with metadata. Metadata are, literally, data about data. For a data commons, it includes information about instrumentation, what data types are available, experimental procedures, and relationships between data elements. For the GDC, each data element includes references to vocabulary and terminology available in the NCI Metathesaurus (https://ncim.nci.nih.gov/ncimbrowser/) and the NCI Enterprise Vocabulary System (EVS) (https://evs.nci.nih.gov). The entire NCI thesaurus is downloadable as an OWL file, with all the rich relationships available for use (https://cbiit.cancer.gov/evs-download/thesaurus-downloads/). One specific example is breast carcinoma (CUI C0678222). There are more than 40 synonyms for breast cancer listed (https://ncim-stage.nci.nih.gov/ncimbrowser/pages/concept_details.jsf?dictionary=NCI%20Metathesaurus&code=C0678222&type=synonym) and many relationships between this concept and either more specific cancers (children) or less specific cancers (parents) (https://ncim-stage.nci.nih.gov/ncimbrowser/pages/concept_details.jsf?dictionary=NCI%20Metathesaurus&code=C0678222&type=relationship). By specifying metadata and referencing resources like EVS, values that are submitted to the data commons can be reviewed for conformance to the element specification. This in turn allows for automated ingestion of data and the scoring of conformance, one measure of data quality. This is one aspect of data harmonization.

Globally Unique Identifiers, Digital Object Identifiers, and Immutability

A critical underlying infrastructure for a data commons is the ability to create an arbitrarily complex collection of data, processes, and outputs and identify it with a globally unique identifier (GUID). That identifier can include complete provenance, ideally that conform to the FDA guidance on electronic signatures and data (see Code of Federal Regulations Title 21 Part 11). Furthermore, a GUID can also be assigned a digital object identifier, or DOI. Once published externally, the DOI needs to always point to the same collection of objects—it needs to be immutable. This is critical to enable reuse, to build trust in the system and the infrastructure, and for publishing results, inferences, visualizations, algorithms, and pipelines. In the GDC, all the publicly available data are assigned immutable GUIDs, and selected collections of data will also be assigned DOIs.

Data Processing Standards

Processing the data submitted to data commons such as the GDC is complex. Broadly speaking, processing steps can be classified as conformance checks, cleaning, quality control (QC), insertion into a data model, standardized processing of primary data, assembly, or creating an ensemble of data. The creation of an ensemble itself may spawn another series of tasks that may also include conformance, cleaning, QC, insertion into an ensemble model, standardized processing of the ensemble, and creating an ensemble of ensembles, potentially triggering yet another layer of tasks. These are usually *pipelines*, and the complexity of the pipelines is limited only by creativity, data size, and processing power. For genomic data, these processes are now very well understood and characterized and have become acceptable by the cancer genomics research community. It is becoming common to use emerging standards such as the Common Workflow Language (CWL) to describe these (Amstutz et al., 2016). For instance, BAM files from whole-genome sequencing (WGS), whole-exome sequencing (WES), RNA-seq, and MethylSeq among other sequencing techniques will spawn a series of processes that import the BAM files into a consistent format, run a series of cleaning and data quality checks, align the data against a reference standard as specified by that submission's metadata, align it to the current reference genome, and generate a series of statistics that provide measures of quality and consistency. Agreeing to a standard series of pipelines for processing data and the expression of these pipelines in a language such as CWL is an important practical step in the harmonization of data. The use of data processing standards and workflow languages such as CWL also increases the reproducibility of the data processing and the reuse of the pipeline across projects and across computing environments.

Comparability across Data Submissions

The ability to analyze data across multiple submissions, multiple projects, and multiple experimental designs is limited not only by the data processing standards, but also by the consistency and completeness of the data annotations and metadata for each data submission. How samples were handled (for instance, storage conditions including warm and cold ischemic time prior to sample preparation) is crucial for reducing batch effects, understanding the potential confounding factors like enzymatic degradation, internal controls, instrument calibration, etc. The failure to accurately record and include these factors limits data reuse (and impact both comparability and reproducibility).

Data Re-analysis

For a data commons like the GDC, as new reference genomes are released and new algorithms and analysis methods are developed, it is critical that the algorithms and references are updated in a managed process, and then all the primary data that is available in the GDC is rerun through the new pipelines, and the results of the new and old pipelines are compared, analyzed, and visualized to characterize and expose the impact of changes in any step of the processing pipeline. Making these processes scalable, repeatable, and robust is an entire discipline unto itself and is critical for reproducible science.

Understanding Cancer

The GDC was designed to incorporate existing well-curated, well-characterized data in cancer. The GDC includes clinical data, imaging, outcomes, and of course, genomic data. Up until the GDC, the primary large genomic dataset was TCGA and TARGET for pediatric cancers. Now, many groups are contributing additional datasets to the GDC, including AACR Project GENIE, Foundation Medicine, and the MMRF CoMMpass Study among others. As information about patient therapies, changes in the tumor genome in response to therapy, and patient response and outcomes data start to flow into the GDC, whole new avenues for understanding cancer biology become possible. Understanding the complex relationships and dynamics of a patient's germline DNA and its contribution to risk, initiation, progression, response to treatment (pharmacogenomics), the tumor microenvironment, and the evolving tumor genome will only be possible with well-characterized, highly curated, consistently analyzed data.

Visualization

Good visualization tools, like cBioPortal, transform our ability as humans to comprehend complex datasets, explore data, make hypotheses, and even test those hypotheses in systems like the GDC. Having well-described metadata and consistent data harmonization practices is critical for providing interpretable analytics and visualization over large datasets. For instance, AACR Project GENIE data from roughly 39,000 patient tumors is now available through cBioPortal at http://www.cbioportal.org/genie/study?id=genie_public#summary. Registration and agreement to terms is required for access (Cerami et al., 2012, Gao et al., 2013, AACR Project GENIE Consortium, 2017).

The Importance of Data Harmonization

One of the well-known challenges of cancer genomics is bridging the gap between molecular signatures and clinical applications. They are related, but different, problems. First, cancer genomic datasets must be large enough to have the statistical power required to model the impact of genomic alterations on clinical outcomes. Second, many molecular signatures are not robust, with the same algorithm on different datasets generating different signatures and with different algorithms with the same target generating different signatures on the same dataset.

The role of harmonization is critical here. If data are harmonized appropriately, then the same analysis on data in two or more different commons can be combined to provide larger datasets with more statistical power. Alternatively, the upstream data, prior to certain steps in data harmonization, can be analyzed with new algorithms that may have not been available when the data were originally harmonized.

It is important to note that validated molecular signatures are robust—the presence of the signature is highly correlated with a clinical outcome. Examples include the presence of BCR-ABL fusion in chronic myeloid leukemia or overall alterations in TP53. The BCR-ABL example is a *gain-of-function* mutation, since this fusion creates a novel tyrosine kinase. Identifying gain-of-function mutations can result in rapid approval of new therapies, as demonstrated by the approval of crizotinib for small population of EML4-ALK fusions in non-small-cell lung cancer patients (Kazandjian et. al., 2014). In the case of genomic measurements, the harmonization strategy deployed by the NCI GDC will allow new prospective datasets that are added to the GDC to be appropriately analyzed with retrospective TCGA datasets.

The NCI GDC has an implemented, portable (using CWL and Docker) data harmonization pipeline for somatic sequencing data. However, other molecular measurements (e.g., proteomics, metabolomics) and phenotypic measurements (e.g., medical imaging or clinical outcomes) will need different processing and harmonizing steps. There is still much to be done, as shown recently by the NCI Clinical Proteomics Tumor Analysis Consortium (CPTAC), where signatures generated using only genomics were augmented with proteomics for only a subset of TCGA patients that allowed unique proteogenomic signatures to be developed (Rodriguez and Pennington, 2018). Such meta-analysis across data commons are still conceptual, as the field is still quite nascent, but given new efforts such as the Blood Profiling Atlas Data Commons (molecular measurements on cancer blood samples; Grossman et. al., 2017) and other still-to-be fully defined data commons, it is critical that all data types and datasets submitted into data repositories have rigorous, well-defined, well-implemented data harmonization strategies. Data and process harmonization across these resources is critical in determining cancer properties across aggregated multi-analyte and multi-scale datasets.

Where Is Data Harmonization Going?

Regulations can make it challenging to move data into a commons, especially if the commons is in another country or region. There is nothing to keep the same data harmonization from being applied to data in geographically distributed commons. In fact, with the containerization of bioinformatics pipelines [O'Connor et al., 2017, da Veiga Leprevost et al., 2017, Tatlow and Piccolo 2016], it is becoming easier to apply the same pipelines to data in different systems. Software containers are a technology to isolate software from its surrounding environment and encapsulate it so that it run in other environments by other users. A good example is provided by the PCAWG (pan-cancer analysis of whole genome) studies that were recently completed (http://docs.icgc.org/pcawg/). With this approach, data harmonization and analysis of distributed data may emerge as an important technique to complement meta-analysis, which is often challenging due to the lack of rigorous data harmonization.

A Platform of Evidence

The ability of a data commons to provide a platform for making carefully curated and well characterized patient-level data available for the identification and publication of evidence of the factors necessary to obtain therapeutic efficacy is one of the end goals for the Research Data Ecosystem as envisioned and described by the Cancer Moonshot Blue Ribbon Panel working group on Enhanced Data Sharing (https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative/blue-ribbon-panel/enhanced-data-sharing-working-group-report.pdf). During the years following the establishment of TCGA and the subsequent move toward precision medicine, it is clear that cancer is no longer a set of anatomically defined diseases, but rather a collection of etiologically and molecularly distinct diseases, making even common cancers more akin to rare disease. The ability to identify effective therapies and avoid toxicities and ineffective therapy is the promise of precision medicine, and in the realm of increasingly stratified cancers data commons defining the Research Data Ecosystem is the current best paved road to assembling the evidence necessary to learn from the biology, clinical presentation, experience and outcomes of every cancer patient.

Conclusion

As we continue to aggregate data from disparate cohort studies, it will be difficult, if not impossible, for data scientists and biomedical researchers to characterize signals and signatures as biologically relevant unless appropriate data harmonization strategies have been employed. This is in part because the signal can be very rare. A recent example is looking for PDL1 amplification in solid tumors. A dataset of 118,187 patients characterized by next-generation sequencing identified only 843 patients with amplified PDL1 (Goodman et al., 2018), or 0.7% of the patients. As more large-scale data commons are built, we must ensure the application of best practices in data harmonization for data submission and processing. This will improve the usability, interoperability, and quality of the data available in a data commons. It will also simplify the process of cross-comons data harmonization, and enable discovery and analysis across the cancer data ecosystem. This in turn will enable new insights into rare events that can only be revealed by combining multiple studies. Finally, data harmonization practices are critical for advancing molecularly driven cancer medicine in a learning health system and translating those advances for patient benefit.

ACKNOWLEDGMENTS

This project has been funded in part with Federal funds from the National Cancer Institute, National Institutes of Health, Task Order No. 17X053 under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does the mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. We would like to thank David Agus, Patrick Gilmore, Norbert Tavares, and Sarah Elder for their inputs in preparing this paper.

REFERENCES

AACR Project GENIE Consortium (2017). AACR Project GENIE: Powering Precision Medicine through an International Consortium. Cancer Discov. *7*, 818–831.

Amstutz, P., Crusoe, M.R., Tijanić, N., Chapman, B., Chilton, J., Heuer, M., Kartashov, A., Leehr, D., Ménager, H., Nedeljkovich, M., et al. (2016): Common Workflow Language, v1.0. figshare. https://doi.org/10.6084/m9.figshare.3115156.v2

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. *2*, 401–404.

da Veiga Leprevost, F., Grüning, B.A., Alves Aflitos, S., Röst, H.L., Uszkoreit, J., Barsnes, H., Vaudel, M., Moreno, P., Gatto, L., Weber, J., et al. (2017). BioContainers: an open-source and community-driven framework for software standardization. Bioinformatics *33*, 2580–2582.

Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. *6*, pl1.

Goodman, A.M., Piccioni, D., Kato, S., Boichard, A., Wang, H.Y., Frampton, G., Lippman, S.M., Connelly, C., Fabrizio, D., Miller, V., et al. (2018). Prevalence of PDL1 Amplification and Preliminary Response to Immune Checkpoint Blockade in Solid Tumors. JAMA Oncol. Published online June 14, 2018. https://doi.org/10.1001/jamaoncol.2018.1701.

Grossman, R.L., Abel, B., Angiuoli, S., Barrett, J.C., Bassett, D., Bramlett, K., Blumenthal, G.M., Carlsson, A., Cortese, R., DiGiovanna, J., et al. (2017). Collaborating to Compete: Blood Profiling Atlas in Cancer (BloodPAC) Consortium. Clin. Pharmacol. Ther. *101*, 589–592.

Kazandjian, D., Blumenthal, G.M., Chen, H.-Y., He, K., Patel, M., Justice, R., Keegan, P., and Pazdur, R. (2014). FDA approval summary: crizotinib for the treatment of metastatic non-small cell lung cancer with anaplastic lymphoma kinase rearrangements. Oncologist *19*, e5–e11.

O'Connor, B.D., Yuen, D., Chung, V., Duncan, A.G., Liu, X.K., Patricia, J., Paten, B., Stein, L., and Ferretti, V. (2017). The Dockstore: enabling modular, community-focused sharing of Docker-based genomics tools and workflows. F1000Res. *6*, 52.

Rodriguez, H., and Pennington, S.R. (2018). Revolutionizing Precision Oncology through Collaborative Proteogenomics and Data Sharing. Cell *173*, 535–539.

Sjöblom, T., Jones, S., Wood, L.D., Parsons, D.W., Lin, J., Barber, T.D., Mandelker, D., Leary, R.J., Ptak, J., Silliman, N., et al. (2006). The consensus coding sequences of human breast and colorectal cancers. Science *314*, 268–274.

Tatlow, P.J., and Piccolo, S.R. (2016). A cloud-based workflow to quantify transcript-expression levels in public cancer compendia. Sci. Rep. *6*, 39259.