Katalyst Publishing Group

Journal of Current Trends in Clinical and Biomedical Research

Open Access

Research Article

The Gap and Opportunity in Connecting Genomic and Phenotypic Data

Peter T Engel*

52 Winthrop St, Newton, MA 02465, USA

Abstract

Correlation between genomic and phenotypic data has the potential to unlock understanding of complex disease and new drug targets. The goal of this work is to analyze cross-correlated data from the International HundredK+ Cohorts Consortium (IHCC) compilation of genome-wide association studies and outline key gaps and opportunities. Out of the more than 35.5 million patient entries, only 17% contain both genomic and clinical data, only 10% contain both genomic and bio-specific sample data, and only 2% contain imaging data. Further, the data are spread across many cohorts and have different data structures and definitions. Of the 7 different combinations of data categories, the area with greatest potential value is the genomic-clinical combination and is thus recommended as the area for greatest focus of further data collection and focus.

Keywords: Genomic, Phenotypic, GWAS, Data Correlation, Disease Targets.

Introduction

Human biology is complex and most diseases stem from that complex biology going awry. Very few diseases stem from a single genetic mutation. The genomic revolution has been moving relatively slowly because even comparing the full genome of a small group of sick people to healthy people rarely reveals obvious druggable targets. The key to unlocking greater therapeutic potential is to develop a full biological picture of critical disease processes, which involves genomics, transcriptomics, proteomics, and metabolomics, followed over a period of time and tied to phenotypic and health outcomes. The discernment of the biological pathways for most diseases is immature at best. Maslove et al., argues that critical illness should be characterized by multi-omic fingerprints [1].

Since the first sequencing of the human genome in 1990, there has been an accelerated effort to obtain larger and larger data sets for exploring the link to disease risk. The company deCODE Genetics was launched in 1996 in Iceland with the aim of using population genetics studies to identify variations in the human genome associated with common diseases, and to apply these discoveries "to develop novel methods to identify, treat and prevent diseases." More recently companies such as 23andMe have launched consumer-oriented business in which people send saliva (containing DNA) samples and in return receive information on ancestry and certain limited genetic predispositions to health-related topics. However, no new drugs or drug targets have so far come from these massive efforts.

This paper explores all of the data that is now being generated from genome-wide association studies (GWAS) around the world and seeks to identify the critical gaps in the data that will be most helpful in order for researchers and drug discovery efforts to find new conclusions about disease and new drug targets.

Materials and Methods

There are now at least 70 large cohort studies with the purpose of gathering data for translational research – in other words translating clinical, genomic, imaging, and diagnostic data into insights to improve care and population health. These cohort studies do not include the hundreds of thousands of clinical trials performed by pharmaceutical companies and academic investigators looking to decipher the impact of specific health interventions such as new drugs, nutritional supplements, exercise, and more. Here the International HundredK+ Cohorts Consortium (IHCC) Atlas, which has summarized the enrollment, geography, and types of data collected from all the major current cohort studies is used as the core analytical dataset [2].

***Corresponding Author:**

Peter T Engel, 52 Winthrop St, Newton, MA 02465, USA
peter.thorv.engel@gmail.com

Received Date: 20 Feb, 2024
Accepted Date: 27 Feb, 2024
Published Date: 15 Mar, 2024

In this work, the type of information from each of the cohorts with other types has been cross-correlated. For example, if a cohort has 100,000 participants where 50% donated genomic information and 50% donated clinical data, then one can expect that the cohort contains 50% x 50% x 100,000 = 25,000 data sets containing both genomic and clinical data. Adding up all of these "complete" datasets over the 70 cohort studies, a total is obtained. For the genomic and clinical data sets, there are about 6 million records or about 17% of the 35,5 million records in total. The four categories of data examined were: 1) genomic data, 2) clinical data, 3) imaging data, and 4) bio-specific data including diagnostic samples.

Results and Observations

Of the 35.5 million cohort data participants, significant gaps exist in obtaining information on both genomic and phenotypic data.

Discussion

Despite a quickly accelerating pace of healthcare data accumulation, surprisingly little of this data is truly facilitating disease understanding and new drug target identification. Less than six million data records with both genomic and clinical data is a wholly inadequate number especially given that these data are not available in one cleaned database. They are in 29 separate databases with different barriers for access, different data coding, and most importantly different basic genomic and clinical definitions and information. The lowest hanging fruit for further research may not be to gather more information, but rather to organize, clean, and congregate the currently available data to enable computational approaches that can more efficiently lead to revolutionary healthcare discoveries.

Of course, obtaining different kinds of information comes with different costs. Figure 2 highlights the value vs the cost of obtaining information where the bubble size indicates the amount of information we have today. Thus, future efforts should be focused on those areas that a) have the greatest value, b) is the easiest (cheapest) to obtain, and c) has the least data today. Clearly genetic information in combination with clinical data and genetic information in combination with data from bio-specific tests are where most efforts should be focused.

	Genomic data	Clinical data	Imaging data
Genomic			
Clinical	17%		
Imaging	2%	2%	
Bio-specific	10%	21%	3%

Table 1: Richness data cross-correlation. Numbers indicate percentages of total data (35,5 million participants) that has two data categories, which can then be cross-correlated. Areas of most data in green with those more sparse in red.

Table 1 summarizes the major findings of this work.

A. Genomic with clinical data (17%). This category is the broadest source for multi-omic disease identification

B. Genomic with imaging data (2%). This category is primarily a source for cancer fingerprinting

C. Genomic with bio-specific data (10%). This category could be a rich source of drug-specific targets

D. Clinical data with imaging data (2%). Low priority category but could have odd case examples that could shed light on certain diseases

E. Clinical data with bio-specific data (21%). Category for identification of diagnostic markers

F. Imaging data with bio-specific data (3%). Low priority category

Only about 0.5 million data sets contain all information (1.4%). This reflects the limited nature of these data sets.

Interestingly, similar low correlation numbers were also found in a recent review of global microbiome studies [3]. Abdill et al conducted a meta-analysis of 2,592 studies, which collected microbiome data from 19 body sites.

Another conclusion from analyzing the cohort data is that the vast majority (90%) of information collected is from populations with European ancestry (counting the U.S. as such). This observation was also made by Fitipaldi and a particularly relevant discussion on impact of innovation for diabetes is included in [4,5].

Further, to truly understand complex chronic diseases such as heart failure or diabetes, information on lifestyle should also be included, which today is increasingly available from step counters, smart phones, etc.

The quickly-expanding healthcare data universe is a treasure trove for humanity. The key to unlocking the understanding of disease and new miracle drugs may exist in the data we already have. Denny et al outline seven ways to begin this journey: starting with huge longitudinal cohorts including routine clinical genomics, electronic health records, and phenomics enriched by diversity and inclusion and protected by privacy and trust, and finally insights unlocked by artificial intelligence [6]. Let's take the next steps.

Conflict of Interest

The author has no conflict of financial interest.The author has no conflict of financial interest.

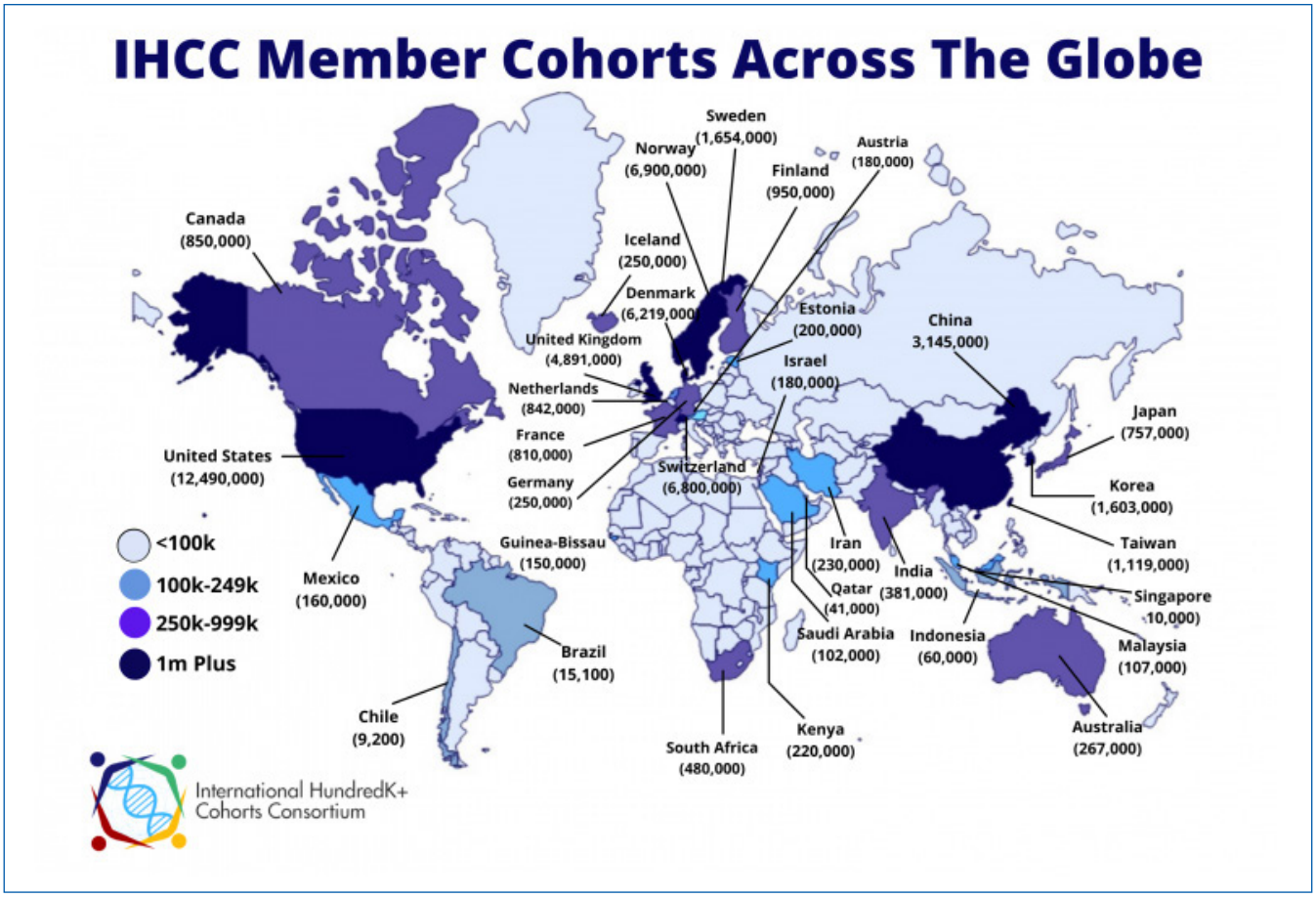

Figure 1: Geographic representation of the IHCC data cohorts

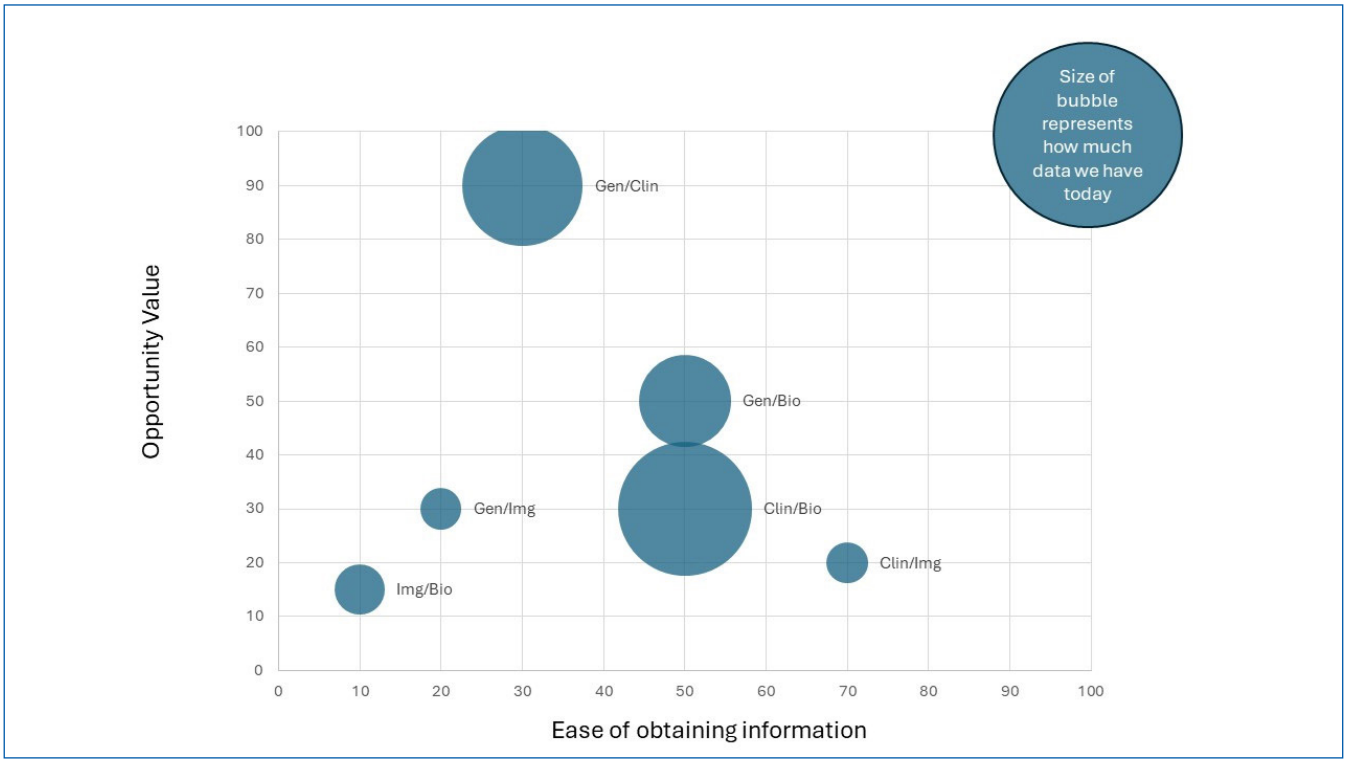

Figure 2: Opportunity vs ease of obtaining information

References

1. Maslove DM, Tang B, Shankar-Hari M, Lawler PR, Angus DC, et al. (2022) Redefining critical illness. Nature Medicine 28:1141-48.
2. IHCC Cohort Atlas [Internet]. Available from: https://atlas.ihccglobal.org/
3. Abdill RJ, Adamowicz EM, Blekhman R (2022) Public human microbiome data are dominated by highly developed countries. Cadwell K, editor. PLOS Biology 20: e3001536.
4. Fitipaldi H, Franks PW (2022) Ethnic, gender and other sociodemographic biases in genome-wide association studies for the most burdensome non-communicable diseases: 2005–2022. Human Molecular Genetics 32: 520-32.
5. Fitipaldi H, McCarthy MI, Florez JC, Franks PW (2018) A Global Overview of Precision Medicine in Type 2 Diabetes. Diabetes 67:1911-22.
6. Denny DC, Collins, FS (2021) Precision medicine in 2030-seven ways to transform healthcare. Cell 184: 1415-9.