The International Hundred Thousand Plus Cohort Consortium (IHCC): Integrating Large-Scale Cohorts to Address Global Scientific Challenges

Teri A. Manolio[1], Philip Awadalla[2], Adam S. Butterworth[3], Robert Califf[4], Zhengming Chen[5], Rory Collins[5], John J. Connolly[6], Philippe Cupers[7], John Danesh[3], Joshua C. Denny[8], Stephanie Devaney[9], Lena Dolman[2,10], Peter Goodhand[2,10], Eric D. Green[1], Hakon Hakonarson[6], David J. Hunter[5], Sekar Kathiresan[11], Norihoro Kato[12], Thomas Keane[13], Gun Peggy Knudsen[14], Rongling Li[1], Andres Metspalu[15], Nicola Mulder[16], Michael D. Musty[17], Matthew Nelson[18], Nancy L. Pedersen[19], Tejinder Rakhra-Burris[20,21], Gad Rennert[22], Dan M. Roden[23], Laura Lyman Rodriguez[1], Camilla Stoltenberg[14], Cathie Sudlow[24], Joyce Tung[25], Walter Willett[26], Teresa Zayas-Cabán[27], Geoffrey Ginsburg[21]

[1] National Human Genome Research Institute, National Institutes of Health, Bethesda, MD

[2] Ontario Institute for Cancer Research, Toronto, Canada

[3] Health Data Research UK-Cambridge, Wellcome Genome Campus, Hinxton, and Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

[4] Duke Forge, the Donald F. Fortin, MD Professor of Cardiology, School of Medicine, Duke University, Durham, NC

[5] Nuffield Department of Population Health, University of Oxford, UK

[6] Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA

[7] Directorate-General for Research and Innovation, European Commission, Brussels, Belgium

[8] Departments of Biomedical Informatics and Medicine, Vanderbilt University, Nashville, TN

[9] All of Us Research Program, National Institutes of Health, Bethesda, MD

[10] Global Alliance for Genomics and Health, Toronto, Canada

[11] Verve Therapeutics, Inc., Cambridge, MA

[12] Medical Genomics Center, National Center for Global Health and Medicine, Tokyo, Japan

[13] European Bioinformatics Institute, Cambridgeshire, UK

[14] Norwegian Institute of Public Health, Oslo, Norway

[15] Estonian Genome Center, Estonian Biobank, Institute of Genomics, University of Tartu, Tartu, Estonia

[16] Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa

[17] Clinical & Translational Science Institute, Duke University, Durham, NC

[18] Genetics, GlaxoSmithKline, Philadelphia, PA

[19] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

[20] Global Genomic Medicine Collaborative, Durham, NC

[21] Center for Applied Genomics and Precision Medicine, Duke University, Durham, NC

[22] Department of Community Medicine and Epidemiology, Carmel Medical Center, Haifa, Israel

[23] Departments of Medicine, Pharmacology, and Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN

[24] Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, UK

[25] Research, 23andMe, Inc., Mountain View, CA

[26] Department of Epidemiology and Nutrition, Harvard T.H. Chan, Boston, MA

[27] Office of the National Coordinator for Health Information Technology, US Department of Health and Human Services, Washington, DC

Correspondence:

Teri A. Manolio, M.D., Ph.D.
Director, Division of Genomic Medicine
National Human Genome Research Institute
6700B Rockledge Drive, Room 3118, MSC 6908
Bethesda, MD 20892-6908
Phone: 301-402-2915
E-mail: manolio@nih.gov

Word count: abstract, 289; text, 3,471; figures 1; tables 5; supplementary tables 2

Abstract

Multiple large cohort studies involving hundreds of thousands of people have recently been launched in several regions worldwide. They are of great value for studying diverse populations and key demographic subgroups, rare genotypes and exposures, and gene-environment interactions. Each cohort is constrained, however, by its size, ancestral origins, and geographic boundaries that limit the subgroups, exposures, outcomes, and interactions it can examine. Combining data across large cohorts to address questions none of them can answer alone enhances the value of each and leverages the enormous investments already made in them to address pressing questions in global health.

Leaders of cohorts assessing a wide range of health and disease measures and aiming to recruit 100,000 participants or more with available biospecimens (or the possibility of collecting them) and the potential for longitudinal follow-up met in March 2018 and again in April 2019 to explore interests in and approaches for forming such a collaboration. The 61 participating cohorts were located in 32 countries, with total current sample sizes across all cohorts of roughly 30 million. Most of the cohort leaders had participants' consent to share data beyond the initial study investigators and were willing to share data more broadly, albeit with some limitations. The group agreed to form the International Hundred Thousand Plus Cohorts Consortium (IHCC) for accelerating the generation and application of population-based scientific knowledge on a global scale. IHCC teams have since been working to develop an interactive cohort registry, outline a scientific research agenda, and detail a policy agenda to make such collaborations possible. Other cohorts, particularly from under-represented areas such as Africa, South America, and South Asia, are invited to join the IHCC in realizing the vast scientific potential of a worldwide collaboration for prospective cohort research.

Introduction

Prospective cohort studies are a crucial epidemiologic tool, particularly for identifying risk factors for disease and measuring their impact [1,2]. Several large cohort studies involving hundreds of thousands of people have recently been established in Canada [3], China [4], Mexico [5], the UK [6], the US [7,8], and elsewhere, and several others are in planning. These studies typically recruit large population samples not selected for disease; measure their physical, medical, behavioral, environmental, and social characteristics and collect biospecimens at entry; and follow participants forward for years or decades to study the development of a variety of diseases. Large cohorts are especially valuable for studying key demographic subgroups; rare genotypes, exposures, and outcomes; and gene-environment interactions [9]. They can also help to refine risk modeling, identify opportunities for improved public health efforts, examine variability in response to therapeutic interventions, and identify new targets for intervention [2,10,11].

As valuable as a cohort of a million or more persons might be, it is still generally constrained by its size and by ancestral origins and geographic boundaries that limit the subgroups, exposures, and interactions it can be used to examine, particularly for rare exposures or outcomes. Combining data from large cohorts to address questions none of them can answer alone enhances the value of each cohort and builds upon the widening scientific culture of data sharing and improved data access [12,13]. Improved computational and bioinformatics capabilities are making such large-scale data sharing efforts feasible [14,15], though these efforts require sophisticated and powerful computing and informatics. They will not be successful without real advances in informatics infrastructure for both clinical care and research. Growing efforts to define responsible policies for data sharing are helping to address privacy concerns and national restrictions on data access [16,17]. Much remains to be done, however, to bring these cohorts together, harmonize their data, and identify efficient methods for answering the many compelling scientific questions that together they are uniquely positioned to examine.

To address these issues, leaders of large-scale cohorts and other scientific experts were invited to a two-day summit in March 2018 sponsored by the National Institutes of Health, the Wellcome Trust, the UK Medical Research Council, the Global Alliance for Genomics and Health (GA4GH, https://www.ga4gh.org/), and the Global Genomic Medicine Collaborative (G2MC, https://g2mc.org/). The meeting was designed to explore improved prospects for harmonizing data standards, information technology, consent, and related aspects; promote data and specimen sharing as well as open access policies; examine the potential for collaborative global genome sequencing and multi "-omics" projects; assess the feasibility of establishing a searchable online global registry of large-scale cohorts; develop federated platforms to integrate individual-level clinical and genomic data; and create a vision for the next decade of collaborative cohort research. Subsequent efforts and a second summit in April 2019 have focused on establishing key working groups and developing a compelling scientific agenda and early work products. This paper summarizes what the cohort leaders propose to do in a cohort consortium and why, describes the proposed structure of such a consortium, outlines considerations for sustainability of these efforts, and invites additional cohorts to participate.

Participating Cohorts

Cohorts meeting four criteria were identified from various compilations [18; Ioannidis JP, personal communication] and knowledge of experts in the field (Z.C., R.C., G.G. T.M.). The criteria were: having or intending to recruit 100,000 or more participants, assessing a wide range of health and disease measures (that is, not being disease-specific), having available biospecimens (or the possibility of collecting them), and having at least the potential for longitudinal follow-up of participants. Additional cohorts not meeting all four criteria but involving under-represented regions such as Africa, the Middle East, South America, and South Asia were also identified. Leaders of these cohorts were invited to participate in a first International Cohorts Summit at Duke University [19]. Of 76 cohorts invited, 59

attended, 3 declined to participate, and 14 did not respond. Two additional cohorts were identified at the meeting and are included in tabulations below.

The 61 participating cohorts were located in 32 countries, including 9 countries that were only represented by being part of at least one of five multinational cohorts (Figure 1). It was not always clear for each cohort whether it represented a true prospective study cohort with systematic sampling and baseline data collection vs. more of a population or health record registry, but each self-identified as a "cohort" so they are referred to as such throughout. Total sample sizes across these cohorts was roughly 30 million, with a targeted total sample size of 37 million. The distribution of sample sizes varied widely by country, with the largest cohorts most commonly drawn from China, Scandinavia, the UK, and the US (Supplementary Table 1). As is evident from the Figure, large cohorts are sorely lacking in Africa, Central and Eastern Europe, Central Asia, and many countries of South America and Southeast Asia. Notably, India— comprising one-sixth of the world's population—is currently represented by a cohort projected to number only 200,000 participants.

Attendees were asked to complete a survey (Supplementary Table 2) describing key characteristics of their cohorts and willingness of their investigators to participate in various aspects of a collaboration such as sharing data and protocols. All but 4 of 61 cohort representatives completed the survey, though not all completed every survey question, and responses were largely free-form so many details remain to be determined. A more systematic approach to collecting and validating these details awaits the development of a cohort registry as described below. The majority of the cohorts had biospecimens available, typically blood, urine and tumor samples; most also had DNA available and had performed at least some genotyping (Table 1). Roughly half had whole exome or genome sequence data generated from at least some samples. Most of the cohorts indicated they had participants' consent to share data beyond the initial study investigators and that the investigators were willing to do so, albeit with some limitations. Most respondents believed data sharing would improve statistical power for identifying associations, advance scientific knowledge, and foster collaborations and new approaches.

Challenges and Opportunities in Developing Cohort Collaborations

Cohort leaders expressed enthusiasm for collaborating but outlined several challenges to integrating large cohorts. These include complexity of available datasets; lack of standardization or harmonization of questionnaires; and inability to move or utilize data due to file sizes, regulatory restrictions, and national legal systems (Table 2). Strategies for addressing these challenges include aligning data standards to encourage and facilitate sharing, standardizing data collection across cohorts prospectively using available tools [20,21], and establishing necessary infrastructures to move analyses to the data rather than vice versa [15].

Registry of cohorts. A critical first step in facilitating collaborations is developing a standardized atlas or registry to share basic descriptive information and metadata about each cohort. Cohort leaders receive myriad requests for such information and would value a central resource to which they could direct subsequent inquiries, but providing this information, distributing it in a user-friendly manner, and keeping it up to date require significant effort. A registry could also increase the international visibility of these cohorts and increase their use, a common metric of their value for continued investment and thus an important factor in their sustainability. A tiered approach to a user-friendly data repository could begin with a registry comprising the most basic information, such as cohort name and website, and progressing in tiers to the most detailed presentation of individual participant-level data (Table 3). A repository of data collection methods and standard operating procedures for sample collection and storage, assays and analyses, quality assessments, etc., could enhance standardization and (if widely available) reduce the cost involved in starting a new cohort. With such a framework, cohorts could participate up to the tier they choose, minimizing their administrative burden as a primary goal while simultaneously increasing their visibility and potential utilization. Harmonization of variables will be a time-intensive but necessary early step to facilitate cross-cohort studies. Automation of some harmonization processes may be possible, but most will require human effort and concomitant funding

support. Building the supporting infrastructure for an interactive, queryable data ecosystem will require considerable technical work. Sharing of individual-level data will be the most challenging step, due not only to privacy and consent concerns but also to the need for harmonization and language translation. Governance of individual-level data sharing could potentially be managed via an independent global federation.

Possible uses for such a system could be to create a "reproducibility network" for rapid validation of scientific findings discovered in a single cohort. In addition, the predictive value and generalizability of polygenic risk scores that incorporate genetic (or multi-omic) risk factors into conventional risk equations could be assessed with greater precision and explored in diverse populations. These cohorts could also provide valuable exploratory data for examining the impact of risk reduction interventions or even potentially as a source of participants for clinical trials (with agreement of the cohort leaders) if such trials would not interfere with their primary observational goals. Other applications include basic research on disease pathogenesis, such as gene-environment interaction, or studies of gene function in persons with unusual genetic variants. As noted above, collaborating cohorts could also focus on rare diseases or rare exposures that require massive sample sizes or substantial genetic and environmental diversity for effective study. Such a network could reduce duplication, improve efficiency, leverage current investments, and possibly provide more sustainable funding models. Access to individual-level data, however, will likely need a more federated model of linked but independent and accessible databases, virtually connected through software interfaces allowing seamless, authorized access [15], that bring researchers and their analytic methods to individual cohorts' data repositories rather than trying to compile all the data at a single site or distribute the data to any and all interested investigators. A blended approach, centralized for the top tiers of the registry and federated for individual-level data, may maximize efficiency, address jurisdictional restrictions on export of data, and promote the creation of analytical and data environments that meet evolving security requirements. Ideal characteristics of such a platform would include use of common interfaces such as DataSHIELD [22] or those from the

GA4GH [15], a standard set of analysis modules, and an accelerated and shared user authorization process. The ability to track and measure wider reuse or impact of particular cohorts could help to demonstrate the value of such a collaborative approach. Such a registry should build upon and unite with other efforts to compile and register cohorts such as Maelstrom Research (https://www.maelstrom-research.org [23]), the Low and Middle Income Longitudinal Population Study Directory (https://www.ifs.org.uk/tools_and_resources/longitudinal) and the European Union's "Common Infrastructure for National Cohorts in Europe, Canada, and Africa (CINECA, https://www.ebi.ac.uk/about/news/press-releases/CINECA-facilitates-transcontinental-human-data-exchange).

Scientific agenda. Compelling scientific questions that could be addressed by collaboration among cohorts (Table 4) are not limited to studying rare exposures and outcomes, though those are obvious and unquestionably valuable early uses. Context-specific analyses of the local relevance of risk factors could better inform global burden of disease estimates and assess what determines "health" in different settings. Multi-national analyses of global health problems such as obesity and exposure to toxic substances such as alcohol and pollutants could identify generalizable approaches for addressing global threats to public health. Country- or cohort-specific risk predictions using standardized methodology could also be compared with a goal not only of producing more generalizable risk estimates but also of recognizing when tailored predictions are more appropriate. Identification and phenotyping of carriers of loss-of-function alleles in nearly every human gene ("human knock-out project") is theoretically feasible if several million genome sequences are available for analysis and linked to detailed genotypic and broad phenotypic data. Assessment of rare genetic variation would be greatly enhanced if the research participants who donated these samples are available for and accepting of re-contact and in-depth study [24].

Invaluable contributions to research methodology could be developed and promulgated by a consortium motivated to develop procedures that are readily disseminated and implemented.

Phenotyping methods for a wide array of health outcomes could be developed using algorithms based on health record systems and other sources. Systems to facilitate and encourage funding for long-term follow-up of health outcomes, particularly in low-resource settings, and ensure access to the widest range of health outcome data would maximize cohorts' ability to achieve long-term, comprehensive follow-up of participants. Such systems are likely to require engagement of research funders and governments, for whom the relevance of health research to improving the effectiveness of healthcare should be emphasized. Novel methods (such as digital health technologies, data linkage, and large-scale imaging) for characterizing exposures, defining outcomes, and visualizing and managing data could also be developed and disseminated. Best practices for communicating results to participants could be shared and optimized by comparing outcomes of differing approaches in different cultures.

Enhancements to existing cohorts that would increase their utility and promote data sharing include collection of biological samples and support for cohort-wide sample analysis and data deposition to minimize sample wastage from inefficient case-control analyses. Generating new single nucleotide polymorphism (SNP)-array genotyping in cohorts without such data and sharing them across all cohorts could be a first step, as much remains to be discovered by genomic studies in under-represented regions. A consortium of cohorts could help develop population-specific genotyping arrays and imputation algorithms based on whole genome sequencing of specific reference populations. Most valuable would be whole genome sequencing data that could be pooled across cohorts, but generating millions of human sequences seems cost-prohibitive in the near term. Per-sample costs of genome sequencing and other 'omics (transcriptomics, proteomics, metabolomics, etc.) could be driven down, however, through efficient processing of millions of samples. Close partnerships are needed with developers of novel assays to determine when assays are ready to be applied to millions of specimens; cohorts can work iteratively with developers to improve these assays.

Policy agenda. Guiding principles for a consortium of cohorts should include pursuit of the best collaborative science in the most ethical and efficient manner possible. Respect for and recognition of

the contributions and rights to privacy of participants and individual investigators are critical, as are fairness and equity in opportunities for cohorts to participate in collaborative ventures and autonomy in choosing whether to participate at all. Sharing of or providing access to de-identified data on individual participants requires a clear understanding of who will be using the data and for what purposes, as well as how that access and use align with the cohorts' consent processes and expectations for participant privacy. Policy considerations will thus need to be discussed concurrently with scientific design aspects as collaborations develop. Policies for consortium governance, data access and use, and appropriate attribution will need to be agreed upon by participating cohorts and be consistent with local regulations and cultural norms. GA4GH has provided a policy framework for genomic data sharing that addresses many of these issues [16].

The European Union (EU)'s General Data Protection Regulation No 2016/679[1] (GDPR) strengthens individuals' rights to privacy and data protection and enhances the transparency and accountability of the data processing. The GDPR lays down specific requirements on scientific research and processing of special categories of sensitive personal data, including health data. Provisions allow processing and re-use for scientific research purposes, subject to specific conditions. The GDPR and other national laws that provide similar or even stricter data protection requirements will need to be considered early in planning collaborations involving cohorts that include research participants in EU member states and the European Economic Area (EEA) states [25].

Near-term policy goals could include creating a governance framework for the registry of cohort studies, defining policy challenges such as incorporating GDPR guidelines and implications, and engaging with primary funders of cohorts to identify potential constraints on participation. Special considerations for involvement of for-profit entities, such as expectations of reciprocity in data sharing or providing other benefits to the consortium, should be laid out as should potential benefits and risks of participation to cohorts and their participants. Current successful collaborations could be identified and useful lessons compiled from what has worked well and what has not. Procedures and formats for

submitting projects and granting data access should be developed and could build on existing exemplary models such as UK Biobank, a cohort that has demonstrated the feasibility and value of making a richly genotyped and phenotyped resource readily accessible to *bona fide* researchers worldwide [26]. Metadata describing specific policy "traits" of cohorts, such as options for access by for-profit entities, requirements for ethics board approval, or requisite fees, would be useful additions to the registry.

Towards an International Hundred Thousand Plus Cohort Consortium (IHCC)

Given the interest among large cohorts in forming a collaboration and the unprecedented scientific opportunities such a collaboration presents, G2MC [27] and the National Institutes of Health (NIH), working with GA4GH and the Wellcome Trust, have begun to pursue the highest priority goals of developing a cohort registry, a research agenda, and a policy framework. The overall goal of the IHCC is to link large cohort studies into a global platform for translational research that would accelerate the generation and application of scientific knowledge and improve the health of individuals throughout the world. A Data Standards and Interoperability team is working to deliver informatics standards and infrastructure to enable accessibility of population-scale genomic and biomolecular data across international borders, building on national investments in health information technology, as well as much ongoing work in this area by GA4GH, the European Union, and related entities. Close integration with these efforts will ensure that IHCC addresses the FAIR principles [28] by using emerging global standards for genomic data sharing. A Scientific Strategy and Cohort Enhancements team has begun to develop a scientific agenda, explore enhancements to existing cohorts, and address gaps in diversity of cohort populations. A Policy and Data Sharing team is developing a policy agenda to facilitate and optimize the value of assembling the cohorts while observing local norms and regulatory constraints. In each of these domains, IHCC will work to identify "early wins" to galvanize the consortium and establish its value proposition, but its full value will unfold over years and decades as its goals are met. Charges and goals of the three teams are outlined in Table 5.

Organizers of the IHCC recognize that initial ascertainment of cohorts was imperfect and outreach incomplete, a situation that needs to be rectified. Leaders of other large cohorts are invited to join the IHCC by contacting info@g2mc.org. Cohort leaders are encouraged to contribute to the registry as it evolves and share their data within the consortium in ways consistent with their participants' consent and local regulations. They are also invited to join the teams and participate in subsequent summits expected to be held roughly annually.

A key precept of the IHCC is that cohort independence and individuality will be respected and viewed as major strengths, and efforts will be made to ensure cohorts in low-income settings have sufficient resources to participate actively while maintaining control and sovereignty over their data. Rather than imposing a single unifying structure, IHCC aims to provide an environment in which all cohorts can learn from each other and share best practices to enable more effective approaches worldwide. Joining the IHCC could provide significant benefits to individual cohorts, such as inclusion in a registry that could relieve them of the burden of responding to repeated inquiries about their design and structure, as well as easier access to larger and more diverse global datasets. With the power of the IHCC community behind them, they could also have stronger voices in negotiating for funding and discounted assay pricing than if speaking alone. The IHCC globally, and cohort leaders locally, should demonstrate the importance of supporting these cohorts and advocate for national funding as part of each country's necessary public health infrastructure. Opportunities and synergies with national funding sources outside of major international funders such as NIH and Wellcome Trust should also be explored, potentially through a model of support for cohorts from within each country similar to that used in the Human Genome Project [29].

The scientific possibilities presented by large cohorts such as those described here are exciting and impressive, but the scientific opportunities presented by combining them and recruiting other large cohorts worldwide are extraordinary. Current cohorts already include substantial geographic and ancestral diversity that has yet to be harnessed effectively for scientific study, while establishing cohorts

in under-represented populations will add immeasurably to the breadth and generalizability of questions that can be addressed. We have the expertise, the tools, and now the will to move forward with such a collaboration; it is largely our own inertia that stands in our way. We owe it to these studies' participants, their funders, and the populations they represent to seize this opportunity for maximizing the substantial ongoing investments in large cohorts and applying them to critical and hitherto intractable problems in global health.

References

1. Kannel WB, Dawber TR, Kagan A, Revotskie N, Stokes J 3rd. Factors of risk in the development of coronary heart disease--six year follow-up experience. The Framingham Study. Ann Intern Med. 1961 Jul;55:33-50. PMID: 13751193

2. Lewington S, Clarke R, Qizilbash N, Peto R, Collins R; Prospective Studies Collaboration. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. Lancet. 2002 Dec 14;360(9349):1903-13. PMID: 12493255

3. Dummer TJB, Awadalla P, Boileau C, Craig C, Fortier I, Goel V, Hicks JMT, Jacquemont S, Knoppers BM, Le N, McDonald T, McLaughlin J, Mes-Masson AM, Nuyt AM, Palmer LJ, Parker L, Purdue M, Robson PJ, Spinelli JJ, Thompson D, Vena J, Zawati M; with the CPTP Regional Cohort Consortium. The Canadian Partnership for Tomorrow Project: a pan-Canadian platform for research on chronic disease prevention. CMAJ. 2018 Jun 11;190(23):E710-E717. PMID: 29891475.

4. Chen Z, Chen J, Collins R, Guo Y, Peto R, Wu F, Li L; China Kadoorie Biobank (CKB) collaborative group. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. Int J Epidemiol. 2011 Dec;40(6):1652-66. PMID: 22158673

5. Tapia-Conyer R, Kuri-Morales P, Alegre-Díaz J, Whitlock G, Emberson J, Clark S, Peto R, Collins R. Cohort profile: the Mexico City Prospective Study. Int J Epidemiol. 2006 Apr;35(2):243-9. PMID: 16556648

6. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med 2015: 12(3): e1001779. PMID:25826379

7. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, Whitbourne S, Deen J, Shannon C, Humphries D, Guarino P, Aslan M, Anderson D, LaFleur R, Hammond T, Schaa K, Moser J, Huang G, Muralidhar S, Przygodzki R, O'Leary TJ. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. J Clin Epidemiol. 2016 Feb;70:214-23. PMID: 26441289

8. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015 Feb 26;372(9):793-5. PMID: 25635347

9. Manolio TA, Bailey-Wilson JE, Collins FS. Genes, environment and the value of prospective cohort studies. Nat Rev Genet. 2006 Oct;7(10):812-20. PMID: 16983377

10. Alegre-Díaz J, Herrington W, López-Cervantes M, Gnatiuc L, Ramirez R, Hill M, Baigent C, McCarthy MI, Lewington S, Collins R, Whitlock G, Tapia-Conyer R, Peto R, Kuri-Morales P, Emberson JR. Diabetes and Cause-Specific Mortality in Mexico City. N Engl J Med. 2016 Nov 17;375(20):1961-1971. PMID: 27959614

11. Barquera S, Schillinger D, Aguilar-Salinas CA, Schenker M, Rodríguez LA, Hernández-Alcaraz C, Sepúlveda-Amor J; Mexico-California Diabetes collaborative group. Collaborative research and actions on both sides of the US-Mexico border to counteract type 2 diabetes in people of Mexican origin. Global Health. 2018 Aug 22;14(1):84. PMID: 30134925

12. Maxson Jones K, Ankeny RA, Cook-Deegan R. The Bermuda Triangle: The Pragmatics, Policies, and Principles for Data Sharing in the History of the Human Genome Project. J Hist Biol. 2018 Dec;51(4):693-805. PMID: 30390178

13. Walport M, Brest P. Sharing research data to improve public health. Lancet. 2011 Feb 12;377(9765):537-9. PMID: 21216456

14. Clarke L, Zheng-Bradley X, Smith R, Kulesha E, Xiao C, Toneva I, Vaughan B, Preuss D, Leinonen R, Shumway M, Sherry S, Flicek P; 1000 Genomes Project Consortium. The 1000 Genomes Project: data management and community access. Nat Methods. 2012 Apr 27;9(5):459-62.PMID: 22543379

15. Global Alliance for Genomics and Health. A federated ecosystem for sharing genomic, clinical data. Science. 2016 Jun 10;352(6291):1278-80. PMID: 27284183

16. Knoppers BM. Framework for responsible sharing of genomic and health-related data. Hugo J. 2014 Dec;8(1):3. PMID: 27090251

17. Contreras JL, Reichman JH. Sharing by design: Data and decentralized commons: Overcoming legal and policy obstacles. Science. 2015 Dec 11; 350(6266): 1312–1314. PMCID: PMC4811371

18. Global Alliance for Genomics and Health. Catalogue of Genomic Data Initiatives. https://www.ga4gh.org/community/catalogue, accessed 12/29/2018.

19. Global Genomic Medicine Collaborative. International Cohorts Summit. https://g2mc.org/events/, accessed 12/30/2018.

20. Evans JP, Smith A, Gibbons C, Alonso J, Valderas JM. The National Institutes of Health Patient-Reported Outcomes Measurement Information System (PROMIS): a view from the UK. Patient Relat Outcome Meas. 2018 Oct 24;9:345-352. PMID: 30498382

21. Pan H, Tryka KA, Vreeman DJ, Huggins W, Phillips MJ, Mehta JP, Phillips JH, McDonald CJ, Junkins HA, Ramos EM, Hamilton CM. Using PhenX measures to identify opportunities for cross-study analysis. Hum Mutat. 2012 May;33(5):849-57. PMID: 22415805

22. Budin-Ljøsne I, Burton P, Isaeva J, Gaye A, Turner A, Murtagh MJ, Wallace S, Ferretti V, Harris JR. DataSHIELD: an ethically robust solution to multiple-site individual-level data analysis. Public Health Genomics. 2015;18(2):87-96. PMID: 25532061

23. Bergeron J, Doiron D, Marcon Y, Ferretti V, Fortier I. Fostering population-based cohort data discovery: The Maelstrom Research cataloguing toolkit. PLoS One. 2018 Jul 24;13(7):e0200926. PMID: 30040866

24. Saleheen D, Natarajan P, Armean IM, Zhao W, Rasheed A, Khetarpal SA, Won HH, Karczewski KJ, O'Donnell-Luria AH, Samocha KE, Weisburd B, Gupta N, Zaidi M, Samuel M, Imran A, Abbas S, Majeed F, Ishaq M, Akhtar S, Trindade K, Mucksavage M, Qamar N, Zaman KS, Yaqoob Z, Saghir T, Rizvi SNH, Memon A, Hayyat Mallick N, Ishaq M, Rasheed SZ, Memon FU, Mahmood K, Ahmed N, Do R, Krauss RM, MacArthur DG, Gabriel S, Lander ES, Daly MJ, Frossard P, Danesh J, Rader DJ, Kathiresan S. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. Nature. 2017 Apr 12;544(7649):235-239. PMID: 28406212

25. Phillips M. International data-sharing norms: from the OECD to the General Data Protection Regulation (GDPR). Hum Genet. 2018 Aug;137(8):575-582. PMID: 30069638

26. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015 Mar 31;12(3):e1001779. PMID: 25826379

27. Manolio TA, Abramowicz M, Al-Mulla F, Anderson W, Balling R, Berger AC, Bleyl S, Chakravarti A, Chantratita W, Chisholm RL, Dissanayake VH, Dunn M, Dzau VJ, Han BG, Hubbard T, Kolbe A, Korf B, Kubo M, Lasko P, Leego E, Mahasirimongkol S, Majumdar PP, Matthijs G, McLeod HL, Metspalu A, Meulien P, Miyano S, Naparstek Y, O'Rourke PP, Patrinos GP, Rehm HL, Relling MV, Rennert G, Rodriguez LL, Roden DM, Shuldiner AR, Sinha S, Tan P, Ulfendahl M, Ward R, Williams MS, Wong JE, Green ED, Ginsburg GS. Global implementation of genomic medicine: We are not alone. Sci Transl Med. 2015 Jun 3;7(290):290ps13. PMID: 26041702

28. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018. PMID: 26978244

29. Green ED, Watson JD, Collins FS. Human Genome Project: Twenty-five years of big biology. Nature. 2015 Oct 1;526(7571):29-31. PMID: 26432225

30. Whiffin N, Minikel E, Walsh R, O'Donnell-Luria AH, Karczewski K, Ing AY, Barton PJR, Funke B, Cook SA, MacArthur D, Ware JS. Using high-resolution variant frequencies to empower clinical genome interpretation. Genet Med. 2017 Oct;19(10):1151-1158. PMID: 28518168

Table 1. Descriptive table from survey responses summarizing cohort sizes, samples and banking, consent, and willingness to share data. (N.B.—not all cohorts responded to all questions.)

Characteristic	Number of Cohorts
Type of samples collected	
Blood	41
Urine	23
Tumor	8
Multiple types (inclusive of above)	32
Central biobank (now or soon)	48
DNA available (now or soon)	
All (or nearly all) participants	44
Subset of participants	10
No	3
Genotyping	
Some or all participants	47
No	5
Genomic sequencing	
Some or all participants	30
No	8
Participant consent to share data	
Yes	51
Varies by subcohort	2
No	6
Study information returned to participant?	
Yes	28
No	27
Investigators willing to share…	
Redacted individual data?	
Yes	20
Yes, with limitations	21
No	2
Summary data	
Yes	40
Yes, with limitations	6
Metadata[1]	
Yes	8
Yes, with limitations	36
Case report forms, data collection materials?	
Yes	34
Yes, with limitations	6
No	1

[1] Data that provide information about and describe the study data.

Table 2. Challenges to establishing and combining large cohorts and potential strategies for addressing them.

Challenge	Potential strategy
Complexity and limited documentation of available data	Align data standards to encourage and facilitate sharing
Lack of standardization and harmonization of questionnaires	Explore potential use of natural language processing to studies' instruments to extract comparable information, similar to what is done with electronic health records Standardize data collection across cohorts prospectively by reusing existing insruments or using tools such as PROMIS and PhenX
Inability to move, send, receive, or utilize data due to size of files, regulatory restrictions, and national legal systems	Use federated data systems to move analysis to data sets rather than sending data to analysts; produce standards to facilitate this Use data platforms based upon open source software and ensure adherence to FAIR principles and emerging data standards
Comprehensive detection and phenotyping of numerous health outcomes in diverse health care settings	Facilitate long-term follow-up of health outcome data Develop automated approaches to phenotyping based on health records Use digital health devices or apps
Potential for sample depletion with repeated subsampling and analysis	Centralize assays and conduct them cohort-wide for maximum efficiency and minimum wastage
Large costs for cohort-wide assays and analyses that facilitate data sharing	Develop close partnerships with assay vendors to secure competitive prices and ensure assay implementation and interpretation are appropriately adapted to work at very large scale
Cross-cultural and inter-individual differences in values, risk tolerance, and privacy perspectives	Use existing frameworks to help address privacy, security, consent, such as the U.S. Precision Medicine Initiative Privacy & Trust Principles and the GA4GH Framework for Responsible Sharing of Genomic and Health-Related Data (available in a dozen languages)

Table 3. Tiered structure for standardized registry of cohorts.

Tier	Contents
0	• Cohort name and website
1 – Cohort Description	• Ascertainment scheme and study design • Sample size and dates of recruitment • Demographics (age, sex, race/ethnicity) • Data types and biospecimens available • Key publications • Mechanisms and requirements for gaining access to data • Data use restrictions (such as use limited to specific condition or to academic users)
2 – Data Description and Methods	• Data collection instruments (exposures, outcomes, etc.) • Description of genomic and other 'omic data • Protocols for sample collection, storage, assays, quality assessment • Data dictionaries and linkings to standardized vocabularies • Data sharing protocol • Consent (structured data use restrictions, full consent form if available)
3 – Counts and Distributions	• Tables of counts (number with particular phenotype, number with DNA samples) • Tables/graphs of distributions of continuous variables (weight, blood pressure, etc.) • Variant-level summary statistics of association data for use in meta-analysis
4	• Individual-level data

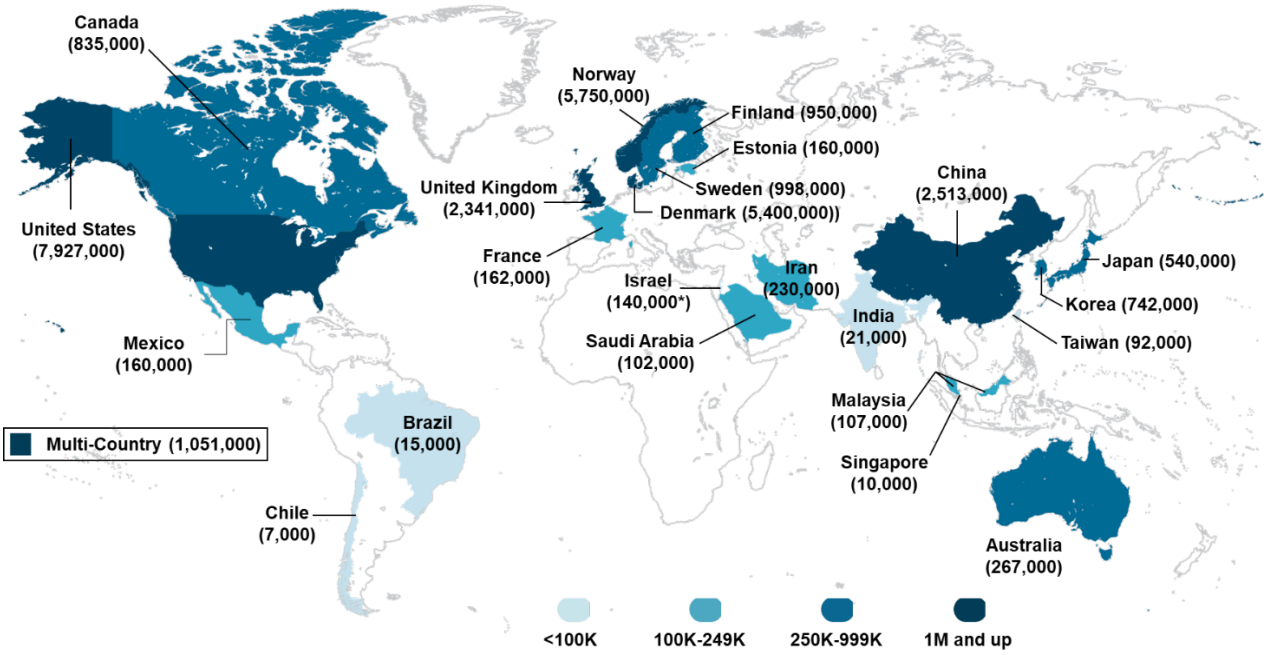
Table 4. Potential contributions of a consortium of cohorts.

Scientific advances
Risks associated with rare exposures and outcomes
Generalizability of risk factors and associations
Population-specific determinants of health
Social or cultural determinants of health
Country- or cohort-specific risk predictions
Impact on health of changing environments through migrant studies
Human knock-out identification and phenotyping
Identification of genetic variants in a specific population present at a frequency too high to be consistent with disease causality [30]
Mitochondrial DNA haplogroups and associated disease susceptibility
Research methods
Phenotyping algorithms for wide array of outcomes using health record systems
Data systems to facilitate long-term follow-up of health outcomes, particularly in low-resource settings
Novel methods for characterizing exposures and defining outcomes
User-friendly approaches to visualizing and managing data
Approaches to communicating genomic and other results to participants in diverse cultures
Population-specific genotyping arrays and imputation algorithms
Novel ‘omic assays optimized for use in millions of participants
Policy tools
Common governance framework to facilitate international collaboration across national cohorts
Clear value statement to share with participants, local stakeholders, and research funders to articulate research and public health benefits to be achieved

Table 5. Teams, charges, and goals [adapted from https://ihccstaging.g2mc.org/].

Data Standards and Interoperability (co-chairs Philip Awadalla, Thomas Keane)
Charge: Deliver IT standards and infrastructure for IHCC to enable population scale genomic and biomolecular data accessible across international borders accelerating research and improving the health of individuals resident across continents
Create a federated solution with one or two central hubs for discovery across the IHCC network that contains cohort-scale genetic data; rich heterogeneous metadata; and Ethical, Legal and Social Implications (ELSI)
Create an atlas for conducting high-level metadata queries with an interactive, searchable database that is updated from cohorts regularly
Create a harmonized cohort metadata using established semantic mapping techniques
Scientific Strategy and Cohort Enhancements (co-chairs Adam Butterworth, Hakon Hakonarson, Gadi Rennert)
Charge: Develop a scientific agenda to identify novel approaches to diagnose and treat genomic conditions, explore enhancements to existing cohorts, and address diversity gaps
Improve the diagnosis, prognosis, and treatment of common rare diseases
Identify high-risk individuals
Improve understanding of variability in response to treatments
Design a strategy for genomic enhancement to cohorts that allows countries/regions that have large cohorts with environmental data samples but limited resources for sequencing to generate genomic information
Explore enhancements to existing cohorts/datasets
Create an integrated ‘omics workspace
Policy and Bio-Data Sharing (co-chairs Gun Peggy Knudsen, Laura Lyman Rodriguez)
Charge: Develop a policy agenda to facilitate and optimize the impact of assembling cohorts; address challenges and identify common needs
Understand the spectrum of legal frameworks governing IHCC cohorts
Understand and apply existing principles to consortium needs wherever possible
Understand the needs of proposed scientific projects regarding policy or data sharing
Develop governance and policy structure to guide cohort activities

Figure 1. Home countries of participating cohorts and registries, with total estimated sample sizes per country.



Supplementary Table 1. Current and targeted sample sizes of participating cohorts, by country.

Country	Cohort Name	Current Size	Target Size
Australia	45 and Up Study	267,153	267,153
Brazil: six cities	ELSA-Brazil Project	15,105	15,105
Canada	Canadian Partnership for Tomorrow Project	315,000	315,000
Canada (Newfoundland and Labrador)	Newfoundland 100K Genome Project / Sequence Bio	520,000	520,000
Chile	Maule Cohort (MAUCO Study)	7,000	10,000
China	China Kadoorie Biobank	512,891	512,891
China	China PEACE (Patient-centered Evaluative Assessment of Cardiac Events) Million Persons Project	2,000,000	4,000,000
Denmark	Danish National Biobank	5,400,000	5,400,000
Estonia	Estonian Genome Project	175,000	200,000
Finland	Finnish Maternity Cohort Serum Bank	950,000	950,000
France	Constances Project	162,000	200,000
India	Barshi Cohort	21,000	200,000
Iran	Golestan Cohort Study	50,000	50,000
Iran	Persian Cohort Study	180,000	180,000
Israel	Israel Clalit Genome Project	-	100,000
Japan	Japan Public Health Center-based Prospective Study (JPHC)	13,000	13,000
Japan	Japan Public Health Center-based Prospective Study for the Next Generation (JPHC-NEXT)	115,405	115,405
Japan	Tohoku Medical Megabank Project	142,000	150,000
Japan	Biobank Japan	270,000	270,000
Korea	Korean Cancer Prevention Study (KCPS-II Biobank)	156,701	156,701
Korea	Korea Biobank Project	585,000	585,000
Malaysia	Malaysian Cohort	106,527	106,527
Mexico	Mexico City Prospective Study	159,755	159,755
Multinational: 7 European; Australian; USA	LIFEPATH (Lifecourse biological pathways underlying social differences in healthy aging)	235,000	235,000
Multinational: South Korea, Vietnam, Cambodia, Japan, China	Korean Genome and Epidemiological Study (KoGES)	245,000	245,000

Multinational: UK, Italy, France, Germany, Norway, Netherlands, Denmark, Spain, Greece, Sweden	EPIC (European Prospective Investigation into Cancer, Chronic Diseases, Nutrition and Lifestyle)	521,000	521,000
Multinational	AstraZeneca integrated genomics initiative	-	500,000
Multinational: Bangladesh, Malaysia, Sri Lanka	Network of South(east) Asian cohorts	50,000	150,000
Norway	Cohort of Norway (CONOR)	200,000	200,000
Norway	Norwegian Mother and Child Cohort Study (MoBa)	284,000	284,000
Norway	Norwegian Family Based Life Course Study	5,266,270	5,266,270
Saudi Arabia	Saudi National Biobank	2,000	2,000
Saudi Arabia	Saudi Human Genome Program	100,000	100,000
Singapore	Singapore National Precision Medicine Program	10,000	1,000,000
Sweden	LifeGene	51,300	300,000
Sweden	Northern Sweden Health and Disease Study	135,000	135,000
Sweden	Apolipoprotein MORtality RISk study (AMORIS)	812,073	812,073
Taiwan	Taiwan Biobank	92,371	300,000
United Kingdom	East London Genes and Health	27,806	100,000
United Kingdom: England	Genomics England / 100,000 Genomes Project	75,000	75,000
United Kingdom: England, Scotland, Wales, Northern Ireland, Isle of Man, Channel Islands	Generations Study (GS)	113,000	113,000
England, Wales, Northern Ireland	UKCTOCS (UK Collaborative Trial of Ovarian Cancer Screening) Longitudinal Women's Cohort – (UKLWC)	202,638	202,638
United Kingdom: England, Scotland, Wales	UK Biobank	502,713	502,713
United Kingdom: England, Scotland	Million Women Study	1,320,000	1,320,000
United Kingdom	UK Blood Donor Cohorts	100,000	350,000

United States (Hawaii, California)	Multiethnic Cohort Study (MEC, NCI)	77,000	77,000
United States	NHSII (Nurses' Health Study II, NCI)	116,430	116,430
United States	NHS (Nurses' Health Study, NCI)	121,700	121,700
United States	PLCO (Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial, NCI)	154,907	154,907
United States	WHI (Women's Health Initiative)	161,808	161,808
United States	MyCode Community Health Initiative	181,117	200,000
United States	Cancer Prevention Study II Nutrition Cohort	185,000	185,000
United States; Northern California members of Kaiser Permanente Health Plan	Kaiser Permanente Research Program on Genes, Environment, and Health	210,000	500,000
United States	BioVU Vanderbilt	244,000	250,000
United States	Children's Hospital of Philadelphia (CHOP) Biorepository	500,000	500,000
United States	Million Veteran Program	650,000	1,000,000
United States	U.S. Precision Medicine Initiative / All of Us	140,000	1,000,000
United States	Cancer Prevention Study II (CPS-II)	1,185,106	1,185,106
United States	23andMe	4,000,000	4,000,000
United States	Environmental influences on Child Health Outcomes (ECHO) Cohort	-	100,000
TOTAL		**30,395,776**	**36,742,182**

Supplementary Table 2. Pre-summit meeting survey (responses available at https://ihcc.g2mc.org/ics2018/).

Questions Relating to Cohort
Name of study
Principal Investigator/lead
Contact email
PubMed ID (or other information) for a protocol/marker paper on this study
Study website
Purpose or major Objectives of study
Disease areas of focus
Is your cohort selected for a specific disease (cancer, diabetes) or unselected for disease?
Current size of population (and target number of participants)
Participating countries
Period of enrollment (and is enrollment on-going?)
Demographic characteristics of participants (age range, proportion male/female, national origin, race)
Major diseases or phenotypes collected to date.
Standardized clinical evaluation components measured, if applicable (e.g. height, weight, blood pressure, exercise testing, spirometry, etc.)
Electronic health/medical records or medical administrative data used to collect clinical phenotypes?
Predominant type of electronic health records (e.g. Epic, Cerner, etc.)
Other sources of clinical data
Environmental exposure data being obtained? What sort?
Other data collected
Biological specimens collected? What sort?
Is there a central biobank?

DNA samples prepared (or available to be prepared) from each participant?
Is genotyping being done on some/all participants?
Is genomic sequencing being done on some/all participants?
Other molecular analyses performed
Did participants provide consent regarding sharing of their data outside the initial study investigators?
How are data or specimens from the cohort made available for research? Any limitations on who can access the data (e.g. by country or sector?)
What study information or data are returned to or accessible by participants?
Follow-up occurring? (years of follow-up). Is recontact possible?
Notes/Comments
Questions Relating to Sharing & Collaboration
May we make the information you provided about your cohort available on an open website?
Are you willing to share data from your cohort? If so, would you share:
a) individual data (redacted to protect confidentiality)?
b) summary data (counts, distributions)?
c) metadata (descriptive information on data collection methods)?
d) case report forms and other data collection materials?
What do you see as the values of sharing?
What challenges do you anticipate with sharing?
What specific legal or regulatory barriers in your country or cohort (aside from ensuring confidentiality and appropriate consent) would prevent you from sharing data with other cohort of cohorts investigators, or openly with anyone who requests them?
What aspects of your cohort are intended for translation to clinical care or population health?
How might genomic sequencing add to/enhance your study objectives?
Might you be willing to contribute funding or other resources to support international collaboration?