HHS Public Access

Author manuscript

Circulation. Author manuscript; available in PMC 2017 April 05.

Published in final edited form as:

Circulation. 2016 April 5; 133(14): 1410–1418. doi:10.1161/CIRCULATIONAHA.115.019506.

Report of the National Heart, Lung, and Blood Institute Working Group: An Integrated Network for Congenital Heart Disease Research

Sara K. Pasquali, MD, MHS[1], **Jeffrey P. Jacobs, MD**[2], **Gregory K. Farber, PhD**[3], **David Bertoch, MHA**[4], **Elizabeth D. Blume, MD**[5], **Kristin M. Burns, MD**[6], **Robert Campbell, MD**[7], **Anthony C. Chang, MD**[8], **Wendy K. Chung, MD, PhD**[9], **Tiffany Riehle-Colarusso, MD, MPH**[10], **Lesley H. Curtis, PhD**[11], **Christopher B. Forrest, MD, PhD**[12], **William J. Gaynor, MD**[13], **Michael G. Gaies, MD, MPH**[1], **Alan S. Go, MD**[14], **Paul Henchey, MS**[15], **Gerard R. Martin, MD**[16], **Gail Pearson, MD, ScD**[6], **Victoria L. Pemberton, RN, MS**[6], **Steven M. Schwartz, MD**[17], **Robert Vincent, MD**[7], and **Jonathan R. Kaltman, MD**[6]

[1]Department of Pediatrics and Communicable Diseases, University of Michigan C.S. Mott Children's Hospital, Ann Arbor, MI

[2]Department of Surgery, Johns Hopkins All Children's Heart Institute, St. Petersburg, FL

[3]National Institute of Mental Health, National Institutes of Health, Bethesda, MD

[4]Children's Hospital Association, Overland Park, KS

[5]Department of Cardiology, Boston Children's Hospital, Boston, MA

[6]Division of Cardiovascular Sciences, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD

Correspondence: Sara K. Pasquali, MD, MHS, University of Michigan C.S. Mott Children's Hospital, 1540 E. Hospital Drive, Ann Arbor, MI 48105, Phone: 734-232-8594, Fax: 734-936-9470, pasquali@med.umich.edu.

Disclosures: Dr. Pasquali receives funding from the National Heart, Lung, and Blood Institute and the Janette Ferrantino Professorship related to congenital heart disease research. Dr. Pasquali is a member of the Society of Thoracic Surgeons Congenital Heart Surgery Database Taskforce, directs the Pediatric Cardiac Critical Care Consortium Data Coordinating Center, and leads the Integrated CARdiac Data and Outcomes (iCARD) Collaborative within the NHLBI-sponsored Pediatric Heart Network. Dr. Jacobs is Chair, Society of Thoracic Surgeons National Database Workforce; Chair, Congenital Heart Surgeons Society Committee on Quality Improvement and Outcomes; Executive Committee Member Pediatric Cardiac Critical Care Consortium; and American College of Cardiology Improving Pediatric and Adult Congenital Treatment Registry Steering Committee Member. Dr. Farber manages the National Institute of Mental Health Data Archives. Mr. Bertoch is Vice President of Comparative Data and Informatics, Children's Hospital Association. Dr. Blume is the co-PI for National Heart, Lung, and Blood Institute Contract #HHSN268200548198C for the Interagency Registry for Mechanically Assisted Circulatory Support. Dr. Chung receives funding from the National Heart, Lung, and Blood Institute related to congenital heart disease research; and is a member of the Pediatric Cardiac Genomics Consortium, and Pediatric Cardiomyopathy Registry. Dr. Curtis receives funding from the US Food and Drug Administration as the Data Core Lead for the Mini-Sentinel Program. Dr. Forrest is the PI of PEDSnet (National Pediatric Learning Health System), funded by the Patient-Centered Outcomes Research Institute, Chair of the PCORnet Research Committee, and Chair of the PEPR Steering Committee (Validation of PROMIS Pediatric Measures in Chronic Disease Consortium). Dr. Gaynor is a member of the Society of Thoracic Surgeons Congenital Heart Surgery Database Taskforce and Public Reporting Committee, and the Pediatric Cardiac Critical Care Consortium Executive Committee. Dr. Gaies receives funding from the National Heart, Lung, and Blood Institute related to congenital heart disease research, and is the Executive Director of the Pediatric Cardiac Critical Care Consortium. Dr. Go receives research funding from the Patient-Centered Outcomes Research Institute, and is the PI of the PORTAL Network Congenital Heart Defect Cohort. Mr. Henchey is an employee of ArborMetrix, Inc. Dr. Schwartz is a member of the Pediatric Cardiac Critical Care Consortium Executive Committee, and the Pediatric Heart Network Executive Committee. Dr. Vincent is the Steering Committee Chair, American College of Cardiology Improving Pediatric and Adult Congenital Treatment Registry, and Pediatric Cardiac Critical Care Consortium Executive Committee Member. The remaining authors have no disclosures.

[7]Department of Pediatrics, Emory University, Atlanta GA

[8]Department of Pediatrics, Children's Hospital of Orange County, Orange, CA

[9]Department of Pediatrics and Medicine, Columbia University, New York, NY

[10]Division of Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention, Atlanta, GA

[11]Duke Clinical Research Institute, Duke University School of Medicine, Durham, NC

[12]Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA

[13]Department of Surgery, Children's Hospital of Philadelphia, Philadelphia, PA

[14]Division of Research, Kaiser Permanente Northern California, Oakland, CA

[15]ArborMetrix Inc., Ann Arbor, MI

[16]Department of Pediatrics, George Washington University School of Medicine, Children's National Medical Center, Washington, DC

[17]Departments of Critical Care Medicine and Paediatrics, The Hospital for Sick Children and The University of Toronto School of Medicine, Toronto, Canada

Abstract

The National Heart, Lung, and Blood Institute convened a Working Group in January 2015 to explore issues related to an integrated data network for congenital heart disease (CHD) research. The overall goal was to develop a common vision for how the rapidly increasing volumes of data captured across numerous sources can be managed, integrated, and analyzed to improve care and outcomes. This report summarizes the current landscape of CHD data, data integration methodologies used across other fields, key considerations for data integration models in CHD, and the short- and long-term vision and recommendations made by the Working Group.

Keywords

congenital heart disease; database; health outcomes

Introduction

As medicine moves into the era of big data, it is important to develop a common vision for how the rapidly increasing volume of data will be managed, integrated, and analyzed to improve care and outcomes. This holds true across a variety of different disciplines and specialties, including the field of congenital heart disease (CHD). Only through coordinated efforts will the CHD community be able to fully leverage available and emerging data sources to support important investigations, and conduct research most efficiently. To facilitate this process, the National Heart, Lung, and Blood Institute (NHLBI) convened a Working Group meeting in January 2015 in Bethesda, MD, to explore issues related to CHD data integration. The goals of the Working Group were to develop a vision for an integrated data network to support CHD research and to identify critical elements as well as potential

barriers. The Working Group consisted of experts in pediatric and adult cardiology, cardiothoracic surgery, health services and outcomes research, epidemiology, informatics, and statistics.

The Era of Big Data

The past several years have been characterized by what has been termed an era of "big data"[1–3]. During this time, the volume, velocity, and variety of data captured across numerous sources have increased exponentially, outpacing traditional techniques for managing and analyzing data. Newer data platforms, computing capabilities, and analytic techniques have been developed to better manage, integrate, analyze, and provide more "real-time" feedback to various industries regarding their data, with the goal of optimizing performance and outcomes[1–3]. For example, the automotive industry is capturing data generated by sensors on electric cars to better understand driving habits such as typical acceleration and braking patterns. These data are merged and analyzed with information on frequency of battery charging, and location of charging stations, to aid in better design of the next generation of vehicles and charging infrastructure[1,4]. In the hotel industry, certain chains merge and analyze weather and airline flight cancellation data, along with information on geographic location of their hotels. These data are used to target mobile ads to passengers likely stranded away from home and enable easy booking of nearby hotels[5].

Data in Medicine

Historically in medicine and in the hierarchy of medical research, the value of databases, registries, and other data sources in the cycle of scientific discovery has not always been recognized[1,6]. "Mining datasets" and "database research" have often been seen as lesser pursuits compared to basic science research or clinical trials. However, several developments have begun to change the way data in medicine are viewed. First, similar to other fields, the volume and granularity of data captured electronically in healthcare has increased exponentially in recent years, including data captured in the electronic health record (EHR), clinical registries, research datasets, monitoring systems, and other sources. With this has come the recognition that analyzing and integrating these datasets can expand the range of questions that can be answered[1]. For example, early results suggest integrating continuous data streams generated by various monitoring systems with clinical outcomes data may enable better prediction and treatment of adverse events in intensive care settings[1,7]. Second, along with this increase in availability of data, there has been a simultaneous decrease in funding to support biomedical research[8]. This has led to further interest in better understanding how to leverage available data to power research more efficiently; for example the use of existing registries as platforms to support clinical trials with the goal of reducing time and costs associated with data collection[9]. Finally, the current national emphasis in healthcare on improving quality and optimizing healthcare value has necessitated analyzing and integrating healthcare quality and cost data across a variety of sources in order to understand the landscape of care delivery and outcomes, to investigate relationships between quality and cost, and to develop strategies for improvement[10]. These and other recent trends have led to a greater recognition in medicine of the value of leveraging the increasing volume of available data. As further evidence of this, the National

Institutes of Health recently launched the Big Data to Knowledge and Precision Medicine Initiatives[11,12]. Both of these programs involve efforts to integrate information across a variety of sources to conduct research more efficiently and improve care.

Current Landscape of CHD Data

Data sources, infrastructure, and collaboration

The current CHD data environment has many assets (Table 1). Numerous existing clinical registries, administrative/billing databases, public health surveillance databases, research datasets, and other sources contain a wealth of important information that can be used to facilitate research, surveillance, and quality improvement activities in the field[13]. In addition, data are being increasingly captured via a variety of newer modalities including the EHR, data generated from medical monitors and devices, and genetic and biomarker data. Some centers are also beginning to capture longer-term outcomes data such as quality of life and neurodevelopmental outcomes as described in more detail in subsequent sections. Most congenital heart programs across the US have existing local infrastructure to support data collection for various registries and other datasets, and infrastructure to integrate data across centers also exists as a part of several national registries, multi-center quality improvement activities, and research efforts[13]. Finally, there is an environment of collaboration among investigators and many congenital heart programs related to participation in these efforts, including the Pediatric Heart Network, National Pediatric Cardiology Quality Improvement Collaborative, and Pediatric Cardiac Critical Care Consortium, among many others[14–17]. Annual meetings of the Multi-societal Database Committee for Pediatric and Congenital Heart Disease further aid in facilitating sharing of ideas and collaboration across different registries and databases[18].

Standardized nomenclature

Another particularly important aspect of the CHD data landscape (Table 1) has been the major effort over the past two decades to develop a standardized nomenclature system[19]. In the 1990s both The European Association for Cardio-Thoracic Surgery (EACTS) and The Society of Thoracic Surgeons (STS) created databases to assess congenital heart surgery outcomes, and established the International Congenital Heart Surgery Nomenclature and Database Project. By 2000, a common nomenclature and common core minimal dataset were adopted by both the EACTS and STS Congenital Heart Surgery Databases. Subsequently, the International Society for Nomenclature of Pediatric and Congenital Heart Disease was formed, and its Nomenclature Working Group cross-mapped the nomenclature developed by the EACTS and STS with the European Pediatric Cardiac Code of the Association for European Pediatric Cardiology, thereby creating the International Pediatric and Congenital Cardiac Code (IPCCC), which is available for free download at [http://www.IPCCC.NET][19]. The IPCCC is currently used by multiple databases that span the spectrum of pediatric and congenital cardiac care including:

- Cardiac surgery (STS Congenital Heart Surgery Database, EACTS Congenital Heart Surgery Database, Japan Congenital Cardiovascular Surgery Database [JCCVSD], United Kingdom Central Cardiac Audit Registry [UKCCAD])

- Cardiology (The IMPACT Registry™ [IMproving Pediatric and Adult Congenital Treatment] of the National Cardiovascular Data Registry® of The American College of Cardiology)
- Anesthesia (The Joint Congenital Cardiac Anesthesia Society [CCAS]–STS Congenital Cardiac Anesthesia Database)
- Critical care (The Pediatric Cardiac Critical Care Consortium [PC^4] and The Virtual Pediatric Intensive Care Unit System [VPS])

Data integration efforts to date

A final strength of the current CHD data environment is the development and implementation of methods over the past several years to support integration and linkages between various CHD data sources[1,13] (Table 1, 2). These methods have capitalized on the strengths and mitigated the weaknesses of different types of datasets, and have enabled research that would have otherwise not been possible with individual datasets alone. In addition, methods to facilitate sharing and integration of data have also begun to support the use of existing data (e.g. clinical registry data) to power studies more efficiently.

Linking on Unique Identifiers—Local medical records, and some larger datasets, contain unique identifiers such as medical record number or social security number that can facilitate linkage with other datasets[1,13,20,21] (Table 2). For example, using these methods investigators have previously linked data from outpatient pediatric cardiology visits for chest pain to the National Death Index and Social Security Death Master File to evaluate for subsequent mortality in this cohort[21]. However, new limitations on the availability of the Social Security Death Master File for research purposes may pose a greater challenge to the use of this dataset in the future.

Linking on Indirect Identifiers—While linkages based on direct or unique identifiers are the easiest way to merge datasets, these identifiers are often not collected in many databases due to a variety of regulatory requirements and privacy concerns, and may only be available at the local level. Thus, methodology has also been developed to link database records through the use of “indirect” identifiers[1,13,22,23] (Table 2). These include variables such as date of birth, date of admission, date of discharge, sex, and center where hospitalized. It has been shown that nearly all records at a given congenital heart center can be uniquely identified using a combination of these indirect identifiers, and that a crosswalk can then be created between two datasets, linking patients on the values of center where hospitalized and the indirect identifiers[23].

This methodology has been utilized in the pediatric cardiovascular population to merge information from a large clinical registry (The STS Congenital Heart Surgery Database) with a pediatric administrative dataset (Pediatric Health Information System Database)[1,13,23–26]. Linking these two datasets allows utilization of the detailed operative and outcomes data from the clinical registry, and the resource utilization data from the administrative dataset. The current linked dataset includes records from >60,000 children undergoing congenital heart surgery at 33 different hospitals from 2004–2010, with

expansion and updating of the dataset underway. Several comparative effectiveness studies and analyses of healthcare costs have been successfully conducted with this dataset, leveraging variables from both data sources to facilitate analyses not otherwise possible with either dataset independently[24–26]. Similar methodology has also been used to merge clinical trial data from the Pediatric Heart Network Single Ventricle Reconstruction Trial with data from Children's Hospital Association Case Mix dataset in order to perform integrated analyses of clinical outcomes and costs[27].

Combinations of indirect and direct identifier linkage methodologies have also been utilized. For example, datasets may use indirect methods to link to a common dataset, and then use the common dataset to identify unique individuals. Using these methods, surveillance data from the Metropolitan Atlanta Congenital Defects Program was indirectly linked to vital records, as was data from the Special Education Database of Metropolitan Atlanta. Then deterministic (direct) linkage was done between the two datasets to evaluate the use of special education services among children with CHD[28].

Center-level Linkages—Linking registry data to other independently collected center-level data through matching on center can be easily accomplished (Table 2). For example, independently collected survey data regarding intensive care unit care models and nursing education and staffing levels have been successfully linked to The STS Congenital Heart Surgery Database[1,13,29]. These linkages enable evaluation of the variables collected in the survey in relation to outcomes data collected in the registry.

Collaboration/Partnering Between Databases—Data can also be shared or integrated through collaboration and partnering between different organizations and datasets (Table 2). These methods can also reduce data entry burden at local sites. For example, the STS and the CCAS recently collaborated to add an anesthesia section to the surgical data collection forms[1,13,30]. Anesthesia data are now collected, harvested, reported, and analyzed at a central data coordinating center along with surgical data for participating centers. This approach is likely more time and cost efficient than creating a separate anesthesia database in which many of the fields regarding patient characteristics and the operative procedure would have been duplicated. Related efforts are underway to incorporate electrophysiology data within the IMPACT Registry[31]. An alternative method involves a more distributed approach with sharing of data definitions and variables between datasets and organizations, information technology solutions allowing single entry of data at the local level, and subsequent submission and distribution of both shared and unique data variables to applicable national datasets and data coordinating centers. An example of this is the shared IPCCC data definitions for certain variables between the STS, PC[4], and IMPACT registries[1,13,15,16,31].

Supplementary Data Modules—Methods have also been developed to create data modules enabling efficient collection of supplemental data points to the primary data source (Table 2). The modules are generally web-based and can be quickly created and deployed to allow "real-time" collection of additional data needed to answer important research questions that may arise. This methodology has been recently successfully used by PC[4] to collect supplemental data to the main registry to study the relationship between Vasoactive-

Inotropic Score and outcome after infant cardiac surgery[1,13,32]. A module allowing for capture of additional data related to inotrope use was created, deployed, and linked to the main registry data for each patient. This facilitated efficient data collection with 391 infants prospectively enrolled across 4 centers in just 5 months[32].

Powering Clinical Trials Through Registries—In recent years it has become increasingly recognized that many variables of interest for prospective investigation, including clinical trials, are already being captured within clinical registries across an engaged group of sites on a routine basis[1,9,13]. It has been proposed that leveraging these existing registry data (and linkage with modules containing additional study-specific data when necessary) may be a more efficient way to power research, avoiding duplicate data collection and minimizing study start-up timelines. To date these methods have been used to facilitate two clinical trials in adult cardiovascular medicine – the SAFE-PCI (Study of Access Site for Enhancement of PCI for Women) and TASTE (Thrombus Aspiration during ST-Segment Elevation Myocardial Infarction) trials, both of which leveraged existing information in two different cardiac catheterization registries[33,34]. In the CHD community, work is currently underway within the Pediatric Heart Network to assess the completeness and accuracy of site's local clinical registry data, and determine whether prospective studies and clinical trials may leverage these data to minimize duplicate data collection and promote greater efficiency.

Current CHD data limitations

While a great deal of progress has been made over the past several years to better integrate and leverage available CHD data sources to more efficiently conduct research, many limitations remain[1,13] (Table 1). As described in the preceding sections, most current methods to integrate data have involved 1:1 linkages of a certain dataset to another to answer a specific question. Regulatory and contracting issues have generally prevented use of these integrated datasets to answer additional research questions after the primary study is completed. More broad and comprehensive strategies for data integration across numerous sources are lacking, and the landscape still consists primarily of individual data silos. There is relatively limited capability from an information technology perspective to broadly share information across datasets, and data governance and collaboration models have yet to be developed across congenital heart centers and national organizations to facilitate such data sharing. Methodology to allow the use of existing data for more efficient prospective studies and clinical trials is just beginning to be developed in the field. Issues regarding consent and confidentiality will also require further evaluation. Finally, important longitudinal outcomes information is lacking, and simply linking existing datasets together will not address this issue.

Data integration models across other fields

Several models supporting more comprehensive data integration exist across other fields, and may be useful examples for the CHD community to consider.

National Institute of Mental Health (NIMH) and National Database of Autism Research

In order to link information across studies enrolling patients with autism, the NIMH uses a global unique identifier (GUID)[1,35,36]. The GUID is a unique code generated in a manner to protect confidentiality. In order to generate a GUID, a combination of patient identifying information (typically name at birth, gender, city of birth, and date of birth) is entered into a software program at the local site[35,36]. Through one-way encryption, these elements are translated into hash codes, which cannot be traced back to the patient. The hash codes are then sent to a central data server which generates a GUID. The GUID is returned to the site where it is entered into the dataset for that patient. The GUID protects privacy in two ways – the identifiable information never leaves the local site, and the GUID cannot be traced back to the patient. Another important feature of the GUID is that the patient or family does not need to remember it. Patients and families only need to remember the individual data elements described above, and from these the same GUID is generated each time, regardless of location or timing of enrollment in a study. Thus, the GUID functions to uniquely identify the patient, is composed of information easily known to the patient that is invariant over a lifetime, and maintains patient privacy[35,36]. GUID's are currently being piloted in the field of CHD for patients enrolled in studies conducted by the Pediatric Cardiac Genomic Consortium.

PEDSnet

PEDSnet, funded by the Patient Centered Outcomes Research Institute (PCORI), is a clinical data research network consisting of eight of the nation's largest pediatric academic medical centers, two existing pediatric consortia/quality improvement collaboratives, and two national data partners (Express Scripts, a national pharmacy benefits management company; and IMS Health, a data aggregator of multi-payor claims data)[37,38]. The goal of PEDSnet is to create a learning health system that integrates research into routine care settings, supports structured data capture, and quality improvement processes to rapidly implement advances in new knowledge[37,38]. To date, PEDSnet has harmonized data from 4.5 million patients captured across member site's EHR systems using a common terminology, and uses open source software to support data submission and aggregation. Analyses are primarily done at a centralized data coordinating center, although distributed queries across sites may also be possible (as discussed in further detail in subsequent sections) for certain research questions[37,38].

The Cardiovascular Research Network (CVRN)

The CVRN consists of 15 geographically distributed health care delivery systems caring for >10 million members that was established through initial funding by NHLBI to conduct large-scale adult cardiovascular research more efficiently, including epidemiologic studies, outcomes research, comparative effectiveness studies, and clinical trials[39]. Within the CVRN, data captured through each health care system's EHR and multiple other electronic databases are linked at the site level using medical record numbers[39]. These data include clinical and resource utilization data across inpatient, emergency department, and outpatient settings, procedures, diagnoses, inpatient and outpatient pharmacy data, and laboratory test results. Data capture and architecture are standardized across each site's Virtual Data

Warehouse using common data elements, naming conventions, and definitions to facilitate combining information in aggregate analyses[39]. Recent CVRN efforts within the CHD population have involved developing natural language text processing algorithms to attempt to identify patients with CHD from unstructured EHR data across multiple health systems[40].

Mini-Sentinel Program of the US Food and Drug Administration (FDA)

The goal of this program is to facilitate active surveillance and monitoring of the safety of medical products across the US. Mini-Sentinel uses a distributed data approach where participating data partners maintain physical and operational control over their own data[41]. A common data model was designed to meet the needs of the program, and each participating organization developed a process to extract, transform, and load its source data, applying the common data model, in order to create the distributed database. These data are then analyzed using programs developed centrally and executed locally by participating organizations[41].

Potential data integration models in CHD

There are several different integration models that may be considered to support CHD research which build on the existing models used across other fields. Two specific models that have received the most attention to date, and their strengths and weaknesses given the needs of the CHD community are discussed below.

Creation of a CHD GUID and data linkages at the national level

One option to support data integration in CHD may build on the work done by the autism research community described in the preceding sections[1,35,36]. This could involve creation of a CHD GUID and collaboration among researchers, professional societies, and other groups to share and merge datasets containing these identifiers at the national level[1]. Potential advantages of this approach include its previous success within the autism research community, the ability for multiple linkages, and ability to maintain privacy in that direct identifiers are not sent outside of local sites. However, there are also certain disadvantages to consider. Some of the data elements needed to generate a GUID in its current form are not necessarily found in the medical record and require direct patient contact each time a GUID is generated – for example the data element of “city of birth”[35,36]. While this may be feasible for certain research studies, there would be several difficulties to consider within the current data collection infrastructure of most large registries and datasets where there generally is not direct patient interaction to capture data elements, such that additional local personnel and/or modification of data collection work flow would be required. There may also be issues to consider regarding consent. In addition, in order to enable linkages, the GUID must not only be generated and incorporated into individual datasets, but professional societies and other organizations must also agree to collaborate and share their datasets for linkage and analysis. Negotiating the various data sharing and governance policies of multiple professional societies and different organizations, who often have a focus primarily on adult cardiology and cardiac surgery, and current policies in some cases prohibiting sharing of data outside of central data coordinating centers, may prove to be challenging.

Supporting local linkages and a distributed data network model

An alternative option involves building on the experience of the CVRN, PEDSnet, and the Mini-Sentinel program described in the preceding sections to support data linkages at the local level, and sharing of these integrated data across heart centers, creating a distributed data network in CHD[1]. Local data linkages are feasible because most often research and registry data also reside locally at each participant site's institution in addition to being aggregated into larger multicenter datasets. Local linkages are relatively easy to perform as direct or unique identifiers are readily available in these datasets as well as in the EHR. Merged local datasets can then be de-identified, and groups of institutions or heart centers can collaborate to share and aggregate information at a central site for analysis[1]. Alternatively, data may be kept at each site and standard algorithms developed to query and analyze the data locally, with results aggregated and combined across sites, similar to the Mini-Sentinel approach[41]. This model addresses some of the limitations identified with the use of GUID, and makes linked information available for both local purposes as well as for aggregate research. Data linkages within congenital heart center data warehouses are already taking place at several centers. For example, at one center participating in the Working Group, local data linkages have supported improved accuracy in determining surgical site infection rates (through merging infection data with CHD clinical registry data), which has in turn aided in efforts to reduce these infections[42]. Further, as discussed in previous sections, sharing data across heart centers for multi-center research is already a common practice, and this methodology could leverage these existing collaborative relationships.

Additional considerations

In addition to strategies for integrating existing data, there are several other related areas of consideration. As described in previous sections, data regarding longer-term outcomes remain very limited, and efforts to promote efficient collection of these data have just recently begun[1]. Preliminary work suggests that engaging with patients and families directly can allow for the successful capture of critical longitudinal outcomes data such as survival, re-hospitalizations (particularly those that occur at institutions other than the surgical center), and important aspects of quality of life and burden of disease. For example, standardized patient-reported outcomes data have been successfully captured across two heart centers participating in the Working Group on >2000 patients to date. Methods are being developed to further the use of web-portals, mobile technology, and social media to allow for more efficient and wide-spread capture of patient-reported data[43,44]. In addition, the Cardiac Neurodevelopmental Outcomes Consortium is also working toward develop methods to capture standardized information obtained during neurodevelopment follow-up clinic visits. It will be important to incorporate these emerging longer-term outcomes data into the overall strategy developed for data integration.

A second area of consideration relates to the EHR. It has been hypothesized that leveraging EHR data can allow for improved efficiency and reduce data collection burden for various registries or research datasets. However these efforts will require additional work to improve the quality and standardization of data currently contained in the EHR[1]. While certain types of structured data may be efficiently captured through the EHR (e.g. lab values, medications), other data critical to CHD research may be more difficult to capture due to the

lack of granularity in the EHR and associated coding schemes (e.g. detailed information regarding anatomic diagnoses and procedures) and the lack of standardized definitions (e.g. for pre-operative comorbidities or post-operative complications).

Finally, with the expansion in the number and types of datasets and opportunities for linkages, it remains important to consider several key factors regarding data collection and analyses in general to ensure that research conducted using these datasets is meaningful[1]. These include issues related to accuracy and completeness of data, standardization (or lack thereof) of data elements and definitions, and the availability of variables within the dataset to perform appropriate risk adjustment or adjustment for differences in case mix across hospitals[1]. The availability and use of linked or integrated data sources does not lessen the importance of these critical factors.

Short and Long-term Vision and Recommendations

The Working Group outlined a vision for how data might be integrated in the CHD community and developed several recommendations to achieve that vision over the short and long-term.

Future vision

The Working Group acknowledged that our current conceptualization of data and data management is largely outdated. In addition, the volume and granularity of available data will continue to increase through more wide-spread capture of genomic and biomarker data, and real-time physiologic data, for example. Our current databases, data structure, and analytics will be insufficient for the task of managing and understanding these data. The Working Group recommended that further collaboration and consultation with data scientists and experts from diverse fields and industries outside of medicine will be important in understanding and incorporating modern data storage, manipulation, and analytic techniques into short and long term data solutions for the CHD community. For example, 3-D graphing techniques of large volumes of data have been shown recently in other industries to identify patterns that would otherwise not be apparent; however, these and other novel techniques have had limited application in medical research to date[45]. The importance of considering these techniques within the context of medical decision making was also discussed, as simple associations found in the data do not necessarily indicate cause and effect, and ensuring accuracy, reliability, and integrity of the data will continue to be important concepts regardless of the data storage and integration techniques used.

Short-term goals

The Working Group recommended that the CHD community take steps in the near term to facilitate more comprehensive integration of information across currently available CHD data sources. The group felt that this could support further research that could not be conducted with individual datasets alone, and could also help to promote efficiency in research. The group recommended that of the data integration strategies discussed in previous sections, methodology supporting local data linkages and a distributed data network across collaborating sites seemed to meet more of the needs of the CHD community

compared with alternative approaches. The group acknowledged work in this area already taking place across several congenital heart center data warehouses as described in previous sections.

The Working Group recommended further efforts toward developing such a data network in CHD through engaging interested sites, developing a common data model, strategies for integration, and data governance policies, identifying potential funding sources, and conducting pilot studies to better understand the value of data integration and demonstrate proof of concept. The Working Group also recognized the lack of important longitudinal outcomes data in the field, and acknowledged that this issue cannot be addressed with linkages of current data sources alone. Further development and testing of strategies to support efficient capture of longer-term outcomes data that may be merged with existing data was recommended and the Working Group recognized ongoing efforts in this area described in the preceding sections. Exploring potential funding for data integration efforts was also recommended including funding through the Pediatric Heart Network, other NHLBI funding opportunities, PCORI, and foundation and philanthropic opportunities.

Data standardization

The Working Group acknowledged the importance of data standardization to facilitate data pooling and analysis, and to decrease data entry burden at sites. The Working Group recognized the IPCCC as the standard nomenclature within the field, and recommended that IPCCC terminology and definitions should be incorporated into all relevant data sources when possible, including CHD registries, clinical data and the EHR, and research datasets, and recognized ongoing work to incorporate IPCCC terms into the International Classification of Diseases coding system. The Working Group recommended that consensus across sites and other stakeholders should be developed in order to standardize data collection in other areas as well, such as the collection of longitudinal follow-up information, neurodevelopmental outcomes, etc. and recognized recent work in this area by the Cardiac Neurodevelopmental Outcomes Consortium and others as described in the preceding sections.

Align goals with stakeholder interests

Moving toward more fully integrated data systems will require engagement with multiple stakeholders, including hospital systems, researchers, national organizations and professional societies, and patients and families. To make a case for change, the field will need to discuss and identify the value of data integration from the perspectives of multiple stakeholders. At the hospital level, one such value point is the optimization of strategic investments already made in the EHR and clinical registries. Data integration will provide a more complete picture of the care provided and thus may enable improved quality, reduction in errors, and increased value. Patients and families are also critical stakeholders, and it will be important for data integration activities to seek to align with their interests, which may include improving care and outcomes, participating in longitudinal patient-reported data collection activities, and engaging in the process of determining quality improvement and research priorities.

Conclusions

Several concurrent trends have provided the opportunity to re-calibrate our approach to data collection, integration, and analytics in biomedical research. In the CHD community, several advances in recent years including nomenclature standardization, development of rich clinical datasets, an environment of multi-center collaboration, and implementation of several data integration techniques, have provided a strong foundation for future work. There is now a need for further integration and collaboration in order to meet present and future challenges, and develop a more efficient and comprehensive research enterprise to improve the care and outcomes of patients with CHD.

Acknowledgments

Working Group Members: Sara K. Pasquali MD MHS (University of Michigan C.S. Mott Children's Hospital), Jeffrey P. Jacobs MD (Johns Hopkins University, All Children's Heart Institute), Gregory K. Farber PhD (National Institute of Mental Health), David Bertoch MHA (Children's Hospital Association), Elizabeth D. Blume MD (Boston Children's Hospital), Kristin M. Burns MD (National Heart, Lung, and Blood Institute), Robert Campbell MD (Emory University), Anthony C. Chang MD MBA MPH (Children's Hospital of Orange County), Wendy K. Chung MD PhD (Columbia University), Tiffany Riehle-Colarusso MD MPH (Centers for Disease Control and Prevention) Lesley H. Curtis PhD (Duke Clinical Research Institute), Sherry Farr PhD (Centers for Disease Control and Prevention), Christopher B. Forrest MD PhD (Children's Hospital of Philadelphia), William J. Gaynor MD (Children's Hospital of Philadelphia), Michael G. Gaies MD MPH (University of Michigan C.S. Mott Children's Hospital), Alan S. Go MD (Kaiser Permanente Northern California), Paul Henchey MS (ArborMetrix, Inc), Gerard R. Martin MD (Children's National Medical Center), Gail Pearson MD ScD (National Heart, Lung, and Blood Institute), Victoria L. Pemberton RN MS (National Heart Lung, and Blood Institute), Steven M. Schwartz MD (The Hospital for Sick Children), Mario Stylianou PhD (National Heart, Lung, and Blood Institute), Robert Vincent (Emory University), Jonathan R. Kaltman MD (National Heart, Lung, and Blood Institute).

Funding Sources: This Working Group was funded by the National Heart, Lung, and Blood Institute. The views and conclusions expressed in this report are those of the authors and do not necessarily represent the official position of the National Heart, Lung, and Blood Institute, the National Institutes of Health, or the Centers for Disease Control and Prevention.

References

1. Pasquali SK, Schumacher KR, Davies RR. Can linking databases answer questions about pediatric heart failure? Cardiol Young. 2015; 25(Suppl 2):160–6. [PubMed: 26377723]
2. [Accessed 5/21/2015] Why "big data" is a big deal. Available at: http://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal
3. Merelli I, Perez-Sanchez H, Gesing S, D'Agostino D. Managing, analyzing, and integrating big data in medical bioinformatics: open problems and future perspectives. Biomed Res Int. 2014; 2014:134023. Epub 2014 Sep 1. doi: 10.1155/2014/134023 [PubMed: 25254202]
4. [Accessed 2/1/2015] Using big data to fight range anxiety in electric vehicles. Available at: http://spectrum.ieee.org/cars-that-think/transportation/sensors/using-big-data-to-fight-range-anxiety-in-electric-vehicles
5. [Accessed 5/21/2015] Big data case study – Hospitality – Red Roof Inn. Available at: http://retailxp.com/hospitality-big-data-case-study-red-roof-inn/
6. Califf RM, Peterson ED, Gibbons RJ, Garson A, Brindis RG, Beller GA, Smith SC. Integrating quality into the cycle of therapeutic development. J Am Coll Cardiol. 2002; 40:1895–1901. [PubMed: 12475447]
7. Sullivan BA, Fairchild KD. Predictive monitoring for sepsis and necrotizing enterocolitis to prevent shock. Semin Fetal Neonatal Med. 2015 Mar 27. pii: S1744-165X(15)00042-6. [Epub ahead of print]. doi: 10.1016/j.siny.2015.03.006
8. Alberts B, Kirschner MW, Tilghman S, Varmus H. Rescuing US biomedical research from its systemic flaws. Proc Natl Acad Sci USA. 2014; 111:5773–7. [PubMed: 24733905]

9. Lauer MS, D'Agostino RB. The randomized registry trial-the next disruptive technology in clinical research? N Engl J Med. 2013; 369:1579–81. [PubMed: 23991657]
10. Porter ME. What is value in health care? N Engl J Med. 2010; 363:2477–2481. [PubMed: 21142528]
11. [Accessed 5/21/2015] Big Data to Knowledge Initiative. Available at: http://bd2k.nih.gov/about_bd2k.html#sthash.GsGUsKgu.c7hNf4zV.dpbs
12. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015; 372:793–795. [PubMed: 25635347]
13. Pasquali, SK.; Jacobs, ML.; Jacobs, JP. Linking Databases. In: Barach, P.; Jacobs, JP.; Lipshultz, SE.; Laussen, P., editors. Pediatric and Congenital Cardiac Disease: Outcomes Analysis, Quality Improvement, and Patient Safety. London, UK: Springer-Verlag; 2015.
14. Mahony L, Sleeper LA, Anderson PA, Gersony WM, McCrindle BW, Minich LL, Newburger JW, Saul JP, Vetter VL, Pearson GD. The Pediatric Heart Network: a primer for the conduct of multicenter studies in children with congenital and acquired heart disease. Pediatr Cardiol. 2006; 27:191–8. [PubMed: 16261271]
15. Gaies M, Cooper DS, Tabbutt S, Schwartz SM, Ghanayem N, Chanani NK, Costello JM, Thiagarajan RR, Laussen PC, Schekerdemian LS, Donohue JE, Willis GM, Gaynor JW, Jacobs JP, Ohye RG, Charpie JR, Pasquali SK, Scheurer MA. Collaborative quality improvement in the cardiac intensive care unit: Development of the Pediatric Cardiac Critical Care Consortium (PC^4). Cardiol Young. 2014; 28:1–7. [PubMed: 24016845]
16. Jacobs ML, Jacobs JP, Franklin RC, Mavroudis C, Lacour-Gayet F, Tchervenkov CI, Walters H, Bacha EA, Clarke DR, Gaynor JW, Spray TL, Stellin G, Ebels T, Maruszewski B, Tobota Z, Kurosawa H, Elliot M. Databases for assessing the outcomes of the treatment of patients with congenital and pediatric cardiac disease – the perspective of cardiac surgery. Cardiol Young. 2008; 18(Suppl 2):101–15. [PubMed: 19063780]
17. Kugler JD, Beekman RH, Rosenthal GL, Jenkins KJ, Klitzner TS, Martin GR, Neish RS, Lannon C. Development of a pediatric cardiology quality improvement collaborative: From inception to implementation. Congenit Heart Dis. 2009; 4:318–328. [PubMed: 19740186]
18. Jacobs JP. Introduction: Databases and the assessment of complications associated with the treatment of patients with congenital cardiac disease. Cardiol Young. 2008; 18(Suppl 2):1–37.
19. Franklin RC, Jacobs JP, Krogmann ON, Beland MJ, Aiello VD, Colan SD, Elliot MJ, Gaynor JW, Kurosawa H, Maruszewski B, Stellin G, Tchervenkov CI, Walters HL, Weinberg P, Anderson RH. Nomenclature for congenital and paediatric cardiac disease: historical perspectives and The International Pediatric and Congenital Cardiac Code. Cardiol Young. 2008; 18(Suppl 2):70–80. [PubMed: 19063777]
20. Jacobs JP, Edwards FH, Shahian DM, Prager RL, Wright CD, Puskas JD, Morales DL, Gammie JS, Sanchez JA, Haan CK, Badhwar V, George KM, O'Brien SM, Dokholyan RS, Sheng S, Peterson ED, Shewan CM, Feehan KM, Han JM, Jacobs ML, Williams WG, Mayer JE, Chitwood WR, Murray GF, Grover FL. Successful Linking of the STS Database to Social Security Data to Examine Survival after Cardiac Surgery. Ann Thorac Surg. 2011; 92:32–37. [PubMed: 21718828]
21. Saleeb SF, Li WYV, Warren SZ, Lock JE. Effectiveness of screening for life-threatening chest pain in children. Pediatrics. 2011; 128:e1062–1068. [PubMed: 21987702]
22. Hammill BG, Hernandez AF, Peterson ED, Fonarow GC, Schulman KA, Curtis LH. Linking inpatient clinical registry data to Medicare claims data using indirect identifiers. Am Heart J. 2009; 157:995–1000. [PubMed: 19464409]
23. Pasquali SK, Jacobs JP, Shook GJ, O'Brien SM, Hall M, Jacobs ML, Welke KF, Gaynor JW, Peterson ED, Shah SS, Li JS. Linking clinical registry data with administrative data using indirect identifiers: Implementation and validation in the congenital heart surgery population. Am Heart J. 2010; 160:1099–1104. [PubMed: 21146664]
24. Pasquali SK, Li JS, He X, Jacobs ML, O'Brien SM, Hall M, Jaquiss RDB, Welke KF, Peterson ED, Shah SS, Gaynor JW, Jacobs JP. Perioperative methylprednisolone and outcome in neonates undergoing heart surgery. Pediatrics. 2012; 129:e385–e391. [PubMed: 22271697]

25. Pasquali SK, Li JS, He X, Jacobs ML, O'Brien SM, Hall M, Jaquiss RDB, Welke KF, Peterson ED, Shah SS, Jacobs JP. Comparative analysis of antifibrinolytic medications in pediatric heart surgery. J Thorac Cardiovasc Surg. 2012; 143:550–557. [PubMed: 22264414]

26. Pasquali SK, Jacobs ML, He X, Shah SS, Peterson ED, Hall M, Gaynor JW, Hill KD, Mayer JE, Jacobs JP, Li JS. Variation in congenital heart surgery costs across hospitals. Pediatrics. 2014; 133:e553–60. [PubMed: 24567024]

27. McHugh KE, Pasquali SK, Hall MA, Scheurer MA. Association of post-operative complications with clinical outcomes and hospital costs following the Norwood operation. Circulation. 2014; 130:A17406.

28. Riehle-Colarusso T, Autry A, Razzaghi H, Boyle CA, Mahle WT, Van Naarden K, Correa A. Congenital heart defects and receipt of special education services. Pediatrics. 2015; 136:496–504. [PubMed: 26283775]

29. Burstein DS, Jacobs JP, Sheng S, O'Brien MS, Rossi AF, Checchia PA, Wernovsky G, Welke KF, Peterson ED, Jacobs ML, Pasquali SK. Care models in congenital heart surgery and associated outcomes. Pediatrics. 2011; 127:6, e1482–e1489.

30. Vener DF, Guzzetta N, Jacobs JP, Williams GD. Development and implementation of a new data registry in congenital cardiac anesthesia. Ann Thorac Surg. 2012; 94:2159–65. [PubMed: 23176940]

31. Martin GR, Beekman RH, Ing FF, Jenkins KJ, McKay CR, Moore JW, Ringel RE, Rome JJ, Ruiz CE, Vincent RN. The IMPACT Registry: Improving Pediatric and Adult Congenital Treatments. Semin Thorac Cardiovasc Surg Pediatr Card Surg Annu. 2010; 13:20–25. [PubMed: 20307857]

32. Gaies MG, Jeffries HE, Niebler RA, Pasquali SK, Donohue JE, Yu S, Gall C, Rice TB, Thiagarajan RR. Vasoactive inotropic score is associated with outcome after infant cardiac surgery: An analysis from the Pediatric Cardiac Critical Care Consortium (PC^4) and Virtual PICU System Registries. Pediatr Crit Care Med. 2014; 15:529–37. [PubMed: 24777300]

33. Frobert O, Lagerqvist B, Olivecrona GK, Omerovic E, Gudnason T, Maeng M, Aasa M, Angeras O, Calais F, Danielewicz M, Erlinge D, Hellsten L, Jensen U, Johansson AC, Karegren A, Nilsson J, Robertson L, Sandhall L, Sjogren I, Ostlund O, Harnek J, James SK. Thrombus aspiration during ST-segment elevation myocardial infarction. N Engl J Med. 2013; 369:1587–97. [PubMed: 23991656]

34. Rao SV, Hess CN, Barham B, Aberle LH, Anstrom KJ, Patel TB, Jorgensen JP, Mazzaferri EL, Jolly SS, Jacobs A, Newby LK, Gibson CM, Kong DF, Mehran R, Waksman R, Gilchrist IC, McCourt BJ, Messenger JC, Peterson ED, Harrington RA, Krucoff MW. A registry-based randomized trial comparing radial and femoral approaches in women undergoing percutaneous coronary intervention: the SAFE-PCI for Women (Study of Access Site for Enhancement of PCI for Women) trial. JACC Cardiovasc Interv. 2014; 7:857–67. [PubMed: 25147030]

35. Hall D, Huerta MF, McAuliffe MJ, Farber GK. Sharing Heterogeneous Data: The National Database for Autism Research. Neuroinformatics. 2012; 10:331–339. [PubMed: 22622767]

36. Johnson SB, Whitney G, McAuliffe M. Using Global Unique Identifiers to Link Autism Collections. J Am Med Inform Assoc. 2010; 17:689–695. [PubMed: 20962132]

37. Forrest CB, Margolis P, Seid M, Colletti RB. PEDSnet: How a prototype pediatric learning system is being expanded into a national network. Health Affairs. 2014; 33:1171–1177. [PubMed: 25006143]

38. Forrest CB, Margolis PA, Bailey LC, Marsolo K, Del Beccaro MA, Finkelstein JA, Milov DE, Vieland VJ, Wolf BA, Yu FB, Kahn MG. PEDSnet: a National Pediatric Learning Health System. J Am Med Inform Assoc. 2014; 21:602–6. Epub 2014 May 12. DOI: 10.1136/amiajnl-2014-002743 [PubMed: 24821737]

39. Go AS, Magid DJ, Wells B, Sung SH, Cassidy-Bushrow AE, Greenlee RT, Langer RD, Lieu TA, Margolis KL, Masoudi FA, McNeal CJ, Murata GH, Newton KM, Novotny R, Reynolds K, Roblin DW, Smith DH, Vupputuri S, White RE, Olson J, Rumsfeld JS, Gurwitz JH. The Cardiovascular Research Network: A new paradigm for cardiovascular quality and outcomes research. Circ Cardiovasc Qual Outcomes. 2008; 1:138–147. [PubMed: 20031802]

40. McGlynn EA, Lieu TA, Durham ML, Bauck A, Laws R, Go AS, Chen J, Feigelson HS, Corley DA, Young DR, Nelson AF, Davidson AJ, Morales LS, Kahn MG. Developing a data

infrastructure for a learning health system: the PORTAL network. J Am Med Inform Assoc. 2014; 21:596–601. [PubMed: 24821738]

41. Curtis LJ, Weiner MG, Boudreau DM, Cooper WO, Daniel GW, Nair VP, Raebel MA, Beaulieu NU, Rosofsky R, Woodworth TS, Brown JS. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. Pharmacoepidemiol Drug Saf. 2012; 21:23–31. [PubMed: 22262590]
42. Atchley KD, Pappas JM, Kennedy AT, Coffin SE, Gerber JS, Fuller SM, Spray TL, McCardle K, Gaynor JW. Use of administrative data for surgical site infection surveillance after congenital cardiac surgery results in inaccurate reporting of surgical site infection rates. Ann Thorac Surg. 2014; 97:651–7. [PubMed: 24365216]
43. Schumacher KR, Stringer KA, Donohue JE, Yu S, Shaver A, Caruthers RL, Zikmund-Fisher BJ, Fifer C, Goldberg C, Russell MW. Social media methods for studying rare diseases. Pediatrics. 2014; 133:e1345–53. [PubMed: 24733869]
44. [Accessed 4/28/2015] Improving patient outcomes after radical prostatectomy. Available at: http://musicurology.com/pro/
45. [Accessed 5/21/2015] The world's top 10 most innovative companies in big data. Available at: http://www.fastcompany.com/most-innovative-companies/2014/industry/big-data

Table 1

Strengths and weaknesses of the current CHD data environment.

Strengths	Weaknesses
Numerous data sources	Data silos
Existing local infrastructure	Lack of comprehensive data integration
Multi-institutional collaboration Standardized nomenclature	Lack of a collaborative data governance model
Application of many data linkage techniques	Limited infrastructure to support wide scale data integration and analysis Limited longer-term outcomes data

CHD = congenital heart disease

Table 2

CHD data integration activities to date.

Method	Examples of linked data
Linking on unique identifiers	clinical data + survival data
Linking on indirect identifiers	registry + administrative/cost data clinical trial + administrative/cost data
Center-level linkages	hospital survey data + registry data
Collaboration/partnering between databases	registries with shared platforms, variables, definitions
Supplementary data modules to main registry	registry data + research modules

CHD = congenital heart disease