


Family studies in the age of big data

Laura Almasy^{a,b,c,1} 

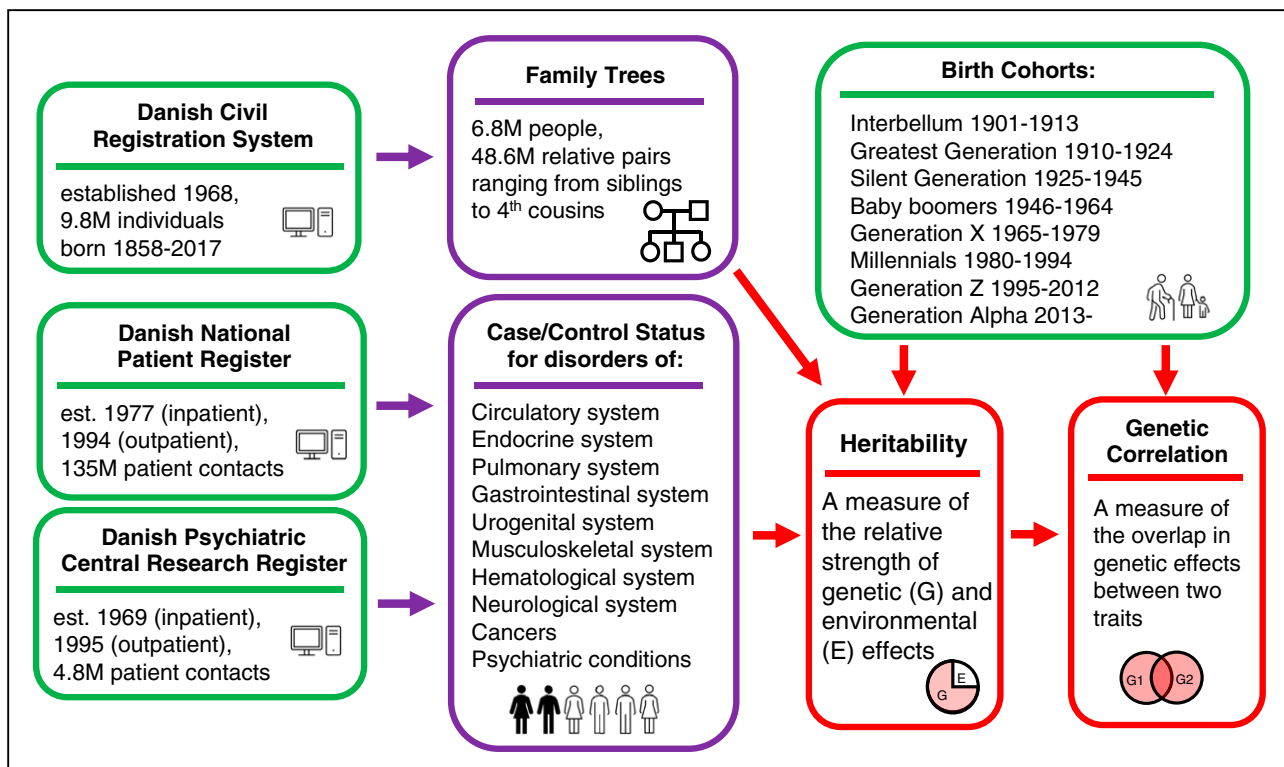


Fig. 1. Overview of the Athanasiadis et al. (1) study. Green boxes represent study inputs, purple boxes represent analytical products, and red boxes represent study outcomes. E, environmental; est., established; M, million; G, genetic.

Family studies are a powerful tool for quantifying the extent to which genetic factors contribute to variation among individuals and for exploring the interaction of genetic and environmental factors. They can also be used to study shared genetic effects across traits, providing clues about etiological pathways. In PNAS, Athanasiadis et al. (1) describe what is one of the largest human family studies to date, including over 6 million individuals who form over 48 million pairs of relatives spanning as many as six generations. They achieved this massive feat by drawing parent-child relationships from the Danish Civil Registration System and connecting overlapping sets of parents and children to identify siblings, aunts, uncles, grandparents, and cousins to build large family trees. This pedigree information was then combined with records from the national health care system to score each individual for the presence or absence of medical disorders in 10 major categories (Fig. 1).

The resource built by Athanasiadis et al. (1) is extremely flexible and will facilitate numerous future genetic studies of both rare and common disorders. In this introductory paper, the authors have used it to examine the relative strength of genetic effects (i.e., heritability) on general categories of disorders and the overlap in genetic effects

between disorders (i.e., genetic correlation). They also explored how these effects changed over time. This analysis had several uncommon features that complement what we know from other studies. First, we typically estimate heritability for a single disorder at a time. It is less common to estimate heritability for general categories of disorders, such as any urogenital disorder or any gastrointestinal disorder. Often, we are concerned with maximizing the genetic signal for gene localization by ensuring that we have a single diagnostic entity, which will hopefully reduce

Author affiliations: ^aDepartment of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA 19104; ^bDepartment of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104; and ^cLifespan Brain Institute, Children's Hospital of Philadelphia and Penn Medicine, Philadelphia, PA 19104

Author contributions: L.A. wrote the paper.

The author declares no competing interest.

Copyright © 2022 the Author(s). Published by PNAS. This article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

See companion article, "A comprehensive map of genetic relationships among diagnostic categories based on 48.6 million relative pairs from the Danish genealogy," [10.1073/pnas.2118688119](https://doi.org/10.1073/pnas.2118688119).

¹Email: almasyl@upenn.edu.

Published March 28, 2022.

the underlying genetic and etiological heterogeneity. In contrast, Athanasiadis et al. (1) are interested in general liability or the idea that heritable factors may affect an entire body system and predispose an individual to multiple disorders within or across body systems. Identifying such cross-system connections may provide valuable clues into the pathophysiology of the associated disorders.

Another unusual aspect of this work is the examination of the change in heritability and genetic correlations across birth cohorts. It is rare for a family study to include individuals spanning more than three generations with large sample sizes in both earlier and more recent birth cohorts. Health information from the Danish National Patient Register and the Danish Psychiatric Central Research Register was available for people born over a period of more than a century, from 1901 to 2017. This led to the intriguing observation that heritability for all disorders increased in

In PNAS, Athanasiadis et al. describe what is one of the largest human family studies to date, including over 6 million individuals who form over 48 million pairs of relatives spanning as many as six generations.

more recent generations. Note, however, that if we divide the sources of variation between individuals into genetic (G) and environmental (E) components, heritability (h^2) is defined as the proportion of the overall variance that is attributable to genetic factors: $h^2 = G/(G + E)$. This means that heritability can change through alterations in the strength of either genetic or environmental factors. If the contribution of environmental factors toward variation in disease risk is decreased through improved nutrition or improved health care, the total variance, which is the denominator in the heritability equation, would decrease, and thus, the heritability could increase without the genetic component changing. This complicates the interpretation of increasing heritability in more recent cohorts. In contrast, for pairs of body systems showing shared genetic effects, Athanasiadis et al. (1) find that the genetic correlations are generally stronger in older generations and decreased in more recent cohorts, although there are some exceptions to this pattern.

Family studies can be difficult to conduct as they place additional limitations on subject recruitment and require specialized approaches to data analysis. Many large-scale gene localization studies these days rely on unrelated individuals either recruited for the presence or absence of a particular disorder or chosen without regard to phenotype in the hopes of capturing the range of variation within a population. In this study design, family relationships are treated as a nuisance, and often, study participants are filtered to remove closely related individuals. Recently, however, investigators have sought to utilize the information in these family relationships and have begun to assemble families from what were originally large population-based studies, such as the UK Biobank, or from registry-based studies.

The Athanasiadis et al. (1) Danish registry cohort joins similar extended genealogy family studies assembled from national population registries or health insurance databases in Taiwan (2–4); Sweden (5–10); Manitoba, Canada (11, 12); Western Australia (13); Norway (14, 15); and Iceland (16). The availability of multiple such cohorts from around the world is important as comparisons between results from different studies will help to control for and disentangle sources of variation that differ between populations, such as cultural or environmental factors and genetic ancestry. These cohorts have been used to study familial risk in heart disease (3, 8), renal disease (2), inflammatory bowel disease (17), cancer (4, 5, 15), fracture risk (11, 12), cardiorespiratory conditions (12), and psychiatric disorders (6, 7, 9, 10, 13, 18).

An advantage of these registry-based studies is that they include nearly everyone born or receiving medical care in the geographic territory covered by the registry.

This represents an important advantage for genetic studies. In a typical study, we are often concerned about the hidden biases in how a sample was selected and the implications of these biases for the generalizability of study findings. Participants in scientific studies may not be representative of the underlying population.

They may be wealthier or more educated than average. Often, females are more likely to participate in research than males. In some cases, individuals with more severe disease may be more motivated to participate in research than individuals with milder symptoms. In other cases, individuals who are the most severely affected may not have the time or ability to participate in research. Study participants also may not represent the full racial, ethnic, and cultural diversity of the population. Registry-based cohorts with essentially complete ascertainment of an entire population avoid these biases and may provide samples in which to examine the effects of ascertainment bias through analyses of subsets of the cohort.

A little more than a century ago in 1918, R. A. Fisher (19) wrote a seminal paper in which he laid out the expected resemblance between relatives of different degrees of relationship presuming a trait was influenced by a large number of genetic factors, each subject to the laws of Mendelian inheritance. Fisher's paper formed the theoretical basis for classical quantitative genetics and family studies, which have been widely used to establish the heritability of medical disorders and anthropometric traits as well as in agriculture and ecology (19). The work of Athanasiadis et al. (1), along with other registry-based genetic studies, brings these classical quantitative genetic approaches into the twenty-first century and the era of big data, allowing us to conduct family studies on a scale undreamt of a hundred years ago.

ACKNOWLEDGMENTS. The research of L.A. is supported by NIH Grants U01MH119690, R01MH119219, U01MH119737, U10AA008401, and U01MH124962 as well as the University of Pennsylvania Autism Spectrum Program of Excellence.

1. G. Athanasiadis et al., A comprehensive map of genetic relationships among diagnostic categories based on 48.6 million relative pairs from the Danish genealogy. *Proc. Natl. Acad. Sci. U.S.A.* **119**, 10.1073/pnas.2118688119 (2022).
2. H. H. Wu et al., Family aggregation and heritability of ESRD in Taiwan: A population-based study. *Am. J. Kidney Dis.* **70**, 619–626 (2017).

3. C.-L. Wang *et al.*, Familial aggregation of myocardial infarction and coaggregation of myocardial infarction and autoimmune disease: A nationwide population-based cross-sectional study in Taiwan. *BMJ Open* **9**, e023614 (2019).
4. H.-T. Lin *et al.*, Familial aggregation and heritability of nonmedullary thyroid cancer in an Asian population: A nationwide cohort study. *J. Clin. Endocrinol. Metab.* **105**, e2521–e2530 (2020).
5. K. Hemminki, K. Czene, Attributable risks of familial cancer from the Family-Cancer Database. *Cancer Epidemiol. Biomarkers Prev.* **11**, 1638–1644 (2002).
6. S. Yao *et al.*, Familial liability for eating disorders and suicide attempts: Evidence from a population registry in Sweden. *JAMA Psychiatry* **73**, 284–291 (2016).
7. A. E. Nordsletten *et al.*, Patterns of nonrandom mating within and across 11 major psychiatric disorders. *JAMA Psychiatry* **73**, 354–361 (2016).
8. M. P. Lindgren *et al.*, Mortality risks associated with sibling heart failure. *Int. J. Cardiol.* **307**, 114–118 (2020).
9. M. J. Taylor *et al.*, Etiology of autism spectrum disorders and autistic traits over time. *JAMA Psychiatry* **77**, 936–943 (2020).
10. R. Zhang *et al.*, Familial co-aggregation of schizophrenia and eating disorders in Sweden and Denmark. *Mol. Psychiatry* **26**, 5389–5397 (2021).
11. S. Yang *et al.*, Objectively-verified parental non-hip major osteoporotic fractures and offspring osteoporotic fracture risk: A population-based familial linkage study. *J. Bone Miner. Res.* **32**, 716–721 (2017).
12. S. Yang *et al.*, Parental cardiorespiratory conditions and offspring fracture: A population-based familial linkage study. *Bone* **139**, 115557 (2020).
13. V. A. Morgan *et al.*, Are familial liability for schizophrenia and obstetric complications independently associated with risk of psychotic illness, after adjusting for other environmental stressors in childhood? *Aust. N. Z. J. Psychiatry* **53**, 1105–1115 (2019).
14. Ø. Næss, D. A. Hoff, The Norwegian Family Based Life Course (NFLC) study: Data structure and potential for public health research. *Int. J. Public Health* **58**, 57–64 (2013).
15. R. Del Risco Kollerud *et al.*, Family history of cancer and risk of paediatric and young adult's testicular cancer: A Norwegian cohort study. *Br. J. Cancer* **120**, 1007–1014 (2019).
16. N. Zaitlen *et al.*, Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet.* **9**, e1003520 (2013).
17. M. Orholm, K. Fonager, H. T. Sørensen, Risk of ulcerative colitis and Crohn's disease among offspring of patients with chronic inflammatory bowel disease. *Am. J. Gastroenterol.* **94**, 3236–3238 (1999).
18. D. Bai *et al.*, Association of genetic and environmental factors with autism in a 5-country cohort. *JAMA Psychiatry* **76**, 1035–1043 (2019).
19. R. A. Fisher, The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* **52**, 399–433 (1918).