Twin Research and Human Genetics (2021), **24**, 103–109
doi:10.1017/thg.2021.18

CAMBRIDGE UNIVERSITY PRESS

Article

Safe Linkage of Cohort and Population-Based Register Data in a Genomewide Association Study on Health Care Expenditure

Eveline L. de Zeeuw[1], Lykle Voort[2], Ruurd Schoonhoven[3], Michel G. Nivard[1,4], Thomas Emery[5], Jouke-Jan Hottenga[1], Gonneke A. H. M. Willemsen[1,4], Pearl A. Dykstra[6], Narges Zarrabi[2], John D. Kartopawiro[3] and Dorret I. Boomsma[1,4,7]

[1]Department of Biological Psychology, Vrije Universiteit, Amsterdam, the Netherlands, [2]SURF, Amsterdam, the Netherlands, [3]Statistics Netherlands, Den Haag/Heerlen, the Netherlands, [4]Amsterdam Public Health Research Institute, Amsterdam, the Netherlands, [5]Netherlands Interdisciplinary Demographic Institute, Den Haag, the Netherlands, [6]Department of Public Administration and Sociology, Erasmus University, Rotterdam, the Netherlands and [7]Amsterdam Reproduction and Development Research Institute, Amsterdam, the Netherlands

Abstract

There are research questions whose answers require record linkage of multiple databases that may be characterized by limited options for full data sharing. For this purpose, the Open Data Infrastructure for Social Science and Economic Innovations (ODISSEI) consortium has supported the development of the ODISSEI Secure Supercomputer (OSSC) platform that allows researchers to link cohort data to data from Statistics Netherlands and run large-scale analyses in a high-performance computing (HPC) environment. Here, we report a successful record linkage genomewide association (GWA) study on expenditure for total health, mental health, primary and hospital care, and medication. Record linkage for genotype data from 16,726 participants from the Netherlands Twin Register (NTR) with data from Statistics Netherlands was accomplished in the secure OSSC platform, followed by gene-based tests and estimation of total and single nucleotide polymorphism (SNP)-based heritability. The total heritability of expenditure ranged between 29.4% (*SE* 0.8) and 37.5% (*SE* 0.8), but GWA analyses did not identify SNPs or genes that were genomewide significantly associated with health care expenditure. SNP-based heritability was between 0.0% (*SE* 3.5) and 5.4% (*SE* 4.0) and was different from zero for mental health care and primary care expenditure. We conclude that successfully linking genotype data to administrative health care expenditure data from Statistics Netherlands is feasible and demonstrates a series of analyses on health care expenditure. The OSSC platform offers secure possibilities for analyzing linked data in large scale and realizing sample sizes required for GWA studies, providing invaluable opportunities to answer many new research questions.

Keywords: Record linkage; register-based data; safely interconnected data; genomewide association study; health care expenditure

(Received 24 March 2021; revise received 25 March 2021)

Data collected for administrative or policy purposes often include detailed information at the individual level and are proving to be of great value for medical and scientific research. The Nordic countries are well known for studies on health outcomes based on register data because these data can be linked across registers with a unique personal identification number (Maret-ouda et al., 2017). Individuals' data on, for example, education or income can be linked to the use of social insurance schemes and health care and also to data on their family members and their place of residence. The advantage of register data is that nonresponse is not a problem and that data are collected in a uniform manner from all individuals. However, register data are generally not collected with a specific research purpose in mind. Cohort studies, on the other hand, employ surveys, collect biological samples and carry out clinical and experimental studies to gather data directly from participants (lifestyle, personality, attitudes, genotypes, biomarkers and clinical diagnoses). Linking individual-level register data to cohort data offers the possibility to enhance existing data resources and address research questions that are of relevance to individuals and to society.

In the Netherlands, population-based register data are collected and analyzed by Statistics Netherlands (CBS; www.cbs.nl/en-gb) to publish reliable statistics on the Dutch economy and society (www.statline.nl). CBS has administrative data for approximately 17 million inhabitants of the Netherlands on pedigree structure and outcomes, including education, income and health, which can all be linked at the individual level. CBS is legally entitled to make these data, under strict terms, available for research purposes as well as link them to external data in a secure remote-access (RA) environment (www.cbs.nl/microdata). However, this RA environment does not offer high-performance computing facilities, which limits the scale of research projects. The Open Data Infrastructure for Social Science and Economic Innovations (ODISSEI; www.odissei-data.nl) has set up a sustainable national data infrastructure for research in the Netherlands by coordinating the integration of data from large cohort studies in the Netherlands and administrative data. ODISSEI has supported the development of a secure platform, called the ODISSEI Secure Supercomputer (OSSC) (Scheerman et al., 2019) on the high-performance computer facility

Author for correspondence: Dorret Boosma, Email: di.boomsma@vu.nl

Cite this article: de Zeeuw EL, Voort L, Schoonhoven R, Nivard MG, Emery T, Hottenga J-J, Willemsen GAHM, Dykstra PA, Zarrabi N, Kartopawiro JD, and Boomsma DI. (2021) Safe Linkage of Cohort and Population-Based Register Data in a Genomewide Association Study on Health Care Expenditure. *Twin Research and Human Genetics* **24**: 103–109, https://doi.org/10.1017/thg.2021.18

at Samenwerkende Universitaire Reken Faciliteiten (SURF; www.surf.nl). SURF hosts and maintains the Dutch national supercomputer Cartesius with 2000 multicore compute nodes, connected with Infiniband for high-speed, low-latency communication and storage access. The OSSC is developed as a 'Platform as a Service' on top of the Cartesius, where users can use the computing resources and software offered on the supercomputer. The platform is based on customizable, virtualized private clusters that are deployed on Cartesius, imposing strict security measures required by data owners, and facilitates access and record linkage (Private Cloud on a Compute Cluster [PCOCC], developed by CEA, https://github.com/cea-hpc/pcocc.). From the CBS RA environment, data can be copied to OSSC through a dedicated VPN connection between CBS and SURF(sara), seamlessly and securely integrating the HPC cluster into their own private network. Stringent automated security controls make sure this VPN is the only path for sensitive data to leave the cluster (see Figure S1).

In the current study, we used the OSSC platform to link CBS register data on health care expenditure to genotype information from the Netherlands Twin Register (NTR; www.tweelingenregister.vu.nl). The NTR is a large twin-family study that has followed thousands of family members longitudinally with surveys and has collected DNA samples from a large number of their participants (Ligthart et al., 2019). The scientific aim of our study was to run a genomewide association (GWA) study and estimate the associations between genetic variants and health care expenditure. While the effect of environmental determinants related to overall health has been extensively studied, less is known about the genetic architecture of individual differences between people. Given that expenditure is directly related to medical conditions that are characterized by genetic contributions, we hypothesize that a substantial contribution of genetic differences to overall health expenditure exists. Knowledge on the genetic contributions to overall health is currently limited to studies that measured it with self-reports. Twin studies indicate that self-rated health is partly due to genetic differences between people, with estimates of a heritability of over 30% in young adulthood (Silventoinen et al., 2007) and almost 50% at older ages (Mosing, Pedersen et al., 2010). The first GWA study (~6700 individuals) on self-rated health did not report any genomewide significant hits (Mosing, Verweij et al., 2010), but a better powered study in almost 112,000 individuals identified 13 independent genomewide significant signals, of which several were in regions previously implicated in specific diseases (Harris et al., 2017). The proportion of variance in self-rated health explained by all the measured common genetic variants was 13%. We argue that investigating overall health, more objectively measured by health care expenditure, will lead to a better understanding of its genetic architecture.

Methods

Participants

The NTR was established around 1987 by the Department of Biological Psychology at Vrije Universiteit Amsterdam (Ligthart et al., 2019). Adult twins, their families and parents of young twins take part in surveys. DNA collection from blood or buccal samples has been done in several large projects (Ligthart et al., 2019). Genotyping has been carried out in subsamples that are, in general, unselected for specific traits. Height measured in centimeters was obtained from clinical and experimental studies or reported by participants in surveys. Many participants filled out more than one survey, and longitudinal height data were checked for consistency. NTR data collection was approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Centre, Amsterdam, an Institutional Review Board certified by the U.S. Office of Human Research Protections (IRB number 00002991 under Federal-Wide Assurance-FWA00017598; IRB/Institute codes, NTR 03-180), and informed consent was obtained from all participants.

Health Care Expenditure

Dutch residents are obliged by law (Health Insurance Act; ZVW) to take out a basic health insurance. Just 0.07% of the population are conscientious objectors and have no insurance for health care expenditure (European Commission, 2017). Statistic Netherlands (CBS) receives, via Vektis (www.vektis.nl), an executive organization of health insurance companies in the Netherlands, health care expenditure as reimbursed under the basic health insurance. Health care expenditure included the annual costs per resident for primary, hospital, mental health, birth and geriatric care, physiotherapy and medication. Not included are costs that (1) fall under a supplementary health insurance, (2) fall outside of the Health Insurance Act and are paid by the patient or (3) fall under long-term care. We analyzed average expenditure costs across all available reporting years (2009–2016) for total health, mental health, primary and hospital care and medication (Statistics Netherlands (CBS), 2019a). Data were log-transformed prior to analyses to correct for the skewness of the data.

Genotype Data

Genotyping was done on several platforms, that is, Affymetrix-Perlegen, Illumina 660, Illumina Omni Express 1M, Affymetrix 6.0, Affimetrix Axiom and Illumina GSA. For criteria on quality control (QC) of the single nucleotide polymorphisms (SNPs) and samples, before and after imputation (see de Zeeuw et al., 2020), data were cross-platform phased and imputed using Mach-admix with GoNL (Francioli et al., 2014) as the reference panel for all SNPs that were, after QC, present for at least one platform (Boomsma et al., 2014). The cross-chip imputed data set was used to calculate genetic principal components with the SmartPCA software (Price et al., 2006). Subsequently, the data set was aligned against the 1000G phase 3 version 5 reference panel and imputed on the Michigan imputation server (Das et al., 2016). Best guess genotypes were calculated for all SNPs in Plink 1.96 (Purcell et al., 2007).

Record Linkage

Personal data (name, address, date of birth) of NTR participants and their Administrative number (A-number) under which Dutch residents are registered in the population register are stored under a pseudonymized NTR identifier on a server that is disconnected from the internet; note that NTR does not have the social security number (BSN) of the participants (Boomsma et al., 2008, 2018). Phenotype data are stored under a different pseudonymized NTR identifier. The genotype data of NTR participants are stored at SURF in binary format and do not contain any identifiers, but the order within the data set is an implicit identifier. At CBS, all individual-level register data (microdata) are stored under a pseudonymized identifier (RIN) and CBS has access to the A-number of Dutch residents.

Figure 1 displays the different steps in the record linkage process. Record linkage between the NTR identifier and CBS identifier

https://doi.org/10.1017/thg.2021.18

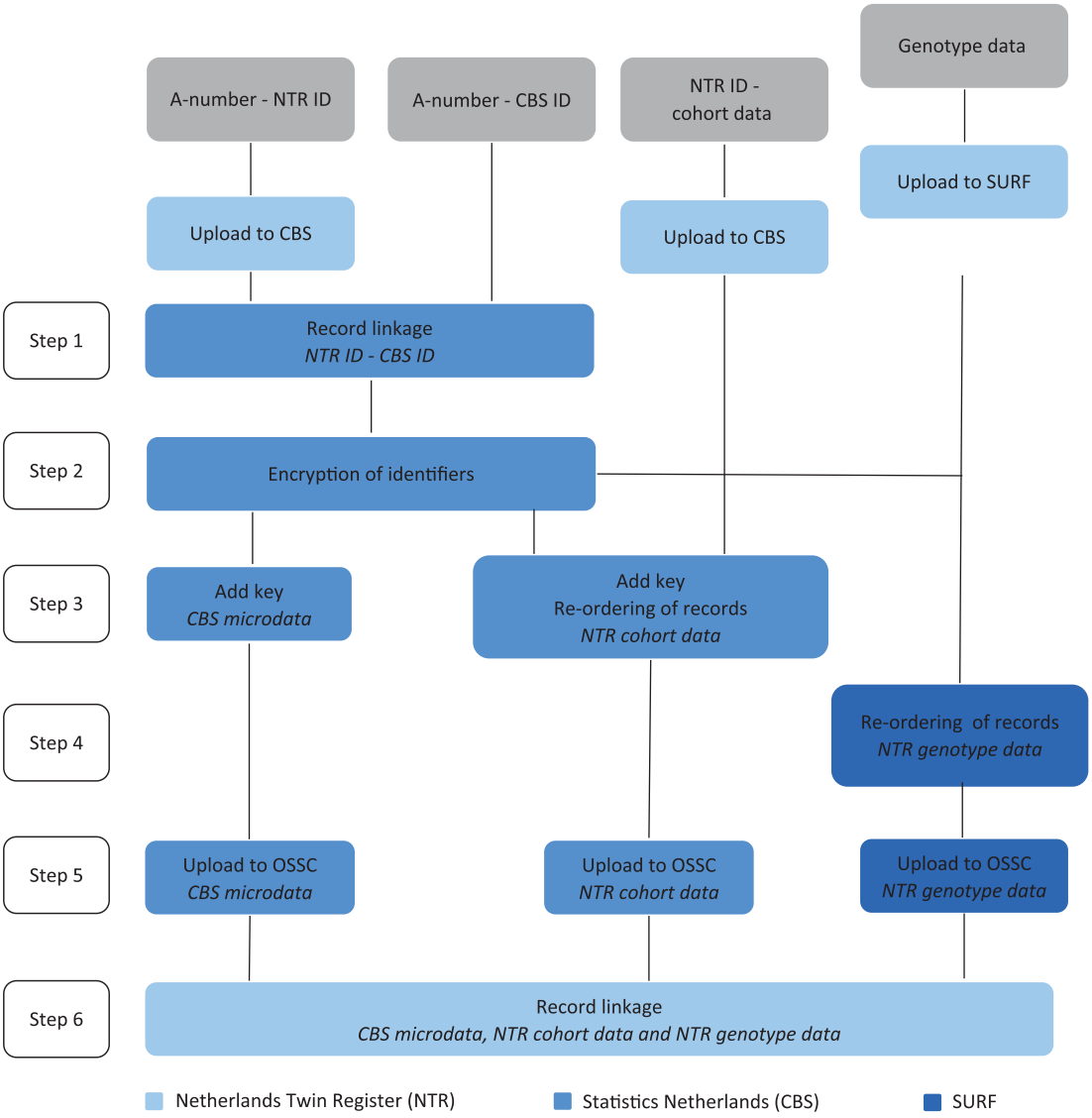

Fig. 1. Flow chart of record linkage between data from the Netherlands Twin Register and data from Statistics Netherlands on the ODISSEI Secure SuperComputer (OSSC) platform. *Note: Step 1:* Record linkage between the NTR identifier and CBS identifier on the basis of the A-number. *Step 2:* NTR and CBS identifiers were encrypted, and the record linkage key was sent to SURF. *Step 3:* The key was added to CBS microdata and NTR cohort data, and NTR cohort data were re-ordered. *Step 4:* SURF used the record linkage key for pseudorandomizing the order of the genotype data. *Step 5:* CBS uploaded CBS microdata and NTR cohort data to the OSSC environment via a secure virtual private network (VPN). SURF placed the re-ordered genotype data and the key of the new order in the OSSC. *Step 6:* NTR linked CBS microdata to NTR cohort data and sorted NTR cohort data in the same new order as the genotype data.

was done by the Central Record Linkage department (CBK) at CBS on the basis of the A-number (step 1). After linkage and removal of the A-number, the NTR and CBS identifiers were encrypted and a record linkage key was sent to SURF (step 2). The key with encrypted identifiers was added to CBS microdata, and NTR cohort data and the NTR cohort data were re-ordered (step 3). SURF, which acted as a trusted third party (TTP), used the record linkage key for pseudo-randomizing the order of the genotype data (step 4). CBS uploaded CBS microdata and NTR cohort data to the OSSC environment via a secure virtual private network (VPN). SURF placed the reordered genotype data and the key of the new order of the genotype data in the OSSC (step 5). NTR linked CBS microdata to NTR cohort data and sorted the NTR cohort data in the same order as the new order of the genotype data (step 6). The complete procedure ensured that none of the parties involved had access to all the record linkage keys and, therefore, could neither identify participants in the cohort nor in the administrative or genotype data.

Statistical Analyses

To verify that the encryption of the NTR identifiers and the pseudo-randomization of the genotype data were correct, we conducted a GWA study on height reported by the NTR participants. The data on height were uploaded to the OSSC together with the other NTR cohort data — that is, data on covariates — and underwent the same procedure of identifier encryption, and the analyses were carried out using the reordered genotype data. A large meta-analysis reported genetic variants associated with height (Wood et al., 2014), and we estimated the genetic correlation between these

https://doi.org/10.1017/thg.2021.18

and our own results with LD-score regression (Bulik-Sullivan et al., 2015). For a successful procedure, we expected a genetic correlation close to 1.

We regressed the log-transformed total health, mental health, primary and hospital care and medication expenditure on all SNPs in linear mixed models in GCTA 1.92.1beta6 (Yang et al., 2011), controlling for genetic relatedness by including a SNP-derived genetic relationship matrix (GRM) with all the off-diagonal elements < .05 set to zero. Population stratification was taken into account by including 10 principal components (PC) from the PCA of the SNP genotypes as fixed effects (Price et al., 2006). Sex, age and age-squared were also included as fixed effects. To declare genomewide significance for SNP-phenotype associations, an alpha level of 5×10^{-8} was adopted (Duggal et al., 2008). Based on the GWA results, we conducted gene-based association analyses in MAGMA (de Leeuw et al., 2015). Genetic variants were assigned to genes on the basis of their position according to the NCBI 37.3 build, resulting in 15,438 genes. The European panel of the 1000 genomes (Auton et al., 2015) data was used as a reference for linkage disequilibrium. In line with the Bonferroni method, a genomewide significance threshold of $.05/15{,}438 = 3 \times 10^{-6}$ was adopted for gene-based association tests.

We estimated family-based heritability by employing the differences in genetic resemblance between all family members in our sample as determined with the GRM in GCTA 1.92.1beta6 (Yang et al., 2011). The SNP-based heritability for each of the outcome measures and the genetic correlation with self-rated health (Harris et al., 2017) were estimated with LD-score regression v1.0.0 (Bulik-Sullivan et al., 2015).

Results

Genotype data were successfully linked to CBS for 16,726 individuals from European ancestry (Galinsky et al., 2016) who had given consent for linking to external databases and for whom the A-number was known. The analysis of adult height data ($N = 12{,}498$) gave a SNP-based heritability of 37.3% (5.9). The genetic correlation between our results and the results from the large meta-analysis on height as reported by the GIANT consortium was .99 (.06), indicating that the record linkage was done correctly.

For 14,572 participants (5842 males and 8727 females; 699 families) with an average age of 45, genotype data and expenditure for total health care in euros were available (males: mean = 1556.7, $SD = 3406.9$; females: mean = 1668.8, $SD = 3432.1$), mental health care (males: mean = 113.2, $SD = 835.1$; females: mean = 158.0, $SD = 1308.7$), medication (males: mean = 175.1, $SD = 784.1$; females: mean = 206.6, $SD = 631.3$), primary care (males: mean = 117.4, $SD = 49.1$; females: mean = 133.7, $SD = 61.1$) and hospital care (males: mean = 813.4, $SD = 2292.9$; females: mean = 1009.4, $SD = 2260.9$; see Figure 2 and Supplementary Figure S2). The unavailability of health care expenditure data might have been due to, for example, residence outside of the Netherlands.

The results of the GWA analyses summarized in Manhattan and Q-Q plots for health care expenditure are depicted in Figure 3 and Supplementary Figures S3–S6. There were no SNPs that were associated with health care expenditure at a genomewide significance level, and gene-based tests did not reveal genes that were significantly associated with health care expenditure, with the strongest associations for the genes *TRPV3* (17p13, $p = 2 \times 10^{-5}$), CAT (11p13, $p = 7 \times 10^{-5}$) and *SSBP2* (5q14, $p = 4 \times 10^{-5}$). Family-based heritability was .319 (.01) for total

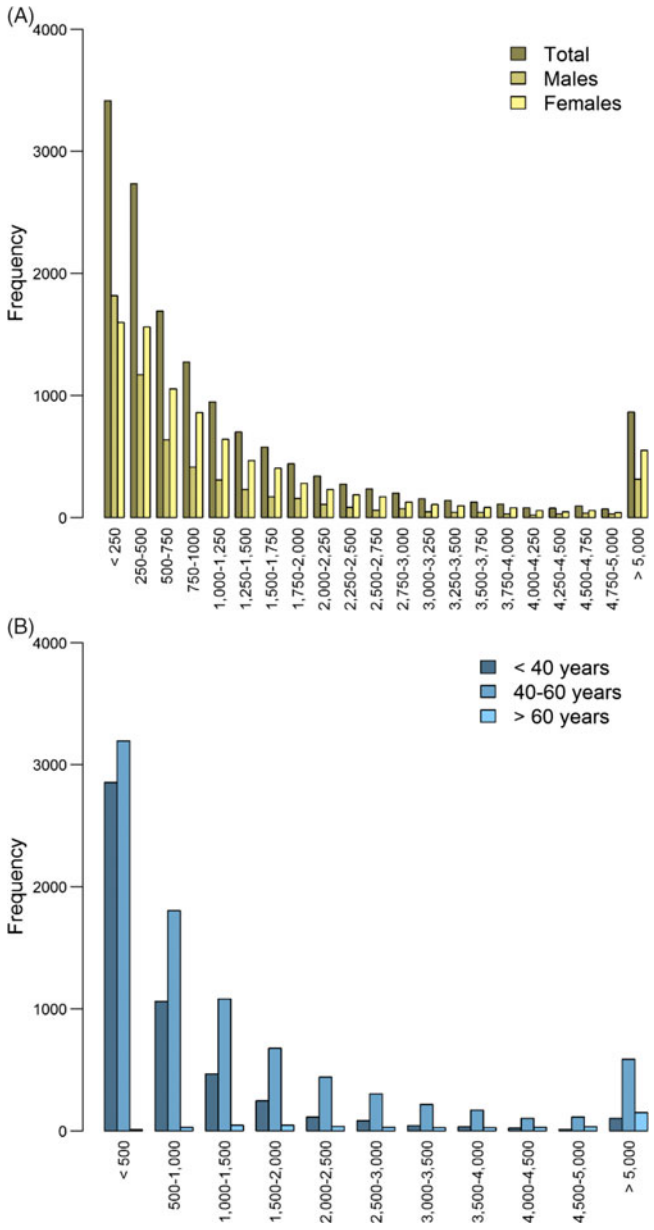

Fig. 2. Histogram of total health care expenditure in euros separately for (A) males ($n = 5845$) and females ($n = 8727$) and (B) for individuals < 40 years ($n = 5059$), between 40 and 60 years ($n = 8706$) and >60 years ($n = 483$).

health care, .302 (.01) for mental health care, .294 (.01) for medication, .375 (.01) for primary health care and .295 (.01) for hospital care, and SNP-based heritability was −004 (.03) for total health care, .054 (.03) for mental health care, −.005 (.04) for medication, .044 (.04) for primary health care and −.002 (.04) for hospital care (see Figure 4). The genetic correlation between the GWA results that showed a significant SNP-based heritability and the results from the meta-analysis on self-rated health (Harris et al., 2017) was .26 (.17) for mental health care expenditure and .70 (.30) for primary care expenditure.

Discussion

The aim of the current study was to demonstrate the feasibility of running a large-scale GWA study for health care expenditure by linking NTR genotype data and CBS administrative data on the

https://doi.org/10.1017/thg.2021.18

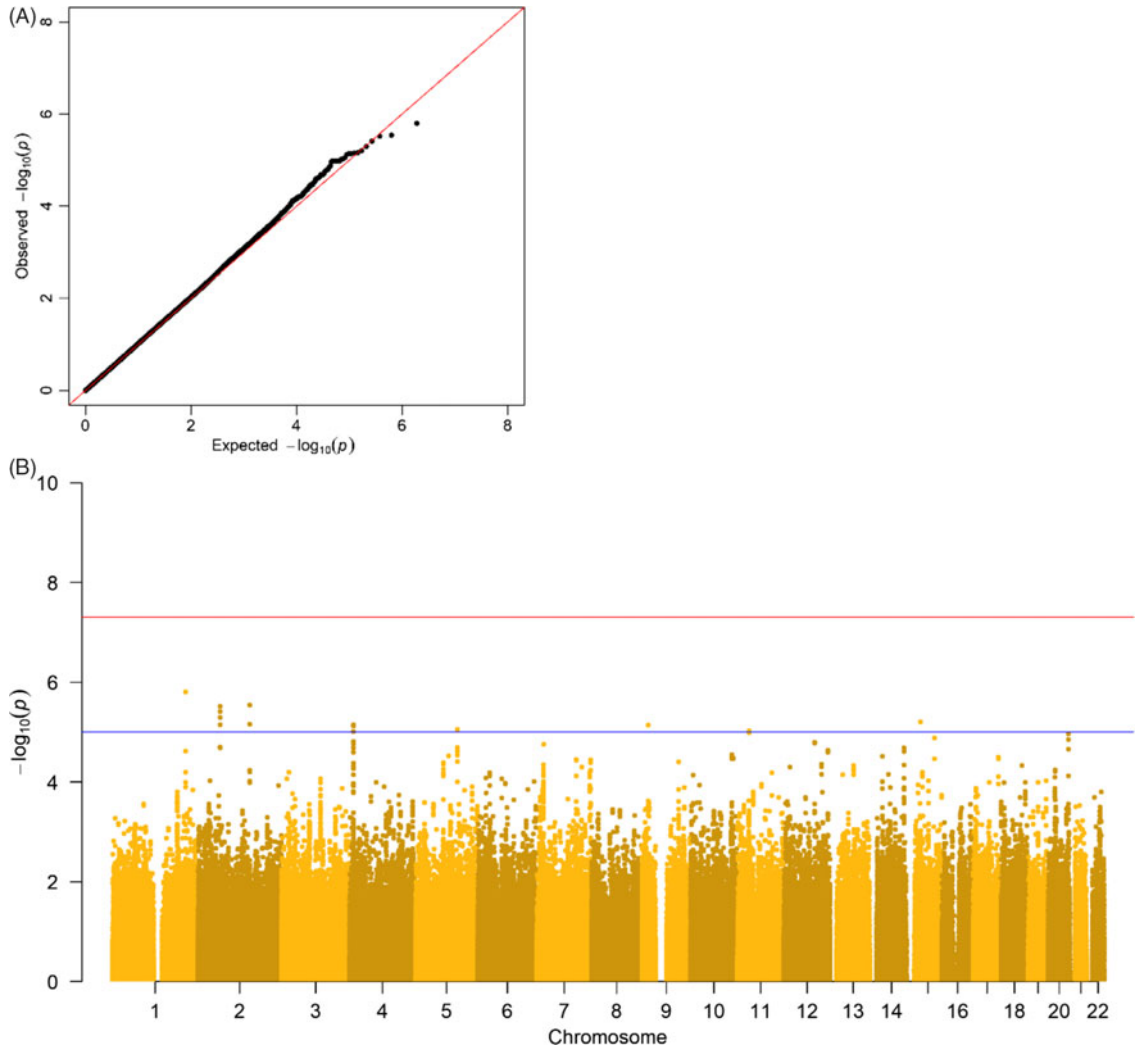

Fig. 3. (A) Q-Q plot and (B) Manhattan plot of p values of the genomewide association analysis for total health care expenditure. *Note:* The top line is the genomewide significance threshold ($p < 5 \times 10^{-8}$), and the bottom line indicates the threshold for suggestive significance ($p < 1 \times 10^{-5}$).

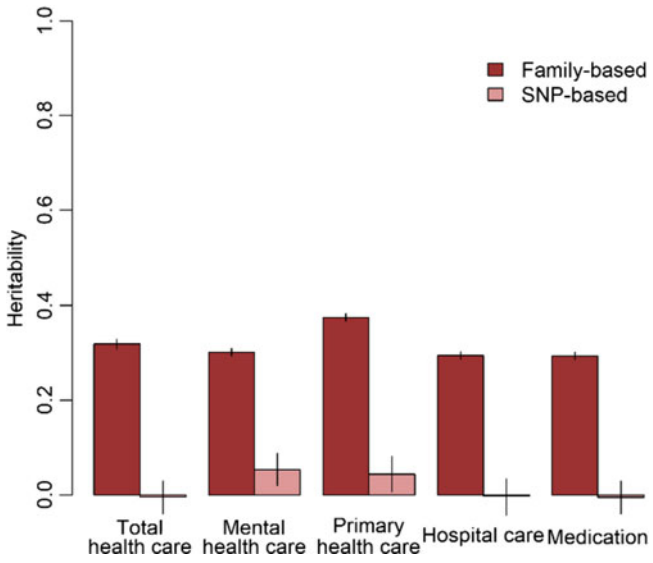

Fig. 4. Family-based and SNP-based heritability of total health, mental health, primary health and hospital care and medication expenditure.

OSSC platform. We successfully linked genotype data to self-reported height from NTR participants in the OSSC after encrypting all identifiers and pseudo-randomization of the order in the genotype data. The finding of a significant SNP heritability for height and a genetic correlation of unity with the GIANT study on height (Wood et al., 2014) confirmed that the encryption and reordering of the data were done correctly. The OSSC thus enabled us to link sensitive data from multiple databases in a privacy protecting way on a secure system that provided the required high-performance computing facilities.

No genetic variants that were genomewide significantly associated with health care expenditure were found. For significantly associated genes at the stringent genomewide level, larger studies are required and this will become feasible when multiple Dutch cohorts collaborate in the OSSC environment. No significantly associated genes were found, but we identified some promising genes — for example, *SSBP2* — that were previously found to be related to diseases (Liu et al., 2008). Overall, the proportion of variance explained by all genotyped autosomal SNPs ranged from zero to 5.4%. The estimate for the SNP-based heritability was different from zero for mental health care expenditure and

https://doi.org/10.1017/thg.2021.18

primary care expenditure. The genetic correlation between self-rated health and health care expenditure was larger for primary care ($r = .70$) compared to mental health care ($r = .26$). Family-based heritability was much larger for all types of health care expenditure (29–38%), indicating that there are possibilities for larger samples to identify genetic variants related to health care expenditure. Health care expenditure may be influenced not only by a large number of genetic variants that all have a very small effect but also by rare genetic variants that, although they can have a large effect, explain only a small part of the differences between people. Larger sample sizes are required to identify the effects of common genetic variants on health care expenditure. The OSSC brings these large sample sizes within reach as it will provide the possibility to link multiple cohorts to the register data, allowing the exact same outcome measure to be analyzed across all these cohorts.

Identifying genetic variants associated with overall health is important as these genetic variants can be employed in bidirectional Mendelian randomization (MR; Smith & Hemani, 2014). MR is a natural experiment that employs genetic variants as instrumental variables, which has been suggested to provide the best opportunity to establish the causal effect of specific traits on overall health (Dixon et al., 2016), since randomized controlled trials are in this case impossible to implement. The identification of the impact of specific traits on overall health will give information needed to estimate the cost-effectiveness of prevention and intervention programs. Such information is essential as health care expenditure in the Netherlands has increased to 76.9 billion euros, 9.9% of the gross domestic product (Statistics Netherlands (CBS), 2019b), only partly explained by the growth of the proportion of the Dutch population over 65 years (Howdon & Rice, 2018).

The OSSC platform presents a novel method for creating secure computing environments on traditional multitenant high-performance computing clusters. The platform as a service provides a customizable virtualized solution using PCOCC to meet strict security requirements for storing, processing and linking highly sensitive data. The OSSC development is moving toward a production-ready version of the platform by automating the manual steps for setting up the virtual environment and creating a scalable and robust solution. The platform allows data- and compute-intensive research projects to be conducted in parallel. The individual-level register data from Statistics Netherlands will offer opportunities for research in public health, health care, economics, education, sociology, bioinformatics, 'omics' and many other fields. The OSSC platform also facilitates multidisciplinary research and promotes open science by allowing for the increased linkage and interoperability of sensitive data. Projects may, for example, link data from MRI scans to data on psychiatric diagnoses as well as to data about urbanization level of the neighborhood to get more insight into the mediating effect of neural processes in the association between the urban environment and psychopathology (Lederbogen et al., 2011). In addition, in light of the current COVID-19 pandemic, the linkage of this population-based register data with cohort data will provide researchers with invaluable opportunities to determine the impact of the pandemic and the lockdown measures on individual outcomes such as mental health and educational outcomes in children, or address questions on host–virus interactions.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/thg.2021.18.

Acknowledgments. We are thankful to the twin families registered with the Netherlands Twin Register for their participation. This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. The results are based on own analyses of the VU Amsterdam researchers based on the nonpublic data from Statistics Netherlands.

Financial support. We gratefully acknowledge the Open Data Infrastructure for Social Science and Economic Innovations (ODISSEI; NWO: NRGWI.obrug.2018.008); 'Netherlands Twin Register Repository: Researching the interplay between genome and environment' (NWO: 480-15-001/674); KNAW Academy Professor Award (PAH/6635) and 'Genetics as a research tool: A natural experiment to elucidate the causal effects of social mobility on health' (ZonMw: 531003014).

Conflict of interest. None.

References

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., … Schloss, J. A. (2015). A global reference for human genetic variation. *Nature*, *526*, 68–74.

Boomsma, D. I., Helmer, Q., Nieuwboer, H. A., Hottenga, J. J., de Moor, M. H., van den Berg, S. M., … de Geus, E. J. (2018). An extended twin-pedigree study of neuroticism in the Netherlands Twin Register. *Behavior Genetics*, *48*, 1–11.

Boomsma, D. I., Wijmenga, C., Slagboom, E. P., Swertz, M. A., Karssen, L. C., Abdellaoui, A., … Van Duijn, C. M. (2014). The genome of the Netherlands: Design, and project goals. *European Journal of Human Genetics*, *22*, 221–227.

Boomsma, D. I., Willemsen, G., Vink, J. M., Bartels, M., Groot, P., Hottenga, J. J., … Van Der Kleij, F. (2008). Design and implementation of a twin-family database for behavior genetics and genomics studies. *Twin Research and Human Genetics*, *11*, 342–348.

Bulik-Sullivan, B., Loh, P. R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., … O'Donovan, M. C. (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, *47*, 291–295.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*, 1284–1287.

de Leeuw, C. A., Mooij, J. M., Heskes, T., & Posthuma, D. (2015). MAGMA: Generalized gene-set analysis of GWAS data. *PLoS Computational Biology*, *11*, 1–19.

de Zeeuw, E. L., Jan, J., Klaasjan, H., Conor, G. O., Erik, V. D., Davies, G. E., … Bergen, E. Van. (2020). Intergenerational transmission of education and ADHD : Effects of parental genotypes. *Behavior Genetics*, *50*, 221–232.

Dixon, P., Davey Smith, G., von Hinke, S., Davies, N. M., & Hollingworth, W. (2016). Estimating marginal healthcare costs using genetic variants as instrumental variables: Mendelian randomization in economic evaluation. *PharmacoEconomics*, *34*, 1075–1086.

Duggal, P., Gillanders, E. M., Holmes, T. N., & Bailey-Wilson, J. E. (2008). Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. *BMC Genomics*, *9*, 1–8.

European Commission. (2017). *State of health in the EU: Country health profile 2017 - Netherlands*. Organisation for Economic Co-operation and Development (OECD). https://www.oecd.org/publications/netherlands-country-health-profile-2017-9789264283503-en.htm

Francioli, L. C., Menelaou, A., Pulit, S. L., Van Dijk, F., Palamara, P. F., Elbers, C. C., … Wijmenga, C. (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*, *46*, 818–825.

Galinsky, K. J., Bhatia, G., Loh, P. R., Georgiev, S., Mukherjee, S., Patterson, N. J., & Price, A. L. (2016). Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. *American Journal of Human Genetics*, *98*, 456–472.

Harris, S. E., Hagenaars, S. P., Davies, G., Hill, W. D., Liewald, D. C. M., Ritchie, S. J., ... Deary, I. J. (2017). Molecular genetic contributions to self-rated health. *International Journal of Epidemiology*, *46*, 994–1009.

Howdon, D., & Rice, N. (2018). Health care expenditures, age, proximity to death and morbidity: Implications for an ageing population. *Journal of Health Economics*, *57*, 60–74.

Lederbogen, F., Kirsch, P., Haddad, L., Streit, F., Tost, H., Schuch, P., ... Meyer-Lindenberg, A. (2011). City living and urban upbringing affect neural social stress processing in humans. *Nature*, *474*, 498–501.

Ligthart, L., van Beijsterveldt, C. E. M., Kevenaar, S. T., de Zeeuw, E., van Bergen, E., Bruins, S., ... Boomsma, D. I. (2019). The Netherlands Twin Register: Longitudinal research based on twin and twin-family designs. *Twin Research and Human Genetics*, *22*, 623–636.

Liu, J. W., Nagpal, J. K., Sun, W., Lee, J., Kim, M. S., Ostrow, K. L., ... Sidransky, D. (2008). SsDNA-binding protein 2 is frequently hypermethylated and suppresses cell growth in human prostate cancer. *Clinical Cancer Research*, *14*, 3754–3760.

Maret-ouda, J., Tao, W., Wahlin, K., & Lagergren, J. (2017). Nordic registry-based cohort studies : Possibilities and pitfalls when combining Nordic registry data. *Scandinavian Journal of Public Health*, *45*, 14–19.

Mosing, M., Pedersen, N., Martin, N., & Wright, M. (2010). Sex differences in the genetic architecture of optimism and health and their interrelation: A study of Australian and Swedish twins. *Twin Research and Human Genetics*, *13*, 322–329.

Mosing, M., Verweij, K., Medland, S., Painter, J., Scott, D., Heath, A., ... Nicholas, G. (2010). A genome-wide association study of self-rated health. *Twin Research and Human Genetics*, *13*, 398–403.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*, 904–909.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., ... Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*, 559–575.

Scheerman, M., Voort, L., & Zarrabi, N. (2019). Secure platform for processing sensitive data on shared HPC systems. *CompBioMed*. https://www.compbiomed-conference.org/wp-content/uploads/2019/07/CBMC19_paper_109.pdf

Silventoinen, K., Posthuma, D., Lahelma, E., Rose, R. J., & Kaprio, J. (2007). Genetic and environmental factors affecting self-rated health from age 16-25: A longitudinal study of Finnish twins. *Behavior Genetics*, *37*, 326–333.

Smith, G., & Hemani, G. (2014). Mendelian randomization: Genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, *23*, 89–98.

Statistics Netherlands (CBS). (2019a). *Zorgkosten van Nederlandse ingezetenen met een basisverzekering*.

Statistics Netherlands (CBS). (2019b). *Zorguitgaven - kerncijfers*.

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., ... Frayling, T. M. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, *46*, 1173–1186.

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, *88*, 76–82.