

Advantages of linking national registries with twin registries for epidemiological research

Line C. Gjerde^{1,2}, Fartein Ask Torvik¹, Kristian Amundsen Østby¹, Gun Peggy Knudsen¹, Nikolai Czajkowski^{1,2}, Ted Reichborn-Kjennerud^{1,3} and Ragnhild E. Ørstavik¹

1) Domain for Mental and Physical Health, Norwegian Institute of Public Health

2) Department of Psychology, University of Oslo

3) Institute of Clinical Medicine, University of Oslo

Correspondence: Line C. Gjerde, Box 4404 Nydalen, 0403 Oslo Telephone: +47 21078379 e-mail: line.gjerde@fhi.no

ABSTRACT

Linking national registries with twin data represents an opportunity to produce epidemiological research of high quality. National registries contain information on a broad array of variables, some of which cannot be measured reliably in regular health surveys. By taking kinship into consideration, twin studies have the benefit of being able to identify confounding stemming from genetic or shared environmental sources. In this paper, we use examples from our own interview and questionnaire-based twin studies from the Norwegian Twin Registry (NTR) on mental disorders, alcohol use and socioeconomic status linked to registry data on medical benefits to demonstrate the value. In the first example, we examined to what extent genetic and environmental factors contributed to sick leave and disability pension and the association between these two types of benefits. In the second example, we explored the genetic and environmental relationship between personality disorders and sick leave. In the third example, a co-twin control design was applied to explore whether there is a true protective relationship between moderate alcohol consumption and health. The fourth example shows to what degree anxiety and depression are associated with later sick leave granted for not only mental disorders, but also somatic disorders, adjusted for confounding by genetic and shared environmental factors. In the fifth example, we address the socioeconomic gradient in sick leave, adjusting for non-observed confounders associated with the family in a co-twin control design. Our examples illustrate some of the potentials obtainable by linking national registries with twin data. The efforts that have been made to create the NTR in Norway and the International Network of Twin Studies (INTR) internationally make these types of linkage studies easier to conduct and available to more researchers. As there are still many areas to explore, we encourage epidemiological researchers to make use of this possibility.

This is an open access article distributed under the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

INTRODUCTION

National registries provide quality controlled information on “hard” outcomes such as employment status, medical diagnoses and treatment, as well as on important confounders such as socioeconomic variables. Data in national registries are updated regularly, and thus provide the opportunity for longitudinal studies. As inclusion is mandatory, the data are, for all practical purposes, complete and not subject to recruitment bias. These types of data are challenging to collect using self-report and interview approaches as invited participants may decline to take part in the study and thereby limiting the representativeness of the data, or responses may be subject to bias due to memory or social desirability. However, having access to “hard” data through a national registry may not be satisfactory for all research purposes. Information on several important health variables is not available in registries. This includes health behavior (such as smoking, alcohol consumption and physical activity) or subjective evaluations of a person’s mental and/or physical health and well-being. One may therefore need to combine registry data with more in-depth data from self-report

or interview measures. As the number and breadth of variables increase, so does the scope of possible research questions. The combination of national registry data and health survey data makes it possible to not only investigate the prevalence of a condition, but also whether this condition is related to work satisfaction, exercise, eating habits, etc.

An important limitation to consider is that data from national registries and health surveys are subject to confounding (see below). The genetically informative data attained using twin registries can create even more valuable research opportunities: Twin data allow for us to investigate etiology and to come closer to inferring causal associations. Linking national registries with twin registries is beneficial both for epidemiological research where access to genetically informative data can be used to create designs that may reveal causal associations, and for twin research, which get access to information that is hard to collect in regular health surveys.

The present paper will give a brief introduction to national registries, confounding, twin designs, and twin registries, before providing five examples from our own research where a national registry (FD-Trygd)

was linked with a twin registry (NTR). The examples may serve as inspiration for using linkage between these two types of registries, and demonstrate that the combination provide opportunities to answer novel research questions.

NATIONAL REGISTRIES

The Nordic countries, and Denmark in particular (Frank, 2000), have succeeded in creating and maintaining large-scale, centralized national registries, encompassing health registries. The latter category is especially suitable for research on various health outcomes, but registries that are not defined as health registries can also provide valuable sources for biomedical research.

A *health registry* can be defined as a collection of systematically stored health information, from which the information on each individual can be retrieved (Lovdata, 2015). The centralized registries are mandatory rather than consent-based, and usually include information that can be traced back to individuals through the national identification number issued to all Norwegians after they are born. These registries are governed by strict regulations, and new registries can only be established after approval from the Parliament.

In addition to the large-scale centralized health registries, there are several national registries that contain valuable information for health research, but which are not regulated by the Act on Personal Health Data Filing Systems (Helseregisterloven, Dahl, 2009). An example is Statistics Norway's Historical Event Database, hereby referred to as FD-Trygd. FD-Trygd is a database containing data from the entire population (1992 and onwards) from several sources: Registries at Statistics Norway; the Norwegian Labor and Welfare Organization and the Employment Directorate; and the Norwegian Tax Administration. The database contains information regarding all social security benefits in Norway, including for instance sickness benefits, social assistance, rehabilitation allowance, disability pension and unemployment benefits (Akselsen *et al.*, 2007). Diagnoses for the benefits are also included in the database. Data from FD-Trygd are often supplemented with information from related datasets, also administered by Statistics Norway, such as the Norwegian Education Registry. Table 1 displays large national registries in Norway.

THE PROBLEM OF CONFOUNDING

In longitudinal observational studies, confounding poses a considerable threat to the validity of studies aiming to identify causal mechanisms. Confounding occurs when an extraneous variable is associated with both the exposure and the outcome. This creates spurious associations. Confounders can be either **measured**, and thereby statistically controlled for in the model, or **unmeasured** (Richmond *et al.*, 2014).

The latter introduces the greatest problems for causal inference. For instance – there is a positive association between smoking and schizophrenia (de Leon and Diaz, 2005). As smoking most commonly precedes a schizophrenia diagnosis (Levander *et al.*, 2007), it is tempting to conclude that smoking may be a causal factor for this disorder. Attempts have been made to explain this hypothesis biologically, for instance that smoking is a way of self-medicating the symptoms of schizophrenia (Dalack *et al.*, 1998). Both smoking and schizophrenia are influenced by genetic factors (Li *et al.*, 2003, Sullivan *et al.*, 2003). If genetic factors influencing smoking also influence liability to schizophrenia, one would observe a correlation between genetic factors influencing both. Previous studies have found that shared genetic risk factors may account for the association between nicotine dependence and schizophrenia (Ferchiou *et al.*, 2012, Lyons *et al.*, 2002). The association may therefore be an example of unmeasured **genetic confounding**, a phenomenon often overlooked in social science. There is also the possibility of unmeasured **environmental confounding**, occurring when there is one or more contextual factors that affect both the exposure and the outcome. For the example above, low socioeconomic status is worth mentioning, as it is associated with both smoking and schizophrenia (de Leon and Diaz, 2005).

One way to evaluate the extent of both genetic and environmental confounding is to study data from genetically informative samples. Although, as described below, such designs do not *rule out* confounding, the possibility of identifying causality is increased. This is especially true for the study of twin samples, and is one of the reasons for the increased popularity of twin studies seen over the past decades.

Table 1. Relevant national registries in Norway.

Health registries
Cause of Death Registry
Cancer Registry of Norway
Medical Birth Registry
Norwegian Immunization Registry
Norwegian Information System for the Nursing and Care Sector
Norwegian Patient Registry
Norwegian Prescription Database
Norwegian Surveillance System for Antimicrobial Drug Resistance
Norwegian Surveillance System for Communicable Diseases
Norwegian Surveillance System for Infections in Hospitals
Registry of Pregnancy Termination
Registry of the Norwegian Armed Forces Medical Services
The Central Tuberculosis Registry
Other registries
Norwegian Education Registry
Norwegian Income Registry
Statistics Norway's Historical Event Database
The National Registry

TWINS – A NATURAL QUASI-EXPERIMENTAL DESIGN

There are two types of twins – monozygotic and dizygotic. Whereas dizygotic (DZ) twins are the result of two sperm cells fertilizing two egg cells, monozygotic (MZ) twins develop from one fertilized egg cell that divides into two embryos during the first two weeks of gestation. As a result, DZ twins resemble ordinary siblings in that they share on average 50% of their genetic material, but unlike ordinary siblings they also share the intrauterine environment and time of birth. Given that MZ and DZ twin grow up in the same family, they are assumed to share the family environment to an equal extent (Plomin *et al.*, 2001). This implies that if MZ twins on average are more similar to each other within pairs than DZ twins on a given trait – the similarity must to some extent be due to genetic influences.

These characteristics of MZ and DZ twins make up a neat, natural quasi-experimental design and renders twin data useful for epidemiological research for at least two reasons. First, twin data make it possible to quantify to what degree genetic and environmental factors influence individual differences in susceptibility to a trait or disorder. This is most often done using the classical twin design (Jinks and Fulker, 1970, Martin and Eaves, 1977). Individual differences in various traits are assumed to originate from three sources; additive genetic (A), shared environmental (C), and non-shared environmental (E) sources. Whereas A make up the genetic part of the equation, C is defined as environmental factors contributing to similarity between twins, and is further assumed to have an equal effect on MZ and DZ twins. E is per definition not shared between twins in a pair, and therefore does not contribute to twin similarity. The E component also includes measurement errors. The amount to which these components contribute to variation in a trait can be estimated using structural equation modelling (Neale and Maes, 2000). Classical twin studies can also be used to investigate more challenging questions, such as why some traits (for instance mental disorders) tend to co-occur, and to what extent genetic and environmental influences can account for stability and change in a given trait or disorder. Twin methodology has also been used to investigate phenotypes that are not traditionally perceived as being influenced by genes, such as life events, and medical benefits.

Second, MZ twin pairs are as close as we can come to an ideal matched-pairs or matched-group design. In the latter design, individuals matched on certain characteristics believed to influence the outcome variable, are assigned to different experimental groups (Bordens and Abbott, 2002). However, there may be more variables, unknown to the researcher that can influence the outcome. In a twin design, the twin pairs are not only matched on genetic factors. Twins are also perfectly matched on common confounders such as age and childhood SES if raised together. Twin data can therefore be used to approximate causal associations. The most used design for this purpose is the discordant twin

design (McGue *et al.*, 2010). Twin pairs that are discordant on a trait, such as alcohol use, enable the researcher to investigate what the counterfactual consequences are for individuals not exposed to a certain condition. Alcohol consumption in moderate amounts has in some studies been linked to beneficial outcomes when compared to abstainers (Goldberg *et al.*, 1999). Although this may seem as a paradox, there are reasonable explanations as to why this pattern emerges. For instance – moderate alcohol consumption is positively associated with socioeconomic status, and negatively associated with disease. In investigating whether alcohol may have beneficial effects, it is important to use a design that allows us to discontinue the influence of confounding. A sample of genetically identical twins that shares rearing environment provides an almost ideal situation for comparing outcomes. If the association between alcohol consumption and an outcome is due to genetic or shared environmental confounding, the alcohol consuming co-twin would be just as likely as the abstaining twin to have a negative outcome.

TWIN REGISTRIES

A twin registry is a collection of data on MZ and DZ twins usually initiated and maintained by an academic institution. Unlike most other national registries, these data are typically consent-based. For broad research aims, the most valuable twin samples are population-based. Until a few years ago, there were several twin cohorts in Norway, distributed over different research institutes (Harris *et al.*, 2002). In 2009, these were incorporated in the Norwegian Twin Registry (NTR) that now consists of approximately 30,000 twins from birth cohorts spanning almost 100 years, and data on a wide range of variables (Nilsen *et al.*, 2012; Nilsen *et al.*, submitted). In other Nordic countries, centralized twin registries have existed longer. Sweden, Denmark and Finland have all established large-scale twin registries (Kaprio *et al.*, 1978, Magnusson *et al.*, 2013, Skytthe *et al.*, 2011). There are also examples of twin registries elsewhere in the world that seek to cover the whole country, such as in Australia (Hopper *et al.*, 2013), the Netherlands (Boomsma *et al.*, 2006), Sri Lanka (Sumathipala *et al.*, 2013) and the United Kingdom (Moayyeri *et al.*, 2013).

Efforts are being made to create twin registries that go across national borders. Through the International Network of Twin Registries (INTR) (Buchwald *et al.*, 2014), the researchers behind this initiative seek to build a platform for global scientific cooperation. The possibilities for scientific benefits are many with a global twin registry network. For instance, it will be possible to harmonize data to increase sample size and gain better understanding of phenotypes across national borders. Meta-analyses can be conducted, and the data will be accessible for a broader array of researchers from different countries.

STUDYING MEDICAL BENEFITS – FIVE EXAMPLES FROM RECENT RESEARCH

Medical benefits, such as economic reimbursement when health problems make a person unable to work, are key concepts of the welfare society. A crucial goal is to keep such benefits at a level that maximizes both the public's view of what is fair, and a level that is "healthy" for the population. With increasing levels of sick leave and disability pensioning in many Western societies the past decades, huge efforts have been made to promote research on risk and protective factors for sick leave and disability pension. As national registries are rare around the world, most of the studies on medical benefits originate from the Nordic countries. Research on medical benefits is challenging, as these are complex phenomena associated with a wide array of risk factors. There are also varying definitions and processes of certifications across countries. Previous research on medical benefits has typically focused on various types of risk factors. Most of the studies on risk factors have not had adequate design or data for revealing possible causal relationships, and many have been characterized as having low quality (Allebeck and Mastekaasa, 2004). Although it is valuable to establish associated risk factors, we also need knowledge on which risk factors that are causing individual differences in these phenomena.

In the remaining part of this paper, we will describe the resulting new knowledge that was gained through the linkage that was made in 2011 between NTR, described below, and FD-Trygd from Statistics Norway. The linkage enabled novel research on the association between personality disorders, alcohol use and medical benefits, among other aims, from a new angle. We will refer to five sub-studies from our research group, including four published papers and one work in progress.

Example 1 – the association between sick leave and disability pension

In the first sub-study, we investigated to what degree genetic and environmental factors contribute to two types of medical benefits that are closely related, namely long-term sick leave and disability pension (Gjerde *et al.*, 2013). Most individuals on sick leave (SL) return to work, but a few will transition to disability pension (DP). There is an increased risk of having DP after one or more episodes of SL (Gjesdal *et al.*, 2005, OECD, 2006), but regular epidemiological studies alone cannot inform us on why this is so. Before the project "Consequences of personality disorders – A population-based study of young adults", there was limited knowledge on the association between SL and DP. Evidence also pointed towards a social transmission mechanism of medical benefits (Kristensen *et al.*, 2004, Rege *et al.*, 2012) – meaning that if you have for instance a parent or a neighbour on a disability pension award, this could increase your

risk of also being awarded disability pension through a social "contagion" mechanism. As explained above, twin data can be used to investigate to what extent family and rearing environment (C) plays a part in this association. The few genetically informative studies that had been conducted on medical benefits also mostly included individuals older than 40 years (Hakonmäki *et al.*, 2008, Narusyte *et al.*, 2011, Svedberg *et al.*, 2012). Knowledge on etiological factors for medical benefits for young adults was therefore lacking.

We used a sample from The Norwegian Institute of Public Health Twin Panel (NIPHTP) (Harris *et al.*, 2006), now included in NTR. The panel contains information on twins that were identified through the Medical Birth Registry (MBR) of Norway. The MBR was established in 1967, and receives mandatory notification of all live- and stillbirths of at least 16 weeks of gestation. The NIPHTP consists of 15,374 like- and unlike-sexed twins that were born in Norway between 1967 and 1979. In 1992 and 1998 a subsample of the participants completed a questionnaire. Here, information on a wide array of variables such as mental and physical health was collected. In addition, an interview study was conducted between 1999 and 2004 to assess lifetime history of psychiatric disorders and substance abuse (Axis I) and personality disorders (Axis II) as diagnosed by the DSM-IV. Zygosity of the twins was determined using questionnaire items previously shown to classify correctly more than 97% of the twin pairs (Harris *et al.*, 2002, Magnus *et al.*, 1983). After linkage with FD-Trygd using National identification numbers, the sample consisted of 7,710 twins that were eligible for sick leave and disability pension allowance.

The disability pension (DP) variable comprised all twins that were on disability pension in the follow-up period (1998 to 2008). Individuals that had been granted DP before 1998 were also included if they were still on DP in the follow-up period. We constructed a variable scored as 0 = "no DP", 1 = "at least one period of graded (40-90%) DP", and 2 = "only full-time (100%) DP".

SL was defined as sickness absence > 16 days, the minimum sick leave period recorded in our dataset. We summed up the total number of sickness absence days, rehabilitation and working days in the 10 year follow up period; either up to the time of granted DP, death or 2008. The SL variable was then defined as a proportion (0-100%) between the cumulative number of sick days and rehabilitation days over the number of potential working days. As the level of employment varied between individuals in our sample, it was necessary to construct the variable as a proportion.

To answer these research questions, we first ran univariate twin models for SL and DP separately. We also included a sibling interaction parameter (Eaves, 1976) in the covariance expressions for the different zygosity groups (MZ twins vs. DZ twins) in the univariate models to investigate whether there was

evidence for twins within pairs affecting each other's liability to SL or DP. We also ran a bivariate Cholesky twin model (Figure 1) to investigate to what extent there were shared genetic and environmental influences on the phenotypes. The Cholesky method (Neale and Cardon, 1992) decomposes the variance-covariance structure of the measures into effects stemming from three latent sources (A, C and E). The first variable is assumed to be a perfect indicator of the latent factors "A₁", "C₁" and "E₁" that can also explain variance in the second variable. The second variable is also explained by a set of latent factors "A₂", "C₂" and "E₂" not shared by the observed variables. Full information maximum likelihood estimation was used to arrive at the parameter estimates, and Akaike's Information Criterion (Akaike, 1987) and chi-square tests were used to select the best fitting model.

We found that the association between SL and DP to a large extent was due to shared etiological factors. Thus, factors that increase risk for SL also tend to increase risk for DP. However, there was evidence of genetic variance not shared between these phenomena. This unique genetic risk factor could for instance reflect liability to more severe mental and somatic disorders that in the end may lead individuals to be excluded from the work force. Most of the variance in SL and DP was also due to genetic influences, which probably reflect liability to genetically transmittable diseases. This finding also implies that the transmission of risk for medical benefits within families is likely due to genetic rather than environmental mechanisms.

Example 2 – The association between personality disorders and sick leave

In the second sub-study, we explored the relationship between personality disorders and long-term sick leave (Gjerde *et al.*, 2014). The background for the paper was that very few studies have investigated how personality disorders affect propensity to work force participation, and none had used a genetically informative design.

The details of the study sample are described under Example 1 and consisted of 2,771 twins eligible for sick leave and with valid scores on the SIDP-IV.

The SL variable is described under Example 1. In addition, personality disorders (PDs) were measured with the Structured Interview for Personality (SIDP-IV), which is a comprehensive semi-structured diagnostic interview for DSM-IV Axis II disorders (Pfohl and Zimmerman, 1995). Each DSM-IV criterion is scored as 0 = "absent", 1 = "sub-threshold", 2 = "present" or 3 = "strongly present". We summed the criteria that were scored as 1 or higher, in accordance with previous studies on the same sample (see e.g. Kendler *et al.*, 2006). In effect, we therefore did not assess PDs per se, but rather dimensional representations of PDs. For convenience, we refer to these variables as PDs.

We first conducted ordinal regression analyses to investigate which PDs were associated with SL. The

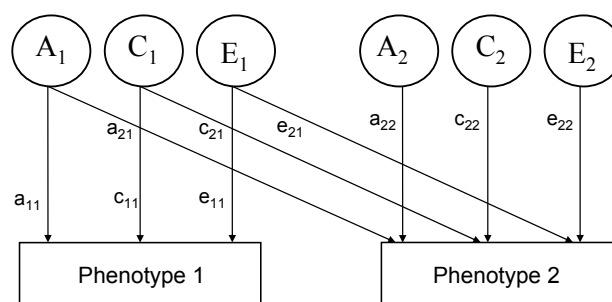


Figure 1. A bivariate twin model, shown for one of the twins in a pair.

Cholesky model described under Paper 1 was then expanded to include four variables. Although multivariate Cholesky models do not permit conclusions regarding causal pathways, some inferences can be made. For phenotypes that are influenced by genetic and environmental factors, finding only a genetic, but not an environmental correlation, would not indicate causality, but rather that the association is due to genetic factors shared between the phenotypes (De Moor *et al.*, 2008).

We found significant associations between schizotypal, borderline and paranoid PD and SL, but 90% of the variance in SL was due to other factors than these disorders. The variance that was shared between the PDs and SL was almost entirely due to shared genetic influences. Together, the genetic contributions from the three PDs could account for 20% of the heritability of SL. The results did not support a causal hypothesis, but rather that the association between PDs and SL was due to genetic confounding. Also, the results imply that environmental factors that increase risk for PDs are not the same as those that increase risk for SL.

Example 3 – alcohol consumption and sick leave

There is a widely held belief that moderate alcohol consumption is health-promoting as compared to not drinking. This assumption is based on the curve-linear relationship that is often observed between alcohol consumption and various health measures, where low drinkers tend to have worse health than moderate drinkers. Some argue that this favourable outcome among moderate drinkers is due to a genuine health-promoting effect of alcohol (Vahtera *et al.*, 2002), while others believe that these findings are more likely due to unmeasured confounding (Andreasson, 2007, Fekjær, 2013, Holder, 2007, Klatsky, 2007). By using twin data, it is possible to investigate this association between alcohol consumption and a chosen health measure to examine whether the association is influenced by genetic confounding.

In 1998, 8,045 twins from the NIPHTP/NTR (described under Example 1) participated in a comprehensive questionnaire based study on mental and somatic health, in which they responded to questions on health measure to examine whether the association is influenced by genetic confounding.

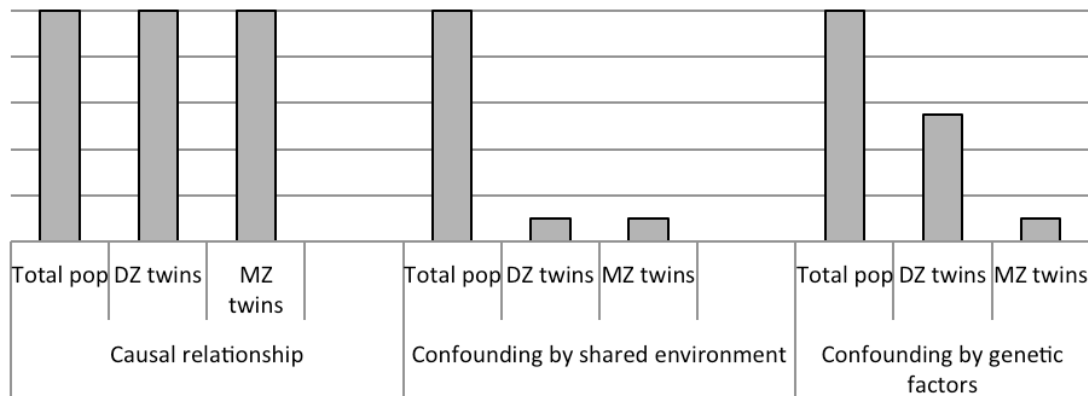


Figure 2. Possible risk patterns that can be expected in a discordant twin analysis.

In 1998, 8,045 twins from the NIPHTP/NTR (described under Example 1) participated in a comprehensive questionnaire based study on mental and somatic health, in which they responded to questions on health behaviour and alcohol use. For each question, the participants were categorized as belonging to one of three consumption groups: low, moderate or high, and in the further analyses we compared the participants with a low level of alcohol consumption with those with a moderate level of consumption. The final sample comprised 6,734 twins.

We performed discordant twin analyses to examine the effect of moderate versus low use of alcohol and the associated effect on sick leave. In a discordant twin analysis, one first estimates an association between an exposure and an outcome (in our case alcohol use and sick leave) in the total sample, for example using linear regression. Then the same association is estimated within twin pairs discordant for the exposure (i.e. alcohol use), for example by using paired sample t-tests. This is done separately for MZ and DZ twin pairs. When discordant twin analyses are performed to investigate whether an association is confounded or not, one of three main patterns of results can be expected. If the relationship between exposure and outcome is causal, and not confounded by childhood environment or genetic factors, the association will be just as strong in the total population as within discordant twin pairs (left model in Figure 2). If the relationship is non-causal and confounded by early shared environment, the association will be weaker within discordant twin pairs than in the total population, and the same within di-zygotic pairs as within monozygotic pairs (middle model in Figure 2). If the relationship is non-causal and confounded by genetic factors, the association will weaken with increased genetic similarity, meaning the association will be weaker within monozygotic pairs than within di-zygotic pairs (right model in Figure 2).

Our preliminary results indicate a risk pattern resembling the rightmost chart in Figure 2, with the association between alcohol and sick leave being weaker within DZ twin pairs than in the total population, and even weaker within MZ twin pairs. This implies that the increased risk for sick leave associated

with low drinking (an association observed in the total population), would be assumed to not be a causal effect from “lack of alcohol”, but rather an effect that is confounded by genetic or familial factors.

Example 4 – Internalizing disorders and sick leave granted for mental and somatic disorders

In the fourth sub-study (Torvik *et al.*, 2014), we investigated to what degree internalizing disorders (anxiety and depression) were prospectively associated with sick leave. It was already known that mental disorders were associated with sick leave, and that mental disorders were associated with disability due to somatic conditions. However, the association between mental disorders and sick leave due to somatic disorders was not known. In addition, we investigated the extent to which common genetic and environmental risk factors influenced these relationships.

We used two different conceptualizations of internalizing disorders: In the large sample consisting of 7,710 twins described under Example 1, we used symptoms of anxiety and depression, assessed with five self-report questions. In a subsample, consisting of 2,770 twins, described under Example 1, we defined diagnostically assessed mood and anxiety disorders, with the exception of specific phobia, as internalizing disorders. The sick leave variables were defined similarly as in the other papers, with the exception that we used two sick leave variables (one for mental and one for somatic diagnoses).

We fitted trivariate Cholesky decompositions (described under Example 1) to the data and ran discordant twin analyses (described under Example 3). The results indicated that internalizing disorders were associated with sick leave granted for both mental disorders and somatic disorders. Both genetic and non-shared environmental factors were found to contribute to the association between internalizing disorders and sick leave granted for mental disorders to an extent that was consistent with a causal effect, as one would expect. However, the association between internalizing disorders and sick leave granted for somatic disorders could be explained by common genetic factors alone, which is not consistent with causality. In line with this, MZ twins discordant for internalizing disorders

differed significantly in rates of sick leave granted for mental but not somatic disorders. The finding that internalizing disorders lead to more sick leave granted for mental disorders is as one should expect, and may be seen as a validation of the method. However, the association between internalizing disorders and sick leave granted for somatic disorders appears to be confounded by genetic factors. Possible explanations for this confounding include comorbidity, possibly due to shared risk factors such as neuroticism, coping skills, and sickness behavior. In conclusion, internalizing disorders in young adults indicated a higher risk of sick leave granted for both mental and somatic disorders, but only the association with sick leave granted for mental disorders was consistent with causal explanations.

Example 5 – Socioeconomic status and sick leave

The fifth sub-study concerned the socioeconomic status (SES) gradient in sick leave (Torvik *et al.*, 2015). There is a well-documented social gradient in health, but there is no consensus on whether this association is causal. There are several causal explanations. Low SES is associated with risky health behaviors and harsher work environments and can affect physiological outcomes, such as blood pressure (Mendelson *et al.*, 2008). On the other hand, there could be selection of healthy individuals into high SES. Parts of the SES gradient in health appear to be confounded by genetic factors (Osler *et al.*, 2007). SES differences in sick leave are not merely a byproduct of the health gradient, but the same mechanisms could be at play for the SES gradient in sick leave. Personality (Roberts *et al.*, 2007), self-control (Moffitt *et al.*, 2011), and general ability (Batty *et al.*, 2006) could represent background risk factors confounding the association between SES and sick leave. There can also be selection into jobs with hard working conditions. In this paper, we estimated the degree to which education and income were prospectively related to sick leave granted for mental and somatic disorders and sick leave granted for any disorder, and investigated whether these associations remained when adjusting for confounding genetic and shared environmental factors, that is, whether the associations were consistent with a causal explanation. We used the questionnaire sample described under Example 1, and used registry data on educational attainment and income at age 30 as exposure variables, and subsequent sick leave as outcome. The average follow-up time was 6.6 years. The results indicated that low education and income were associated with sick leave granted for both mental and somatic disorders, and with sick leave granted for any disorder. Education was more strongly related to sick leave than was income, possibly because the sample was young. Associations were attenuated within dizygotic twin pairs and reduced to non-significance within monozygotic twin pairs, suggesting influence of familial factors on the associations between SES and sick leave. Therefore, the association between SES and

sick leave was at least strongly, and possibly fully, confounded by familial factors. Non-significant findings cannot be taken as evidence that causal effects between SES and sick leave do not exist. However, considering the sizes of the confidence intervals, possible causal effects were clearly smaller than the crude associations between SES and sick leave. Of course, it may be that other indicators of SES, such as social capital or occupation are more important with regard to sick leave. Some individuals hold jobs where they cannot fully utilize their education, whereas other individuals with low educational level achieve high-status positions. Further, and importantly, if health differences between SES groups are due to lifestyle, diet, substance abuse, health literacy or wear and tear injuries, the effects of these are likely to accumulate over time. The effects of SES on sick leave may therefore be stronger among older individuals. Nevertheless, education and income per se were not likely to strongly affect sick leave in young adulthood.

DISCUSSION

In the current paper, we have used the examples of personality disorders, internalizing disorders, alcohol use and socioeconomic status and their possible influence on medical benefits to illustrate how the linking of a health registry with a twin registry can contribute to maximizing the potential inherent in both data sources.

The first study example – namely on the genetic and environmental association between sick leave and disability pension, was the first to investigate this association using a twin design. Only two Swedish (Narusyte *et al.*, 2011, Svedberg *et al.*, 2012) and a Finnish study (Harkonmäki *et al.*, 2008) had been conducted before on the phenotypes separately using a twin design. These countries were most likely the first to look into these phenotypes using twins due to their excellent population- and twin registry data. Twin studies generally require large samples in order to attain sufficient statistical power. When investigating rare conditions such as disability pension, the sample size becomes vital. The study described in example 1 would be possible to do without registry data. Sweden has a large twin study where self-report data on sick leave and disability pension was obtained for over 29,000 twins (Svedberg *et al.*, 2010). When comparing the self-report data with registry data, the authors found good agreement between self-report and registry data on disability pension, but only moderate agreement for sick leave (Svedberg *et al.*, 2010). A study based on self-report would therefore be an alternative if registry data were not available, but the quality of the data would have been different. Although the Swedes have the opportunity to use the rich data source from their twin studies, such a study would be difficult to do in Norway without the registry data on medical benefits, as after linkage with the twin registry, only 253 individuals had disability pension. Doing

twin modelling on so few data points was challenging, and impossible for all zygosity groups. Consequently, we were unable to investigate sex differences.

Much of the same arguments as above can be used for the second example on personality disorders and sick leave. The study was the first of its kind, and would be difficult to do without registry data and the fairly large twin sample inherent in NTR for which we have diagnostic data on. The number of twins endorsing personality disorders symptoms was however low, and this is a limitation of the study. It is possible that linking the entire NTR with the Norwegian Patient Registry would improve matters somewhat, but then we may encounter another problem, as health registries depend on information from health services. Studies using health registry data are therefore, like studies with clinically recruited samples, not necessarily representative for all individuals with a condition, but rather for subjects receiving treatment for the condition. The bias increases when we take into consideration the fact that individuals with personality disorders come into contact with treatment services less often than individuals with other mental disorders. Moreover, subjects in treatment usually have more severe disorders and higher rates of comorbidity, a pattern that may contribute to Berkson's bias (Berkson, 1946).

In all examples but the first and second, we could profit from the large amount of twins participating in the questionnaire study. The linkage with the FD-Trygd registry opened up possibilities to investigate phenomena for which there exists a large phenotypic literature, such as alcohol consumption, internalizing disorders and socioeconomic status. By using discordant twin analyses to investigate associations we placed ourselves in a stronger position to delineate the causes behind the phenotypic links. The preliminary indications of familial confounding between alcohol and sick leave may be a welcomed contribution to an inconclusive literature. Furthermore, it may add nuance to the positive image of drinking that is often portrayed in the media. The study on internalizing disorders and mental and somatic sick leave showed that there is most likely a causal link between internalizing disorders and sick leave granted for mental disorders, but that the association between internalizing disorders and sick leave granted for somatic disorders most likely was due to genetic confounding. The first part may not be surprising, and is also supported by previous observational studies (e.g. Stansfeld *et al.*, 2011). The use of a discordant twin design rendered us able to make a stronger claim that the association is true than has previously been possible. The study on socioeconomic status and sick leave was the first of its kind, and profited from the detailed and annually updated data on education and income in the Norwegian Education Registry and the Norwegian Income Registry. The study found evidence for the associations being due to familial confounding. Thus, other factors may be involved,

such as personality or specific disorders, in explaining the associations.

The study on internalizing disorders and sick leave may be the best example included here to demonstrate how one can maximize the value that lies within both national registries and data available from previous twin studies. Here, diagnoses were extracted from two independent sources – FD-Trygd and the interview study in NTR. In addition, symptoms of internalizing disorders were available from the questionnaire study in NTR. The registry diagnoses were used to separate sick leave due to mental and somatic disorders. Twin analyses were then conducted on the two NTR sources of internalizing symptoms, which made a validity-check of the results possible.

Limitations

It would be beyond the scope of this paper to go into details on the limitations in the various examples described. These are described in the previously published papers (Gjerde *et al.*, 2013, Gjerde *et al.*, 2014, Torvik *et al.*, 2014, Torvik *et al.*, 2015). It is, however, important to note that twin studies have several important limitations, whether data are linked to public registries or not. Most important is the issue of statistical power. As only a fraction of the population are twins – and thus eligible for inclusion in population based twin studies, it is hard to obtain large enough data sets to study disorders or traits of low prevalence. The analyses are further restricted by the requirement of a sufficient number of discordant pairs in all zygosity groups to obtain reliable data from behavioral genetic and co-twin control analyses. Some of these limitations are described in more details in other papers in the current issue of *Norsk Epidemiologi*.

Nevertheless, we argue that there is much to be gained from linking health surveys from twins with population registries, and most importantly a deeper understanding of the association between exposure and outcomes first obtained from traditional epidemiologic studies. Our examples above illustrate some of these potentials. The efforts that have been made to create the NTR in Norway and the INTR internationally make these types of linkage studies easier to conduct and available to more researchers. As there are still many areas to explore using the potential within a linkage between national and twin registries, more epidemiological researchers should make use of this possibility.

ACKNOWLEDGEMENTS

The work was supported by a grant from The Sickness Absence Research Programme at The Norwegian Research Council. The collection of twin data was also in part supported by previous grants from The Norwegian Research Council, The Norwegian Foundation for Health and Rehabilitation, and The Norwegian Council for Mental Health. We are very grateful to the twins for their participation.

REFERENCES

- Akaike, H. (1987). Factor analysis and AIC. *Psychometrika* **52**, 317-332.
- Akselsen, A., Lien, S. & Siverstøl, Ø. (2007). FD-trygd. List of variables. Statistics Norway: Oslo.
- Allebeck, P. & Mastekaasa, A. (2004). Chapter 5. Risk factors for sick leave – general studies. *Scand J Public Health* **32**, 49-108.
- Andreasson, S. (2007). The health benefits of moderate consumption called into question. *Addict Res Theory* **15**, 3-5.
- Batty, G. D., Der, G., Macintyre, S. & Deary, L. (2006). Does IQ explain socioeconomic inequalities in health? Evidence from a population based cohort study in the west of Scotland. *BMJ* **332**, 580-583.
- Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics* **2**, 47-53.
- Boomsma, D. I., de Geus, E. J. C., Vink, J. M., Stubbe, J. H., Distel, M. A., Hottenga, J. J., et al. (2006). Netherlands Twin Register: From twins to twin families. *Twin Res Hum Genet* **9**, 849-857.
- Bordens, K. S. & Abbott, B. B. (2002). *Research design and methods: A process approach (5th ed.)*. McGraw-Hill: New York, NY, US.
- Buchwald, D., Kaprio, J., Hopper, J. L., Sung, J., Goldberg, J., Fortier, I., et al. (2014). International Network of Twin Registries (INTR): Building a platform for international collaboration. *Twin Res Hum Genet* **17**, 574-577.
- Dahl, C., Stoltenberg, C. (2009). Gode helseregistre – bedre helse. Strategi for modernisering og samordning av sentrale helseregistre og medisinske kvalitetsregistre 2010-2020. Folkehelseinstituttet: Oslo.
- Dalack, G. W., Healy, D. J. & Meador-Woodruff, J. H. (1998). Nicotine dependence in schizophrenia: Clinical phenomena and laboratory findings. *Am J Psychiatry* **155**, 1490-1501.
- de Leon, J. & Diaz, F. J. (2005). A meta-analysis of worldwide studies demonstrates an association between schizophrenia and tobacco smoking behaviors. *Schizophrenia Res* **76**, 135-157.
- De Moor, M. H. M., Boomsma, D. I., Stubbe, J. H., Willemsen, G. & de Geus, E. J. C. (2008). Testing causality in the association between regular exercise and symptoms of anxiety and depression. *Arch Gen Psychiatry* **65**, 897-905.
- Eaves, L. (1976). A model for sibling effects in man. *Heredity* **36**, 205-214.
- Fekjær, H. O. (2013). Alcohol – a universal preventive agent? A critical analysis. *Addiction* **108**, 2051-2057.
- Ferchiou, A., Szoke, A., Laguerre, A., Meary, A., Leboyer, M. & Schurhoff, F. (2012). Exploring the relationships between tobacco smoking and schizophrenia in first-degree relatives. *Psychiatry Res* **200**, 674-678.
- Frank, L. (2000). Epidemiology – When an entire country is a cohort. *Science* **287**, 2398-2399.
- Gjerde, L. C., Knudsen, G. P., Czajkowski, N., Gillespie, N., Aggen, S. H., Røysamb, E., et al. (2013). Genetic and environmental contributions to long-term sick leave and disability pension: a population-based study of young adult Norwegian twins. *Twin Res Hum Genet* **16**, 759-766.
- Gjerde, L. C., Røysamb, E., Czajkowski, N., Knudsen, G. P., Østby, K., Tambs, K., et al. (2014). Personality disorders and long-term sick leave: A population-based study of young adult Norwegian twins. *Twin Res Hum Genet* **17**, 1-9.
- Gjesdal, S., Haug, K. & Ringdal, P. R. (2005). The risk of disability pension among young persons on long-term sick leave. *Tidsskr Norske Lægeforen* **125**, 1801-1805.
- Goldberg, D. M., Soleas, G. J. & Levesque, M. (1999). Moderate alcohol consumption: The gentle face of Janus. *Clin Biochem* **32**, 505-518.
- Harkonmäki, K., Silventoinen, K., Levälähti, E., Pitkaniemi, J., Huunan-Seppälä, A., Klaukka, T., et al. (2008). The genetic liability to disability retirement: A 30-year follow-up study of 24,000 Finnish twins. *PLoS One* **3**, e3402.
- Harris, J. R., Magnus, P. & Tambs, K. (2002). The Norwegian Institute of Public Health Twin Panel: A description of the sample and program of research. *Twin Res* **5**, 415-423.
- Harris, J. R., Magnus, P. & Tambs, K. (2006). The Norwegian Institute of Public Health twin program of research: an update. *Twin Res Hum Genet* **9**, 858-864.
- Holder, H. D. (2007). A warning shot across the bow? *Addict Res Theory* **15**, 6-8.
- Hopper, J. L., Foley, D. L., White, P. A. & Pollaers, V. (2013). Australian Twin Registry: 30 Years of Progress. *Twin Res Hum Genet* **16**, 34-42.
- Jinks, J. L. & Fulker, D. W. (1970). Comparison of biometrical genetical, MAVA, and classical approaches to analysis of human behavior. *Psychol Bull* **73**, 311-349.
- Kaprio, J., Sarna, S., Koskenvuo, M. & Rantasalo, I. (1978). The Finnish Twin Registry: formation and compilation, questionnaire study, zygosity determination procedures, and research program. *Prog Clin Biol Res* **24 Pt B**, 179-184.
- Kendler, K. S., Czajkowski, N., Tambs, K., Torgersen, S., Aggen, S. H., Neale, M. C. & Reichborn-Kjennerud, T. (2006). Dimensional representations of DSM-IV Cluster A personality disorders in a population-based sample of Norwegian twins: A multivariate study. *Psychol Med* **36**, 1583-1591.
- Klatsky, A. L. (2007). Errors in selection of "error-free" studies. *Addict Res Theory* **15**, 8-16.
- Kristensen, P., Bjerkedal, T. & Brevik, J. I. (2004). Long term effects of parental disability: A register based life course follow-up of Norwegians born in 1967-1976. *Norsk Epidemiologi* **14**, 97-105.
- Levander, S., Eberhard, J. & Lindstrom, E. (2007). Nicotine use and its correlates in patients with psychosis. *Acta Psychiatr Scand* **116**, 27-32.
- Li, M. D., Cheng, R., Ma, J. Z. & Swan, G. E. (2003). A meta-analysis of estimated genetic and environmental effects on smoking behavior in male and female adult twins. *Addiction* **98**, 23-31.

- Lovdata (2015). Helseregisterloven [Act on Personal Health Data Filing Systems]. <https://lovdata.no/dokument/NL/lov/2014-06-20-43?q=helseregister>: Norway.
- Lyons, M. J., Bar, J. L., Kremen, W. S., Toomey, R., Eisen, S. A., Goldberg, J., et al. (2002). Nicotine and familial vulnerability to schizophrenia: A discordant twin study. *J Abnorm Psychol* **111**, 687-693.
- Magnus, P., Berg, K. & Nance, W. E. (1983). Predicting zygosity in Norwegian twin pairs born 1915-1960. *Clin Genet* **24**, 103-112.
- Magnusson, P. K. E., Almqvist, C., Rahman, I., Ganna, A., Viktorin, A., Walum, H., et al. (2013). The Swedish Twin Registry: Establishment of a biobank and other recent developments. *Twin Res Hum Genet* **16**, 317-329.
- Martin, N. G. & Eaves, L. J. (1977). The genetical analysis of covariance structure. *Heredity* **38**, 79-95.
- McGue, M., Osler, M. & Christensen, K. (2010). Causal inference and observational research: the utility of twins. *Perspect Psychol Sci* **5**, 546-556.
- Mendelson, T., Kubzansky, L. D. & Thurston, R. C. (2008). Affective and cardiovascular effects of experimentally-induced social status. *Health Psychology* **27**, 482-489.
- Moayyeri, A., Hammond, C. J., Hart, D. J. & Spector, T. D. (2013). The UK Adult Twin Registry (TwinsUK Resource). *Twin Res Hum Genet* **16**, 144-149.
- Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., et al. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proc Natl Acad Sci USA* **108**, 2693-2698.
- Narusyte, J., Ropponen, A., Silventoinen, K., Alexanderson, K., Kaprio, J., Samuelsson, A. & Svedberg, P. (2011). Genetic liability to disability pension in women and men: a prospective population-based twin study. *PLoS One* **6**, e23143.
- Neale, M. C. & Cardon, L. R. (1992). *Methodology for genetic studies of twins and families*. Kluwer Academic Publishers: Dordrecht, the Netherlands.
- Neale, M. C. & Maes, H. H. (2000). *Methodology for genetic studies of twins and families*. Kluwer Academic Publishers: Dordrecht, the Netherlands.
- Nilsen, T. S., Brandt, I., Magnus, P. & Harris, J. R. (2012). The Norwegian Twin Registry. *Twin Res Hum Genet* **15**, 775-780.
- OECD (2006). *Sickness, Disability and Work: Breaking the Barriers – Norway, Poland and Switzerland Vol. 1*. OECD Publishing: Paris.
- Osler, M., McGue, M. & Christensen, K. (2007). Socioeconomic position and twins' health: a life-course analysis of 1266 pairs of middle-aged Danish twins. *Int J Epidemiol* **36**, 77-83.
- Pfohl, B. B. & Zimmerman, M. (1995). *Structured Interview for DSM-IV Personality (SIDP-IV)*. University of Iowa, Department of Psychiatry: Iowa City.
- Plomin, R., DeFries, J. D., McClearn, G. E. & McGuffin, P. (2001). *Behavioral Genetics*. Worth Publishers: New York.
- Rege, M., Telle, K. & Votruba, M. (2012). Social interaction effects in disability pension participation: evidence from plant downsizing. *Scand J Econ* **114**, 1208-1239.
- Richmond, R. C., Al-Amin, A., Smith, G. D. & Relton, C. L. (2014). Approaches for drawing causal inferences from epidemiological birth cohorts: A review. *Early Hum Dev* **90**, 769-780.
- Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A. & Goldberg, L. R. (2007). The power of personality: the comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspect Psychol Sci* **2**, 313-345.
- Skytthe, A., Kyvik, K. O., Holm, N. V. & Christensen, K. (2011). The Danish Twin Registry. *Scand J Public Health* **39**, 75-78.
- Stansfeld, S. A., Fuhrer, R. & Head, J. (2011). Impact of common mental disorders on sickness absence in an occupational cohort study. *Occup Environ Med* **68**, 408-413.
- Sullivan, P. F., Kendler, K. S. & Neale, M. C. (2003). Schizophrenia as a complex trait – Evidence from a meta-analysis of twin studies. *Arch Gen Psychiatry* **60**, 1187-1192.
- Sumathipala, A., Siribaddana, S., Hotopf, M., McGuffin, P., Glozier, N., Ball, H., et al. (2013). The Sri Lankan Twin Registry: 2012 Update. *Twin Res Hum Genet* **16**, 307-312.
- Svedberg, P., Ropponen, A., Alexanderson, K., Lichtenstein, P. & Narusyte, J. (2012). Genetic susceptibility to sickness absence is similar among women and men: Findings from a Swedish twin cohort. *Twin Res Hum Genet* **15**, 642-648.
- Svedberg, P., Ropponen, A., Lichtenstein, P. & Alexanderson, K. (2010). Are self-report of disability pension and long-term sickness absence accurate? Comparisons of self-reported interview data with national register data in a Swedish twin cohort. *BMC Public Health* **10**, 763.
- Torvik, F. A., Gjerde, L. C., Røysamb, E., Tambs, K., Kendler, K. S., Czajkowski, N. O., et al. (2014). Genetic and environmental contributions to the relationship between internalizing disorders and sick leave granted for mental and somatic disorders. *Twin Res Hum Genet* **17**, 225-235.
- Torvik, F. A., Ystrom, E., Czajkowski, N., Tambs, K., Røysamb, E., Ørstavik, R., et al. (2015). Socioeconomic status and sick leave granted for mental and somatic disorders: a prospective study of young adult twins. *BMC Public Health* **15**, 9.
- Vahtera, J., Poikolainen, K., Kivimäki, M., Ala-Mursula, L. & Pentti, J. (2002). Alcohol intake and sickness absence: A curvilinear relation. *Am J Epidemiol* **156**, 969-976.