BMJ Innovations - *Commentary*

TITLE

Why democratize bioinformatics?

Gabriella Captur[1,2] MD PhD MRCP MSc, Rodney H Stables[3] DM FRCP, Dennis Kehoe[4] PhD BSc, John Deanfield[5] BA FRCP, James C Moon[2,6,7] MD MBBS MRCP (corresponding author).

1 UCL Biological Mass Spectrometry Laboratory, Institute of Child Health and Great Ormond Street Hospital, 30 Guilford Street, London, UK
2 NIHR University College London Hospitals Biomedical Research Centre, London, UK
3 Liverpool Heart and Chest Hospital, Liverpool, UK
4 Aimes Grid Service Providers Ltd, Fairfield, Liverpool UK
5 Farr Institute of Health Informatics Research at London, London, UK; National Institute of Cardiovascular Outcomes Research, University College London, London, UK
6 UCL Institute of Cardiovascular Science, University College London, Gower Street, London, UK
7 Barts Heart Centre, The Cardiovascular Magnetic Resonance Imaging Unit and The Centre for Rare Cardiovascular Diseases Unit, St Bartholomew's Hospital, West Smithfield, London, UK

CORRESPONDING AUTHOR
Prof James C Moon
Institute of Cardiovascular Science,
University College London,
Gower Street,
London WC1E 6BT, UK

E-mail: j.moon@ucl.ac.uk Phone No: +44 2034563081 Fax No: +0203 456 3086

WORD COUNT 1,900

KEY WORDS
Bioinformatics, data-sharing, standardisation, cardiovascular magnetic resonance, cardiac imaging

ABSTRACT
Network bioinformatics and web-based data collection instruments have the capacity to improve the efficiency of the UK's appropriately high levels of investment into cardiovascular research. A very large proportion of scientific data falls into the long-tail of the cardiovascular research distribution curve, with numerous small independent research efforts yielding a rich variety of specialty data sets. The merging of such myriad datasets and the eradication of data silos, plus linkage with outcomes could be greatly facilitated through the provision of a national set of standardised data collection instruments—a shared-cardioinformatics library of tools designed by and for clinical academics active in the long-tail of biomedical research. Across the cardiovascular research domain, like the rest of medicine, the national aggregation and democratization of diverse long-tail data is the best way to convert numerous small but expensive cohort data sources into big data, expanding our knowledge-base, breaking down translational barriers, improving research efficiency and with time, improving patient outcomes.

MAIN TEXT

Background

Within clinical research institutions across the UK currently, only a small proportion of generated data is effectively being captured and safely stored long term; research efforts are fragmented and the challenges of multi-centre collaboration are not yet overcome. A shared national initiative of accessible and secure bioinformatics solutions tailored to the needs of junior and senior clinical-academics has the potential to address this unmet need and cardiovascular research provides a clear example.

Cardiovascular disease is a leading public health problem and a number one killer in the UK accounting for 40% of all national deaths and costing the UK economy £29bn a year in healthcare expenditure and lost productivity. The UK spends more of its healthcare budget on cardiovascular disease and research than any other EU economy[1,2]. Over the last 20 years, there has been an explosive growth in cardiovascular investigations, imaging and therapies across the National Health Service (NHS) underpinning clinical care but also the >£117 million annual research investment[3] that creates expensive clinical cohorts[4]. There is a pressing need to merge and curate (for at least 10 years) not only the large well-organised big cardiac science datasets[5–8] but also the richly diverse and heterogeneous smaller cohort data sets produced by small groups and individual cardiologists, the so-called long-tail data[9] (**Figs. 1 and 2**)—the large proportion of scientific data that falls into the long-tail of the distribution curve[10]; a product of the numerous small independent research efforts yielding a rich variety of specialty cardiac research data sets. The extreme right portion of the long-tail includes unpublished dark data: siloed databases locked up in applications, null findings, laboratory notes, log archives, untagged image files, animal care records, etc.[9] Dark data in cardiology can be illuminating but it is often inaccessible to the outside world. The merging of such myriad datasets and the eradication of data silos, plus linkage with outcomes could be greatly facilitated through the provision of a national set of standardised data collection instruments—a shared-cardioinformatics library of tools designed by and for clinical academics active in the long-tail of cardiovascular research. Such bioinformatics set-up costs are high, usually placing them beyond a single centre's capabilities, which is why a national cross-centre initiative is required.

Large national initiatives aggregating registry data like that led by the National Institute for Cardiovascular Outcomes Research (NICOR)[11] are testament to the fact linkage of national cardiovascular databases is feasible and has the potential for increased international comparative data analysis. However, NICOR infrastructure is not tailored to serve the needs of individual researchers aiming to conduct small-to-medium scale cardiovascular research in the cloud. The doctoral student with a sample size of 100 curating a 3-year project with finite funding needs accessible bioinformatics tools that he/she can customise and control. Electronic bioinformatics tools for these groups are usually limited to those provided locally by universities but such institutional databases are not easily accessible to collaborators in other centres. These investigators (sometimes junior staff) need access to secure but intuitive electronic data collection solutions that they can customize to the needs of their niche project. They need a simple but hierarchal way of controlling access, freedom to edit instruments and ease of data export to permit local statistical analysis.

Advantages of shared infrastructures for research in the long tail

Web-based data collection instruments have the capacity to improve the efficiency of the UK's appropriately high levels of investment into cardiovascular research. A national initiative, as opposed to segregated single-centre university-based infrastructures, automatically creates dissemination standards not through imposition, but because tools will be genuinely good, easy to use, accessible and practically helpful.

Cardiac research in the UK requires and receives high levels of funding to create expensive patient cohorts but these cohorts are typically non-standardized, partitioned to reflect the group's niche expertise and data is rarely curated long term nor integrated with outcomes. From concept to guideline and then through to clinical practice, takes many steps. Disseminated cloud-based bioinformatics broadens the range of translation that any individual research group can singly perform, facilitating the transition of ideas along the translational pathway (e.g. from single-centre cohort, to multi-centre, to outcome-studies, to standardization, to guidelines). Standardized data collection, growing sample size, linking to other domains of science and then trickling results between groups, suddenly become easy and information governance

strengthened. Infrastructure re-use becomes possible and new areas of research are spawned through linkages, previously unachievable, leading to diffused benefits.

The release of a browser-based, flexible and secure electronic-data capture (EDC) infrastructure automatically encourages research groups to share (data, instruments and dictionaries). Expensive multi-centre UK cohort datasets may be securely accessed from any part of the country and robustly de-identified, standardized, curated, and merged with other sorts of data for maximum scientific yield. With this infrastructure of "connectedness", collaboration is suddenly easier providing a sustainable route to creating large-scale cardiac data[12] and increasing the yields from UK research investment by accelerating the transition of a scientific idea into a new biomarker, clinical test or patient therapy. From scientific concept to societal benefit is a multistep process and network bioinformatics are needed along the pathway to impacts—academics may conceive ideas but teams, small-medium enterprises and pharmaceutical companies need to input and connect as ideas evolve.

For research exploring the development of novel cardiac biomarkers, data sharing in the long-tail is key as it ensures research transparency, mitigates against known biases in publication and increases data reuse by third parties[10]. Effect sizes need to be measured in Phase-I/II drug studies, but real world disease and real world biomarker performance needs to be measured for Phase-III and this is where unexpected trial futility is often discovered (globally, the last 20 Phase-III trials in heart failure have been negative[13] at a waste of billions)—a potential "regression to the truth" as cardiac biomarkers exit the expert centres and real world data handling commences. Understanding and anticipating the size of this "real world effect" is hard without access to multi-centre, unselected pan-UK patient cohorts managed and curated using standardised bioinformatics tools at national level.

The time is right—Important developments in UK health informatics

The promotion of bioinformatics assets to support long-tail research in the UK coincides with the NHS' growing appetite for information technology (IT) innovation and its growing focus on the procurement of smarter health informatics strategies. Several trusts are currently undergoing major transformational change and investing in 'Health Clouds' as funding constrictions drive health services to seek more efficient paperless reconfigurations, reduce complexity, improve data security, and drive up the quality of patient services. National bodies like the Commissioning Support Units and the Health and Social Care Information Centre (HSCIC) have been established to support this process. Healthcare clouds permit efficient electronic health information exchange (HIE) allowing providers to rapidly and securely access and share a patient's medical information electronically but this process is dependent on data standardization[14]. Once standardized, the data transferred can seamlessly integrate into a recipients' Electronic Health Record (EHR). The Open EHR vision for UK healthcare aims to create life-long interoperable patient EHRs, a key-stone component of which is semantic interoperability[15] made possible through the CEN/ISO EN13606—a European norm for semantic interoperability in the EHR communication, approved by the European Committee for Standardization (CEN) and by the International Organization for Standardization (ISO)[16]. Open EHR and NHS health cloud technologies have major research ramifications—long-tail research instruments will be able to piggy-back onto this broader, evolving national infrastructure, permitting flexible spin up of resources as and when needed ("power-by-the-hour") and self-provisioning (studies can be containerised and rapidly deployed and re-used).

Potential caveats of sharing in the long-tail

Merging myriad datasets potentially introduces the risk of re-analysis of poor quality datasets or analysis of excellent datasets by non-experts using inappropriate applications, thus flooding the field with conflicting results[17]. There is also the financial cost and time investment involved in preparing data and data collection instruments to permit their use by others but shared standardized tools once developed will avoid this issue and deliver superior research network intensity. Cardiac researchers will dedicate enormous time preparing papers for publication driven by citation and H-index incentives, to the satisfaction of funders and to ensure survival of their teams and centres but the career yields from large-scale data sharing (especially of dark data) are not that explicit[18]. Investigators are at the mercy of the work ethic and replication etiquette[19] of analysing third parties, co-authorship on downstream publications may be sporadic, there is commonly a sense of loss of control. Furthermore, data sharing could expose

data errors or suboptimal reporting practices in high-impact studies many years after their original publication. If clinical guidelines had incorporated such data as evidence for patient care the implications could be devastating[20].

Example solution for shared cardioinformatics and future directions

In a pan-UK effort to tackle the barrier to multi-centre data integration in the long-tail, starting with cardiology, our group has previously partnered with IT architects (not-for-profit organisation AIMES Grid Service Providers, www.aimes.uk) to deploy a cloud hypervisor pilot that provides easy-to-access bioinformatics tools for UK academics in the cardiovascular sciences. The primary EDC instrument used was REDCap (Research Electronic Data Capture [21] distributed non-commercially by Vanderbilt University for academia)—a simple proprietary, user-friendly, no-cost, browser-based, metadata-driven system for data collection and management available to academics. It has several obvious advantages over competing infrastructures like standard office applications (Microsoft Excel and Access) or other EDC systems like Open Source OpenClinica™. Learning OpenClnica is more difficult than for REDCap with fewer online training modules pages and no international Consortium to turn to for support; there is no project development mode so undoing or replacing fields during set-up is cumbersome. REDCap permits advanced customization through the use of hooks, hacks, application programming interfaces (API) and plugins offering a flexible way of adding micro-features and widgets to research projects with specific user-driven requirements (e.g. our in-house 17-segment cardiac bulls-eye plot for efficient regional wall motion scoring: http://www.cardioproject-redcap.org/MAPSTER/cardiocalc_wma.htm). Another benefit is the ability to combine REDCap directly with R[22] through APIs which can easily export and import data into R, reducing the burden of data transformation and the possibility of human errors while streamlining the entire process of data collection, cleaning and analysis[23].

This UK model, committed to the usual STAndards for the Reporting of Diagnostic accuracy studies (STARD)[24], has been designed around the baseline skillset, aptitudes and needs of the everyday principal investigator and his/her junior/senior research team. Its uptake has been exponential due to its ease of customization and set-up efficiency particularly for junior academics starting a new project. It currently supports 250 UK researchers, with a total of 105 projects actively recruiting. It is provisioned in a safe-haven environment consisting of a G-Cloud-assured high-availability cluster with a disaster recover element existing in a separate G-Cloud set up. Both are located within a highly secured data centre information security management system (ISMS) managed by AIMES and in compliance with the standards of the ISO/International Electrotechnical Commission and with the Health Insurance Portability and Accountability Act (HIPAA). AIMES is an accredited N3 cloud provider on the UK Government's G-Cloud procurement framework. The platform is designed to accept the upload of data free of personally identifiable information deriving from research projects that already have the necessary ethical approvals in place. Backup power, backup servers and data restore facilities provisioned are fully compliant to NHS Information Governance (IG) requirements. In handling research data, the platform is aligned with good clinical practice and the UK Data Protection Act (1998). Investigators applying for grants and research ethics approval, interested in using this infrastructure, are provided with boilerplate verbiage outlining its data security features and with access to online and face-to-face EDC training sessions. Investigators using the platform retain responsibility for obtaining appropriate consent from participants and they are asked to verify this in the mandatory User Responsibility Document when registering for access to the system. In collaboration with professional ISMS managers at AIMES experienced in patient data security, IS027001 standards and the UK NHS IG Governance ISMS, we have established processes to ensure full respect for ethics and research governance across the pilot, relevant to participants and participating researchers. The pilot is currently awaiting registration as a database platform with the UK Research Ethics Committee.

The vision for this pilot has grown out of extensive local experience, particularly in cardiac imaging—a field that is facing barriers to the clinical delivery of biomarkers because doing this properly requires multi-centre collaboration and integration with other data types. The on-going Open EHR developments coupled with advancements in Hadoop, other Apache open-source projects, and cloud computing,[25] offer huge opportunities to the research community—EDC research tools such as REDCap can be integrated and long-tail cardiac data sets mined from within the EHRs using big data tools rather than simply limiting the research model to data

collection by individual groups. It will become possible to capture niche cohort data from out of the larger routine clinical record but further development of electronic patient record systems in the NHS is required—the ultimate research objective is to permit flawless mining of the entire EHR in a national, secure real-time web solution that also offers complete universal follow-up of outcomes (linked to hospitalisation and death records etc.) by electronic surveillance.

Conclusion

Biomedical research costs will spiral in the UK if individual centres continue to build their own individual bioinformatics clouds instead of sharing these in a national resource, ideally with funding by the NHS. Expensive research eventually translates into increased tariffs for new therapies—reflecting a lack of understanding of basic biology, or at least the transition of that understanding into clinical practice.

We are convinced that across the cardiovascular research domain, like the rest of medicine, the national aggregation of diverse long-tail data is the best way to convert numerous small but expensive cohort data sources into big data for improved knowledge. This practical and structured integration is achievable through a sustainable, common platform of network bioinformatics, breaking down translational barriers, improving research efficiency and with time, patient outcomes.

ACKNOWLEDGEMENTS

The authors are indebted to the staff of the AIMES Grid Service Providers Plc. for the establishment of the UK pilot for cardioinformatics.

FUNDING

GC is supported by the National Institute for Health Research Rare Diseases Translational Research Collaboration (NIHR RD-TRC) and by the NIHR University College London Hospitals Biomedical Research Centre. JCM is directly and indirectly supported by the University College London Hospitals NIHR Biomedical Research Centre and Biomedical Research Unit at Barts Hospital, respectively. This cloud-based EDC pilot has been funded by Barts Charity Grants #MGU0305 and #1107/2356/MRC0140 to GC and JCM. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

CONTRIBUTORSHIP STATEMENT

GC, JCM and DK planned the pilot infrastructure and wrote the manuscript. RS and JD provided expert support and review of the manuscript. JCM is responsible for the overall content.

COMPETING INTERESTS

GC, RS, JD and JCM declare no competing interests. DK is the Cheir Executive Officer of the AIMES Grid Service Providers Plc., a commercial data centre service provider based in the North West of England.

REFERENCES

1 Leal J, Luengo-Fernández R, Gray A, *et al.* Economic burden of cardiovascular diseases in the enlarged European Union. *Eur Heart J* 2006;**27**:1610–9.

2 Luengo-Fernández R, Leal J, Gray A, *et al.* Cost of cardiovascular diseases in the United Kingdom. *Heart* 2006;**92**:1384–9.

3 Wellcome Trust. Medical research: what's it worth? Aust. Stud. 2008.http://www.wellcome.ac.uk/stellent/groups/corporatesite/@sitestudioobjects/documents/web_document/wtx052111.pdf (accessed 17 Feb2016).

4 Cardiovascular imaging at the crossroads. *JACC Cardiovasc Imaging* 2010;**3**:316–24.

5 Kramer CM, Appelbaum E, Desai MY, *et al.* Hypertrophic Cardiomyopathy Registry: The rationale and design of an international, observational study of hypertrophic cardiomyopathy. *Am Heart J* 2015;**170**:223–30.

6 Patel AR, Steel K, Daly CA, *et al.* Evidence from a multicenter CMR registry indicates that stress CMR imaging provides highly effective risk stratification in patients suspected to have myocardial ischemia. *J Cardiovasc Magn Reson* 2014;**16**:M1.

7 Bruder O, Wagner A, Lombardi M, *et al.* European cardiovascular magnetic resonance (EuroCMR) registry – multi national results from 57 centers in 15 countries. *J Cardiovasc Magn Reson* 2013;**15**:1–9.

8 Fonseca CG, Backhaus M, Bluemke D a, *et al.* The Cardiac Atlas Project--an imaging database for computational modeling and statistical atlases of the heart. *Bioinformatics* 2011;**27**:2288–95.

9 Ferguson AR, Nielson JL, Cragin MH, *et al.* Big data from small data: data-sharing in the 'long tail' of neuroscience. *Nat Neurosci* 2014;**17**:1442–7.

10 Wallis JC, Rolando E, Borgman CL. If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS One* 2013;**8**:e67332.

11 Gale CP, Weston C, Denaxas S, *et al.* Engaging with the clinical data transparency initiative : a view from the National Institute for Cardiovascular Outcomes Research (NICOR). 2006.

12 Wong TC, Captur G, Valeti U, *et al.* Feasibility of the REDCap platform for Single Center and Collaborative Multicenter CMR Research. *J Cardiovasc Magn Reson* 2014;**16**:P89.

13 Krum H, Tonkin A. Why do phase III trials of promising heart failure drugs often fail? The contribution of 'regression to the truth'. *J Card Fail* 2003;**9**:364–7.

14 Hersh WR, Totten AM, Eden KB, *et al.* Outcomes from health information exchange: systematic review and future research needs. *JMIR Med informatics* 2015;**3**:e39.

15 Garde S, Knaup P, Hovenga E, *et al.* Towards semantic interoperability for electronic health records. *Methods Inf Med* 2007;**46**:332–43.

16 Păun ID, Sauciuc DG, Iosif NO, *et al.* Local EHR management based on openEHR and EN13606. *J Med Syst* 2011;**35**:585–90.

17 Turner CF, Pan H, Silk GW, *et al.* The NIDDK Central Repository at 8 years--ambition, revision, use and impact. *Database (Oxford)* 2011;**2011**:bar043.

18 Ioannidis JPA. Measuring co-authorship and networking-adjusted scientific impact. *PLoS One* 2008;**3**:e2778.

19 Grens K. The rules of replication. *SCIENTIST* 2014;**28**:71–3.

20 Cole GD, Francis DP. Perioperative β blockade: guidelines do not reflect the problems with the evidence from the DECREASE trials. *BMJ* 2014;**349**:g5210.

21 Harris PA, Taylor R, Thielke R, *et al.* Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009;**42**:377–81.

22 R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

23 Nutter B, Lane S. REDCapAPI: Accessing data from REDCap projects using the API. 2015.

24 Bossuyt PM. The STARD Statement for Reporting Studies of Diagnostic Accuracy: Explanation and Elaboration. *Clin Chem* 2003;**49**:7–18.

25 Zou Q, Li X-B, Jiang W-R, *et al.* Survey of MapReduce frame operation in bioinformatics. *Brief Bioinform* 2014;**15**:637–47.

26 Cummings MP, Temple GG. Broader incorporation of bioinformatics in education:

opportunities and challenges. *Brief Bioinform* 2010;**11**:537–43.

FIGURES

Figure 1 The spectrum of research in cardiology

Most cardiology research projects are between 20 and 1,000 subjects, typically representing also the middle of the translational pathway (red discontinuous box). The smallest studies may not necessarily need bioinformatics; the largest have funding already but are outnumbered 34:1 by the smaller studies. Creating cohort studies is expensive. Little bioinformatics exists to support them. *Plot (2015) summarises lists study sizes in 300 consecutive cardiac trials registered with: www.clinicaltrias.gov.*

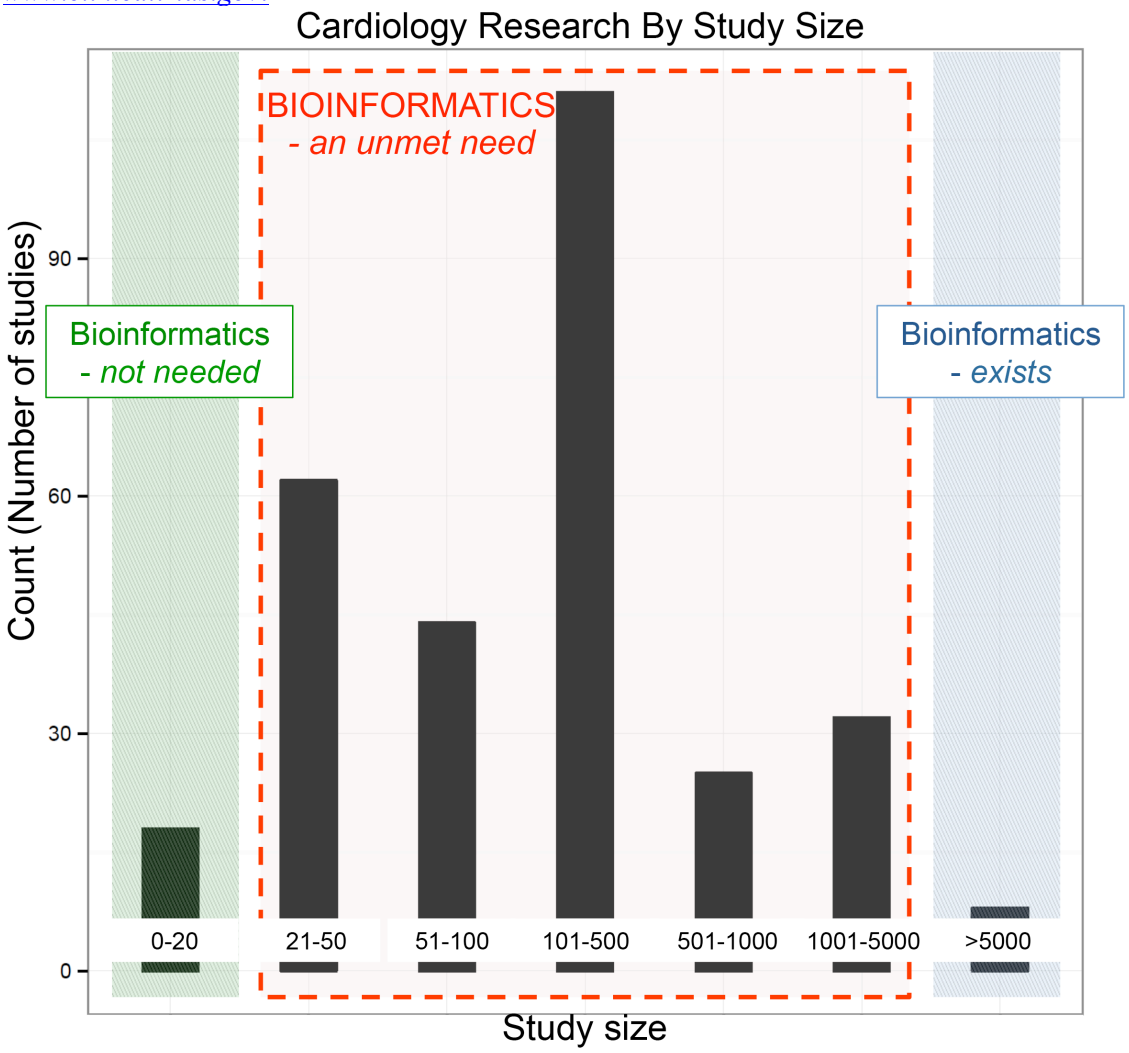

Figure 2 Cardiac-bespoke bioinformatics platform.
Searching the Bioinformatics Links Directory (www.bioinformatics.ca), >3,000 biomedical data archival platforms can be found. This plot shows the number of bioinformatics Web-servers by domain: absolute levels and growth over time. Using search terms for the cardiac domain ('cardiology', 'cardiac' and 'cardiovascular') it transpires that there are no servers dedicated to cardiac research between 2006 and 2015 (plot adapted from Cummings et al.[26]).