Title

A comprehensive map of genetic relationships among diagnostic categories based on 48.6 million relative pairs from the Danish genealogy

Journal

Proceedings of the National Academy of Sciences of the United States of America, 119(6)

ISSN

0027-8424

Authors

Athanasiadis, Georgios
Meijsen, Joeri J
Helenius, Dorte
et al.

Publication Date

2022-02-08

DOI

10.1073/pnas.2118688119

Peer reviewed

A comprehensive map of genetic relationships among diagnostic categories based on 48.6 million relative pairs from the Danish genealogy

Georgios Athanasiadis[a,b,c,1], Joeri J. Meijsen[a,b], Dorte Helenius[a,b], Andrew J. Schork[a,b,d], Andrés Ingason[a,b,e], Wesley K. Thompson[a,b,f], Daniel H. Geschwind[g,h,i], Thomas Werge[a,b,e,j,1,2], and Alfonso Buil[a,b,1,2]

[a]Institute of Biological Psychiatry, Mental Health Services Capital Region of Denmark, Roskilde 4000, Denmark; [b]Lundbeck Foundation Initiative for Integrative Psychiatric Research, Aarhus University, Aarhus 8210, Denmark; [c]Department of Evolutionary Biology, Ecology and Environmental Sciences, University of Barcelona, Barcelona 08028, Spain; [d]Neurogenomics Division, The Translational Genomics Institute, Phoenix, AZ 85004; [e]Lundbeck Foundation GeoGenetics Centre, Natural History Museum of Denmark, University of Copenhagen, Copenhagen 1350, Denmark; [f]Population Neuroscience and Genetics Lab, University of California San Diego, La Jolla, CA 92093; [g]Neurogenetics Program, Department of Neurology, David Geffen School of Medicine at University of California, Los Angeles, CA 90095; [h]Center for Autism Research and Treatment, Semel Institute, David Geffen School of Medicine at University of California, Los Angeles, CA 90095; [i]Department of Human Genetics, David Geffen School of Medicine at University of California, Los Angeles, CA 90095; and [j]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen 2200, Denmark

Edited by Mary-Claire King, Departments of Medicine and Genome Sciences, University of Washington, Seattle, WA; received October 11, 2021; accepted December 23, 2021

For more than half a century, Denmark has maintained population-wide demographic, health care, and socioeconomic registers that provide detailed information on the interaction between all residents and the extensive national social services system. We leverage this resource to reconstruct the genealogy of the entire nation based on all individuals legally residing in Denmark since 1968. We cross-reference 6,691,426 individuals with nationwide health care registers to estimate heritability and genetic correlations of 10 broad diagnostic categories involving all major organs and systems. Heritability estimates for mental disorders were consistently the highest across demographic cohorts (average $h^2 = 0.406$, 95% CI = [0.403, 0.408]), whereas estimates for cancers were the lowest (average $h^2 = 0.130$, 95% CI = [0.125, 0.134]). The average genetic correlation of each of the 10 diagnostic categories with the other nine was highest for gastrointestinal conditions (average $r_g = 0.567$, 95% CI = [0.566, 0.567]) and lowest for urogenital conditions (average $r_g = 0.386$, 95% CI = [0.385, 0.388]). Mental, pulmonary, gastrointestinal, and neurological conditions had similar genetic correlation profiles.

heritability | genetic correlation | human disease | register data | Denmark

Denmark, like other Nordic countries (1–4), has maintained for more than half a century population-wide demographic, health care, and socioeconomic registers that provide detailed information on the interaction between all residents and the extensive national social services system (5, 6), including familial information via parental links (7, 8). This has allowed population-based studies of the causes and consequences of disease at an unprecedented scale and detail (9).

Several studies in the Nordic countries have leveraged diagnostic information from cross-referenced civil and health care registers on pairs of close relatives for quantitative genetic studies—that is, co-occurrence and familial coaggregation, heritability and genetic correlation, and nonrandom mating (10–14). However, the dynamics of a population (e.g., changes in mating patterns and family structure, health care provision, clinical practice, and diagnostic systems) may compromise such initiatives and bias quantitative genetic estimates and inference on human behavior. Thus, realizing the potential of Nordic population and health care registers depends on insights into the structure and network properties of the entire genealogy and accounting for underlying changes in the frequencies of human traits, notably in population demographics and disease frequencies.

Here, we reconstruct the Danish genealogy using the population-wide Danish Civil Registration System that holds information on family relationships for all individuals with at least 1 d of legal residence in Denmark since 1968 (7, 8). We describe the size, structure, and network properties of the genealogy along 116 y. We leverage the cross-reference to the nationwide, public, and health care registers to estimate occurrence, heritability, and genetic correlations for 10 broad categories of medical conditions across eight consecutive demographic cohorts.

Results

Overview of the Analysis. The Danish Civil Registration System (7, 8) has been registering all people legally residing in Denmark since 1968, and it includes information about sex, date of birth, parental links, and life events (e.g., migration or death). By April 2017 (time of data freeze for this report), 9,851,330 individuals were registered in the Danish Civil Registration System. The system is linked via anonymized identification numbers to the Danish National Patient Register (6) and the Danish Psychiatric Central Research Register (5) that include all diagnostic information regarding general medical conditions

Significance

The ability to extract multigenerational family relationships from large-scale population cohorts provides a powerful means to understand the heritability of a wide range of diseases and their genetic relationships to each other. By showing how the heritability of broad diagnostic categories changes over time and how said categories are related on the genetic level, our analysis of the Danish genealogy and linked national patient registers illustrates the vast potential of this resource in current biomedical research.

Author contributions: G.A., T.W., and A.B. designed research; G.A., J.J.M., D.H., A.J.S., A.I., W.K.T., D.H.G., T.W., and A.B. performed research; G.A., T.W., and A.B. analyzed data; and G.A., T.W., and A.B. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

See online for related content such as Commentaries.

[1]To whom correspondence may be addressed. Email: georgios.athanasiadis@regionh.dk, alfonso.buil.demur@regionh.dk, or thomas.werge@regionh.dk.

[2]T.W. and A.B. contributed equally to this work.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2118688119/-/DCSupplemental.

Published February 7, 2022.

and specific psychiatric conditions, respectively, including all inpatient and outpatient contacts.

By use of the parental links, we reconstructed the Danish genealogy, which we then leveraged together with the diagnostic information from the Danish National Patient Register to study the genetic architecture of medical conditions as defined in the 8th and 10th Revisions of the International Classification of Diseases (ICD-8 and ICD-10, respectively). Inspired by the extensive comorbidity (15) between disorders affecting the same organ or characterized by the same pathology, we analyzed 10 broad diagnostic categories of medical conditions—that is, nine somatic and one mental (*SI Appendix*, Table 1).

To study changes in genetic architecture in time and identify possible epidemiological biases such as truncation and censoring, we carried out the analyses in eight consecutive demographic cohorts with characteristic cultural, political, and economic features of Western societies (*SI Appendix*, Table 2; https://www.careerplanner.com/Career-Articles/Generations.cfm). The eight cohorts are the Interbellum Generation (birth year: 1901 to 1913), the Greatest Generation (1910 to 1924), the Silent Generation (1925 to 1945), the Baby Boomers (1946 to 1964), Generation X (1965 to 1979), Millennials (1980 to 1994), Generation Z (1995 to 2012), and Generation Alpha (2013 to 2025). We note that there is a 4-y overlap between the Interbellum and Greatest Generation. Individuals born outside the eight cohorts were not considered in the analyses. In addition, we discarded individuals that died before January 1, 1977 (date on which the Danish National Patient Register was established).

Genealogy Network Structure. To inform downstream heritability and genetic correlation analyses, we initially determined the size and structure of the Danish genealogy by examining its network properties (Fig. 1).

Of the 9,851,330 registered individuals, 6,801,107 (69.04%) had at least one registered relative, while 3,050,223 (30.96%) were unconnected singletons and were therefore excluded from further analysis (Fig. 1*A*). The single largest pedigree includes 5,396,661 individuals—that is, 54.78% of all registered individuals and 79.35% of the individuals with at least one relative (Fig. 1*A*). The genealogy also includes 251,513 smaller unconnected pedigrees (n = 1,404,446), among which there are 100,400 trios and 58,804 quartets (Fig. 1 *A* and *B*).

The 6,801,107 connected individuals span only six generations and include 2,377,043 founders—that is, individuals with no parental links. It is expected that some of the founders are closely related (e.g., siblings or cousins), but, in the lack of parental links or genetic information, we are unable to consider this in our analyses. The narrow generation span combined with the high number of founders has implications in the ascertainment of distant relative pairs (Fig. 1*C*). As a result, 29,739,188 out of 41,798,152 annotated relative pairs (71.15%) are concentrated within a radius of three meioses, encompassing parent–offspring, full siblings, half siblings, grandparent–grandchild, avuncular, half avuncular, and great grandparent–great grandchild pairs (Fig. 1*D*).

Disease Prevalence in the Danish Genealogy. The two oldest (Interbellum and Greatest) and the one youngest (Generation Alpha) demographic cohort had considerably fewer individuals (N~45,103 to 325,066) and were therefore expected to be less informative than the five larger cohorts (N~1,020,953 to 1,489,329) (*SI Appendix*, Table 2).

Disease prevalence for all 10 diagnostic categories peaks in the Greatest and the Silent Generation and declines to a minimum in Generation Alpha (*SI Appendix*, Table 3). Circulatory conditions constitute the most frequent category, affecting 61.9% of the individuals in the Greatest Generation, whereas hematological and musculoskeletal conditions are the least frequent categories, affecting at their peak 12.11 and 12.71% of individuals in the Greatest and Silent Generations, respectively (*SI Appendix*, Table 3 and Fig. 1).

While the decline in relative frequency is very similar across diagnostic categories, consistent with a uniform age-dependent effect on the age of onset of disease, mental and pulmonary conditions are characterized by distinct profiles, remaining at elevated frequency until the two youngest cohorts—that is, Generation Z and Generation Alpha (*SI Appendix*, Fig. 2).

Heritability. We estimated heritability (h^2) of the 10 diagnostic categories (15) by applying the latent correlation of relative pairs to Falconer's method (16). In our analysis, we considered all family relations within a radius of three meioses (i.e., all up to second degree and great grandparents/great grandchildren) because these were abundant enough to yield accurate estimates.

For all 10 diagnostic categories, heritability increases across demographic cohorts and peaks in Generation Z. Due to truncation, censoring, and data scarcity that characterize the oldest and youngest generations, we consider estimates from the four midmost and largest cohorts—that is, Silent Generation, Baby Boomers, Generation X, and Millennials—to be a priori more reliable (Figs. 2 and 3).

Estimates of heritability varied notably between diagnostic categories (Figs. 2 and 3) and consistently across demographic cohorts as reflected in their cross-cohort weighted average estimates (Table 1). Heritability estimates for mental disorders were consistently the highest across demographic cohorts (average h^2 = 0.406, 95% CI = [0.403, 0.408]), whereas estimates for cancers and neurological conditions were the lowest (average h^2 = 0.130, 95% CI = [0.125, 0.134] and average h^2 = 0.154, 95% CI = [0.151, 0.157], respectively).

Heritability could not be estimated for some diagnostic categories in the two oldest and the two youngest demographic cohorts due to data scarcity (Figs. 2 and 3 and Table 1). In addition, most heritability estimates were similar when analyses were restricted to full sib pairs only, although the consideration of multiple sib pairs from the same family resulted in wider CIs (*SI Appendix*, Figs. 3 and 4 and Table 4). A similar trend was observed when analyses were restricted only to individuals born in Denmark (n = 6,017,195) as reflected in the high correlation between measures (r = 0.94; *SI Appendix*, Fig. 5).

Finally, we note that with the notable exception of cancers and conditions of the hematological system, no single disease seems to dominate the 10 diagnostic categories under study (*SI Appendix*, Fig. 6)—and consequently, the corresponding heritability estimates.

Genetic Correlations. To understand the mutual relationships between the 10 broad diagnostic categories (15), we estimated their genetic correlations (r_g) by combining within- and between-category estimates of the latent correlation into Falconer's method (16). We considered all family relations within a radius of three meioses and restricted the analyses to the four most data-rich demographic cohorts mentioned in *Genealogy Network Structure* (Fig. 4 and Dataset S1).

All r_g except two were positive, and all of them except one were also significantly different from zero. Overall, r_g were highly consistent between consecutive cohorts, thus further boosting confidence in the estimates (*SI Appendix*, Fig. 7). This trend was more marked for certain diagnostic categories such as mental, pulmonary, and neurological than others. In all 10 diagnostic categories, younger cohorts showed lower r_g than older generations, whereas the opposite trend was observed for heritability that consistently increased in younger cohorts (Fig. 4 and *SI Appendix*, Dataset S1). The average r_g of each of the 10 diagnostic categories with the other nine categories was highest for gastrointestinal conditions (0.567; SE = 0.0005) and lowest for urogenital conditions (0.386; SE = 0.0008).

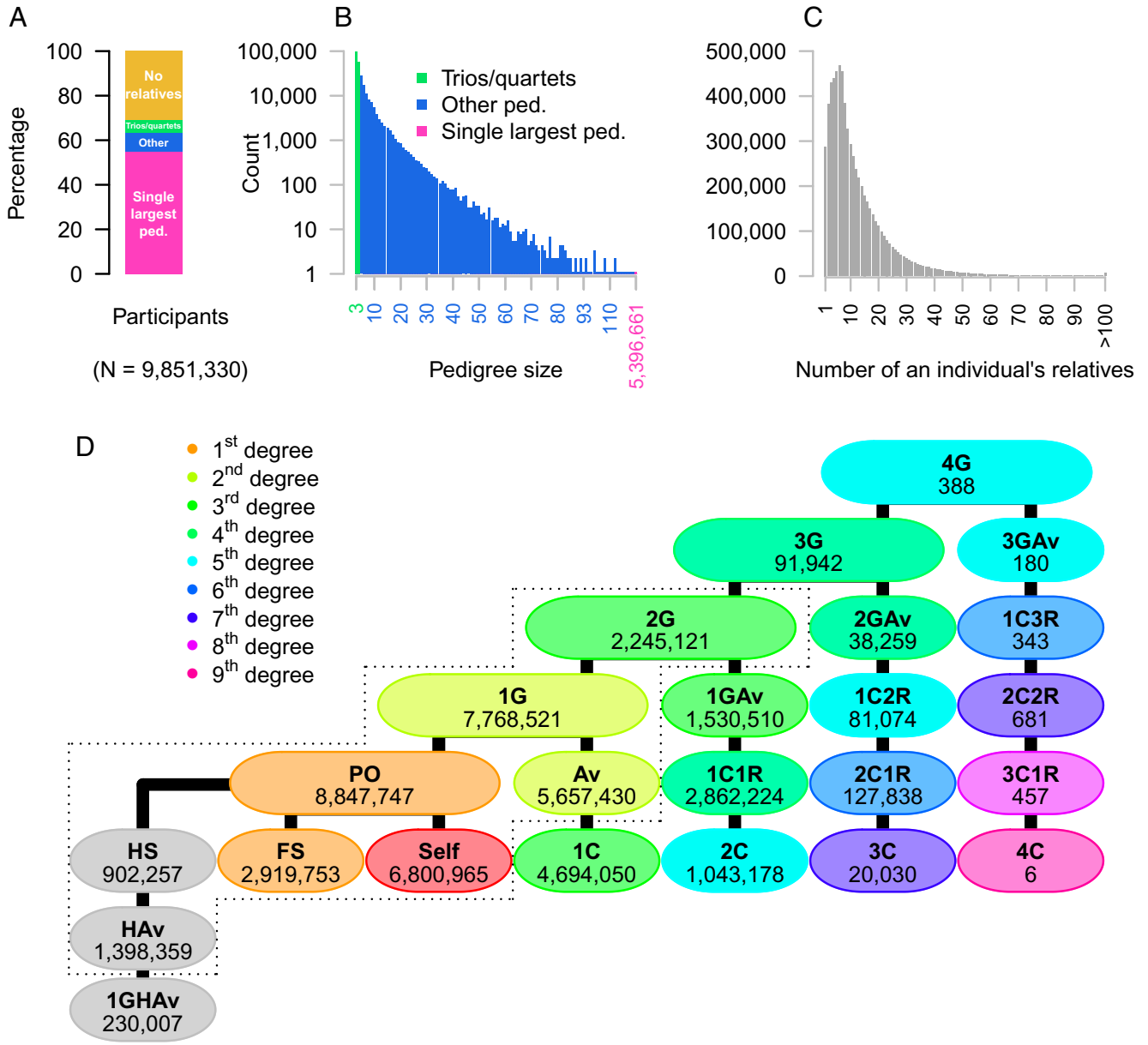

Fig. 1. Size and structure of the Danish genealogy. (*A*) Membership of the 9,851,330 registered participants in the identified network components. The vast majority of the participants (69.04%) had at least one known relative (pink, blue, and green). (*B*) Frequency of network components ordered by pedigree size. One component with size n = 5,396,661 (pink) includes 79.35% of the connected individuals. (*C*) Frequency of number of an individual's relatives. The Danish genealogy is dominated by individuals with few relatives. (*D*) Frequency of familial relationships and relative position to "Self." Color-coding corresponds to degree of relationship. FS, full siblings; HS, half siblings; 1C, first cousins, etc.; PO, parent–offspring; 1G, grandparent–grandchild, etc.; Av, avuncular; 1GAv, grand–avuncular, etc.; 1R, once removed, etc. The structure is enriched for close familial relationships (outlined by the dotted line).

We further used 45 cross-generation weighted average genetic correlations to hierarchically cluster the 10 diagnostic categories (Fig. 5). We observed three major clusters in the dendrogram: one including mental, pulmonary, gastrointestinal, and neurological conditions; another one involving musculoskeletal conditions and cancers; and a third one involving urogenital, hematological, circulatory, and endocrine conditions.

Finally, genetic correlation estimates based on full sib pairs alone, in which most pairings are not intergenerational, are shown in *SI Appendix*, Figs. 8–10 as well as Dataset S2 and were generally consistent with analyses based on all family relations.

Discussion

In this study, we present the Danish genealogy constructed from the Danish Civil Registration System, which holds information on all individuals born or with residence in Denmark since 1968. The genealogy extends back up to six generations, with the oldest connected individuals being born in 1872 and the youngest in 2017. We partitioned 6,691,426 Danish citizens into eight demographic cohorts based on year of birth. Notably, by cross-linking the Danish Civil Registration System with hospital discharge diagnoses from the public and egalitarian Danish health care system, we were able estimate heritability and

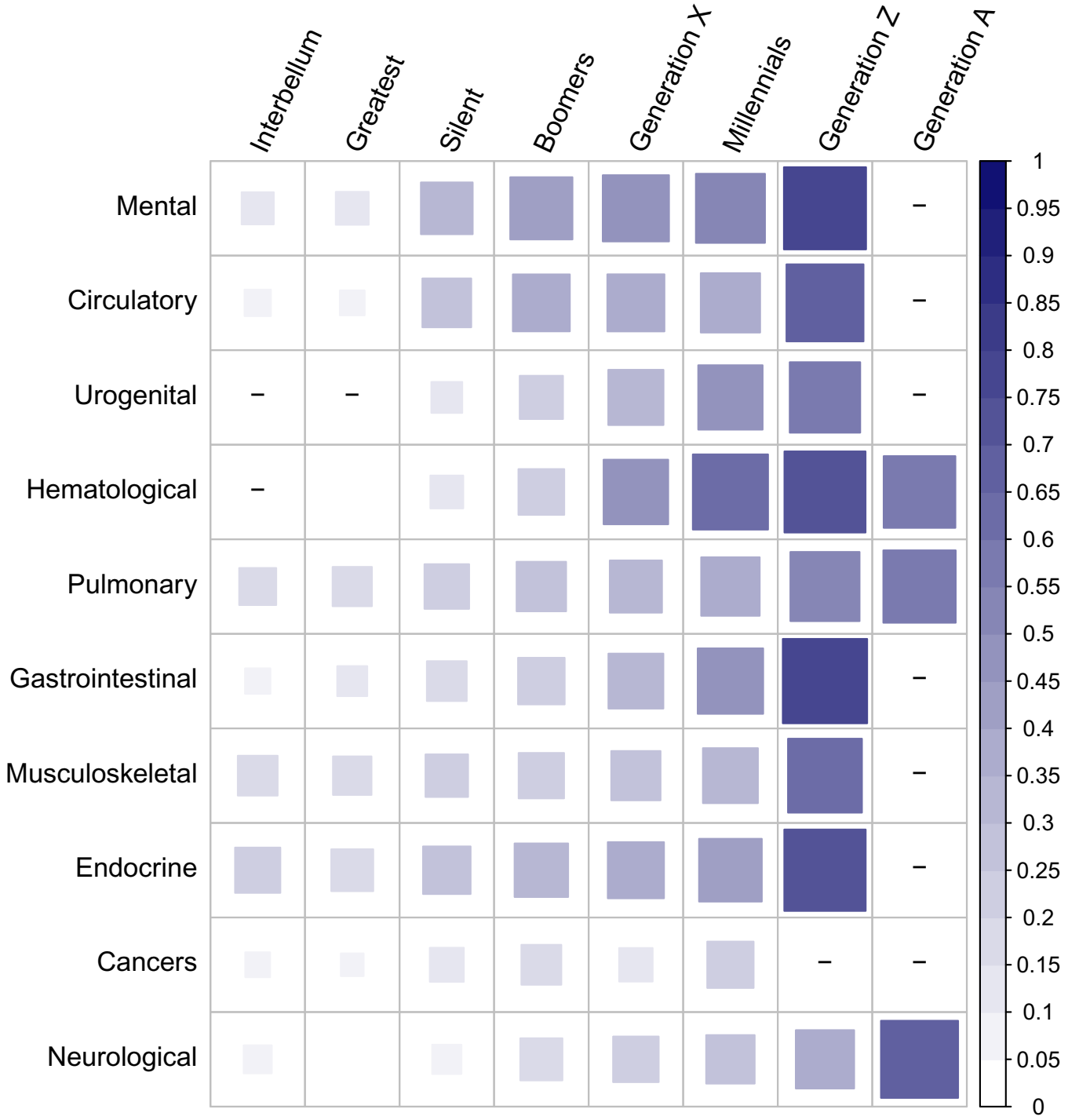

Fig. 2. Heritability estimates of 10 broad diagnostic categories by demographic cohort. Most estimates correspond to average values from all available relative pairs weighted by sampling variance. Least-squares estimates are reported for the hematological category in Generation Z and Generation Alpha. Both tile size and shade intensity are proportional to heritability values. All estimates were significantly different from zero.

genetic correlations for 10 broad diagnostic categories encompassing all major organ systems and most ICD-8/ICD-10 codes while describing the epidemiological biases of truncation and right censoring in the oldest and youngest demographic cohorts, respectively.

The heritability of single diseases and genetic correlations between them have been studied extensively not only in family data but also thanks to the development and application of genome-wide association studies to thousands of human traits (17). In a few instances (e.g., for mental disorders), genetic risk variants shared across diagnoses with clearly distinct clinical characteristics and age of onset have been identified (18). However, neither the heritability nor the genetic correlations have been systematically studied for organ-defined disease categories as grouped by 10 chapters of ICD-10. In addition, such studies have never been carried out within a single population such as the Danish, serviced and monitored uniformly for decades by an egalitarian health care system.

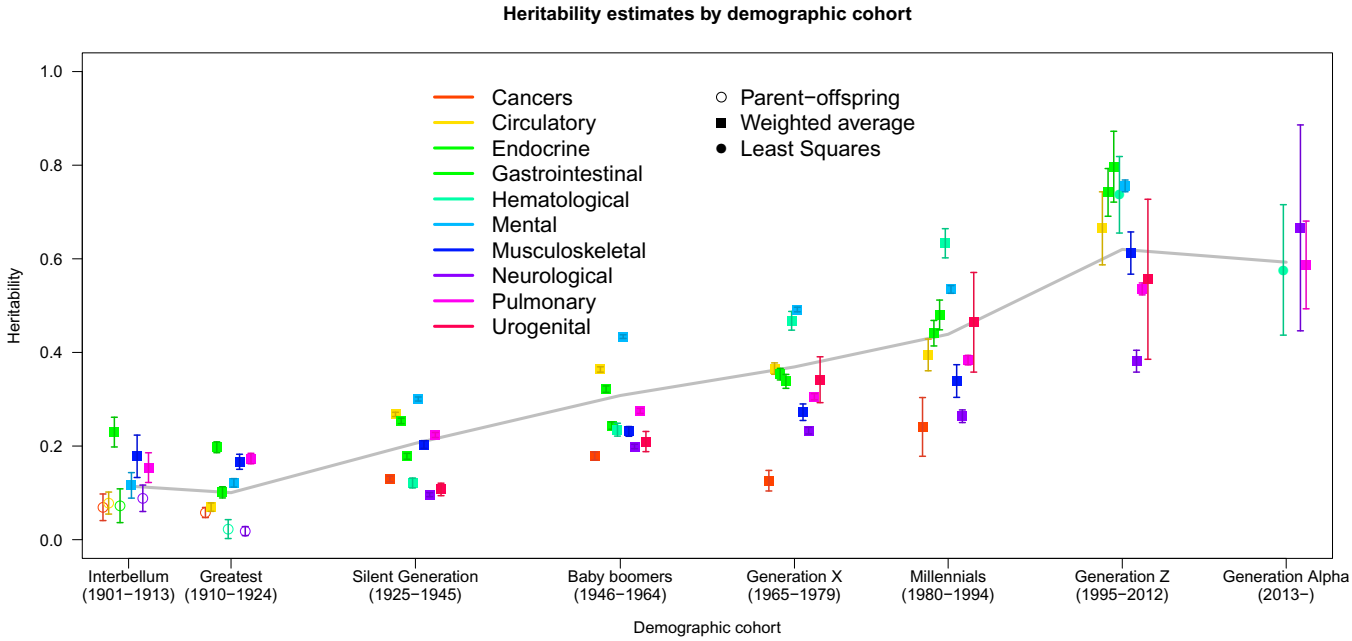

Fig. 3. Heritability estimates (and 95% CIs) of 10 broad diagnostic categories by demographic cohort. Estimates were derived from parent–offspring pairs alone (empty circles), averages from all available relative pairs weighted by sampling variance (filled squares), or least-squares regression (filled circles). The gray line corresponds to cross-category weighted average estimates.

We estimate the heritability to be high for several of the 10 disease categories. This is consistent with high genetic correlation between individual diagnoses within each category as reported for mental disorders (18) and more broadly for brain disorders (19) as well as with the broader notion that genetic liability is generally organ specific. For mental conditions in particular, heritability point estimates reach 0.7, which is higher than reported for the common and less heritable mental disorders such as depression (0.4) (20) and anxiety (0.3~0.4) (21) and similar to those for highly heritable, rare illness, such as schizophrenia (0.81) (22) and bipolar disorder (0.6~0.8) (23).

Moreover, the lower heritability estimates in older cohorts and the higher heritability estimates in younger cohorts might be because disease risk is generally plateauing in the former, whereas accumulation of diagnoses in the latter is an ongoing process, interrupted by right censoring. Younger cohorts are therefore enriched for younger ages of onset, which in many instances go along with stronger genetic signals and higher heritability estimates as known for mental disorders in which early onset disorders such as autism and attention-deficit/hyperactivity disorder are commonplace. It could also be posited that the accumulation of environmental exposures throughout life renders nongenetic factors more important in aging-related conditions, thus resulting in overall lower heritability estimates in older cohorts. On the other hand, stronger genetic correlations in older cohorts might be due to the accumulation of comorbidities in older cohorts compared to younger cohorts.

The fact that genetic correlations were almost exclusively positive across all cohorts probably reflects how diseases, at least in the broad composite definitions we use in this work, are problems of the normal functioning of organs and systems, whereby the disorganization of one or more of them should be detrimental for others, ultimately resulting in further pathology. The positive genetic correlations match comorbidity observations in the clinical domain.

Notably, we observe that the ranking of heritability and average genetic correlation estimates compare for most of the 10 diagnostic groups, although there are also marked exceptions. Mental, gastrointestinal, and circulatory conditions rank high both for heritability and average genetic correlation, whereas neurological conditions, despite showing the lowest heritability estimates, are genetically highly correlated with the other

Table 1. Heritability of 10 broad diagnostic categories across eight demographic cohorts

	Interbellum			Greatest			Silent			Baby boomers			Generation X			Millennials			Generation Z			Generation Alpha			Cross-cohort average	
	h^2	SE	Method	h^2	SE	Method	h^2	SE	Method	h^2	SE	Method	h^2	SE	Method	h^2	SE	Method	h^2	SE	Method	h^2	SE	Method	h^2	SE
Cancers	0.0692	0.0145	PO	0.058	0.0055	PO	0.1297	0.003	WA	0.1787	0.0044	WA	0.126	0.0112	WA	0.2409	0.032	WA	—	—	—	—	—	—	0.1295	0.0022
Circulatory	0.0782	0.012	PO	0.0695	0.0046	WA	0.2676	0.0023	WA	0.3636	0.0029	WA	0.3655	0.0062	WA	0.3947	0.0173	WA	0.6651	0.0398	WA	—	—	—	0.2777	0.0016
Endocrine	0.2298	0.0162	WA	0.1974	0.0059	WA	0.2537	0.003	WA	0.3216	0.0035	WA	0.3536	0.0063	WA	0.4411	0.0139	WA	0.7419	0.026	WA	—	—	—	0.2850	0.002
Gastro-intestinal	0.0725	0.0184	PO	0.1011	0.0061	WA	0.1781	0.0032	WA	0.2438	0.004	WA	0.3382	0.0076	WA	0.4801	0.0162	WA	0.7968	0.0386	WA	—	—	—	0.2067	0.0022
Hemato-logical	—	—	—	0.0226	0.0103	PO	0.1211	0.0054	WA	0.235	0.007	WA	0.4676	0.0102	WA	0.6332	0.0159	WA	0.7368	0.0417	LS	0.5763	0.0711	LS	0.2128	0.0036
Mental	0.116	0.0139	WA	0.1214	0.0046	WA	0.3005	0.0022	WA	0.4338	0.002	WA	0.4917	0.0025	WA	0.5345	0.0042	WA	0.7558	0.0064	WA	—	—	—	0.4057	0.0012
Musculo-skeletal	0.1781	0.0231	WA	0.1664	0.0082	WA	0.2033	0.0042	WA	0.2313	0.0053	WA	0.2723	0.009	WA	0.3389	0.0178	WA	0.6123	0.023	WA	—	—	—	0.2227	0.0028
Neurological	0.0884	0.0144	PO	0.0183	0.0049	PO	0.0961	0.0022	WA	0.199	0.0024	WA	0.2329	0.0036	WA	0.2637	0.0069	WA	0.3813	0.0119	WA	0.6661	0.1122	WA	0.1538	0.0014
Pulmonary	0.1539	0.0162	WA	0.1733	0.0057	WA	0.2236	0.0027	WA	0.275	0.0027	WA	0.3054	0.0033	WA	0.3836	0.0052	WA	0.5357	0.0065	WA	0.5869	0.0478	WA	0.2804	0.0015
Urogenital	—	—	—	—	—	—	0.1073	0.0068	WA	0.2096	0.011	WA	0.3417	0.025	WA	0.4644	0.0543	WA	0.5563	0.0872	WA	—	—	—	0.1511	0.0056
Cross-trait average	0.1145	0.0054		0.1004	0.0019		0.2058	0.0009		0.308	0.001		0.3691	0.0015		0.4388	0.0027		0.6201	0.0041		0.5928	0.0374			

PO, parent–offspring; WA, weighted average; LS, least-squares regression.

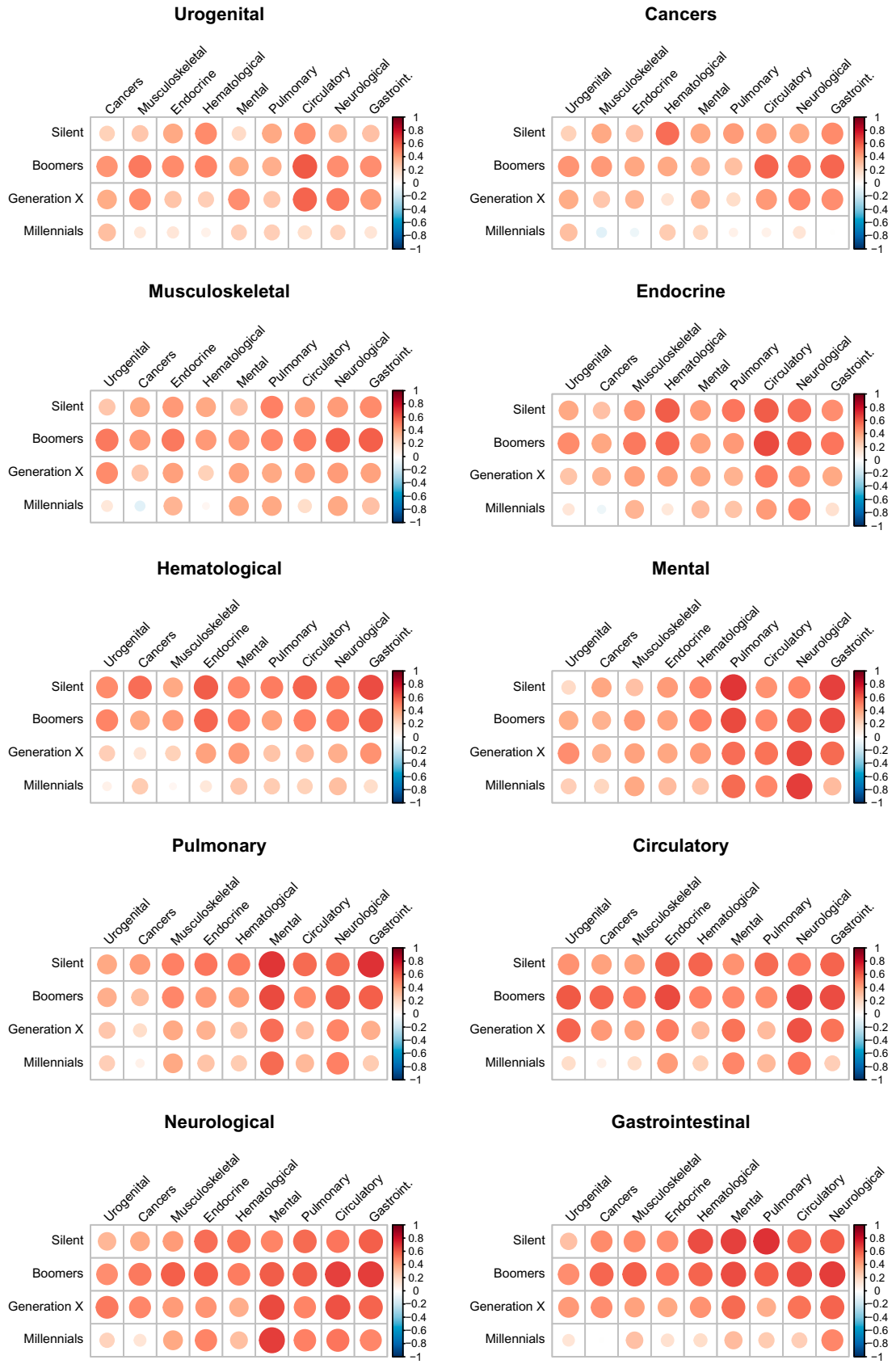

Fig. 4. Genetic correlations of each of 10 broad diagnostic categories with the remaining nine by demographic cohort. Only the four most data-rich cohorts—Silent Generation, Baby Boomers, Generation X, and Millennials—were considered. Estimates were based on averages from all available relative pairs within a radius of three meioses weighted by sampling variance. Blank cells correspond to correlations not significantly different from zero.

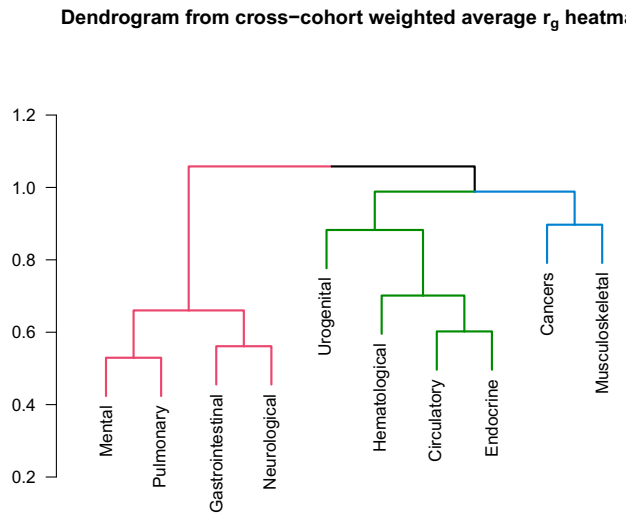

Fig. 5. Dendrogram of 10 broad diagnostic categories based on cross-cohort weighted average genetic correlations. Only the four most data-rich cohorts—Silent Generation, Baby Boomers, Generation X, and Millennials—were considered. Estimates were based on averages from all available relative pairs within a radius of three meioses weighted by sampling variance.

diagnostic groups, implicating broadly the etiology of disease affecting the nervous system in disorders of most other organ systems. Contrary, other low-heritability groups, such as cancer and musculoskeletal conditions, have low genetic correlations suggestive of their etiologies being dominated by disease-specific, environmental exposures and somatic mutations for the former and accidents for the latter. Similarly, endocrine conditions, dominated by type 2 diabetes, have relatively low heritability, possibly reflecting behavioral causes.

While circulatory and gastrointestinal conditions are the most heritable and genetically correlated diagnostic categories, their patterns of genetic correlation with other diagnostic categories are nonetheless highly diverse. In fact, gastrointestinal conditions were clustered with neurological and mental disorders, and while the clustering of the two latter disease categories of the nervous system could be anticipated and possibly reflects organ-specific components of their heritability, their proximity to gastrointestinal conditions is notable and may stem from the extensive innervation that underlies the gut–brain axis and the proposed relation between gut microbiota for brain functioning and mental health (24). Contrary to the proximal clustering of brain and gut disorders reflecting shared organ specificity or functionalities, that of endocrine with circulatory conditions as well as that of cancers with hematological illnesses more likely reflects sequelae in which one illness is a consequence or complication of a prior and otherwise, unrelated condition, in case, diabetes leading to circulatory complications and cancers to anemia because of bleeding from internal organs.

Although the reconstructed Danish genealogy is limited to six generations and thus dates back in time considerably less than the genealogy of Iceland (25), we note that most diagnostic categories include between a quarter- and a-half-million individuals, making this genealogy dataset highly apt for studies of heritability, genetic correlations, and the impact of behavioral and environmental changes over time. Also in comparison with Iceland, the relative shallowness of the reconstructed Danish genealogy, compared to, for instance, the much deeper Icelandic pedigree dating back to the 11th century (25), renders linking distant relatives a challenging task and supports the use of classical relative pair-based methods rather than linear mixed models. Furthermore, truncation and censoring biases in the oldest and youngest cohorts, as well as changes in the environment and clinical practices over time, favor the use of horizontal over vertical familial relationships and justify the stratification of the analysis by demographic cohort rather as opposed to a single analysis across the entire genealogy.

While this dataset is ideally poised for quantitative genetic analyses, it also presents limitations. As already discussed, truncation in the older demographic cohorts and right censoring in the younger ones can introduce bias to heritability and genetic correlation estimates. In order to explore the effects and biases of time, we split the available data into eight demographic cohorts and show that the four midmost cohorts—that is, the ones least affected by truncation and censoring—yield consistent estimates.

In addition, given the lack of genetic data, we have no means to safeguard our analysis from false paternities and adoptions. As a result, a small portion of the ascertained familial relationships may be overstated, affecting our heritability and genetic correlation estimates. Nevertheless, given the high abundance of relative pairs, we believe that the effect of these biases is limited. Similarly, the lack of parental links before the timeframe of the registries will lead to understating distant familial relationships, which could bias heritability estimates based on frameworks that utilize the entire relationship matrix such as linear mixed models. However, because our estimates are based on known family pairs, we believe that issues coming from an underestimation of familial relationships are limited in our analysis.

Furthermore, modifications in the diagnostic classification system, which changed from ICD-8 to ICD-10 in 1995, and the registration of outpatient contacts that began in the same year (9) may complicate precise tracking across demographic cohorts, although the focus on broad diagnostic categories in this study is expected to reduce this bias.

Finally, our analyses make no attempt to distinguish a priori between genetic correlation resulting from pleiotropy and co-occurrence of disease in relatives because of sequelae as discussed in the seventh paragraph of *Discussion* for cancers and anemia.

For mental disorders, the relatively high frequency in the younger cohorts coincides with the introduction of novel child and adolescent disorders in ICD-10—that is, attention-deficit/hyperactivity disorder and autism. Similarly, pulmonary conditions show increasing frequencies in younger generations consistent with increasing worldwide prevalence of smoking and asthma in young people (26). While potentially biasing our findings, changes in disease frequency across time also constitutes an entirely novel research field opening for the identification of nongenetic factors independently or through gene-environment interactions influencing risk of disease. In fact, as the habit of smoking spreads and increases during the middle of the 20th century (26) and the prevalence of pulmonary and circulatory conditions increases correspondingly, the heritability is expected to decrease; thus, modeling a shared environment in households will allow for studies seeking to identify nongenetic factors that impact disease. Such analyses can be empowered by the knowledge of geographical (co)location of the residence of Danish citizens from cradle to grave as a proxy for shared environment.

In conclusion, here we presented the Danish genealogy as a resource that, in combination with the National Health Registers, allows whole-population, quantitative genetic analysis with applications to health sciences. The presented resource and analytical framework will contribute to the advancement of precision medicine, allowing the systematic mapping of heritabilities and genetic correlations of comorbidity patterns and sub-diagnostic traits such as age of onset and treatment response and to inform on clinically relevant phenomena such as assortative mating, nonadditive genetics, and shared environment. While this and similar genealogies from the Nordic countries represent unique resources (1–4), the changes in biases, environment, and clinical practices necessitate the integration of time-dependent and survival analysis

frameworks. Explicit modeling of biases is warranted to fully exploit the oldest and youngest generations.

Materials and Methods

Danish Civil Registration System. The Danish Civil Registration System was established in 1968, registering all people alive and living in Denmark since then (7, 8). The Danish Civil Registration System includes personal identification number, sex, date of birth, and continuously updated information on vital status (e.g., migration or death). The personal identification number is virtually immutable, thus enabling accurate links across different registers. As of April 2017, the system contained 9,851,330 individuals born between January 1, 1858, and April 21, 2017.

Danish National Patient Register. The Danish National Patient Register (6) includes the medical records of all patients treated in Danish general hospital inpatient departments since January 1, 1977, as well as in outpatient clinics since 1 January 1994 (or occasionally since 1995). Since 2002, the Register also includes Danish patients treated in hospitals outside Denmark and activities in specialist medical practices not paid by the health insurance agreement. As of April 2017, the register contained 287,593,154 records with diagnostic information for the 135,070,194 patient contacts available in the dataset.

Danish Psychiatric Central Research Register. The Danish Psychiatric Central Research Register (5) was first computerized in 1969 and includes admissions to psychiatric inpatient facilities up to and including 1994. Since 1995, the Register also contains outpatient contacts to psychiatric departments. As of April 2017, the register contains 7,298,910 records with diagnostic information for the 4,826,984 psychiatric hospital contacts.

Ethics Approval. This study was approved by the Danish Health Data Authority (project no. FSEID-00003339) and the Danish Data Protection Agency. By Danish law, informed consent is not required for register-based studies.

Data Cleaning. The most important requirement for accurately establishing pairwise familial relationships is that any given individual has either no register links to their parents—that is, they are a founder—or both register links to their parents. This is to guarantee that familial relationships are not underestimated (e.g., incorrectly ascertaining half siblings instead of full siblings). Bearing this in mind, the 2017 Danish Civil Registration System data freeze includes 1) 198,892 individuals with only one parental link, 2) five individuals with two identical parental personal identification numbers, 3) 880 individuals that are adopted, 4) 3,000 individuals with same-sex parents, and 5) 123,331 individuals belonging to twin pairs/multiple births. There is overlap in the above five categories. In order to yield as many pairwise relationships as possible, instead of eliminating the aforementioned individuals, we converted them into founders—that is, we eliminated their parental links. Thus, if said individuals have descendants that meet our two-parent criterion, we can include their pairwise familial relationships in our analyses.

Connectivity. Genealogies can be analyzed as graphs—that is, a set of nodes (individuals) that are joined by edges representing parent–offspring relationships (27). Bearing this in mind, we used the networkx module in Python (28) to explore network connectivity in our data.

After eliminating invalid parental links, we converted the data into an edge list and loaded it as an undirected graph. Each edge in the graph represents a parent–offspring relationship between two nodes. If the parents of an individual are known, then two edges are added to the list (one for each parent). If no parental information is available—that is, in the case of founders—no edge is added to the list. Individuals can be entirely unconnected (singletons)—that is, they present no parental or offspring links.

The list consisted of 8,848,128 edges involving 6,801,107 individuals—that is, ∼69.04% of all available individuals in the Register. The remaining 3,050,223 individuals (30.96%) were singletons. A bit over half of those singletons (1,753,057 or 57.47%) were born in Denmark or Greenland, whereas the rest were born elsewhere. The distribution of the singletons by demographic cohort is shown in *SI Appendix*, Fig. 11. Overall, singletons born in Denmark belong to older demographic cohorts and represent childless individuals with no parental links, whereas singletons born outside of Denmark belong to younger demographic cohorts and represent immigrants without familial links in Denmark.

networkx computes the number and size of components—that is, the network subsets that are completely unconnected from all other subsets. This process returned one large component (n = 5,396,661) and 251,513 significantly smaller ones (n = 1,404,446), among which there were 100,400 trios and 58,804 quartets (Fig. 1 *A* and *B*). The single largest network component includes 54.78% of all registered individuals and 79.35% of the individuals with at least one relative (Fig. 1*A*). The overwhelming majority of the connected individuals (88.47%) were born in Denmark or Greenland. The distribution of the connected individuals by demographic cohort is shown in *SI Appendix*, Fig. 11.

Graph topology also indicated that the 6,801,107 connected individuals span only six generations; of those individuals, 2,377,043 (∼34.95%) are founders—that is, they have no parental links. The narrow generation span combined with the high number of founders has implications in the ascertainment of distant relative pairs.

Relatedness. We used the pydigree module in Python (29) in order to estimate all nonzero pairwise coefficients of expected relatedness $\hat{\pi}$ for the 6,801,107 connected individuals. pydigree reads a file in pedigree (PED) format as a directed acyclic graph and enumerates all legitimate paths connecting a given pair of individuals. From any given starting point, only paths toward previous generations are allowed as well as one change of direction at most. The lengths $g \in \mathbb{G}$ of the paths connecting a pair of individuals are used to estimate their kinship coefficient φ (30, 31):

$$\phi = \sum_{g \in \mathbb{G}} \frac{1}{2^{g+1}}.$$

We note that $\hat{\pi}$ is twice the kinship coefficient φ.

To avoid looping over unconnected individuals, we applied the procedure only to each of the 2,377,043 founders with their corresponding descendants (easily identified with pydigree). Because different founders can share descendants, we removed duplicate estimates with a Python script. Kinship coefficients for unreported pairs were assumed to be 0.

As a result of the above procedure, we obtained 44,099,130 pairs of familial relationships from the large pedigree and 4,522,710 from the rest of the smaller pedigrees, totaling 48,621,840.

Apart from the value of $\hat{\pi}$ for any given pair of individuals, we registered the number of all possible connecting paths and their corresponding length as well as node depth of each individual in the path. Combined with $\hat{\pi}$, this additional topological information allowed us to annotate the familial relationships with great accuracy (*SI Appendix*, Table 5).

The distribution of number of an individual's relatives is heavily right skewed with a long tail (mean = 12.3; median = 9; mode = 6; Fig. 1*C*). Moreover, the distribution of number of meioses between connected individuals, considering the shortest path per pair, is also right skewed with mean = 2.7, median = 3, and mode = 2. This implies that the ascertained relative pairs in the Danish genealogy are dominated by close familial relationships.

Only a negligible fraction (0.03%) of the annotated familial relationships were connected by more than two paths, consistent with very few consanguineous relationships or marriage loops in the population, and these pairs were discarded from further analyses.

Phenotypes. In this work, we focused on 10 broad diagnostic categories that correspond to the definitions used in a recent publication (15). These were conditions of the 1) circulatory system, 2) endocrine system, 3) pulmonary system including allergies, 4) gastrointestinal system, 5) urogenital system, 6) musculoskeletal system, 7) hematological system, and 8) neurological system as well as 9) cancers and 10) mental conditions. Each of these broad diagnostic categories is a composite measure of presence or absence of any disease falling within the specific diagnostic category (*SI Appendix*, Table 1).

In general, if an individual has an in- or outpatient hospital admission or contact for one of the above medical conditions in the Danish National Patient Register and/or the Danish Psychiatric Central Research Register, we ascertain said individual as a case for said condition, with no respect to contact frequency or comorbidities—that is, diagnostic categories were not mutually exclusive. We considered both ICD-8 and ICD-10 codes for the ascertainment of a given phenotype, even though it is important to note that there is not always a 1-to-1 correspondence between the two coding systems. Only diagnoses coded as "main" or "auxiliary" were considered for the phenotyping (as opposed to "basic," "referral," "temporary," and "complication").

In general, this study considered all diagnoses assigned in relation to an in- or outpatient hospital admission or contact as recorded systematically in the Danish National Patient Register and/or the Danish Psychiatric Central Research Register.

Individuals with no entries for a given condition were treated as controls for said condition. However, this strategy is vulnerable to truncation and censoring biases because health records are not quantitatively or qualitatively homogeneous across demographic cohorts. To minimize the risk of including too many false controls in the control group, we only

studied individuals who were alive and living in Denmark after January 1, 1977 (date on which the Danish National Patient Register was established) or born in the interval (January 1, 1977, to January 1, 2017). As a result, we ended up with a subset of 6,691,426 individuals for all our quantitative analyses.

Heritability and Genetic Correlations. We used a classical approach for the estimation of total narrow-sense heritability and genetic correlations (16). For phenotype x—and given a familial relationship R (e.g., parent–offspring, full siblings, etc.)—if r_{x_1,x_2} is the correlation coefficient between two paired variables ($\mathbf{x_1}$, $\mathbf{x_2}$) holding the phenotypic observations for pairs of related individuals, the corresponding heritability is:

$$h_x^2 = \frac{r_{x_1,x_2}}{2\varphi_R}.$$

Similarly, the genetic correlation between phenotypes x and y, for a given familial relationship R, is:

$$r_{g,xy} = \frac{r_{x_1,y_2} + r_{y_1,x_2}}{2\sqrt{r_{x_1,x_2} r_{y_1,y_2}}}.$$

Because disease phenotypes are binary—that is, case control—we applied the latent correlation coefficient (also known as tetrachoric correlation coefficient), which measures agreement between two raters. In its simplest form, latent trait modeling assumes that the observed binary variables result from the discretization (at a given threshold) of unobserved (latent) variables that are normally distributed. The correspondence to the liability threshold model (32, 33) is apparent. In the case-control context, raters are vectors of binary phenotypes corresponding either to within- [($\mathbf{x_1}$, $\mathbf{x_2}$) and ($\mathbf{y_1}$, $\mathbf{y_2}$)] or between-phenotype [($\mathbf{x_1}$, $\mathbf{y_2}$) and ($\mathbf{y_1}$, $\mathbf{x_2}$)] paired data. We note that one rater corresponds to the genealogically older member of a familial relationship (e.g., father), whereas the other rater corresponds to the genealogically younger one (e.g., daughter). In the case of "genealogically contemporary" relationships such as siblings or cousins, relatives in the raters are sorted by age.

For the estimation of latent correlation coefficients, we used a standard maximum likelihood procedure from the polycor package in R.

In the case of heritability, valid estimates were those 1) with a positive value and 2) significantly different from zero. Moreover, when heritability estimates from multiple familial relationships were available, we combined them by computing their weighted average and weighted SE.

We computed average heritability values (h^2) and SE (s) weighted by sampling variance:

$$h^2 = \frac{\sum_{i=1}^{n} \frac{h_i^2}{s_i^2}}{\sum_{i=1}^{n} \frac{1}{s_i^2}}, \quad s = \frac{1}{\sqrt{\sum_{i=1}^{n} \frac{1}{s_i^2}}}.$$

We also used weighted least squares to estimate the slope β (corresponding to h^2) of the model

$$\mathbf{R} = \boldsymbol{\mu} + 2\Phi\beta + \boldsymbol{\varepsilon},$$

where $\mathbf{R}$ is a vector of correlation coefficients, Φ is a vector of kinship coefficients, $\boldsymbol{\mu}$ is the intercept vector, and $\boldsymbol{\varepsilon}$ is the error vector with $\boldsymbol{\sigma}^2(\boldsymbol{\varepsilon}) = \mathbf{W}^{-1}$. $\mathbf{W}$ is a diagonal matrix of weights used in the regression.

We carried out the analysis for all available pairs with no regard to sex. For estimates from horizontal familial relationships—that is, siblings and cousins—both individuals had to be from the same generation. For estimates from the rest of the relationships, only relatives from previous generations were considered. We did not consider heritability estimates when the correlation coefficient was negative or when the CIs fell outside [0, 1].

In the case of genetic correlations, valid estimates were those whose 95% CIs were contained within [−1, 1]. When genetic correlation estimates from multiple familial relationships were available, we combined them by computing their weighted average and weighted SE as above.

We note that estimates of heritability and genetic correlations depend on the definitions of the traits under study and that heritability of broadly defined traits will also reflect genetic correlations between the narrowly defined traits included in each broad trait category.

Scripts. The scripts used for data analysis can be found on GitHub at https://github.com/yorogosu/genealogy/.

Data Availability. This work is based on Danish register data that are not publicly available due to privacy protection, including General Data Protection Regulation (GDRP). Only Danish research environments are granted authorization. Foreign researchers can, however, get access to data. Further information on data access can be found at https://www.dst.dk/en/TilSalg/Forskningsservice or by contacting the senior corresponding authors.

ACKNOWLEDGMENTS. This study was supported by grants from European Union's Horizon 2020 Research and Innovation Programme: the "predicting comorbid cardiovascular disease in individuals with mental disorder by decoding disease mechanisms" project (CoMorMent, grant number 847776) and the "using real-world big data from eHealth, biobanks and national registries, integrated with clinical trial data to improve outcome of severe mental disorders" project (REALMENT, grant number 964874). The study was also supported by the Danish National Research Foundation (grant number DNRF148).

1. I. J. Bakken, A. M. S. Ariansen, G. P. Knudsen, K. I. Johansen, S. E. Vollset, The Norwegian Patient Registry and the Norwegian Registry for Primary Health Care: Research potential of two nationwide health-care registries. *Scand. J. Public Health* **48**, 49–55 (2020).
2. B. Gudbjornsson *et al.*, Rofecoxib, but not celecoxib, increases the risk of thromboembolic cardiovascular events in young adults-a nationwide registry-based study. *Eur. J. Clin. Pharmacol.* **66**, 619–625 (2010).
3. J. F. Ludvigsson *et al.*, External review and validation of the Swedish National Inpatient Register. *BMC Public Health* **11**, 450 (2011).
4. R. Sund, Quality of the Finnish Hospital Discharge Register: A systematic review. *Scand. J. Public Health* **40**, 505–515 (2012).
5. O. Mors, G. P. Perto, P. B. Mortensen, The Danish Psychiatric Central Research Register. *Scand. J. Public Health* **39**, 54–57 (2011).
6. M. Schmidt *et al.*, The Danish National Patient Registry: A review of content, data quality, and research potential. *Clin. Epidemiol.* **7**, 449–490 (2015).
7. C. B. Pedersen, The Danish Civil Registration System. *Scand. J. Public Health* **39**, 22–25 (2011).
8. C. B. Pedersen, H. Gøtzsche, J. O. Møller, P. B. Mortensen, The Danish Civil Registration System. A cohort of eight million persons. *Dan. Med. Bull.* **53**, 441–449 (2006).
9. M. Schmidt, L. Pedersen, H. T. Sørensen, The Danish Civil Registration System as a tool in epidemiology. *Eur. J. Epidemiol.* **29**, 541–549 (2014).
10. P. Lichtenstein *et al.*, Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: A population-based study. *Lancet* **373**, 234–239 (2009).
11. A. E. Nordsletten *et al.*, Patterns of nonrandom mating within and across 11 major psychiatric disorders. *JAMA Psychiatry* **73**, 354–361 (2016).
12. B. B. Trabjerg *et al.*, ALS in Danish registries: Heritability and links to psychiatric and cardiovascular disorders. *Neurol. Genet.* **6**, e398 (2020).
13. N. Zaitlen *et al.*, Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet.* **9**, e1003520 (2013).
14. R. Zhang *et al.*, Familial co-aggregation of schizophrenia and eating disorders in Sweden and Denmark. *Mol. Psychiatry* **26**, 5389–5397 (2020).
15. N. C. Momen *et al.*, Association between mental disorders and subsequent medical conditions. *N. Engl. J. Med.* **382**, 1721–1731 (2020).
16. D. S. Falconer, *Introduction to Quantitative Genetics* (Longman, Scientific & Technical; Wiley, 1989).
17. J. Zheng *et al.*, Early Genetics and Lifecourse Epidemiology (EAGLE) Eczema Consortium, LD Hub: A centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).
18. A. J. Schork *et al.*, A genome-wide association study of shared risk across psychiatric disorders implicates gene regulation during fetal neurodevelopment. *Nat. Neurosci.* **22**, 353–361 (2019).
19. V. Anttila *et al.*, Brainstorm Consortium, Analysis of shared heritability in common disorders of the brain. *Science* **360**, eaap8757 (2018).
20. P. F. Sullivan, M. C. Neale, K. S. Kendler, Genetic epidemiology of major depression: Review and meta-analysis. *Am. J. Psychiatry* **157**, 1552–1562 (2000).
21. M. G. Craske *et al.*, Anxiety disorders. *Nat. Rev. Dis. Primers* **3**, 17024 (2017).
22. P. F. Sullivan, K. S. Kendler, M. C. Neale, Schizophrenia as a complex trait: Evidence from a meta-analysis of twin studies. *Arch. Gen. Psychiatry* **60**, 1187–1192 (2003).
23. N. Craddock, P. Sklar, Genetics of bipolar disorder. *Lancet* **381**, 1654–1662 (2013).
24. M. Clapp *et al.*, Gut microbiota's effect on mental health: The gut-brain axis. *Clin. Pract.* **7**, 987 (2017).
25. J. Taylor, Iceland's genealogy database. *Circulation* **114**, f103–f104 (2006).
26. B. Lundbäck, H. Backman, J. Lötvall, E. Rönmark, Is asthma prevalence still increasing? *Expert Rev. Respir. Med.* **10**, 39–51 (2016).
27. W. De Nooy, A. Mrvar, V. Batagelj, *Exploratory Social Network Analysis with Pajek: Revised and Expanded Edition for Updated Software* (Cambridge University Press, ed. 2, 2018).
28. A. A. Hagberg, D. A. Schult, P. J. Swart, "Exploring network structure, dynamics, and function using NetworkX" in *Proceedings of the 7th Python in Science Conference*, G. Varoquaux, T. Vaught, J. Millman, Eds. (2008), pp. 11–15.
29. J. E. Hicks, Pydigree: A Python library for manipulation and forward-time simulation and of genetic datasets. https://www.biorxiv.org/content/10.1101/213413v1(5 November 2017).
30. G. Malécot, *The Mathematics of Heredity* (W. H. Freeman, 1970).
31. S. Wright, Coefficients of inbreeding and relationship. *Am. Nat.* **56**, 330–338 (1922).
32. E. R. Dempster, I. M. Lerner, Heritability of threshold characters. *Genetics* **35**, 212–236 (1950).
33. D. S. Falconer, The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann. Hum. Genet.* **29**, 51–76 (1965).