HHS Public Access

Author manuscript

Curr Opin Pediatr. Author manuscript; available in PMC 2019 April 01.

Published in final edited form as:

Curr Opin Pediatr. 2018 April ; 30(2): 269–275. doi:10.1097/MOP.0000000000000602.

An ECHO Viewpoint of Data Analysis Centers for Collaborative Study Designs

Lisa P. Jacobson[1], **Bryan Lau**[1], **Diane Catellier**[2], and **Corette B. Parker**[2]

[1]Johns Hopkins University Bloomberg School of Public Health, Baltimore Maryland 21205

[2]RTI International, Research Triangle Park, North Carolina 27709

Abstract

Purpose of review—A highly complex collaborative study design that pools and extends existing studies, such as the Environmental influences on Child Health Outcomes (ECHO) Program, requires a Data Analysis Center (DAC) with resources and expertise to create a secure environment for housing and analyzing the shared data, harmonize and structure the shared data for different purposes, and apply appropriate and innovative designs and analytic methods. The DAC, in partnership with cohort investigators, must ensure that results from ECHO-wide cohort analyses are appropriately interpreted and reproducible.

Recent findings—Understanding the cohorts contributing to ECHO is critical for developing a collaborative environment and the methods to best analyze the data without bias. We further describe the development of the ECHO-wide cohort Metadata Catalog, the architecture of the ECHO-wide cohort data platform, and analytical approaches to facilitate early productivity.

Summary—The ECHO DAC has established a secure environment for the transfer and storage of ECHO cohort data and information, and initiated processes to promote productive collaborations. Understanding the ECHO DAC responsibilities and assets will help to overcome communication and trust challenges encountered in the initiation of this complex ECHO-wide cohort collaborative research study.

Keywords

data analysis center; collaborative study design; data platform; meta-data catalog; collective analysis

Introduction

Varying models of data centers exist; here we present the model that we are using for the Data Analysis Center (DAC) of the Environmental influences of Child Health Outcomes (ECHO) Program, and the motivations for choosing this model. The DAC's mission is to create a collaborative, secure environment with the appropriate methods for valid, reproducible scientific investigations to improve child health outcomes. These activities are

Corresponding Author: Lisa P. Jacobson, Johns Hopkins University Bloomberg School of Public Health, 615 N. Wolfe Street, Baltimore, Maryland 21205, Ljacobs1@jhu.edu, Phone: 410-502-9770.

Conflicts of interest: None

not accomplished in isolation, but rather in collaboration with the other ECHO investigators. In complex programs that leverage multiple existing study cohorts such as ECHO, success depends on an experienced and multidisciplinary DAC that can build effective relationships with scientists across the collaboration and provide methodologic expertise to appropriately handle varying study designs. An organized, documented and secure database facilitates the use of data administratively and analytically. Successful data centers establish the data infrastructure and operations to ensure validity of study findings, provide leadership in statistical analysis and study design, and contribute as scientific partners in evaluating research hypotheses. Expertise in data centers facilitates maximum extraction of information through the application or development of appropriate statistical analyses that can meet the challenges specific to the collaborative research study.

Here we present our initial methods used to meet some of the challenges encountered with starting the DAC. We describe the secure ECHO-wide cohort data platform, and approaches to data analysis to accelerate early research productivity. To facilitate implementation of new data processes, the DAC presents the broad approach to the ECHO Steering Committee, and provides documentation and hosts webinars for the relevant members of the larger ECHO community.

Data Center Models

The mission of a DAC for a multi-cohort collaboration is very different than that of a data repository, whose primary function is to warehouse and distribute data. Serving as a data repository does not require scientific involvement, i.e., providing data management or statistical guidance to the study investigators, and as such can be funded completely independently of the collaborative study. It functions as a library and is often the place where versions of study data are sent for public access; examples of data repositories include the National Technical Information Service (NTIS, https://www.ntis.gov/), dbGaP (https://www.ncbi.nlm.nih.gov/gap) for genotype and phenotype data and analyses, and BioLINCC (https://biolincc.nhlbi.nih.gov/home/) that includes the repository for data from studies funded by the National Heart Lung Blood Institute (NHLBI). A list of discipline-specific and general repositories is periodically updated by *Scientific Data* [1*].

In contrast, the data center for a multicenter study is responsible for receiving, processing, harmonizing, and analyzing the data, in addition to providing publically accessible data. In clinical trials, the data center establishes the data collection and transfer methods for use by the field sites, performs data quality assurance, including creating checks during the data entry process and on the batched data, and designs and conducts the statistical analyses. The Data Coordinating Center (DCC) model has the added responsibilities of overseeing protocol development and administration, and providing logistical support to the study [2]; NHLBI compiled a best practice checklist for successful DCCs [3]. Although clinical trial designs may be complex, each trial is established with few specified hypotheses and statistical methods to apply to a relatively small closed set of focused data elements.

Data centers are also integral in multicenter observational studies. Unlike clinical trials designed to only test a few specific hypotheses, productive long-running large multicenter cohort studies also establish a data platform for subsequent investigations and nesting

additional studies. Examples of such studies include the Atherosclerosis Risk in Communities Study (ARIC, https://www2.cscc.unc.edu/aric/desc) [4], the Multicenter AIDS Cohort Study (MACS, http://aidscohortstudy.org/) [5], Chronic Kidney Disease in Children (CKiD, https://statepi.jhsph.edu/ckid/) [6], and Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-Be (nuMoM2b, https://www.nichd.nih.gov/research/supported/Pages/nuMoM2b.aspx) [7]. Data center responsibilities in such studies typically include all those of the clinical trial DCC with the exception of the trial-specific activities such as randomization and treatment allocation. Also, often in a large multicenter observational study, committees or working groups of experts from across the centers with the DCC develop the study protocols. In this model, the DCC investigators have expertise in study coordination, data management, and statistics, and usually some knowledge in key content areas of interest. This content area knowledge with methodologic expertise bridges the communication between investigators across the centers and leads to the correct development and application of research methods.

The ECHO DAC model is a modified version of the observational study DCC, without coordinating and regulatory oversight responsibilities. In ECHO, these latter responsibilities are performed by the ECHO Coordinating Center. Successful achievement of the scientific aims of the ECHO-wide cohort necessitates familiarity and collaboration between disciplines. Housed in Johns Hopkins University and RTI International, the DAC capitalizes on the breadth of expertise available in the two institutions. The DAC scientific team comprises investigators with complementary multidisciplinary expertise. In addition to the epidemiologists, statisticians, analysts, data managers, and informatics technologists to build the computing and data systems, and develop analytical methods to deal with the expected complexities of the data, the DAC includes investigators with expertise in environmental exposures, geospatial sciences and technology, epigenetics, genetics, and pediatric outcomes, allowing the DAC to participate fully as a scientific partner. A good example for the utility of this expertise occurred with a recent “collective analysis,” a type of disseminated analysis (described below) whereby the DAC works with a writing team to develop the concept and analysis plan, and distributes the statistical code to the cohorts for implementation. A writing team recently wished to examine community-level parameters in an analysis. The DAC investigator on the writing team reached out to both an outcome expert at Johns Hopkins Hospital, Dr. Jeanne Sheffield, director of Fetal/Maternal Medicine and DAC co-investigator, for input on variables, and a geospatial scientist at RTI, William Wheaton, who is also the co-chair of the ECHO Geospatial Working Group for assistance with standardizing the geocoding and database linkage. A geocoding tool known as DeGauss (Decentralized Geomarker Assessment for Multi-Site Studies) was identified for possible use on the project. DeGAUSS is a freely available, open source tool [8*] that facilitates distributed, reproducible address geocoding. After testing DeGAUSS, the DAC team suggested a means by which it could be included in the workflow for the collective analysis so that spatial characteristics of the cohorts could be analyzed at the census tract level. Besides demonstrating the multidisciplinary utility of the DAC, this exemplifies the DAC scientific partnership – collaborating with a team of cohort investigators, searching for a solution to a problem, providing an innovative, useful tool with instruction to the cohorts, and implementing its use in an ECHO-wide cohort analysis.

The collaborative partnership is critical to the ECHO DAC. From the first meeting of the ECHO Steering Committee, we worked with teams of cohort investigators to develop concepts for collective analyses and review papers. Each ECHO-wide Cohort Working Group has at least one DAC representative, whose role and responsibilities vary according to group. For the Outcome Working Groups, the DAC investigators assist the co-chairs and facilitator with meeting preparation, provide methodologic guidance for new research, and are liaisons between the DAC and the Working Group when research ideas and concepts are proposed and need analytical plans. The DAC members are individuals with strong methodologic skills and backgrounds in the content area. DAC investigators also have leadership roles on Working Groups for activities closely related to DAC responsibilities, including the Data Sharing, Data Harmonization, Innovations in Data Analysis, Geospatial, Epigenetics and Genetics Working Groups. With multidisciplinary expertise, DAC investigators are also members, and subgroup chairs to other cross-cutting Working Groups.

Understanding the Cohorts that Contribute to ECHO is High Priority

In collaborative study designs, a term to capture collaboration across a network of observational studies with heterogeneous designs [9*], understanding the data derivation and potential biases is necessary for applying methods appropriate for proposed data analyses. Key aspects to glean about the cohorts include their target populations and selection criteria, calendar time periods of enrollment and follow-up, the frequency and modes of data collection, and whether any underlying outcome or condition influenced the data collection. For example, if data collection in a cohort relied on electronic medical records, sicker individuals, e.g., children with obesity-related conditions, may have contributed more data than other children, e.g., non-obese. In this cohort with increased frequency of observation based on health, health-related outcomes, such as asthma, may be diagnosed more than in a different cohort where children were seen annually regardless of their medical history. If these cohorts disproportionately contribute numbers of obese and non-obese children, ignoring differences in data collection when pooling the data for analysis may lead to a biased estimate of the association between obesity and asthma. These types of study design characteristics are known as metadata.

The heterogeneity of the ECHO cohorts provides a richness of host characteristics for risk stratification and comparisons, elongated time-period coverage for examination of temporal trends, and variability in exposures for multifactorial examinations, so that harmonizing extant and new data will yield an invaluable resource for child health research. However, naively pooling the data in analyses may lead to inaccurate conclusions. Expertise at the DAC, in collaboration with complementary expertise across the ECHO components, will enable robust and valid inference from data collected with disparate methods and frequencies. We will also maximize extraction of information through the application of appropriate statistical analyses that can incorporate temporally varying exposures and account for staggered entries and informative attrition, and develop methods to meet identified gaps and analytical challenges.

Methodologic expertise and support for analysis also varies across the contributing cohorts' teams. The ECHO DAC offers equitable access to sophisticate and specialized methodologic

expertise to all ECHO investigators. The Innovations in Data Analysis Working Group, co-chaired by a DAC investigator, was formed to discuss methodologic issues, and disseminate new methods and software as they become available. The bidirectional exchange of information and tools will benefit both cohort-specific and ECHO-wide cohort research and strengthen the collaborative relationship.

In addition to collecting metadata for the purpose of understanding the cohorts, it is also to obtain information about the extant data. Whereas all collaborative study designs have some commonality across their contributing studies [9*], there are two primary goals for implementing a collaborative study design. One is to address a common research question. For this goal, everyone agrees on the data needs to address the hypothesis – outcome, exposure, confounders, and possibly modifiers and mediators. The Cohort Consortium Vitamin D Pooling Project of Rarer Cancers is an example of this type of collaborative study design [10]. The second goal is to establish a platform for answering questions as they arise, requiring close collaboration with contributors, experts in team science and developing the appropriate methods to appropriately work with the data. Examples of this collaborative study design include the Healthy Birth, Growth, and Development–Knowledge Integration (HBGDki) initiative sponsored by the Bill and Melinda Gates Foundation, (http://hbgdki.org/) [11*], the North American AIDS Cohort Collaboration on Research and Design (https://statepiaps7.jhsph.edu/NAaccord/) [12], and the Chronic Kidney Disease Prognosis Consortium (www.jhsph.edu/ckdpc) [13]. The ECHO-wide cohort consortium fits into this latter category of collaborative study designs. Therefore, getting information about the extant data was crucial for developing the ECHO-wide cohort data protocol, and for developing ECHO-wide research questions.

As the ECHO DAC, we consider establishing and strengthening collaborative relationships with each cohort preeminent - to understand the study designs, scientific aims, available expertise, and existing data. Below we provide some of the methods that we have initiated to accomplish this goal.

Gathering Cohort-Specific Information and Materials—One of the first introductions to the ECHO cohorts was to participate on calls between the NIH project director and the Principal Investigators of the ECHO cohort awards. To further capture data systematically, the DAC distributed surveys that covered: 1) Contact Information; 2) General Information and Study Design; 3) Data Elements and Domains; 4) Biospecimen Collection, Assays, Storage; and 5) Data Management. In addition, the DAC established a secure FTP site where cohort investigators could share cohort study materials (e.g., protocol, manuals of operations, data dictionaries, forms) with the DAC. We continue to create and disseminate surveys to the cohorts to gather more information as needed.

Establish Metadata Catalog—The DAC used the information from the surveys to develop a searchable Metadata Catalog, which we placed on the secured ECHO-wide cohort website available through a web portal, ECHOPortal (described below). All ECHO investigators are provided with password-protected accounts to access the Metadata Catalog. The DAC demonstrated its features at an in-person meeting of the Steering Committee and hosted a webinar; its features and content increase with each new phase. In addition to

describing the study designs and available data according to the survey information, the Metadata Catalog also provides access to study materials for those cohorts where the contact Principal Investigator and the lead investigator for the cohort have agreed to general sharing of these materials with the ECHO Program. Investigators use the Metadata Catalog to assess feasibility for ECHO-wide cohort research, and contact information to communicate with other cohort investigators with similar data to develop research collaborations. As the ECHO-wide individual-level data become available via the ECHOPortal, the Metadata Catalog will expand with details on exactly what data are available on the platform, facilitating broad collaboration on the development of research concepts and analysis plans.

Cohort Advocate Teams—With 84 cohorts, it is difficult for all DAC investigators to get a deep understanding of every cohort, and for a couple of people at the DAC to be able to respond to questions which may arise from this many cohorts. Therefore, we have assigned groups of individuals from data management, harmonization and analysis to each cohort. The goals are to develop a deeper understanding of the cohort, i.e., to become the go-to person for other DAC personnel and investigators, and to provide the cohorts with individuals at the DAC to contact when they have additional questions.

The ECHOPortal

ECHOPortal (ECHOPortal.org) is the highly secure computational environment being developed by the DAC for data transfer, storage and analysis. In the first year, we established and implemented FISMA [14]. Moderate security requirements, including security control documentation, vulnerability scanning, and security certification and accreditation review, leading to an Authority to Operate ECHOPortal within RTI's onsite network.

With ECHOPortal we have deployed functionality to register users, host the Metadata Catalog, establish the ECHO-wide cohort data lake, and begin ingesting data through secure data file transfer. In addition, we have laid the groundwork for ECHOPortal to evolve into a hybrid cloud-based environment, expanding functionality through numerous service offerings, and providing elastic storage and computational power. Developing ECHOPortal in phases has allowed rapid deployment, and continual refinement and inclusion of features. Figure 1 shows the high-level component architecture envisioned for ECHOPortal. The design supports data capture, receipt, storage, tracking, editing, harmonization, and analysis needs for ECHO-wide cohort research (including a reproducible workspace for analysts) within a secure platform. ECHOPortal users will be given controlled, time-limited access to tailored views and functions based on their current role(s) and responsibilities on ECHO.

Approaches to Analyses

Due to the complexity and scope of ECHO, analyses will be conducted in several different manners. These analyses are in addition to review papers which have been initiated to increase public understanding of the ECHO cohorts, and identify knowledge gaps in relevant content areas where ECHO may contribute. Review-plus papers are review papers supplemented with ECHO cohort metadata.

Collective Analyses—To speed early research productivity, before the establishment of the ECHO-wide cohort data collection protocol and platform, we initiated collective analyses, an innovative adaptation of meta-analysis to address core questions and accelerate the development of a collaborative research environment. In this approach the DAC creates the statistical code for the cohorts to analyze their own data and provide their results to the DAC to compile with meta-analysis. Unlike traditional meta-analyses that rely on published results from independent methods, this approach standardizes the statistical methods used by each contributing cohort. There are other meta-analysis hybrids that have been used in collaborative study designs; for example, one group conducts meta-analyses using both individual level data and estimates [15].

Individual-level Analyses—Full implementation of the ECHO-wide cohort data platform awaits implementation of the ECHO-wide cohort protocol, and once established, will be used to address many research questions. As described above, some research efforts are underway through collective analyses. Another approach involves the submission of harmonized data from self-selected groups of cohorts, according to a data dictionary specific to a proposed analysis plan, and uses Data Use Agreements between the cohorts and the DAC. This approach, as all ECHO-wide cohort analyses, are performed after the Steering Committee approves the scientific proposal with the DAC-developed Analysis Plan, as specified by the Publication Policy.

DAC analysts document all analyses, whether using cohort results from the disseminated approach or individual level data, which are then subject to review by DAC investigators, to enhance reliability, or reproducibility. The DAC continues to develop internal policies for conducting quality-assured ECHO cohort analyses; all analysts working on the ECHOPortal will be trained on these practices.

Conclusion

The DAC model offers more than a data repository for the ECHO cohorts. The DAC resources and expertise provide a collaborative environment for understanding and appropriately utilizing the complex ECHO data to conduct rigorous research on pediatric health. Essential DAC initial activities include methods to understand the study designs, investigator expertise, and extant data across the cohorts, planning and developing the secure platform and systems for data transfer, management and analysis, and working with all component investigators to develop the policies that enhance the successful conduct of the collaboration.

Acknowledgments

Funding: This work was funded under NIH grant U24OD023382

The authors thank the many members of the ECHO Data Analysis Center faculty and staff for their contributions to the vision for the Center. Special thanks to the following individuals and group.

Computing infrastructure team at RTI: Grier Page, PhD, Rebecca Boyles, MSPH, Chris Siege, MS, Evan Patterson, BS, Ying Qin, MS, Leena Dave, MS, and Fred Huebner, MS.

Project director at JHU: Monica McGrath, PhD

Content expert co-investigator at JHU: Jeanne S. Sheffield, M.D.

Geospatial Technologist at RTI: William Wheaton, MA

Financial support: This work was supported by funding from NIH for the Environmental influences of Child Health Outcomes (ECHO) Cohort Data Analysis Center (U24OD023382).

References

1. Scientific Data ISSN 2052-4463 (online). (http://www.nature.com/sdata/policies/repositories). This is a list of recommended data repositories for depositing data sets associated with peer-reviewed publications. The list is date-stamped, and described by special content as appropriate.
2. Meinert, CL. Clinical Trials Design, Conduct and Analysis. 2. Oxford University Press, Inc; New York: 2012.
3. National Heart Lung and Blood Institute. Compendium of Best Practices for Data Coordinating Centers - NHLBI, NIH. 2011. (https://www.nhlbi.nih.gov/research/resources/compendium)
4. The ARIC investigators. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. Am J Epidemiol. 1989 Apr; 129(4):687–702. [PubMed: 2646917]
5. Kaslow RA, Ostrow DG, Detels R, Phair JP, Polk BF, Rinaldo CR Jr. The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants. Am J Epidemiol. 1987 Aug; 126(2):310–8. [PubMed: 3300281]
6. Furth SL, Cole SR, Moxey-Mims MM, Kaskel F, Mak RH, Schwartz GJ, Wong CS, Muñoz A, Warady BA. Design and methods of the Chronic Kidney Disease in Children (CKiD) prospective cohort study. Clin J Am Soc Nephrol. 2006; 1:1006–15. [PubMed: 17699320]
7. Haas DM, Parker CB, Wing DA, Parry S, Grobman WA, Mercer BM, Simhan HN, Hoffman MK, Silver RM, Wadhwa P, Iams JD, Koch MA, Caritis SN, Wapner RJ, Esplin MS, Elovitz MA, Foroud T, Peaceman AM, Saade GR, Willinger M, Reddy UM. NuMoM2b study. A description of the methods of the Nulliparous Pregnancy Outcomes Study: monitoring mothers-to-be (nuMoM2b). Am J Obstet Gynecol. 2015 Apr; 212(4):539e1–539.e24. DOI: 10.1016/j.ajog.2015.01.019 [PubMed: 25648779]
*8. Brokamp C, Wolfe C, Lindgren T, et al. Decentralized and reproducible geocoding and characterization of community and environmental exposures for multisite studies. J Am Med Inform Assoc. 2017 Nov 8. Epub ahead of print This paper describes the features of a standalone containerized software for standardized geocoding and geomarker coding at local sites in a multicenter study, and demonstrates its reliability compared to other software. doi: 10.1093/jamia/ocx128
*9. Lesko CR, Jacobson LP, Althoff KN, et al. Collaborative, Pooled, and Harmonized Study Designs for Epidemiologic Research: Challenges and Opportunities. Int J Epidemiol. (IN PRESS). In this paper, the collaborative study design is described in detail, together with opportunities and challenges associated with these collaborations. Methods to overcome some of the biases are discussed.
10. Helzlsouer KJ, Committee VS. Overview of the Cohort Consortium Vitamin D Pooling Project of Rarer Cancers. Am J Epidemiol. 2010; 172(1):4–9. DOI: 10.1093/aje/kwq119 [PubMed: 20562193]
*11. Jumbe NL, Murray JC, Kern S. Data sharing and inductive learning — toward healthy birth, growth, and development. N Engl J Med. 2016; 374:2415–2417. The Healthy Birth, Growth, and Development–Knowledge Integration (HBGDki) initiative sponsored by the Bill and Melinda Gates Foundation offers one example of how data sharing from multiple observational studies and clinical trials can be used to improve public health. In this paper, the authors describe some of the challenges they faced in gathering the data and broadly describes their approaches to effectively use the data. DOI: 10.1056/NEJMp1605441 [PubMed: 27168111]
12. Gange SJ1, Kitahata MM, Saag MS, Bangsberg DR, Bosch RJ, Brooks JT, Calzavara L, Deeks SG, Eron JJ, Gebo KA, Gill MJ, Haas DW, Hogg RS, Horberg MA, Jacobson LP, Justice AC, Kirk GD, Klein MB, Martin JN, McKaig RG, Rodriguez B, Rourke SB, Sterling TR, Freeman AM, Moore RD. Cohort profile: the North American AIDS Cohort Collaboration on Research and Design

(NA-ACCORD). Int J Epidemiol. 2007 Apr; 36(2):294–301. DOI: 10.1093/ije/dyl286 [PubMed: 17213214]

13. Matsushita K, Ballew SH, Astor BC, Jong PE, Gansevoort RT, Hemmelgarn BR, et al. Cohort profile: the chronic kidney disease prognosis consortium. Int J Epidemiol. 2013; 42(6):1660–8. DOI: 10.1093/ije/dys173 [PubMed: 23243116]
14. An Act to strengthen Federal Government information security, including through the requirement for the development of mandatory information security risk management standards (Federal Information Security Management Act of 2002). Public Law 107–347; DOCID: f:publ347.107 (https://www.gpo.gov/fdsys/pkg/PLAW-107publ347/html/PLAW-107publ347.htm)
15. Matsushita K1, Coresh J2, Sang Y1, et al. Estimated glomerular filtration rate and albuminuria for prediction of cardiovascular outcomes: a collaborative meta-analysis of individual participant data. Lancet Diabetes Endocrinol. 2015 Jul; 3(7):514–25. Epub 2015 May 28. DOI: 10.1016/S2213-8587(15)00040-6 [PubMed: 26028594]

Key Bullet Points

- The ECHO Data Analysis Center offers a secured collaborative environment with expertise to maximize the utility of the ECHO cohort data to conduct reproducible research on pediatric health.
- Understanding the study designs, populations and extant data in the ECHO cohorts is critical for the Data Analysis Center to develop the appropriate data management and analytical methodology for this complex study, and is being accomplished with many methods, including surveys and direct communications with the cohort investigative teams.
- Bidirectional communication, understanding the Data Analysis Center model, and collaborative productivity will help assuage the concerns encountered with bridging data in a collaborative study design.

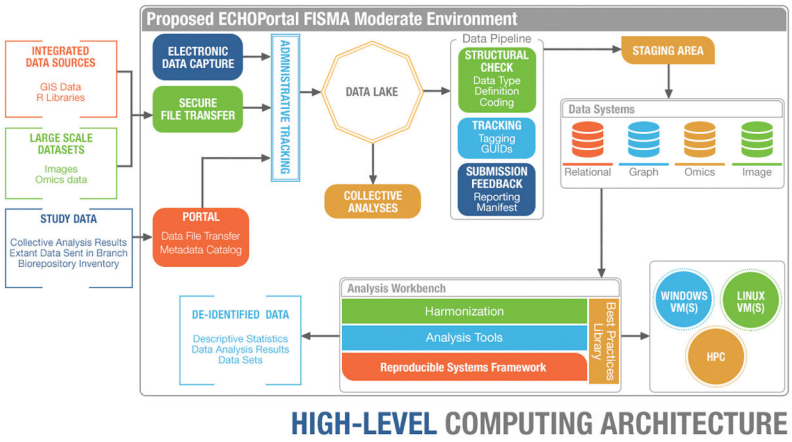

Figure 1.
An overview of the ECHOPortal Architecture