Big Heart Data: Advancing Health Informatics through Data Sharing in Cardiovascular Imaging

Avan Suinesiaputra¹, Pau Medrano-Gracia¹, Brett R. Cowan¹, and Alistair A. Young¹

¹Dept. of Anatomy with Radiology, University of Auckland, New Zealand

The burden of heart disease is rapidly worsening due to increasing prevalence of obesity and diabetes. Data sharing and open database resources for heart health informatics are important for advancing our understanding of cardiovascular function, disease progression and therapeutics. Data sharing enables valuable information, often obtained at considerable expense and effort, to be re-used beyond the specific objectives of the original study. Many government funding agencies and journal publishers are requiring data re-use, and are providing mechanisms for data curation and archival. Tools and infrastructure are available to archive anonymous data from a wide range of studies, from descriptive epidemiological data to gigabytes of imaging data. Meta-analyses can be performed to combine raw data from disparate studies to obtain unique comparisons or to enhance statistical power. Open benchmark datasets are invaluable for validating data analysis algorithms and objectively comparing results. This review provides a rationale for increased data sharing and surveys recent progress in the cardiovascular domain. We also highlight the potential of recent large cardiovascular epidemiological studies enabling collaborative efforts to facilitate data sharing, algorithms benchmarking, disease modeling and statistical atlases.

Index Terms—Cardiac MRI, cardiac atlas, anatomical models, data sharing.

I. INTRODUCTION

CARDIOVASCULAR disease (CVD) is the world’s leading cause of morbidity and mortality. Healthcare costs from CVD are rising due to the increasing prevalence of obesity, diabetes and metabolic syndrome. The results of the Global Burden of Disease Study show that ischemic heart disease was the leading cause of disability-adjusted life years in 2010 [1]. In many countries, obesity is likely to overtake tobacco as the leading risk to health by 2016 [2].

The ability to integrate data from multiple sources and across many variables has significant potential for the evaluation and treatment of patients. These methods enable deeper characterization of a particular patient, and more precise mapping to similar patients in pertinent subpopulations [3]. Efficient characterization of subpopulations requires data harmonization, new analytical algorithms and analysis of heterogeneous data. This approach relies on data sharing infrastructure and high-throughput database resources in large epidemiological studies as well as in small experimental studies, which would otherwise require a large cost or effort to reproduce.

Data sharing for scientific advancement is well established in many fields. In genomics, databases have existed for many years for archiving and curating data [4]. In neuroscience, large repositories are available for specific disease groups or general atlases [5]. Efforts are now being made to facilitate data sharing in the cardiovascular domain [6]. This review highlights the rationale and the need for data sharing in the cardiovascular research community, and outlines current efforts to provide infrastructure and tools for data sharing. Examples are given by atlas-based analysis of cardiac remodeling and ongoing community efforts to establish benchmarking platforms for cardiac image analysis.

II. LARGE-SCALE CARDIOVASCULAR POPULATION STUDIES

A large part of our current understanding of multivariable risk factors in the etiology of CVD arose from the Framingham Heart Study, which began in 1948 and is still continuing with over 1,200 publications. The Framingham scores assess cardiovascular risk factors of hypertension, smoking, lipid profile, obesity, diabetes and inactivity; the pathophysiological progression from hypertension to heart failure; the relationship between atrial fibrillation and stroke; and the value of population-based longitudinal studies [7]. With the advent of large-scale databases and data mining methods, it has recently become possible to identify relationships across many different types of information [8].

A wealth of data is now available from non-invasive cardiac imaging examinations. Prospective longitudinal studies derived from large-scale imaging data have been established to investigate the pathogenesis of cardiac diseases [11], [16]. Longitudinal studies, which follow patients over the years, enable scientists to understand the evolution of cardiac disease from sub-clinical manifestations to clinical symptoms. The combination of imaging data with other diagnostic information and biomarkers gives a new rich field of big heart data (defined for the purposes of this review as imaging studies combined with clinical assessments in large study cohorts on an unprecedented scale). These present significant opportunities for new discoveries to reduce the burden of CVD.

Table I lists existing large-scale cardiovascular studies with collection of gigabytes of imaging data. The Multi-Ethnic

---
TABLE I

MAJOR CARDIOVASCULAR EPIDEMIOLOGICAL STUDIES WITH THE INCLUSION OF IMAGING DATA.

<table>
<thead>
<tr>
<th>Study</th>
<th>Country</th>
<th>Images</th>
<th>Start</th>
<th>Size</th>
<th>Group</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>MESA [9]</td>
<td>USA</td>
<td>ECG, CT, MRI</td>
<td>2000</td>
<td>6,814</td>
<td>Hispanic, Chinese, White, African-American</td>
<td>45–84</td>
</tr>
<tr>
<td>JHS [10]</td>
<td>USA</td>
<td>ECG, CT, MRI</td>
<td>2000</td>
<td>5,302</td>
<td>African-American</td>
<td>21–84</td>
</tr>
<tr>
<td>UK Biobank [11]</td>
<td>UK</td>
<td>MRI, ECG, DEXA</td>
<td>2006</td>
<td>100,000</td>
<td>Multi</td>
<td>40–69</td>
</tr>
<tr>
<td>CPTP [12]</td>
<td>Canada</td>
<td>MRI</td>
<td>2009</td>
<td>10,000</td>
<td>Multi</td>
<td>35–69</td>
</tr>
<tr>
<td>Framingham Offspring [14]</td>
<td>USA</td>
<td>MRI</td>
<td>2002</td>
<td>1,707</td>
<td>Multi</td>
<td>NA</td>
</tr>
</tbody>
</table>

Study of Atherosclerosis (MESA) study was initiated in 2000 to focus on the manifestation of subclinical to clinical CVD before signs and symptoms develop in the heterogeneous population of the US [9]. The study has sampled 6,814 men and women aged 45–84 years old across six centers. The analysis of the 10 year follow-up has recently been completed for around 3,000 participants [17]. The MESA study was the first major population-based study to use cardiac MRI [18]. Data sharing is provided by the MESA study through their ancillary studies and publication protocols.

The Jackson Heart Study (JHS) was designed to determine the root causes of CVD in 5,302 African-American individuals aged 21–84 years living in the southeastern USA (Jackson, Mississippi) [10], [19], [20]. This group experiences increased mortality from CVD as well as higher incidence of hypertension, obesity and diabetes. A unique aspect of JHS was the association of neighborhood disadvantages in terms of economic, sociocultural, behavioral, dietary and physical activity measures with cardiometabolic risk factors [21], [22]. The JHS provides a rare insight into the interactions between genotype and phenotype in a high-risk population. In terms of imaging data, the JHS has collected 3,000 CT and MRI examinations containing heart function and calcium scores [23].

Other studies such as the Canadian Partnership for Tomorrow Project (CPTP) with 10,000 participants, ICELAND MI with 936 patients with myocardial infarction, and the Framingham Offspring study with 1,707 follow-up cardiac MRI scans added more imaging data for cardiac research. While not a study per se, the EuroCMR registry has the main goal to evaluate the diagnostic potential of cardiac MRI as well as its cost-effectiveness. More than 27,000 consecutive patients have been enrolled from 57 centers in 15 countries. Similarly, the Global CMR registry has been established to collate MRI patient data from around the world with 44,000 cases contributed to date.

III. SHARING MEDICAL DATA

With the ability to mine multidimensional relationships in diverse datasets, sharing patient medical records across clinical providers and researchers becomes a key factor to improve quality of care, to reduce healthcare costs and to minimize human errors [24]. However, sharing medical data remains a big challenge due to the extremely granular nature of healthcare data. Hospitals, imaging centers, healthcare institutions and clinics must enforce strict guidelines to uphold patient data privacy and security. Therefore, central to enable medical data sharing is a secured protection of patient data, such as mandated by the HIPAA (Health Insurance Portability and Accountability Act) in the US or the Data Protection Act in the UK. Ensuring patient data privacy is vital to sharing medical records to a wider community.

In general, there are three entities in data sharing: data contributors, organizers and users (see Fig. 1). The data contributor is a provider who wants to release their data to a wider research community without breaching agreements with patients and other clinical interests. Data contributors must ensure informed consent is obtained, which is compatible with data sharing, in conjunction with local ethics committee or institutional review board (IRB) approval. They can only contribute anonymized or de-identified data. The de-identification process removes any protected health information that can be used to identify an individual, e.g. name, dates (except...
TABLE II
EXISTING MEDICAL DATA SHARING INFRASTRUCTURES THAT INCLUDE CARDIAC DATA

<table>
<thead>
<tr>
<th>Name</th>
<th>Domain</th>
<th>Accessibility</th>
<th>Studies (size)</th>
<th>Data type</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAP</td>
<td>Cardiac</td>
<td>Registration</td>
<td>Asymptomatic (2,450)</td>
<td>MRI</td>
<td><a href="http://www.cardiacatlas.org">www.cardiacatlas.org</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Guest</td>
<td>Myocardial infarction (465)</td>
<td>Finite element</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Normal human heart atlas (1)</td>
<td>Contours</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mixed normal, hypertrophy, heart failure (45)</td>
<td>Binary images</td>
<td></td>
</tr>
<tr>
<td>AMDB</td>
<td>Cardiac</td>
<td>Registration</td>
<td>Breast MRI (153)</td>
<td>MRI</td>
<td><a href="http://www.amdb.isd.kcl.ac.uk">www.amdb.isd.kcl.ac.uk</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Guest</td>
<td>Clinical logs (62)</td>
<td>Test files</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Radiology reports (2,363)</td>
<td>CT</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CT colonography (103)</td>
<td>laboratory test</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Diabetes mellitus data (17)</td>
<td>data (13,000)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lung CT (398+)</td>
<td>ICU record (53)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Coronary artery disease data (20,000)</td>
<td>Motion sensor data</td>
<td></td>
</tr>
<tr>
<td>CVRG</td>
<td>Cardiac</td>
<td>Open</td>
<td>Normal canine heart (7)</td>
<td>MRI</td>
<td>cvgrid.org</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Failing canine heart (5)</td>
<td>CT</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Normal human (1)</td>
<td>Laboratory test</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Canine heart atlas (1)</td>
<td>data (13,000)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ischemic human CT (13)</td>
<td>ICU record (53)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Non-ischemic human CT (12)</td>
<td>Motion sensor data</td>
<td></td>
</tr>
<tr>
<td>PhysioNet</td>
<td>Cardiac</td>
<td>Open</td>
<td>European ST-T database (79)</td>
<td>ECG</td>
<td><a href="http://www.physionet.org">www.physionet.org</a></td>
</tr>
<tr>
<td></td>
<td>Brain</td>
<td></td>
<td>Long-term ST database (43)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Biomedical signals</td>
<td></td>
<td>Arhythmia database (23)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Noise stress test (14)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Heart rate database (10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Congestive heart failure (15)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Partial epilepsy (7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VIP</td>
<td>Various organs</td>
<td>Registration</td>
<td>Simulated data</td>
<td></td>
<td><a href="http://www.creatis.insa-lyon.fr/vip">www.creatis.insa-lyon.fr/vip</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unspecified</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The organizer may also be involved in a process to enrich the original patient data to make the data more useful for broader research activities. Such data enrichment activities include ontological annotations, manual labeling, model fitting and population analysis (see Fig. 1).

User responsibility can be defined by a data distribution agreement, which dictates the specific type of research, commitment to secure the data within their host computer, and a potential intellectual property agreement. All conditions or restrictions on data use should be specified in the data distribution agreement.

IV. EXISTING CARDIAC DATA SHARING

In the cardiac domain, several efforts have been initiated to establish infrastructures for data sharing by providing anonymized baseline examinations, imaging data and other derived computations, such as anatomical models and statistical shape analysis. Table II summarizes existing medical data sharing infrastructures, which also include cardiovascular data.

Established in 2010, the Cardiac Atlas Project (CAP) is a worldwide consortium to host large cardiac image data with derived finite element models of the heart and associated diagnostic information [6]. Over 3,000 cardiac MRI cases have been contributed to the database, which are being used by more than 20 research groups worldwide for various research activities. CAP is a registered ancillary study of MESA and has developed methods to translate results between MESA and other studies using atlas-based bias correction methods [32]. Patient-specific models of the heart provide a standard coordinate system, which map the heart according to anatomical location. CAP has developed methods to pool data from different sources in a standardized manner, and...
to correct bias arising from imaging or analysis protocol. CAP is endorsed by the Society of Cardiovascular Magnetic Resonance, which maintains an upload site where cases can be contributed to the atlas project [33].

The Anatomical Models Database (AMDB) is a web accessible framework to share and reuse cardiovascular models [30], [31]. AMDB stores various cardiac geometry models, accessible for any researchers to perform a simulation or a benchmarking study in cardiac electrophysiology and mechanics. The aim of AMDB (formally known as euHeart) is to allow multi-scale computational modeling of the heart [34]. AMDB employs a web service tool that personalizes a geometric heart model given binary images of the heart and cardiac parameters [35].

The iDASH (integrated data analysis, anonymization and sharing) is a general framework to provide a scalable tool to share and access medical data, including cardiac diseases [36]. The iDASH stores and shares heterogeneous data from multiple domains (see Table II). A specific infrastructure that shares only ECG signals is provided by PhysioNet. Currently, PhysioNet stores and freely shares large scale de-identified recorded physiologic signals, time series and other related biomedical data [37]. These are healthy subjects and patients with various diseases, including sudden cardiac death, congestive heart failures, sleep apnea, aging, epilepsy and gait disorder. The CardioVascular Research Grid (CVRG) provides an infrastructure to securely and seamlessly access complex data of cardiovascular studies, including an automated ontological labeling of anatomical differences [38]. CVRG is currently establishing an easy access platform with cloud-based and browser-based tools without the need to install complex software. The Virtual Imaging Platform (VIP) is another openly accessible online data sharing platform, which focuses more on computationally extensive biomedical simulation processes [39].

V. APPLICATION: POPULATION-BASED CARDIAC REMODELING

The human heart is continuously remodeling (changing shape) in response to pathology, aging, environmental and genetic factors. Cardiac remodeling can be maladaptive when linked to heart failure progression [40], [41], but remodeling can be adaptive during normal growth or intensive physical exercises. Even in the early stages of heart failure, adaptive remodeling can be observed because the heart maintains its function in spite of pressure or volume overloading in the acute phase of cardiac injury [42]. There is a transition from adaptive to maladaptive remodeling in the progression of heart disease, but when and how this transition occurs still remains unknown [43]. As illustrated in Fig. 2, population-based studies that combine cardiac shape, function and other clinical data can characterize these remodeling processes.

As the heart remodels, its geometry, mass, composition and volume changes. The shape of the heart can become less elliptical and more spherical [44]. Ventricular sphericity (width to height ratio) has been observed in symptomatic patients and associated with decreased survival [45] and adverse remodeling [46]. Increased LV chamber dimension [47], lower systolic dimension change [48] and hypertrophy [49] have also been observed in asymptomatic individuals. Fig. 3 shows how LV size and sphericity were increased from normal volunteers to patients with heart failure in a subset of cases from CAP.

The understanding of cardiac remodeling processes is particularly important in quantifying effects of treatment and reverse remodeling [50]. Several studies have been actively investigating how remodeling occurs in the population by using statistical shape and function tools [51]–[56]. For instance, principal component analysis (PCA) on the cardiac shapes of 102 adults born preterm (30 weeks) and 102 age matched healthy volunteers have shown different shape indices in adults who were born preterm, compared with normal term controls [53]. A statistical cardiac atlas built from 1991 asymptomatic individuals has shown a relationship between traditional risk factors (hypertension, smoking, sex and diabetes) with LV remodeling [56]. These studies have demonstrated the power of geometrical heart shape models to identify cardiac remodeling in certain sub-populations or in later life.

VI. APPLICATION: BENCHMARKING

A major advantage of data sharing is the availability of standardized datasets that can be used to develop, validate and compare (benchmark) new automated methods. Benchmarking is important to evaluate the relative efficacy of the growing number of proposed automated methods. Objective comparisons are more difficult to perform if each published method presents results using private data and incompatible platforms.

Community efforts have been made to provide platforms for benchmarking on existing and emerging algorithms through a series of “challenges”. A left ventricular segmentation challenge was held in 2009 using 45 cases from a mixed patient dataset (normal, heart failure, myocardial infarction and hypertrophy) [57]. These data are in the public domain and can be accessed directly via the Cardiac Atlas Project website. A similar challenge was run in 2011 with more cases (200 patients with myocardial infarction) [58], which subsequently established a platform for collaboration between research groups to define a common consensus ground truth for the myocardial segmentation [27]. Other benchmarking frameworks utilizing cardiac data sharing include cardiac motion [59], segmentation of scar tissue in the left atrium [60], and automated extraction of centerlines in coronary arteries [61]. Future challenges and benchmarking studies are planned, including cardiac mechanics, electrophysiology and correcting breathing motion artifacts in perfusion MR images.

VII. CHALLENGES

The main challenge in big heart data studies lies primarily on the willingness of providers to release data anonymously to a wider community. Fortunately, there is an ongoing progression towards open access data required by national funding agencies for new research projects, such as the National Institutes of Health in the US and the National Health Services in the UK. Reproducibility of research outcomes is
the main driving force of this research direction. The onus is now on organizers to provide robust and user-friendly data sharing platforms. They must guarantee to both providers and the general public that patient data have been de-identified properly, data integrity are maintained, and data dissemination occurs in the most secure way.

Standardization is a key technical challenge in data sharing. The DICOM standard for storing pixel and non-pixel data (patient information, 3D image geometry, examination records, acquisition parameters, etc.) can change from vendor to vendor or even software version. Picture Archiving and Communication System (PACS) allows radiographic images being stored and transferred electronically within and between hospitals. Open source PACS architectures, such as dcm4chee [62], enable quick deployment of a web-accessible imaging data server for public use. Data standardization is even more difficult for non-radiographic image data. Initial efforts have been proposed, including the standardization of a modeling language (FieldML) to link different organs and tissues [63], an open electrocardiogram (ECG) format [64] and cardiac electrophysiology (EP) key data elements [65].

VIII. FUTURE PERSPECTIVES

A. Novel geometry-based cardiovascular risk predictors

Of all various pathways to end-stage CVD, the primary preventive measure is to reduce risk factors and to prevent early disease progression. Improving major risk factors has led to considerable reduction in morbidity and mortality due to CVD, but yet CVD remains the leading cause of death and disability in developed countries. Traditional CVD risk assessment currently disregards the geometrical shape of the heart, which leads to incomplete information about how the heart remodels over time in a matched sub-clinical group, global and regional functional parameters and other biomechanical properties. Early diagnosis and risk stratification of individuals at risk can be improved by augmenting current risk factors predictors [66], a process which can be accelerated by sharing cardiac data.

B. Disease modeling

The availability of cardiac models enables development of computer-aided diagnosis applications for cardiac diseases [32], [67]. Subtle differences in cardiac motion, which are inherently difficult to assess, can be objectively quantified by a statistical model of cardiac motion in the population. Such applications will allow better tracking and monitoring of the progression of a cardiac disease over time. This becomes essential in a rapidly progressive disease such as pulmonary arterial hypertension, where regular assessment is needed before tailoring the treatment in each patient [68].

C. Personalization

Collecting individual data gains statistical knowledge about the population; a valuable resource to be brought back to individuals [69] by means of a personalized treatment [34], [70], [71]. Preliminary studies in this frontier include the prediction of fiber orientation in the left ventricle [72], heart
geometric personalization [73], [74], predicting cardiac resynchronization response [75] and generating an artificial patient-specific model for pre-surgical planning [76]. Personalized medicine will gain more attention in the future and will be more tightly connected with the availability of big heart data.

IX. Conclusion

Data sharing is becoming essential to the further development of the field of cardiovascular research. Large databases are now being developed which bring together disparate types of data and enable multi-variate greater-depth analysis of information on each patient. In particular, population-based tools have proven useful for comparing one individual’s anatomy to a population or between sub-populations, and in the detection, quantification and monitoring the disease progression. Data-sharing initiatives will aid in the validation and further automation of data analysis methods.

REFERENCES


