From the DEPARTMENT OF MEDICAL EPIDEMIOLOGY
AND BIOSTATISTICS
Karolinska Institutet, Stockholm, Sweden

TWIN AND FAMILY STUDIES ON THE DEVELOPMENT OF COGNITIVE AND EXTERNALIZING PROBLEMS

Ralf Kuja-Halkola

Stockholm 2014

Published by Karolinska Institutet. Printed by Universitetsservice US-AB.

ISBN 978-91-7549-460-9

ABSTRACT

Cognitive and externalizing problems are responsible for much personal suffering, as well as large monetary costs for society. Intervention and prevention efforts have often failed in reduction of unwanted behaviors, perhaps due to lack of understanding of the development of these complex traits. Studies on risk factors often treat the associations naively by not considering potential unmeasured common causes of the risk factor and the outcome. These common causes may be shared within families; i.e., the association is subject to familial confounding. Analyses informed of family belonging can help to further the knowledge regarding causality. Therefore, in this thesis, I used existing, and developed novel, methodologies to assess familial confounding. Models to adjust for familial confounding, as well as models identifying sources of familial confounding, were implemented. Further, I developed a genetically sensitive longitudinal design with multiple raters and time-points.

In study I, advancing paternal age was associated with offspring violent offending. Advancing paternal age was found to increase the incidence of violent criminal convictions among re-offending offspring when siblings were compared; a result congruent with causal inference. Contrary, and congruent with non-causal inference, advancing paternal age did not increase the probability to ever be convicted a violent criminal offence

Study II identified an association between maternal smoking during pregnancy (SDP) and offspring stress coping in late adolescence. However, the association did not persist when exposure-discordant siblings were compared. This result is compatible with a non-causal interpretation, and suggests that the association is due to familial confounding; other factors shared between siblings accounts for the association. In a quantitative genetic analysis these factors were found to be of genetic origin.

Study III continued the investigation of SDP, this time as a risk factor for cognitive outcomes (general cognitive ability and poor academic achievement), externalizing outcomes (criminal convictions, violent criminal convictions, and drug misuse), and pregnancy related outcomes (birth weight, preterm birth, and born small for gestational age). SDP was associated with all outcomes, but within-sibling analyses found that the association persisted only for pregnancy related outcomes, and disappeared for cognitive and externalizing outcomes. Quantitative genetic analyses found that genetics accounted for the majority of the association between SDP and cognitive and externalizing outcomes.

In study IV externalizing traits in mid-childhood were found to predict ADHD-like traits in adolescence when pre-existing associations were adjusted for. Further, ADHD-like traits in late adolescence predicted externalizing behavior in young adult age. The two traits were correlated in mid-childhood (age 8-9), and become even more correlated through early (age 13-14) and late (age 16-17) adolescence and young adult age (age 19-20). Stable and new factors accounted for approximately half of the

correlation between the traits each, throughout development. Genetic variation explained two thirds of the correlations.

In conclusion, advancing paternal age might increase the rate of violent offending among violent re-offenders. SDP does not seem to be causing offspring cognitive and externalizing problems in adolescence and adulthood; the observed associations are better explained by shared genetic factors. Externalizing behavior predicts ADHD early in development, and ADHD predicts externalizing behavior late in development. And, although new sources of covariance arise throughout development, ADHD-like and externalizing traits become even more correlated from childhood to young adulthood.

SAMMANFATTNING PÅ SVENSKA

Kognitiva problem och externaliserande (utåtagerande) beteenden har potentiellt väldigt skadliga effekter, både på individen och på samhället i stort. Genom att identifiera kausala samband (orsakssamband) mellan riskfaktorer och kognitiva och externaliserande problem ökar möjligheten för preventiva åtgärder. Omvänt så har preventiva insatser baserade på skensamband haft begränsade, eller till och med skadliga, effekter för, till exempel, antisocialt beteende. Problematiken i att identifiera kausala riskfaktorer är som störst då det inte finns möjlighet att utföra experimentella studier på samband, utan forskningen måste förlita sig observationsstudier. I observationsstudier har forskaren ingen kontroll på vem i studiepopulationen som blir exponerad. Detta är problematiskt då det oftast inte är slumpen som avgör vem som blir exponerad för, till exempel, rökning under graviditet. Blivande mödrar som röker under graviditeten är potentiellt annorlunda än de som inte gör det, och orsakerna till detta kan vara associerat med ett studerat utfall, till exempel förhöjd risk för antisocialt beteende. Barnet kan ärva både benägenhet att röka och förhöjd risk för antisocialt beteende. En association mellan mammans rökning under graviditeten och antisocialt beteende hos barnet skapas således, oberoende på om rökningen i sig påverkar risken för antisociala handlingar. Dessa typer av skensamband har jag valt att kalla familjär confounding (ungefär sammanblandning eller förväxling). I de fyra studier i denna avhandling använder jag, och utvecklar, metoder för att (1) justera för familjär confounding, och (2) skatta troliga faktorer som orsakar sambandet (genetiska och miljömässiga).

I första studien fann vi ett samband mellan äldre fäder och fler våldsbrott hos barnen. Tidigare forskning har visat på ett samband mellan äldre pappor och risk för allvarliga mentala sjukdomar som schizofreni och bipolär sjukdom, och dessa sjukdomar är associerade med högre risk att begå våldsbrott. Vi delade upp våldsbrottslighet i två mått; sannolikhet att bli dömd för minst ett våldsbrott, och antalet våldsbrott hos återfallsförbrytare. Då vi justerade för familjära faktorer, genom att jämföra inom syskon med samma pappa, fann vi att sambandet mellan äldre fäder och sannolikheten att begå minst ett våldsbrott försvann. Dock kvarstod sambandet för antalet våldsbrott hos återfallsförbrytare. Ungdomsbrottslingar (individer som begår antisociala handlingar under ungdomen men sedan slutar) och återfallsförbrytare antas ha olika etiologi, där återfallsförbrytare antas vara mer biologisk drivna. En hypotes som föreslagits förklara sambandet mellan äldre fäder och mentala sjukdomar är *de novo*-mutationshypotesen. Mannens könsceller delas kontinuerligt efter puberteten, och vid varje delning finns viss risk för nya mutationer; *de novo*-mutationer. Vårt fynd indikerar en biologisk förändring som påverkar våldsbrottsligheten, detta är samstämmigt med *de novo*-mutationshypotesen.

I andra och tredje studien var vi intresserade av hur mammans rökning under graviditet påverkade barnets kognitiva problem (här fångat med mått på stresstålighet, intelligens och akademisk framgång) och externaliserande beteenden (brottsdomar, våldsbrottsdomar och drog- och alkoholmissbruk) under ungdom och ung vuxen ålder.

Vi kontrasterade detta med samband mellan rökning under graviditet och födelserelatera utfall (födelsevikt, född för tidigt, och född liten för graviditetslängd). Vi fann associationer mellan rökning under graviditet och samtliga utfall. Då vi jämförde inom syskon kvarstod sambandet för de födelserelatera utfallen, men det försvann för de kognitiva och externaliserande utfallen. Detta indikerar en skillnad mellan typerna av utfall; resultatet för födelserelaterade utfallen är samstämmigt med en kausal tolkning, medan resultatet för kognitiva och externaliserande utfallen förklaras bättre av familjär confounding än kausala effekter. I analyser där familjära confoundingen skattades fann vi att genetiska effekter förklarade merparten av samvariationen. Detta kan tolkas som att det finns genetisk predisposition för mamman att röka under sin graviditet vilken är associerad med kognitiva och externaliserande beteenden. Denna genetik ärvs av barnen, som då uppvisar dessa kognitiva och externaliserande problem.

I fjärde studien undersökte vi samutvecklingen av ADHD och externaliserande beteende från barnaår (8-9 års ålder) till ung vuxen ålder (19-20 år) i en födelsekohort av tvillingar. Skattningar av ADHD och externaliserande beteende hade gjorts vid fyra tillfällen under denna tid, både av föräldrar och av tvillingarna själva. Vi fann att externaliserande beteende predicerade ADHD från barndom (8-9 år) till tidig ungdom (13-14 år), men inte *vice versa*. ADHD predicerade dock externaliserande beteende från sen ungdom (16-17 år) till tidig vuxen ålder (19-20 år). ADHD och externaliserande beteende samvarierade till en betydande grad i barndomen, och blev sedan mer och mer samvarierande fram till ung vuxen ålder. Två tredjedelar av denna ökande samvariation kunde tillskrivas genetiska effekter. Tidigare ADHD och externaliserande beteende förklarade ungefär hälften av samvariationen vid varje tidpunkt, detta innebar att ungefär hälften förklarades av nytillkommen samvariation. Samutvecklingen av ADHD och externaliserande beteende är således ytterst dynamisk, detta utmanar den rådande tron att ADHD föregår externaliserande beteende.

Sammantaget visar avhandlingen på vikten av att testa alternativa hypoteser för associationer där samma familjära faktorer kan orsaka både exponering och utfall. Med metoder för att kontrollera för sådana familjära faktorer fann vi att ökande ålder hos fäder ökar antalet våldsbrott hos återfallsförbrytare. Vidare fann vi att associationen mellan rökning under graviditet och kognitiva och externaliserande problem hos barnen bättre förklaras av familjära faktorer än som kausala samband. Associationerna mellan släktingar var sådana att genetisk likhet bäst förklarade sambanden. Vi fann även att externaliserande beteende predicerar ADHD från barndom till tidig ungdom, och att ADHD predicerar externaliserande beteende från sen ungdom till ung vuxen ålder. ADHD och externaliserande beteende samvarierar i barndomen, trots att nya källor till variation i båda måtten tillkommer så ökar samvariationen sedan succesivt upp till ung vuxen ålder.

LIST OF PUBLICATIONS

I. Kuja-Halkola R, Pawitan Y, D'Onofrio BM, Långström N, Lichtenstein P
Advancing paternal age and offspring violent offending: A sibling-comparison study
Development and Psychopathology, 2012, 24(3), 739-753

II. Kuja-Halkola R, D'Onofrio BM, Iliadou AN, Långström N, Lichtenstein P
Prenatal smoking exposure and offspring stress coping in late adolescence: no causal link
International Journal of Epidemiology, 2010, 39, 1531-1540

III. Kuja-Halkola R, D'Onofrio BM, Larsson H, Lichtenstein P
Maternal smoking during pregnancy and adverse outcomes in offspring: What are the underlying mechanisms?
Manuscript

IV. Kuja-Halkola R, Larsson H, D'Onofrio BM, Lichtenstein P
The mechanisms behind the co-development of ADHD and externalizing behavior from childhood to adulthood
Manuscript

CONTENTS

LIST OF ABBREVIATIONS

A	Additive genetic component
ACE	Twin model of additive genetic, shared and non-shared environment
ADHD	Attention-deficit/hyperactivity disorder
AP	Attention problems scale
C	Shared environmental component (sibling/twin)
CBCL	Child Behavior Checklist
DAG	Directed acyclic graph
DNA	Deoxyribonucleic acid
DZ	Dizygotic, fraternal, non-identical, twin
E	Non-shared, unique, environmental component / residual variance
Ext	Externalizing behavior scale
GEE	Generalized estimating equation
GLM	Generalized linear model
GLMM	Generalized linear mixed model
MSE	Mean squared error
MZ	Monozygotic, identical, twin
RNA / mRNA	Ribonucleic acid / messenger ribonucleic acid
SEM	Structural equation model
SDP	Maternal smoking during pregnancy
TCHAD	Swedish Twin Study of Child and Adolescent Development

1 INTRODUCTION

Cognitive and externalizing problems are obviously of great importance for the individuals, as well as for the society. Thus, reductions of consequences of such traits are imperative, and identifying risks is a first step to finding modifiable factors suitable for interventions and preventions. Standard epidemiological approaches are often used to investigate associations between exposures and outcomes, i.e. to identify potential risk factors. However, mere associations do not mean causality – a well-known fact in epidemiological research.[1] A large problem is that individuals are not randomly assigned to exposures; rather, they often select themselves to specific exposures (e.g., some people are more prone to smoke). Thus, even though a specific exposure and a specific outcome may be associated, the exposure might not cause the outcome. A common source for such selection is familial background (including both genetic and familial environmental factors), consequently methodology oblivious of such familial factors may produce biased inferences;[2] a situation henceforth referred to as familial confounding. In the current thesis I have utilized relatives to identify mechanisms underlying associations and therefore increased the understanding of causality, and thus, hopefully, provided novel insights in development of cognitive and externalizing problems.

An example of familial confounding is if the exposure stems from a trait of an individual (or her relative) and the outcome also is a trait of said individual. This might cause spurious associations, for example if the exposure and outcome share the same underlying cause. As an example, consider maternal smoking while pregnant and criminality in offspring.[3] Suppose there are factors making a mother more prone to smoking which are related to criminality, e.g. impulsivity. The mother may pass some of these factors to her offspring, thus increasing the offspring's liability to smoke and, importantly, to commit criminal acts. If the factor (e.g., impulsivity) is under genetic influence then a way for mother to pass the factor to her offspring is by genetic inheritance. However, it may also be passed down by environmental factors, e.g. if the offspring learns a behavior from her mother. The association in these cases is confounded by familial factors; the association between being subjected to maternal smoking *in utero* and criminal offending is present, regardless if there is a causal effect of smoking *per se* or not.

Methods for analyzing family data exists (e.g., between-within method,[4] further described below), but are sometimes not well developed. First, many different approaches exist to reach the same conclusions, often under different names, thus systematic handling of analyses of family data is desirable for broader understanding and comparisons of results. Second, twin analyses have been the focus of much development in the behavioral genetic field while models for family data have not been so well investigated. Thus, methodology for estimation of genetic and environmental sources for association using family data are lacking for many types of analytic problems.

Therefore, in this thesis I have developed, and used, different methodologies to address familial confounding in the study of risk factors for developing cognitive and externalizing problems. In study I-III I have used family data to address whether the associations between risk factors and outcomes are consistent with causal explanations or whether the associations are better explained by familial confounding. Additionally, a different approach has been used in study II-IV, where the genetic and environmental sources of associations have been estimated, using both twin and family data.

1.1 EXTERNALIZING PROBLEMS

Consequences of externalizing behavior may be detrimental. For example, the Word Health Organization has declared that the externalizing behavior violence is a leading worldwide public health problem, and 28.8 deaths per 100,000 person years occurred world-wide because of violence in 2002.[5] The monetary cost is counted in billions of US dollars each year,[6] and the burden of personal suffering is inestimable.

In the fifth edition of the Diagnosis and Statistical Manual of Mental Disorders (DSM-5) one cluster of disorders is called externalizing disorders and include disruptive, impulse control, conduct, attention-deficit/hyperactivity (ADHD), substance-related, and addictive disorders.[7] The concept of externalizing behavior includes these disorders, but also a wider attention to normal-variation versions of these, and similar, traits. This broad definition of externalizing behavior, including anti-social behavior, substance abuse, and disinhibitory personality, has sometimes been referred to as the "externalizing spectrum".[8,9] Indeed it has been shown that these traits share an underlying (i.e., externalizing) factor.[8] In the current thesis the concept of externalizing problems is indeed wide, it includes serious traits such as criminality, violence, drug abuse, and aggression, as well as disinhibitory traits such as rule-breaking behavior, temper tantrums and lying.

1.2 COGNITIVE PROBLEMS

Similarly to externalizing problems, cognitive problems are costly, both because of personal suffering and in monetary terms. It has been estimated that learning and developmental disabilities accounts for some 10% of the global burden of disease.[10] Further, intellectual disabilities has a world-wide estimated prevalence of 10.4 in 1000 individuals.[11] The numbers refer to a variety of disorders, ranging from mild mental retardation to down syndrome, in the current thesis, however, the focus is on more normal variation of cognitive traits; stress coping, general cognitive ability, and low academic achievement.

1.3 IDENTIFYING RISK FACTORS

Although a wide variety of risk factors have been identified for development of cognitive and externalizing behavior, many have not been properly assessed for causality. Among these risk factors are those which cannot be manipulated by researchers in randomized trials, often due to ethical considerations (e.g., assign mothers to smoke, or not smoke, while pregnant). In these situations the researcher has to rely on observational studies; this presents some problems since exposure level may

not be regarded as randomly distributed in the population. Causality is important if, for example, the goal is to introduce interventional efforts to avoid unwanted outcomes. If it is possible to intervene on causal agents, rather than their correlates, the possibility to prevent negative outcomes is improved. Indeed many interventional efforts to reduce anti-social behavior has had limited success in the past, and some interventions have even had harmful effects.[12]

In addition to observing that the associations are complex, and that there potentially are important confounding factors, the notion of traits "running in families" may be added. Familial recurrence risk (familial aggregation) is the risk in relatives to individuals with a specific trait (or disease) to also have the trait. For many phenotypes and diseases familial recurrence risks are significantly greater than the risk in the general population. Some examples of studies estimating recurrence risks in relatives are violent criminality,[13] bipolar disorder,[14] and schizophrenia.[15] Another way to quantify this association is to estimate a correlation between relatives in families; Pearson correlation for continuous variables, tetrachoric correlation for binary variables, or polychoric for ordered categorical variables. In the current thesis, rather than estimating recurrence risks, estimates of sources of familial similarity are presented (based on correlation between relatives).

1.3.1 Measured confounders

The standard approach to adjust for confounding is by including a measure of the specific confounding variable into analyses and adjust for it statistically. Examples in the current thesis are socio-economic status in the analysis of paternal age and offspring violent criminality, maternal age when pregnant in the analyses of maternal smoking during pregnancy (SDP) and cognitive and externalizing outcomes, and gender in the analysis of ADHD and externalizing behavior. In these situations a clear idea of what the confounding variable is exists, as well as a readily available measure of it.

1.3.2 Unmeasured confounders

Although adjustment of a measured confounding factor generally is preferable, it is not always feasible. To try and get around this problem and take into account potential confounders shared between relatives, an approach where relatives are compared is applied in this thesis. Indeed, usage of this approach has seen a recent increase in popularity, and it has a potential to address situations where confounding variables are unavailable, or even when it is uncertain which the confounding variables are.

As an example, consider the exposure "parental height" and the outcome "offspring weight". The association will definitely be observed, however the characterization of it as an exposure-outcome association might be faulty. Probably the situation is better described by assuming that there exist factors, shared to some extent between relatives (i.e., parents and offspring), that are responsible for the association.

1.3.2.1 Genes

A common situation is that of "genetic confounding". Generally, in this thesis, this refers to passive gene-environment correlation, further described below.

To continue the example of parental height and offspring weight above; genes affecting general growth would be excellent candidates for the common cause of the two phenotypes. Offspring would inherit these genes from her parents, influencing her weight to be at a certain level. If measures of these growth genes are unavailable, an analysis utilizing relatives could help to shed light on the mechanisms.

1.3.2.2 Environments

Another possible way that familial confounding might work is through sharing of environments. Generally the environment thought of is childhood-environment.

An example; parental rearing regimes might cause offspring in a family to be similar with regards to an outcome, e.g. academic achievement, if the rearing regime also is associated with an exposure, e.g. parental age, an association may be introduced.

1.3.2.3 Adjust or estimate?

Depending on the research question at hand, interest might be in family-adjusted models (i.e., controlling for familial confounding), quantitative genetic models (i.e., estimating the degree of correlation that is due to genetic and environmental effects, also known as biometric models), or both. In the current thesis both family-adjusted models and quantitative genetic models are provided for exposure-outcome associations. The general approach employed in this thesis is to find family-adjusted estimates, and then perform quantitative genetic analyses. The quantitative genetic estimates should then be interpreted in light of the family-adjusted results; when family-adjusted estimates suggest total familial confounding of an association the quantitative genetic models may provide an estimation of the source of confounding. When familial confounding is not explaining the entire association interpretation of quantitative genetic estimates should be made more carefully.

1.3.3 Development is not stable

Although longitudinal analyses of a childhood phenotype's effect on an outcome in, say, late adolescence indeed may be informative, it may not capture developmentally important processes. For example, one might hypothesize that ADHD is predictive of externalizing behaviors in early adolescence, but not in late adolescence, if properly controlled for previous existing associations. A result confirming this hypothesis would imply that interventions aimed at externalizing behaviors might have more to gain from being informed about ADHD in early adolescence than in late. Therefore, employing methods which are sensitive to these types of questions is imperative in developmental studies.

1.4 FAMILIAL CONFOUNDING

1.4.1 Familial risk

A surprisingly large proportion of studies comparing exposure-discordant relatives observe familial confounding, rendering associations congruent with non-causal explanations. Some examples are early alcohol debut and alcoholism,[16] birth weight and cardiovascular disease,[17] and maternal alcohol use disorder and offspring ADHD.[18] Even in associations where the results are not congruent with non-causal associations it is often found that familial confounding may account for some proportion of the association.

It is not a coincidence that much research using extended families to investigate familial confounding uses data from Sweden (or the other Nordic countries). Sweden has a long history of keeping registers of the population, and since 1994 Statistics Sweden has had centrally kept registers connecting every Swede to her parents. From 2000 this register is known as the Multi-Generation Register and includes all people registered in Sweden any time since 1961.[19] This register allows researchers to study a wide variety of relatives; siblings, cousins, offspring-parents, offspring-aunt/uncle, grandchildren-grandparents, etc. By estimating, and comparing, associations between different relatives the sources of familial confounding may be examined.

1.4.2 Methodology in familial confounding

Although this great data source is available, the methodology for analyses is somewhat lacking in coherence. Using the same data many different methods of analysis may be applied; e.g., adjusting for measured traits in relatives, calculating recurrence risk in relatives, estimating correlations in relatives, adjusting for factors assumed shared between relatives, and using quantitative genetic methods to estimate univariate or multivariate statistics. The analyses make different assumptions, and some designs may be used in combination.[2] The choice of method of analysis depends, at least in part, on traditions in different areas of research.

In estimation of familial recurrence risks all types of relatives are traditionally used. While, in quantitative genetic analyses (e.g., estimating heritabilities), there is a long tradition of using twins. A reason for this tradition is that analyzing twins has some advantages; however, there is no general reason why other types of relatives should not be used for quantitative genetic modeling. This thesis includes both twin analyses and analyses using other types of relatives, and tries to systematize the usage of extended families in quantitative genetic modeling. Studies using extended families in a similar, albeit not equivalent, fashion as in the current thesis have named the designs differently; e.g., the stealth design,[20] the cascade design,[21] extended family design,[22] children of twins design,[23] and children of siblings design.[24]

2 AIMS

The aims of the work included in this thesis were twofold; (1) To gain insight into development of cognitive and externalizing problems by using genetically informed samples. (2) To apply existing and develop new methodological approaches which allowed answering questions regarding causality and etiology of the associations. The four studies aimed at answering the following specific questions:

Study I: Is advancing paternal age associated with offspring violent criminality? If so, does the association persist when adjusting for factors shared between siblings?

Study II: Is exposure to maternal smoking *in utero* associated with worse stress coping ability? Does this potential effect persist when offspring of mothers discordant for smoking between pregnancies are compared? What is the genetic and environmental contribution to the association?

Study III: Does maternal smoking during pregnancy predict cognitive and externalizing problem outcomes if adjusted for familial effects? Do the results differ for birth-related outcomes? What are the genetic and environmental contributions to the associations?

Study IV: How does ADHD and externalizing behavior co-develop from childhood into adulthood? Is there a trend of one trait predicting the other, and how do genetic and environmental contributions affect the comorbidity?

3 MATERIAL

3.1 SWEDISH POPULATION REGISTER DATA

Studies I-III is based on nationwide Swedish registers. The use of unique civic registration number (*personnummer*) assigned to every Swede makes it possible to link the registers. Study I and II utilizes a linkage up until 2005, while study III utilizes an extended linkage until 2010.

The *Total Population Register* is held by the Swedish Tax Agency.[25] Among other variables the register contains gender, place of birth, immigration dates, emigration dates, and death dates.

The *Multi-Generation Register* contains a linkage between offspring and parents and is held by Statistics Sweden (*Statistiska centralbyrån; SCB*).[19] Using this register, it is possible to find who is related to whom. For example, if two individuals share the same mother and father it can be deduced that they are full siblings or twins. If they share father, but does not have the same mother, they are paternal half siblings. Similarly, multiple generations may be used to find any degree of relatives, only limited by the coverage of the registers. In the current thesis first-, second-, and third-degree of relatives are used, such as siblings, offspring, parents, uncles, aunts, cousins and grandparents. The register covers all Swedes born 1932 and later, who were alive and living in Sweden in 1947, and linkage to parents is almost complete for all individuals born in Sweden since 1947. The coverage is slightly worse for linkage to fathers than to mothers (98% for fathers, and 100% for mothers, for offspring born 1960 or later) and even worse for immigrants (e.g., around 50% for children born in the 1980's).[19]

The *Swedish Twin Register* was established in the 1950's and is held at the Department of Medical Epidemiology and Biostatistics, Karolinska Institutet. Data have been collected in multiple waves of questionnaires. It currently has over 75,000 pairs where zygosity has been determined.[26] In study III the Swedish Twin Register was used to identify monozygotic (MZ; identical) and dizygotic (DZ; non-identical) twin pairs who were mothers. In study IV a subset of the twin population was used; all twins born between May 1985 and December 1986 as collected in the TCHAD study (described below).[27]

The *Medical Birth Register* is held at the National Board of Health and Welfare (*Socialstyrelsen*).[28] The register contains detailed information on perinatal variables such as maternal smoking during pregnancy, birth weight, gestational age, and congenital malformations. This register contains close to all births in Sweden, only approximately 1-2% are missing.[28] Most variables have acceptable low levels of missingness (e.g., smoking in early pregnancy where 4-9% is missing) while some have an unacceptable missing proportion (e.g., smoking during weeks 30-32 where 67-89% is missing during 1990-1998).

The *Crime Register* contains all convictions in Sweden, and is held at the National Council for Crime Prevention (*Brå*). Recorded offenses pertain to sentences in lower

court where individuals were found guilty of the offence, whether issued prison sentence, probation, or fined. Fines issued directly by police are not included in the Crime Register. The age of criminal responsibility in Sweden is 15; hence only sentences after this age exist.

In studies II and III data from the *Military Conscription Register* is used. Military conscription was mandatory for men until 2007 and it has been estimated that more than 95% of young males attended conscription in the 1990's in Sweden.[29]

Diagnoses from all inpatient care instances requiring over-night hospitalization are found in the *Patient Register*, held at the national Board of Health and Welfare. From 2001 admissions in some outpatient settings are included.

Swedish authorities performed *National Population and Housing Censuses* every fifth year between 1960 and 1990, the registers are held at Statistics Sweden. The censuses aimed at capturing financial and social aspects of Swedish citizens' lives. In study I and II data from these censuses has been used to assess socio-economic status during childhood, as measured by occupation and income of parents.

The *Education Register* is held by Statistics Sweden,[30] it contains the highest achieved education for every Swede.

The *Grade-9 Register* is held by Statistics Sweden and contains grades for Swedes, who has finished the compulsory first 9 years of school, at approximately 15 years of age.

3.2 SWEDISH TWIN STUDY OF CHILD AND ADOLESCENT DEVELOPMENT

Study IV uses a series of questionnaires sent out to a birth cohort of twins; the Swedish Twin study of Child and Adolescent Development (TCHAD).[27] All twins born between May 1985 and December 1986 were invited to participate in the study. The first invitation was sent out to the parents in 1994 when the twins were aged 8-9 year, they then received questionnaires again in early (13-14 years) and late (16-17 years) adolescence and early adulthood (19-20 years). The response rates were 75%, 73%, 74% and 78% for parents (ages 8-9, 13-14, 16-17 and 19-20) and 78%, 82% and 59% for the twins (ages 13-14, 16-17 and 19-20).[27,31]

3.3 MEASURES

In this thesis, a wide definition of cognitive and externalizing problems is used, ranging from normal-variation cognitive ability to violent criminal convictions. In study III perinatal outcomes are included as a contrast to the cognitive and externalizing problem outcomes.

Study I-III has standard epidemiological exposures, or risk factors, while study IV has a non-standard epidemiological design and considers each of two phenotypes as both

exposure and outcome. In Table 1 the exposures and outcomes are described “at a glance”.

In addition to these variables, the following covariates have been used: gender (from the Total Population Register), maternal age at childbirth (Medical Birth Register), parity (Medical Birth Register and Multi-Generation Register), birth year (Total Population Register), bipolar diagnosis (Patient Register), schizophrenia diagnosis (Patient Register), parental criminal convictions (Crime Register and Multi-Generation Register), immigrant status (Total Population Register), socio-economic status (register of National Population and Housing Censuses), parental income (register of National Population and Housing Censuses), and mother’s country of birth (Total Population Register and Multi-Generation Register).

Table 1. Main measures used in current thesis.

Measure	Exposure / Outcome	Data source	Comment	Used in study
ADHD-like traits	Exposure, Outcome	TCHAD questionnaire; CBCL scale	The Attention Problems scale	IV
Birth weight	Outcome	Medical Birth Register	Measured by midwives after delivery	III
Born small for gestational age	Outcome	Medical Birth Register	Having birth weight in the lowest 10% compared within gender and gestational day	III
Criminal convictions	Outcome	Crime Register	Any registered criminal conviction	I, III
Drug misuse	Outcome	Crime Register, Patient Register	Drug-related convictions and/or diagnoses of alcohol/drug dependence	III
Externalizing traits	Exposure, Outcome	TCHAD questionnaire; CBCL scale	The Externalizing scale	IV
General cognitive ability	Outcome	Military Conscription Register	From the Swedish National Defense's Enlistment Battery	III
Low academic achievement	Outcome	Grade-9 Register	The 10% with lowest grades and the non-completers of the compulsory first nine years of school	III
Maternal smoking during pregnancy	Exposure	Medical Birth Register	Answered at first antenatal visit (approximately week 15), used as a binary variable (No or Yes >0 cigarettes per day)	II, III
Paternal age at birth	Exposure	Total Population Register, Multi-Generation Register	Difference between birth date in father and offspring	I
Preterm birth	Outcome	Medical Birth Register	Born earlier than gestational day 259	III
Stress coping	Outcome	Military Conscription Register	Ability to cope with stress during wartime	II
Violent criminal convictions	Outcome	Crime Register	Homicide, assault, robbery, threats and violence against an officer, gross violation of a person's/woman's integrity, unlawful coercion, unlawful threats, kidnapping, illegal confinement, arson, and intimidation	I, III

4 METHODS

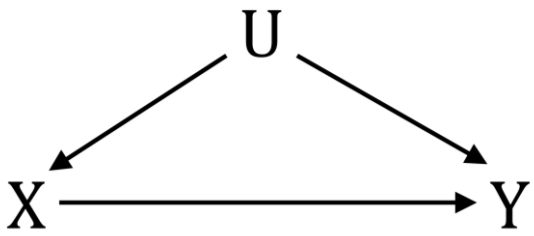
All studies in this thesis are observational. In observational studies the researcher does not intervene on who gets an exposure, however they may investigate causes for getting the exposure, as well as estimating the exposure-outcome association. It is possible that the causes for being exposed also are causes for having certain levels of the outcome, or are associated with causes for the outcome. This situation is discussed below under the phrase "confounding".

In epidemiological terminology the current studies are *cohort*-studies; i.e., a population is assessed for the exposure and then followed over time to observe levels of the outcome.[1] All data in the registers have been collected prospectively.

4.1 CONFOUNDING

Confounding of an association may be defined as when there is a common cause of exposure and outcome, though this is a slightly simplified definition.[1] The definition is easy to grasp, and is applicable in a wide range of scenarios. In Figure 1 this is shown in a Directed Acyclic Graph (DAG).[32]

Figure 1. A confounder, U, has a causal effect both on exposure, X, and outcome, Y.



Box 1. *A note on DAGs*

In a DAG the direction of causation is encoded as arrows; X→Y reads "X may cause Y". A causal path between two variables is the path from one variable to the other, always following the direction of the arrows. A DAG may not be cyclic, i.e. it is not allowed to have causal paths from one variable and back to itself. A path between two variables is *open* (i.e., an association between the two variables may exist) if you can trace from one variable to the other, through arrows and variables, without passing any place where two arrow heads meet, called *inverted fork* (e.g., →X←). Statistically adjusting for a variable in an open path will close it, adjusting for a variable in an inverted fork will open any path through it.

In Figure 1 the association between the exposure, X, and outcome, Y, is confounded by the variable U. Two examples of this situation from this thesis are

- The exposure paternal age, the outcome violent crime, and the confounder socio-economic status.

- The exposure maternal smoking during pregnancy, the outcome low academic achievement, and the confounder "genes for impulsivity".

In the first example a measure of the confounder can relatively easy be found. In the second example the confounder could theoretically be measured, however doing so would require prior knowledge of the genetic loci as well as difficult and expensive genotyping.

4.2 ADDRESSING MEASURED CONFOUNDING

In the situation when an exposure-outcome association is confounded by a variable of which a measure is available, it is possible to adjust the association for this variable. Standard methods for this statistical adjustment are stratification (i.e., estimating the association separately for different levels of the confounder) and regression model adjustment (used to find estimates averaged over the different levels of the confounder). In this thesis the general adjustment method for observed potential confounders has been regression model adjustment using Generalized Linear Models (GLM).

4.3 THE GENERALIZED LINEAR MODELS FRAMEWORK

The GLM provides a widely usable approach for estimation of associations and statistical adjustment for other covariates (which may be confounders). In study I-III GLM has been used to find ordinary epidemiological estimates of studied associations (i.e., linear regression coefficients, odds ratios, risk ratios [ratios of probabilities], and incidence rate ratios). In a GLM an outcome is associated with covariates (including the exposure) through a link function, this function determines on which scale inferences are made. A general way of writing this is

$$g\big(E(Y_i|\boldsymbol{x}_i)\big) = \boldsymbol{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots, \quad (1)$$

where $g(\cdot)$ is the link function, $E(Y_i|\boldsymbol{x}_i)$ is the expected value of the outcome, Y_i, for observation i, given the vector of covariates $\boldsymbol{x}_i$, and $\boldsymbol{\beta}$ is a vector of regression coefficients. Examples of link functions used in present thesis are unity (ordinary linear regression, for continuous outcomes), logistic (logistic regression for binary outcomes, estimating odds ratios), complementary log-log (for binary outcomes, suits well with time-at-risk analyses), and log (for count outcome, used in truncated Poisson regression).

An assumption in the GLM is that of independent observations of the outcome, given the covariates included. This assumption may be violated when using family or twin data, where relatives potentially are correlated in many variables, including the exposure and outcome. One of the consequences of violating this assumption is that the standard errors of regression coefficients may be under-estimated. To solve this problem Generalized Linear Mixed Models (GLMM) may be used. In GLMM the GLM has been expanded to further include random effects, $\boldsymbol{b}$;

$$g\left(E\left(Y_{ij}|\boldsymbol{x}_{ij},\boldsymbol{z}_{ij}\right)\right)=\boldsymbol{x}_{ij}^{T}\boldsymbol{\beta}+\boldsymbol{z}_{ij}^{T}\boldsymbol{b}_{ij}, \quad (2)$$

where Y_{ij} is the observation of subject j in cluster i, and $\boldsymbol{z}_{ij}$ is a design vector indicating presence of random effects. The random effect may potentially come from any distribution; it is, however, common to assume the random effect to be normally distributed with mean zero. To adjust for correlation of outcome in clusters (twin pairs or families) it is common to include a random effect shared between subjects in the cluster ($b_i \sim N(0,\sigma_b^2)$);

$$g\left(E\left(Y_{ij}|\boldsymbol{x}_{ij},\boldsymbol{z}_{ij}\right)\right)=\boldsymbol{x}_{ij}^{T}\boldsymbol{\beta}+b_i=\beta_0+\beta_1 x_{ij1}+\beta_2 x_{ij2}+\cdots+b_i. \quad (3)$$

In study II, and to some extent in study III, GLMM has been used to get valid standard errors for the regression estimates.

Another approach to try and get unbiased standard errors of regression estimates in the situation of clustered data is Generalized Estimating Equations (GEE). In GEE a function for the mean is specified, as is a function for the covariance within clusters. This may be expressed as

$$\begin{aligned} E(\boldsymbol{Y}_i|\boldsymbol{x}_i)&=\boldsymbol{\mu}_i(\boldsymbol{\beta}),\\ Var(\boldsymbol{Y}_i|\boldsymbol{x}_i)&=\boldsymbol{V}_i(\boldsymbol{\beta}), \end{aligned} \quad (4)$$

where $\boldsymbol{Y}_i$ is the vector of outcomes in cluster i, $\boldsymbol{x}_i$ is the matrix of covariates for all subjects in the cluster, and $\boldsymbol{\mu}_i(\boldsymbol{\beta})$ and $\boldsymbol{V}_i(\boldsymbol{\beta})$ are functions for the mean and variance, respectively. The functions may be taken from a general distribution (e.g., in study I, binary with cloglog link, and truncated Poisson with log link), and the variance function may incorporate correlation between subjects in a cluster. The functions do not, however, need to represent a specific distribution; e.g., the mean and variance may be specified as if from different distributions. To find the estimates of $\boldsymbol{\beta}$ the following estimating equation is solved:

$$\sum_{i}^{n}\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_i)=0, \quad (5)$$

where $\boldsymbol{y}_i$ is the observed $\boldsymbol{Y}_i$. The advantage is that each cluster is considered together, and a post-estimation correction of the standard errors may be done using the so-called sandwich formula, or robust variance formula.[33] This formula adjusts the standard errors produced from model fitting for misspecification of the model, hence they are *robust* to these potential misspecifications.[33] In study I GEE has been used to get valid standard errors.

Both these methods (GLMM and GEE) will, as noted above, potentially provide unbiased estimates of the standard errors of the fixed effect regression parameters $\boldsymbol{\beta}$. The parameter estimates may also change somewhat in the estimation procedure compared to a GLM estimate.

A way to assess precision (i.e., significance) of parameter estimates without using standard errors is by profile likelihood intervals.[33] Suppose the parameter θ is to be assessed for precision. The idea is that parameters in the model, other than θ, might influence the precision of the estimate of θ. These other parameters may include, e.g., variance of shared effects, making twins in a pair similar. A profile likelihood-interval estimates all parameters except θ freely for different values of θ. By finding the values for θ which produces significant reductions in likelihood a profile likelihood interval may be obtained. This likelihood interval is unbiased of other parameters modeled, e.g. variance shared. In study IV profile likelihood intervals have been used to assess significance of parameter estimates.

In the situation when the clustering variable (e.g., twin pair or family) is associated with the exposure as well as the outcome, the regression estimates obtained using any of the above described methods will be biased.[34] Another way of stating the situation is to say that there are (unmeasured) confounders, which potentially are shared between members of a twin pair or a family. To adjust for these confounding variables other techniques are required.

4.4 ADDRESSING UNMEASURED CONFOUNDING

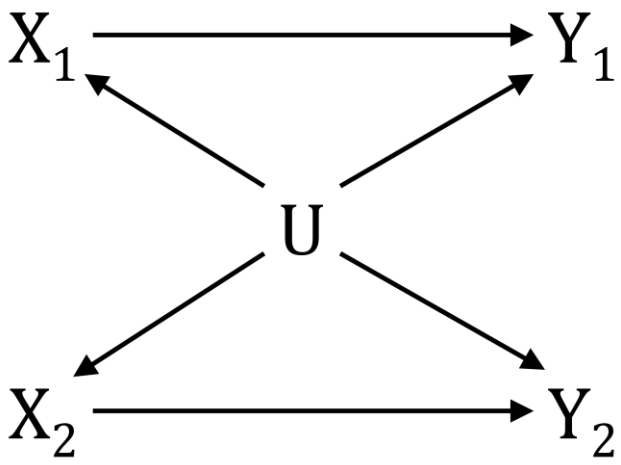
Unmeasured variables inducing a spurious, i.e. non-causal, association between two variables can lead to incorrect inferences. A simple case is that of "genetic confounding"; a set of genes might cause a level of a trait interpreted as an exposure (e.g., ADHD) as well as an outcome (e.g., criminality). In an analysis unadjusted for this set of genes an association would be observed and, in the case where the researcher does not consider any alternative explanations, may be interpreted as a causal effect of the exposure on the outcome even if it is not. The first step is of course to acknowledge that confounders might be present, and the second step to adjust for the confounders. This can be problematic, however, since it might not be possible to measure the confounding variable (e.g., current smoking status of a large cohort) or specifically known what the confounding variable actually is (e.g., the contributing loci are not identified).

When the confounder is shared within the twin pair or family, as is the case with genetic alleles (however shared to different degrees depending on the relation between individuals assessed), a possibility for better inferences arise. Relatives may be used to (1) get a less biased estimate and/or (2) estimate genetic and environmental contributions.

4.5 WITHIN-FAMILY ESTIMATES – ADJUSTING FOR UNMEASURED CONFOUNDING

A key feature in study I-III is the use of family-adjusted estimates; within-cousin, and within-sibling estimates. The basic notion is that there are unmeasured confounders of the association studied, and that these are shared between subjects in a family. In Figure 2 a DAG representing a simple situation of familial confounding for two siblings is shown. Here sub-index 1 represent sibling 1 and sub-index 2 represent sibling 2. In this situation estimation of the association between X and Y, without adjustment for U, produces a biased estimate. If U is an unmeasured confounder within-family adjustments may be used. A, perhaps, common misconception is that GLMM with a random effect representing the family belonging would produce estimates adjusted for associations within family members. Although that is not the case, there are other possibilities to adjust for such confounders. In the simple situation depicted in Figure 2, adjustment of confounder U without actually measuring it may be done in several ways, two of which are described below.[34]

Figure 2. A shared confounder, U, influences both exposure, X, and outcome, Y, for two subjects in a family.



4.5.1 Adjustment by cluster mean of exposure

A mean value of the exposure within a cluster (e.g., a family, or twin-pair) is included in the regression model. In a simple linear model with unity link, no other covariates than the exposure, and two observations per cluster it looks as (let $\bar{x}_i = (x_{i1} + x_{i2})/2$)

$$\begin{aligned} E(Y_i|x_{i1}, x_{i2}) &= \beta_0 + \beta_W x_{i1} + \beta_B \bar{x}_i \text{ , or} \\ E(Y_i|x_{i1}, x_{i2}) &= \beta_0 + \beta_W (x_{i1} - \bar{x}_i) + \beta_B^* \bar{x}_i. \end{aligned} \quad (6)$$

Both model specification produces the same estimate of β_W, but estimates of β_B and β_B^* are different. This approach is often labelled the "between-within" method, since it produces a between-cluster estimate (regression coefficient from the cluster mean; β_B or β_B^*) as well as a within-cluster estimate (regression coefficient from the exposure, adjusted for the cluster mean; β_W). In study I, II and III this method has been used.

4.5.2 Adjustment using conditional likelihood

In conditional regressions, regressions based on the conditional likelihood, the aim is to estimate the same parameter of interest as the within-parameter in the between-within model. These conditional likelihood techniques, however, generally make less distributional assumptions, and produce no between-cluster estimates. In this scenario the conditioning of likelihood is made on the sufficient statistic for cluster-specific intercepts,[34] e.g. the intercept for a cluster (cluster number *i*) can be found if $\sum_j y_{ij}$ is known. In study III this method has been used.

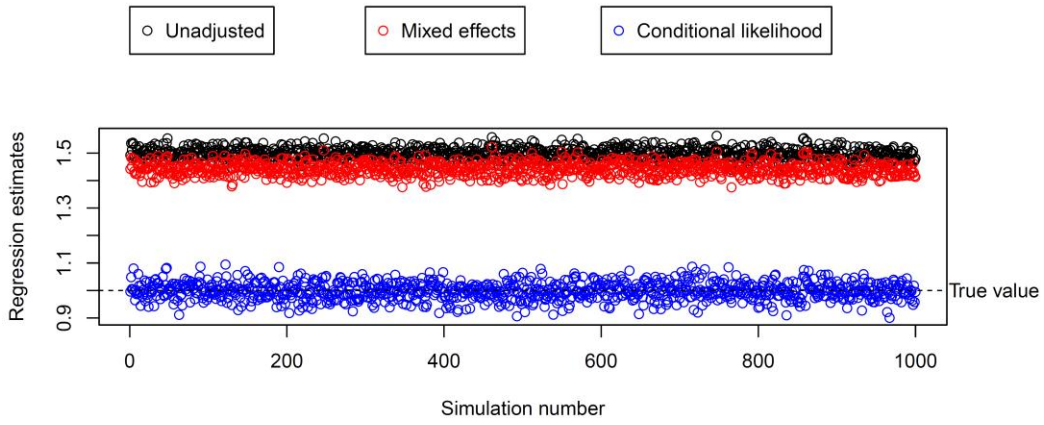
4.5.3 Example of within-pair estimation

To exemplify this, and re-iterate that inclusion of random effects shared within a cluster, in the situation described for Figure 2, does not yield unbiased estimates; Assume that Figure 2 represents a linear model where each arrow represent a regression coefficient with value 1. Further, assume that U is a random variable, continuous and normally distributed with mean 0 and variance 1, that X is the sum of U and a normally distributed random error (mean = 0 and variance = 1), and that Y is the sum of X and U and has a normally distributed random error term (mean = 0 and variance = 1):

$$\begin{aligned} X_{ij} &= 1 \cdot U_i + \varepsilon_{Xij}, \\ Y_{ij} &= 1 \cdot X_{ij} + 1 \cdot U_i + \varepsilon_{ij}, \\ \text{where } i &= 1, \dots, N;\ \text{j} = 1,2; U_i \sim N(0,1); \varepsilon_{Xij} \sim N(0,1); \text{and } \varepsilon_{ij} \sim N(0,1). \end{aligned} \tag{7}$$

One way to phrase this situation is that the random effect, U, causing Y, is also causing a predictor, X. I simulated data from this set-up for 1000 pairs, and I did it 1000 times, each time calculating the unadjusted estimate, the mixed effects estimate, and the conditional estimate. The result of the simulation is shown in Figure 3, and it is clear

Figure 3. Results from simulation.



that both the unadjusted and the mixed effects estimates are biased while the conditional estimates are close to 1, i.e. the regression value from which the data was simulated. To put a number on the bias in each approach I estimated the mean squared error (MSE = $\frac{1}{n}\sum\left(\hat{\beta} - \beta_{true}\right)^2$; n = number of samples, $\hat{\beta}$ = estimate, β_{true} = true

parameter value) of the estimates for the three approaches. The unadjusted approach had MSE = 0.249, the mixed effects approach had MSE = 0.195, and the conditional approach had MSE = 0.001. In the specific example of linear regression, approach 1 above, including pair-specific mean of exposure, produces the same estimate as the conditional regression approach. However in a general GLM setting this is not true, and approach 1 has been named the "poor man's" approximation of approach 2.[34]

This example is of course a simplification of reality. First, there are possibly many confounders of the association shared between family members. Second, the sharing between family members is probably not 100% for all confounders. Third, this example disregards other sources of bias, such as measurement error in the exposure and correlation between non-confounder variables which affects the exposure.

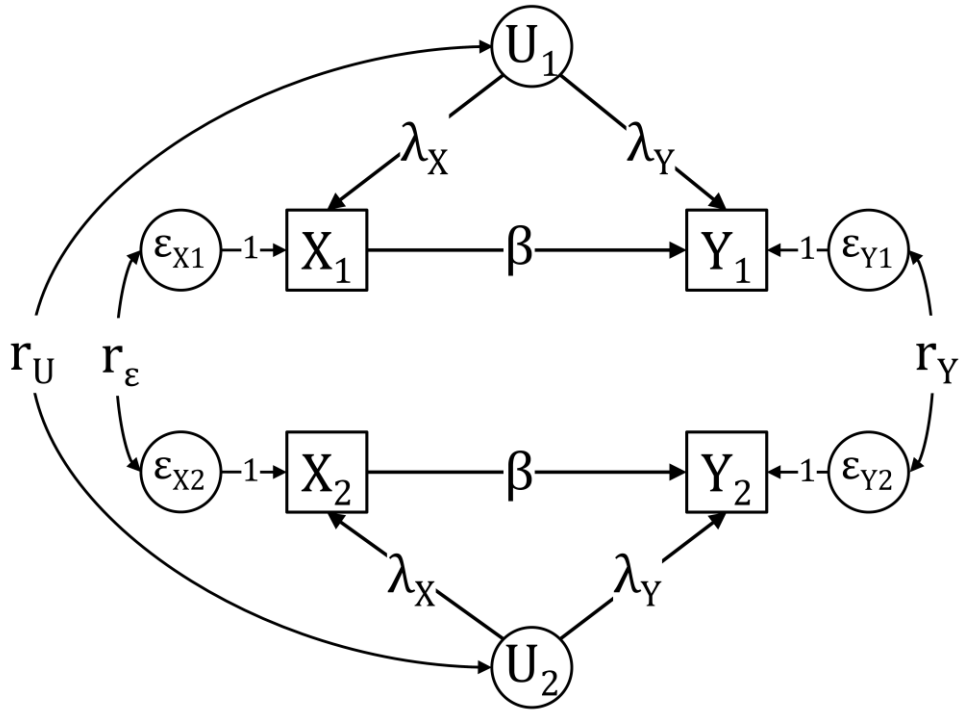
Box 2. *A note on path diagrams*

The path diagram[35] is a representation of a set of equations; each single headed arrow represents a regression, and each double headed arrow represents a correlation/covariance (including variances). The boxes represent measured variables, while the circles represent non-observed, latent, variables. The covariance/correlation between two variables can be deduced: Trace against the direction of an arrow and then with the arrow, or simply with the arrow, from one variable to another. Only pass through each variable once in each traced path, trace through at most one double headed arrow in each path. Multiply all coefficients traced through to get the covariance/correlation.

In Figure 4 a slightly more complex situation is shown as a path diagram, for simplicity all variances are assumed to be 1. The X:s are exposures, the Y:s outcomes, the U:s are confounding variables, and the ε_X:s are variables potentially making the exposure similar, and the ε_Y:s are variables potentially making the outcomes similar. The correlations, r_U, r_ε and r_Y, determine how similar U:s, X:s, and Y:s are.

This model is, of course, also a simplification of reality. The model assumes that there is only one confounder, and one source of variance to the exposure that is not a confounder. What happens if more potential sources of variance and covariance are included? To assess this, probably, more realistic scenario I simulated data as in Figure 4, only I allowed for 100 U-pairs and 100 X-pairs. I let the correlation between U, r_U, be a random variable uniformly distributed in the range [0.5,1] and the correlation between random variables ε_X, r_ε, be in the range [0,0.5]. I further let λ_X and λ_Y take on random values from a normal distribution with mean 1 and variance 1, and r_Y was set to 0. The parameter of interest was β (still equal to 1), and I simulated 1000 data sets

Figure 4. Path diagram representing familial confounding in a pair of family members.



of 1000 pairs each. The observed MSEs favored the conditional approach; the MSE was 0.114 for the unadjusted approach, 0.071 for the mixed effects approach, and 0.041 for the conditional approach.

Even in this semi-complex scenario at least one major issue is left out; measurement error in the exposure. It is clear that measurement error in the exposure will bias the regular estimate of any analysis toward a null effect. Unfortunately the bias in the within-estimates is even greater.[36,37]

4.5.4 Theoretical bias

The within-family estimate, if estimated as above described, will indeed be closer to the true, causal, effect under certain conditions. The change in bias may be expressed as a function of r_U, r_ε, λ_X, and λ_Y and the variance of U (σ_U^2) and ε_X (σ_ε^2):[37]

$$bias(\beta_W) = \frac{\lambda_Y \lambda_X \sigma_U^2}{\lambda_X \sigma_U^2 + \sigma_\varepsilon^2 \frac{1 - r_\varepsilon}{1 - r_U}}. \quad (8)$$

For the bias in an ordinary estimate, substitute r_U and r_ε for 0 (making the bias indifferent to correlation in variables between subjects in the pair). Thus, when the correlation r_U between twins in a pair is greater than the correlation r_ε, the within-estimate is less biased than an ordinary estimate. If the correlation r_U equals 1, and no other confounders or random errors exist, the within-estimate is unbiased. It should be

noted that generally no empirical estimates of r_U and r_ε can be found since the main reason for performing within-sibling analysis is absence of measurements of U.

In the situation when siblings are compared, and the exposure may be considered a trait of a parent (e.g., SDP), the r_U is a *within-individual* correlation. In this sense the siblings, from the parent's point of view, may be considered repeated measurements (for the mother in the case of SDP). Thus this analysis, although here called within-sibling analysis, have the potential to be at least as "powerful" as within-twin-pair analyses on MZ twins.

4.6 QUANTITATIVE GENETIC MODELS – ESTIMATING UNMEASURED CONFOUNDING

Study II-IV uses relation between study subjects to estimate quantitative genetic measures, as well as environmental measures that are assumed to be shared or not shared between different types of relatives. The basis of these models has been developed within, e.g., the behavioral genetic scientific community,[38-40] and are based on work, starting in the early 20th century with Ronald A. Fisher[41] and Sewall Wright.[42] In the behavioral genetics field much focus has been on the usage of twins to estimate heritabilities, genetic correlations, and testing causal inferences.

The basis of performing analyses using relationships, rather than measured molecular genetic variation, is the assumption of multiple genetic loci having an effect on the measured value of a trait. In combination with the central limit theorem, and some assumptions discussed below, estimates of genetic contribution to the variation of a trait in a population may be calculated. This proportion is referred to as the heritability. This concept may also be expanded to the proportion of covariance between measures explained by genetic contribution, leading to multivariate quantitative genetic models and concepts such as gene-environment correlation.

4.6.1 The human genome - what is a gene?

Box 3. *Some words used in genetics*

Allele:	A variation of a gene
Dominance:	Within-locus interaction
Epistasis:	Between-locus interaction
Exons:	The protein-coding part of the genome
Gene:	A segment of DNA that carries a unit of information
Locus:	A specific region of the genome where a gene is located, plural loci
Mutation:	Change in the DNA, may be anything from change in one SNP (point mutation) to large structural changes
Pleiotropy:	When one gene has effects on more than one phenotype
SNP:	Single nucleotide polymorphism, one base in the genome

The human DNA (deoxyribonucleic acid) is found in every cell of the body (with some exceptions, such as the erythrocytes, the red blood cells), and are collected in 23 pairs

of chromosomes (22 autosomal and on pair of gender-chromosomes). The human DNA consists of approximately 3 billion base pairs; there are four different bases (adenine, thymine, guanine, and cytosine), and base-triplets (codons) code for sequences of amino acids. DNA may be transcribed to RNA (ribonucleic acid), or mRNA (messenger ribonucleic acid), in the nucleus of a cell. mRNA is then transported out of the nucleus and translated into amino acid sequences, i.e. proteins.[39]

In 1958 Crick stated the central dogma of molecular biology: DNA transcribes to RNA which produces a protein, proteins cannot carry information to RNA/DNA but RNA may carry information to DNA.[43] One historical interpretation of this dogma is that a gene is a protein-coding sequence of DNA. This interpretation is, however, losing its validity with recent advances in molecular genomics, where the understanding of the function of RNA is being revised.[44] Only about 1.2% of the genome codes for proteins (exons) in roughly 20,000 loci, but it has been shown that about 80% of the genome has some biochemical function.[45] Indeed the complexity of the genomic functioning is far from being understood. Therefore, rather than thinking of genes as "recipes for proteins", a gene may be thought of as "A segment of DNA that carries a unit of information",[44] be it a protein-recipe or something else which might affect a phenotype.

4.6.2 Adding up effects at many loci – the polygenic model

According to the *polygenic model*[38,39] the phenotypic value, P, may be stated as

$$P = G + E, \tag{9}$$

where G is the genetic, and E environmental, contribution to the observed phenotypic value. This polygenic model assumes that genes' effect on a phenotypic value is caused by different alleles at a large number of loci. In other words; G is the sum of the effect of many alleles at many loci:

$$G = \sum_{i \in L} G_i \cdot l_i, \tag{10}$$

where L is the set of loci contributing to the genetic variation of effects on the phenotype, G_i is the effect of allele i, and l_i is 1 if the allele is present and 0 else. It is assumed that l_i is randomly distributed in the population. If there are no interactions between, or within, loci then G is the sum of many independent random variables, and thus the central limit theorem tells us that

$$G \sim N(\mu_G, \sigma_G^2). \tag{11}$$

A similar reasoning may be employed for the environmental contribution to the observed phenotypic value, thus

$$E \sim N(\mu_E, \sigma_E^2). \tag{12}$$

Therefore the phenotypic value has the distribution

$$P = G + E \sim N(\mu_G + \mu_E, \sigma_G^2 + \sigma_E^2) = N(\mu, \sigma_G^2 + \sigma_E^2)\,, \tag{13}$$

where G and E are assumed independent.

This is, of course, a rough simplification of reality. First, the included G represents simple linear effects of all alleles, known as additive genetic contribution (A), this is a bit simplistic. Genetic effects are not independent of each other; there are intra-locus interactions (i.e., interactions between alleles at the same locus in the two chromosomes in a pair) known as dominance effects (D), as well as between-locus interaction known as epistasis (I). Second, it is unlikely that genetic and environmental effects are independent; genes may be expressed differently under different environmental conditions, and *vice versa*. Therefore, a description of the phenotypic value of a trait that is closer to reality might be expressed as

$$P = A + D + I + E + (A + D + I) * E, \tag{14}$$

where D captures the deviation from additive effects within loci, I the deviation from additive effects from between loci interactions, and "$*$" represents interaction between effects. It may be noted that D (and I) is not the effect of the loci where dominance (epistatic) effects are present, rather it is how much the effects differ from regular additive effects at these loci, as captured in A.

Turning to estimation and usage of quantitative genetic models; the focus is on determining the proportion of variability in a population explained by each of the contributing sources to P. In statistical language: explain the variation in P. Note that the "effect of a gene" now refers to the effect of one allele compared to another allele at the same locus. Indeed if no variation exists at a locus, no variation may be exerted on the phenotype. The variance of P may be expressed as

$$\begin{aligned} Var(P) = {} & Var(A) + Var(D) + Var(I) + Var(E) + 2Cov(A, D) \\ & + 2Cov(A, I) + \cdots, \end{aligned} \tag{15}$$

where the summation continues for all covariances within genetic effects, between genetic and environmental effects, and for interaction effects. These covariances are generally assumed to be zero; for the genetic covariances the effects are defined to be independent, making the the covariance necessarily zero.[39] However, the assumption that gene-environment covariances are zero is not as robust. Further, epistatic interactions are often assumed to not be present, as are gene-by-environment

interactions. Under these assumptions, the heritability is defined as proportion variance explained by genetic factors

$$\frac{Var(A) + Var(D)}{Var(P)}. \tag{16}$$

This is often referred to as the broad sense heritability; it is however common to only include the additive effect and estimate what is known as the narrow sense heritability:

$$\frac{Var(A)}{Var(P)}. \tag{17}$$

In many applications this is the reported estimate for genetic influence on a trait, and in study III and IV this narrow sense heritability is reported, albeit not as main findings.

It is also commonplace in twin studies that the environmental effect is divided into two distinct sources; shared environment (C) and non-shared environment (E; note that this is not the same E as above). Shared and non-shared generally means between twins in a pair, as it is used in study IV. Together with the additive genetic effects, A, this is often referred to as the *ACE*-model, and is used in a majority of quantitative genetic twin studies.

4.6.3 Structural equation modeling and the *ACE*-model

The structural equation model (SEM) framework is very flexible modeling of the covariance matrix.[46] Many standard methods may be described within the SEM framework, e.g. linear regression, factor analysis and ANOVA. And, more importantly, these methods may be modeled in combination simultaneously. The basic idea is to express a set of equations in terms of a covariance matrix

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}), \tag{18}$$

where $\boldsymbol{\Sigma}$ is the observable covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the modeled, or expected, covariance matrix, which depends on parameters $\boldsymbol{\theta}$. A model may be postulated and, if its implied covariance matrix may be expressed as a function of the parameters in this model, SEM can be utilized. An example is the *ACE*-twin model (as discussed in above), which can be written in equation form as

$$T_{ij} = \mu + A_{ij} + C_i + E_{ij}, \tag{19}$$

where T_{ij} is the phenotypic value of twin j in twin pair i, μ is a fixed effect and A_{ij}, C_i and E_{ij} are random effects, as used in a mixed effects linear model. Further, assume that

$$A_{ij} \sim N(0, \sigma_A^2), \qquad C_i \sim N(0, \sigma_C^2), \qquad E_{ij} \sim N(0, \sigma_E^2). \tag{20}$$

The unknown parameters are thus $\boldsymbol{\theta} = (\sigma_A^2, \sigma_C^2, \sigma_E^2)$. Let g_i be a known value of the genetic relatedness between twins in pair i (= 1 for MZ twins and = .5 for DZ twins). Further, the random parameter C_i is shared between twins in a pair, and the random parameter E_{ij} is unique to each twin in a pair. Following these assumptions the expected covariance matrix for twin pair i is

$$\boldsymbol{\Sigma}_i(\sigma_A^2, \sigma_C^2, \sigma_E^2) = Cov\left(\begin{bmatrix} T_{i1} \\ T_{i2} \end{bmatrix}\right) = \begin{bmatrix} \sigma_A^2 + \sigma_C^2 + \sigma_E^2 & g_i\sigma_A^2 + \sigma_C^2 \\ g_i\sigma_A^2 + \sigma_C^2 & \sigma_A^2 + \sigma_C^2 + \sigma_E^2 \end{bmatrix}. \tag{21}$$

Using likelihood techniques a solution for the unknown parameters may be found.

Following the nomenclature of Bollen[46] and predecessors, the set of equations in a SEM may be divided into three distinct equations, the structural equation and two measurement equations; in mentioned order:

$$\begin{aligned} \boldsymbol{\eta} &= \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}, \\ \boldsymbol{x} &= \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\varepsilon}_x, \\ \boldsymbol{y} &= \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon}_y. \end{aligned} \tag{22}$$

Any SEM can be expressed in this form,[46] and there is always a theoretical solution. Although this theoretical solution may be expressed, it is not certain that it may be numerically found, i.e. the model may not be identifiable. In equations (22) $\boldsymbol{\eta}$ are latent (unobserved) variables which may be influenced by other latent variables, and are called endogenous variables. The latent variables in $\boldsymbol{\xi}$ may not be influenced by other variables, and are called exogenous variables. The observed variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are indicators of exogenous and endogenous variables, respectively. $\mathbf{B}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Lambda}_{\boldsymbol{x}}$, and $\boldsymbol{\Lambda}_{\boldsymbol{y}}$ are matrices of regression coefficients. Included in the model are also the covariance matrices of $\boldsymbol{\xi}$, $\boldsymbol{\zeta}$, $\boldsymbol{\varepsilon}_{\boldsymbol{x}}$, and $\boldsymbol{\varepsilon}_{\boldsymbol{y}}$.

4.6.4 Bivariate associations and passive gene-environment correlation

If interest is in estimating genetic and environmental sources of unmeasured confounding, rather than adjust as discussed in section 4.5, a different methodological approach is needed.

In the *ACE*- twin model; by assuming covariance between A, C and E in two different traits in an individual it is possible to draw conclusions regarding potential sources of association. The aim might be to estimate how much of an association between two traits that is due to the same set of genes effecting each of the traits, known as

pleiotropy. In what is referred to as the "correlated factors model" the within-individual covariance matrix of trait X_1 and trait X_2 may be expressed as

$$\begin{aligned} &Cov\left(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}\right) \\ &= \begin{bmatrix} \sigma_{A1}^2 + \sigma_{C1}^2 + \sigma_{E1}^2 & r_A\sigma_{A1}\sigma_{A2} + r_C\sigma_{C1}\sigma_{C2} + r_E\sigma_{E1}\sigma_{E2} \\ r_A\sigma_{A1}\sigma_{A2} + r_C\sigma_{C1}\sigma_{C2} + r_E\sigma_{E1}\sigma_{E2} & \sigma_{A2}^2 + \sigma_{C2}^2 + \sigma_{E2}^2 \end{bmatrix}, \end{aligned} \tag{23}$$

where sub-index 1 refer to trait X_1 and sub-index 2 to trait X_2, sub-index A, C and E defined similar as previously. Here the potential pleiotropic effect is captured by r_A, and other potential sources of association due to environment is captured by r_C and r_E.

There are other ways to model the bivariate genetic and environmental overlap between two traits; the most common is called the "Cholesky model". The name comes from André-Louis Cholesky who showed that any symmetric and semi-definite matrix may be decomposed to the product of a (lower) triangular matrix and its transpose.[47] In the Cholesky model each of A, C and E are defined multivariately as lower triangular matrices;

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & 0 \\ a_{12} & a_{22} \end{bmatrix}, \boldsymbol{C} = \begin{bmatrix} c_{11} & 0 \\ c_{12} & c_{22} \end{bmatrix}, \boldsymbol{E} = \begin{bmatrix} e_{11} & 0 \\ e_{12} & e_{22} \end{bmatrix}. \tag{24}$$

The bivariate association, within an individual, may thus be expressed as

$$\begin{aligned} &\boldsymbol{Cov}\left(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}\right) = \boldsymbol{A}\boldsymbol{A}^T + \boldsymbol{C}\boldsymbol{C}^T + \boldsymbol{E}\boldsymbol{E}^T \\ &= \begin{bmatrix} a_{11}^2 & a_{11}a_{12} \\ a_{11}a_{12} & a_{12}^2 + a_{22}^2 \end{bmatrix} + \begin{bmatrix} c_{11}^2 & c_{11}c_{12} \\ c_{11}c_{12} & c_{12}^2 + c_{22}^2 \end{bmatrix} + \begin{bmatrix} e_{11}^2 & e_{11}e_{12} \\ e_{11}e_{12} & e_{12}^2 + e_{22}^2 \end{bmatrix} \\ &= \begin{bmatrix} a_{11}^2 + c_{11}^2 + e_{11}^2 & a_{11}a_{12} + c_{11}c_{12} + e_{11}e_{12} \\ a_{11}a_{12} + c_{11}c_{12} + e_{11}e_{12} & a_{12}^2 + c_{12}^2 + e_{12}^2 + a_2^2 + e_2^2 + e_2^2 \end{bmatrix}. \end{aligned} \tag{25}$$

The Cholesky and correlated factors models' solution fits any data equivalently well, and are exchangeable, i.e. any parameter in one of the models may be derived from parameters in the other.[48] An advantage of the Cholesky model is that it cannot produce negative definite matrices (an advantage since a negative definite matrix in this situation would indicate covariances greater than the product of the square roots of corresponding variances; the matrix would not be a covariance matrix). In study II a version of the Cholesky approach has been used, and in study III and IV the correlated factors approach has been used.

Gene-environment correlation is the situation where one of the traits is viewed as a measure of an environment for the individual. Usually three different types of gene-environment correlations are assumed to exist; passive, active, and evocative.[39] In the current thesis the focus is on the passive type. *Passive gene-environment correlation*

occurs when an individual is exposed to an environment, e.g. by their parents, which is correlated with their genetic predisposition for other traits. An example is exposure to the environment of maternal smoking while *in utero* and cognitive and externalizing outcomes, as analyzed in study II and III. The bivariate models express the covariation between the exposure, SDP, and outcomes. Thus, an estimation of the passive gene-environment correlation is found in r_A; the correlation between A in mothers, expressing their genetic propensity to smoke while pregnant, and A in offspring outcomes, the genetic contribution to their studied traits.

4.6.5 Twin modeling

In study IV all data are from questionnaires sent out to twins and their parents. In twin models the differing genetic similarity and the equal environment assumption makes it possible to estimate A, C and E effects. MZ twins share all of their genes while DZ twins on average share 50% of their segregating genes. The assumption of 50% shared for DZ twins might be violated if the population is not mating at random, i.e. there is assortative mating, meaning that the phenotypic value is associated to individuals having offspring together.

4.6.6 Family modeling

The quantitative genetic models in study II and III utilizes other relatives than twins. The genetic correlation may be theoretically derived, under assumption of no assortative mating and disregarding the sex-chromosomes. Since, after meiosis, an offspring gets one of each chromosome from the mother and from the father. An offspring thus share 50% of the genetic material *identically by descent* with either parent ("identical by descent" left out below). Further, calculations show that, on average, full siblings share 25% of the genetic material that came from the mother, and 25% from the father. The total amount of shared genetic material is therefore 25%+25%=50%. Similar calculations may be made for any type of relation; on average, an individual shares 25% of genetic material with her aunt if her parent and aunt are full siblings or DZ twins. On average, she shares 50% of genetic material if her parent (mother in this case) and aunt are MZ twins, etc. The theoretical derived amount of shared genetic variation has been experimentally confirmed; for sib pairs the average proportion shared, adjusting for assortative mating, is 50%, with a standard deviation of 4%. For half sibs the mean and standard deviation, adjusting for assortative mating, are 25% and 3%, respectively.[49]

Regarding environmental effect on traits, it is not as straight forward as in the twin model. More shared environments may be postulated; shared between cousins, shared between siblings, shared between twins etc. It gets even more an empirical question when arguing about what happens across generations; what environment makes children similar to their parents? This is an issue which historically has been discussed, see e.g. Rao et al.[50] and Eaves et al.[51]

5 STUDY SUMMARIES

5.1 STUDY I – ADVANCING PATERNAL AGE AND VIOLENT CRIME

Background. Evidence of advancing paternal age being a causal risk factor for severe offspring mental health problems have accumulated over the last decade.[52-54] And some mental health problems are associated with criminality.[55,56] A proposed mode of causation is mutational errors in the spermatogenesis, *de novo* mutations, which accumulate as males age, and has effects on children that are worse the older the father is at conception. Indeed, in humans the majority of mutations originate in the male germ line,[57] and most mutations are not advantageous.

A well-known theory of development of anti-social/criminal behavior, which appeals to common sense and fit well with observations, is that of adolescence-limited versus life-course-persistent behavior.[58] In this study we wanted to take this into account by looking at two measures of violent offending; (1) ever convicted of a violent crime, and (2) number of violent crimes among violent criminals.

To our knowledge, there are no previous studies of the potential effects of advancing paternal age on criminal or anti-social behavior.

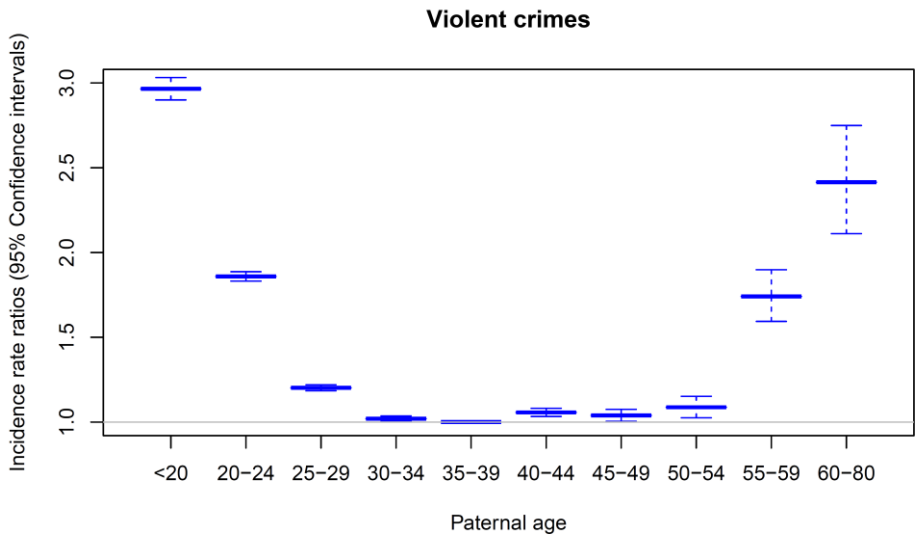
Aim. We wanted to estimate associations between advancing paternal age and the two operationalizations of violent criminality. Further, we wanted to estimate within-sibling associations to adjust for factors shared by offspring of the same father.

Methods. We obtained all convictions, and dates of sentences, for each subject in the cohort. This allowed us to calculate the two measures of violent criminality; ever committing a crime (binary) and number of crimes committed. Importantly we calculated the time at risk for conviction after the individual's 15th birthday. This allowed us to account for differential follow up time, and to try and capture the increased risk for violent criminality during adolescence.

To respond to the aims stated above we used the hurdle Poisson model,[59] where count outcomes are analyzed in separate models; one for the probability of observing a count greater than 0 (binary outcome) and another for counts among those who had a count of 1 and above (truncated [at 0] Poisson outcome). To account for the correlated data structure we wanted to use GEE, since no such procedure was available for truncated Poisson outcome we had to develop it. We analyzed the data using paternal age as a categorical variable. We then proceeded to fit within-family models where paternal age was modeled using a spline function to account for potential non-linearity in the associations.

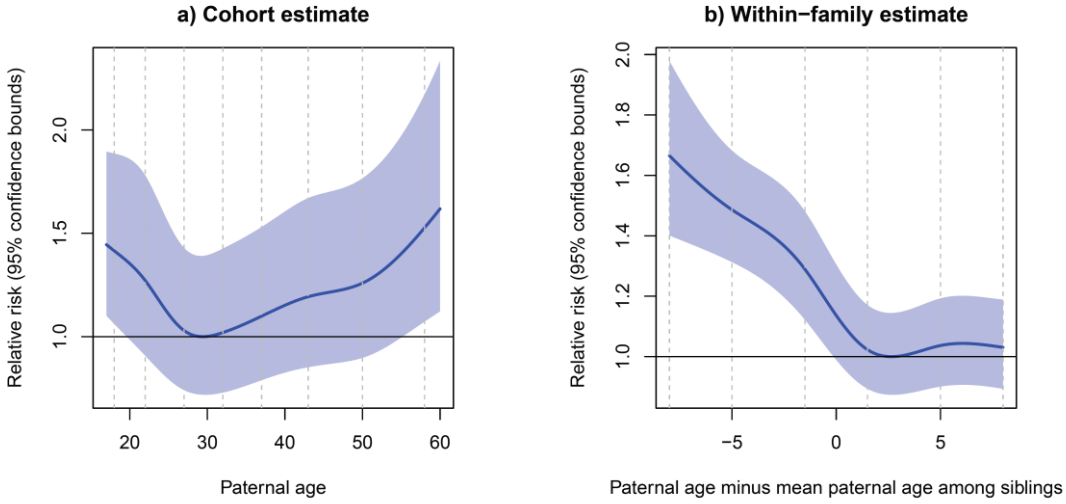
Results. Paternal age showed a U-shaped association with number of violent crimes; offspring of both young and old fathers had higher incidence of convictions of violent crimes compared to offspring of fathers aged 35-39 (Figure 5). When analyzing the data using the two different measures of the outcome the same association was seen; both lower and higher paternal age increased the risk of being convicted of at least one

Figure 5. Crude incidences of violent crime.



violent crime and increased the incidence among those convicted of at least one violent crime (panel a in Figure 6 for at least one violent crime, and panel a in Figure 7 for number of violent crimes). When we performed within-sibling analyses the results differed; the association of advancing paternal age increasing the risk of ever being convicted for a violent crime completely disappeared (panel b in Figure 6) while the association with number of crimes prevailed (panel b in Figure 7). A quantification of this prevailing result is; the younger of two siblings born 10 years apart (i.e., the one born when the father was older) was on average convicted for 15% more violent crimes than his older sibling.

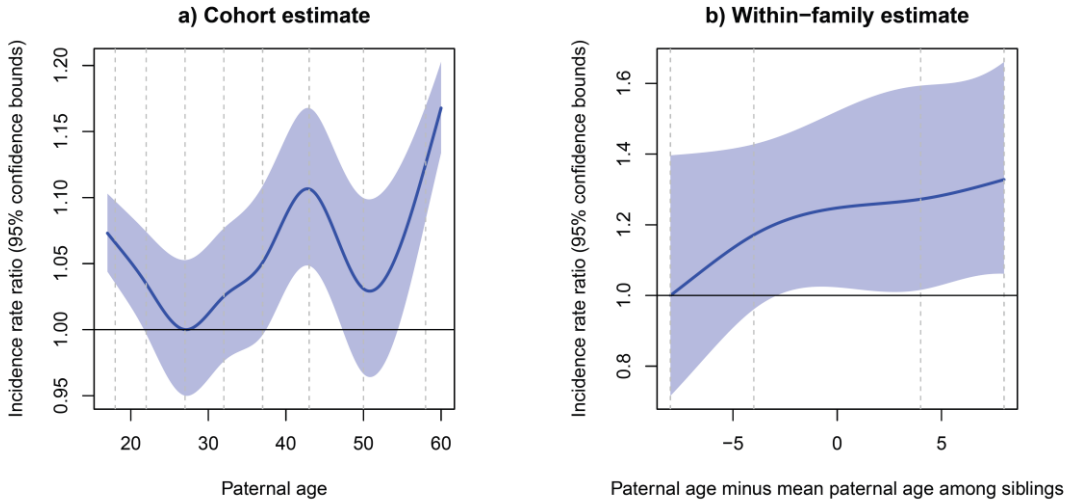
Figure 6. Analysis of probability to ever be convicted of a violent crime.



Note: Cohort estimate refers to estimate from ordinary exposure parametrization, as opposed to between-within parametrization.

In sensitivity analyses with the outcome being nonviolent crimes instead of violent crimes, the results were similar, with the notable exception that in the analysis of the

Figure 7. Analysis of number of violent crimes in violent offenders.



Note: Cohort estimate refers to estimate from ordinary exposure parametrization, as opposed to between-within parametrization.

association with number of nonviolent crimes (the truncated Poisson analysis) where the association disappeared when siblings were compared.

Discussion. Although an association between advancing paternal age and offspring ever being convicted of a violent crime was observed, the association seemed to be attributable to familial factors shared between offspring of the same father. In contrast, the observed association of increased incidence of violent criminal convictions among those convicted of at least one violent crime persisted in within-family analysis. The difference in inference, regarding familial factors being responsible for the associations, indicates potentially different etiologies for the two constructs of violent crime. The persistent violent criminal is suggested to be more biologically driven[58,60,61] while the one-time offenders follow a societal norm (adolescent limited anti-social behavior).[58] If the *de novo* mutations hypothesis is true, effects of new mutations would be biological. Thus, we would expect to see a larger, persisting, effect in the re-offending analysis as found in this study.

Methodological considerations. One main problem in analyzing paternal age as a risk factor for violent criminality is that it is inherently correlated with many other potential risk factors; maternal age, birth order, calendar time, socio-economic status, etc. We were able to adjust for some of the potential confounders (e.g., maternal age and birth order), but not others. The most problematic potential confounder is birth period effects; in the within-sibling analysis differences in paternal age between siblings are necessarily perfectly correlated with calendar time. To attempt to avoid bias from this source we used a birth cohort (1958-1979) which has very similar pattern of incidence of convictions of violent criminal offences between birth periods and over covered follow up ages.

For the within-sibling estimates to be less biased than the ordinary estimate the assumption is that there exists confounders which are stable in the fathers and between each offspring. Examples of such hypothesized confounders are; socio-economic status,

income, education, mental illnesses, and genes for impulsivity. Further, these assumed confounders should be more correlated than other factors making the exposure, paternal age, correlated. We cannot specify these factors, and it is possible that there are such factors which correlate more highly than factors that are confounders, making the within-family estimates more biased than the ordinary estimate.

5.2 STUDY II – SMOKING DURING PREGNANCY AND STRESS COPING IN OFFSPRING

Background. Exposure to nicotine *in utero* is associated with a wide range of disadvantageous outcomes.[3,62-74] However, animal and human studies have reached conflicting results in some areas; cognitive functioning being one. Animal models indicate a causal association between nicotine exposure and e.g. learning problems,[67] memory problems,[67] and, notably, stress vulnerability.[74] In humans associations between SDP and, e.g., poor academic achievement[65,72] and IQ[73] has been observed; however, for both outcomes the association disappeared when siblings were compared, a result congruent with non-causal associations.[2]

To further the knowledge in the field we investigated whether exposure to SDP was associated with poorer stress coping in late adolescent Swedish males.

Aim. We wanted to estimate association between SDP and stress coping, and, to assess causality, estimate within-sibling and within-cousin effects. Further, we wanted to estimate the source of association using quantitative genetic models.

Methods. At the military conscription the conscriptees were assessed for stress coping ability during wartime. A clinical psychologist rated each conscriptee, yielding a score from 1 to 9 on a Likert-type scale, higher value indicating better coping, with stipulated mean 5 and standard deviation 2.

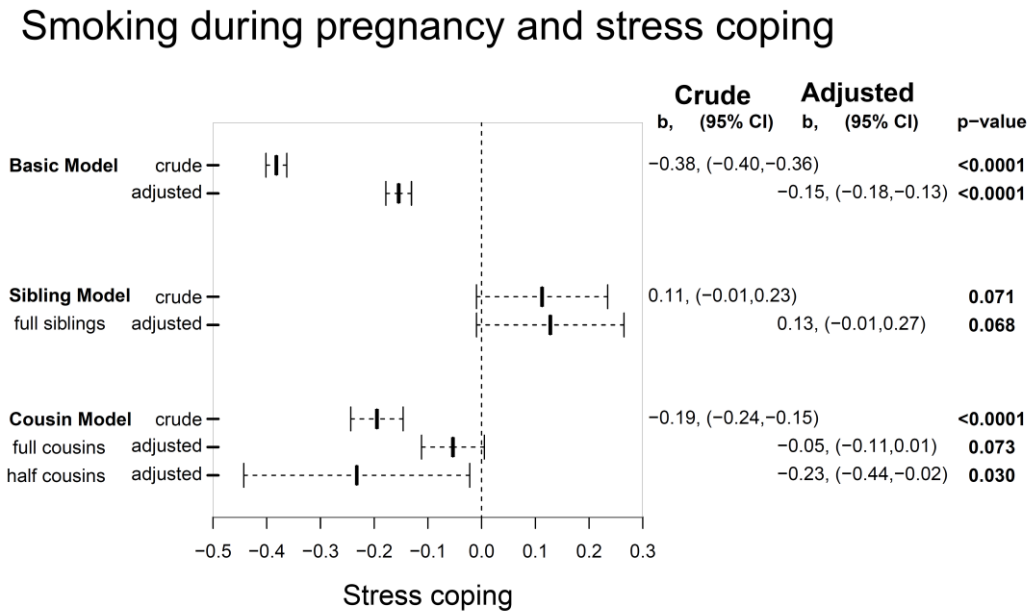
To respond to the aims we adopted a similar approach as in D'Onofrio et al.[66] We estimated the association between SDP and the stress coping measure. We compared siblings and cousins differentially exposed to SDP to get estimates adjusted for shared factors. Finally we fitted quantitative genetic models to assess whether genes and/or environments were responsible for the association.

Results. Offspring who were exposed to SDP scored on average 0.38 points lower on the stress coping scale (Figure 8). The association persisted when adjusting for potential confounders (-0.15; Figure 8). When differentially exposed full cousins were compared the association was reduced to -0.05 (Figure 8), and completely disappeared when siblings were compared (0.13; Figure 8). The quantitative genetic model showed that additive genetic effects were responsible for the majority of the association.

Discussion. This study showed that an association between SDP and stress coping exists in humans. However, the association disappeared when siblings differentially exposed to SDP were compared, a result congruent with non-causal inference. This association was shown to be mainly of genetic origin. An interpretation of the results is that mothers who have a genetic liability to smoke during pregnancy also pass genetic risks of poor stress coping to her offspring.

Methodological considerations. Potential systematic misclassification of exposure, e.g. mothers not reporting smoking while pregnant more often for later-born siblings, may have biased estimates of the within-family associations. We did not investigate

Figure 8. Regression results.



this potential source of bias any further in this study, nor did we investigate carry-over or sibling contagion effects; however, they are thoroughly investigated in study III in this thesis, albeit not for the outcome of this study.

For the interpretation of the within-sibling estimate as less biased than the ordinary estimate there must exist confounders shared between siblings. In the current study we hypothesize that the most important confounder is found in maternal genetic variation, which is perfectly shared between siblings.

The SEM used for the quantitative genetic analysis was adapted from D'Onofrio et al.,[66] and fitted in the software Mplus.[75] Mplus did not allow for combined analysis of categorical (i.e., SDP) and continuous (i.e., stress coping) variables in the current analysis, thus we had to treat SDP as continuous. However, the modeled association between SDP and stress coping was at the mothers level, using mean SDP over all included pregnancies. The modeled covariance was thus between the mean of binary variables and a continuous variable. Whether this is a correct analysis to make is an empirical question, there is, however, no particular theory supporting why the distribution of the exposure should be defined as it was.

5.3 STUDY III – SMOKING DURING PREGNANCY AND ADVERSE OUTCOMES IN OFFSPRING

Background. Findings congruent with non-causal effects of SDP on long-term cognitive and externalizing outcomes in offspring in humans have come from sibling-studies and in-vitro fertilization studies.[2,71,76] Although the evidence has accumulated over some years, the scientific community is far from consensus. For example, an editorial accompanying an article on SDP and conduct disorder[77] concluded that "… there is widespread consensus that maternal smoking during pregnancy has adverse, long-term effects on neurobehavioral development in the offspring…", and that "prenatal tobacco smoke exposure contributes significantly to subsequent conduct disorder in the offspring" was the take home message from the study.[78] This conclusion was drawn despite the authors of the article being much more cautious; "The causal explanation for the association between smoking in pregnancy and offspring conduct problems is not known but may include genetic factors and other prenatal environmental hazards, including smoking itself."[77] Misunderstandings like this, and some conflicting evidence from animal studies and human studies,[79] call for a broader take on the issue. We decided to perform family-studies across different domains of outcomes, and investigate mechanisms by which the associations may arise.

Aim. We wanted to assess the association between SDP and several outcomes in cognitive and externalizing areas, crudely and adjusted for covariates. We also wanted to contrast results in these areas with results with pregnancy related outcomes. By comparing siblings and cousins in within-family designs we wanted to assess whether the associations were consistent with causal interpretations. Finally we wanted to estimate genetic and environmental sources of variation in SDP and outcomes, and, more importantly, decompose the covariation between SDP and each outcome into these sources.

Methods. We used eight different outcomes in three different areas of interest; pregnancy related outcomes (birth weight, preterm birth, and being born small for gestational age), long-term cognitive problems (low academic achievement and general cognitive ability), and externalizing problems (criminal convictions, violent criminal convictions, and drug misuse). We used gender of offspring, maternal age at childbirth and birth year as covariates in the associations studied.

In response to our aim we systematized within-family analyses. Further, we developed a SEM for analyzing the intergenerational associations, an extension of the SEM used in study II. The model developed was named the *ACMPE*-model, it allows for estimation of more environmental sources of covariance between exposure and outcome than, e.g., the *ACE*-model does. In Table 2 the potential sources of (co)variance are listed. Binary variables, such as SDP and preterm birth, were modeled with the liability-threshold approach.[38] Using the liability-threshold approach corresponds to assuming that each subject has a liability distributed according to a normal distribution. A 1 is observed if the liability is above some estimated threshold,

otherwise a 0 is observed. This is equivalent to a GLM with binary outcome and a probit link. Continuous variables were assumed to come from the normal distribution.

Table 2. Sources of variance and covariance in the *ACMPE*-model.

Parameter	Parent generation	Offspring generation
A (*Additive genetic*)	Additive genetics	Additive genetics
C (*Common environment*)	Environment unique to one mother	Environment shared between siblings
M (*Maternal environment*)	Environment shared between sisters who are mothers	Environment shared between cousins
P (*Paternal environment*)	Spouse effect	Paternal effect
E (*Non-shared environment*)	Environment unique to each pregnancy	Unique individual environment

Analyses were performed in three steps; all associations were analyzed using the exact same methods:

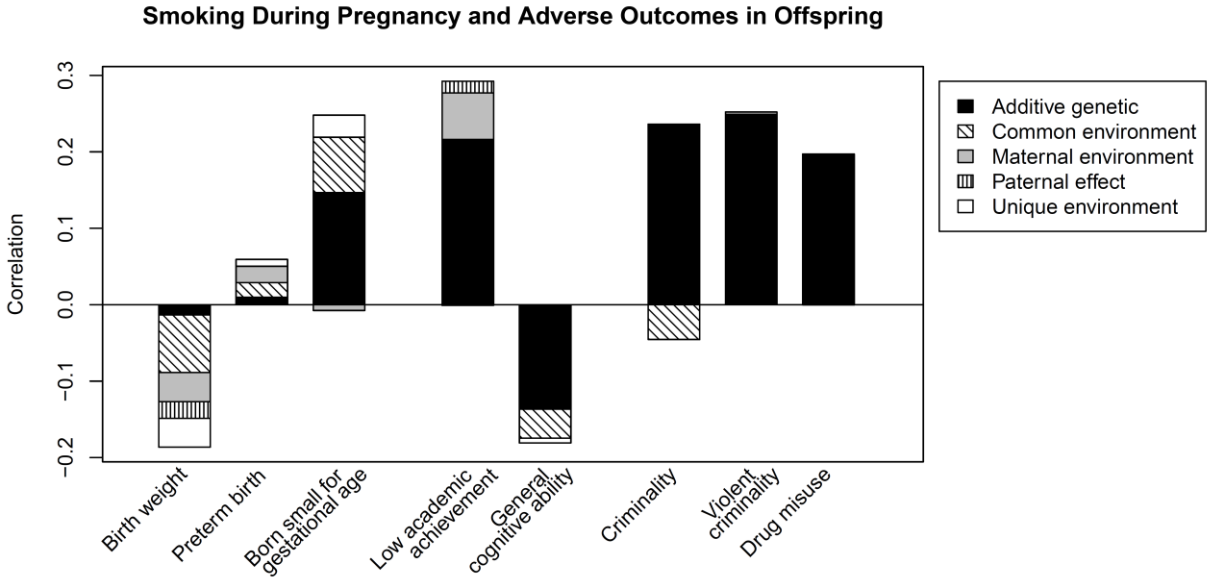
1. Regular epidemiological estimates; regression coefficients for continuous outcomes and odds ratios for binary outcomes were estimated.
2. Within-family estimates; half- and full-cousins, and half- and full-siblings were compared to get within-relative estimates. The continuous outcomes were analyzed using the between-within method, the binary outcomes were analyzed using conditional likelihood.
3. Quantitative genetic estimates; each association was analyzed using the *ACMPE*-model where the associations were split up into genetic and environmental sources of covariance.

Results. SDP was associated with each of the outcomes, both crudely and when adjusted for covariates. The associations between SDP and pregnancy outcomes persisted when comparing relatives, albeit somewhat reduced. For the long-term cognitive and externalizing problem outcomes it diminished when comparing cousins, and disappeared completely when comparing siblings.

The quantitative genetic models showed some difference in the pattern of overlap between exposure and outcomes. For the pregnancy related outcomes non-shared environmental influences were important, explaining 12-20% of the covariance, for the long-term cognitive and externalizing outcomes it was not, explaining 0-4% of the covariance (Figure 9). For the cognitive and externalizing outcomes genetics explained over 74% of the covariance.

Discussion. This study confirmed earlier observations of the association between SDP and pregnancy related outcomes persisting in within-sibling analyses, thus being consistent with causal interpretations. The study also was in line with observations of associations being non-significant in within-sibling analyses of SDP and long-term cognitive and externalizing problem outcomes, consistent with a non-causal

Figure 9. Results for the quantitative genetic model.



interpretation. Understanding of the underlying mechanisms responsible for the associations between SDP and outcomes is necessary for appropriate prevention, intervention and future research. Thus, the study took the analyses one step further than earlier and partitioned the associations into genetic and environmental sources.

Methodological considerations. Both results where the associations did (i.e., pregnancy related outcomes) and did not (i.e., cognitive and externalizing outcomes) persist in exposure discordant siblings were observed. Hence, the previously mentioned problem of measurement error in the exposure, where within-estimates are biased towards null effects,[36,37] is very unlikely to account for the findings.

Similarly as in study II we rely on the stable genetic variation in mothers between pregnancies as the major confounder of the association to be able to interpret the within-sibling estimates as less biased than the ordinary estimates.

In this study a model, named the *ACMPE*-model, was introduced. The aim of this model is to utilize extended pedigrees to quantify relative contributions of various sources of (co)variance in a parental trait (SDP) and outcomes in the offspring generation. The model makes assumptions about sharing of environments which may be erroneous, however. No attempts to justify these assumptions were made.

We addressed two critical assumptions in the sibling comparison design, these assumption and their potential biases have been identified before (e.g. in D'Onofrio et al.);[2]

(1) Carry-over effects. If the exposure in siblings born in the first pregnancy influences the outcome in second-born siblings directly, the assumptions in within-sibling design are violated. An example of this is if the smoking during the first pregnancy causes a biological alteration in the mother, which persists and thus affects the outcome for the second sibling regardless of smoking status in second pregnancy.

(2) Sibling contagion effects. If the outcome of first-born sibling influences the outcome in second-born sibling the assumptions are violated. An example of this is if the behavior of first offspring influences the behavior of second offspring directly. This might happen, for instance, if the first-born sibling is involved in anti-social behavior and influence the second-born to be involved as well.

This can be represented in DAGs; Figure 10 to 12 shows first-born and second-born offspring of a mother. SDP_1 is smoking status during pregnancy of first-born sibling, SDP_2 smoking status for second-born, OUT_1 is outcome in first-born, OUT_2 outcome in second-born. U is shared confounders, V_1 and V_2 are non-shared confounders, and L represents non-confounders that make the exposure status similar between pregnancies. The DAG represents a similar situation as in Figure 4; the path diagram of familial confounding.

To see if we can draw any conclusions as to whether these effects are present, we considered the association between SDP_1 and OUT_2 and the association between SDP_2 and OUT_1. In Table 3 the open paths between SDP_1 and OUT_2 and between SDP_2 and OUT_1 are listed for each of the scenarios described above. The main point to notice here is that both carry-over and sibling contagion effects introduces additional open paths between SDP_1 and OUT_2 but not between SDP_2 and OUT_1. Under the assumption that the association within each of the siblings are the same (i.e., $SDP_1 \rightarrow OUT_1$ represent the same statistical association as $SDP_2 \rightarrow OUT_2$ does; $SDP_1 \leftarrow U \rightarrow OUT_1$ represents the same as $SDP_2 \leftarrow U \rightarrow OUT_2$ does, etc.) we may

Figure 10. No additional effects.

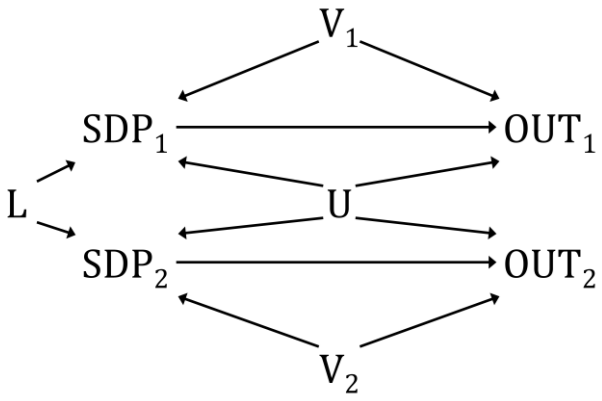


Figure 11. Carry-over effect (dashed line).

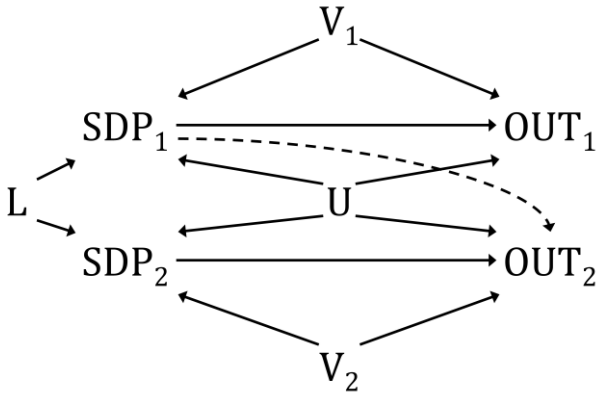


Figure 12. Sibling contagion effect (dashed line).

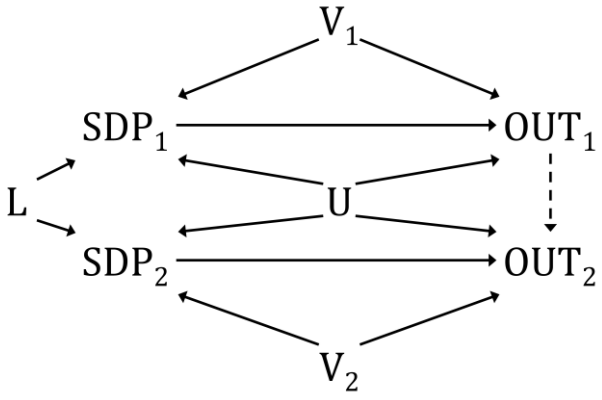


Table 3. Open paths.

	Open paths from SDP_1 to OUT_2	**Open paths from SDP_2 to OUT_1**
No additional effects	$SDP_1 \leftarrow U \rightarrow OUT_2$ $SDP_1 \leftarrow U \rightarrow SDP_2 \rightarrow OUT_2$ $SDP_1 \leftarrow L \rightarrow SDP_2 \rightarrow OUT_2$	$SDP_2 \leftarrow U \rightarrow OUT_1$ $SDP_2 \leftarrow U \rightarrow SDP_1 \rightarrow OUT_1$ $SDP_2 \leftarrow L \rightarrow SDP_1 \rightarrow OUT_1$
Carry-over effect	$SDP_1 \leftarrow U \rightarrow OUT_2$ $SDP_1 \leftarrow U \rightarrow SDP_2 \rightarrow OUT_2$ $SDP_1 \leftarrow L \rightarrow SDP_2 \rightarrow OUT_2$ $SDP_1 \rightarrow OUT_2$	$SDP_2 \leftarrow U \rightarrow OUT_1$ $SDP_2 \leftarrow U \rightarrow SDP_1 \rightarrow OUT_1$ $SDP_2 \leftarrow L \rightarrow SDP_1 \rightarrow OUT_1$
Sibling-contagion effect	$SDP_1 \leftarrow U \rightarrow OUT_2$ $SDP_1 \leftarrow U \rightarrow SDP_2 \rightarrow OUT_2$ $SDP_1 \leftarrow L \rightarrow SDP_2 \rightarrow OUT_2$ $SDP_1 \rightarrow OUT_1 \rightarrow OUT_2$ $SDP_1 \leftarrow V_1 \rightarrow OUT_1 \rightarrow OUT_2$ $SDP_1 \leftarrow U \rightarrow OUT_1 \rightarrow OUT_2$	$SDP_2 \leftarrow U \rightarrow OUT_1$ $SDP_2 \leftarrow U \rightarrow SDP_1 \rightarrow OUT_1$ $SDP_2 \leftarrow L \rightarrow SDP_1 \rightarrow OUT_1$

test the presence of a carry-over and/or sibling contagion effect. If a test reveals that SDP_1 has a significantly stronger association to OUT_2 than SDP_2 has to OUT_1 it is indicative of either carry-over and/or sibling contagion effect being present.

For the present study we investigated these effects by analyzing the association between exposure in one offspring and outcome in the other, and included an interaction term between birth-order and effect size. If this interaction term is significant, and the SDP_1-OUT_2-association is stronger than the SDP_2-OUT_1-association, the result would indicate that carry-over and/or sibling contagion effects were present. The results are found in Table 4, and show that, although there were some significant interaction effects for the pregnancy outcomes, there were no support for the SDP_1-OUT_2-association being stronger than the SDP_2-OUT_1-association for the long-term cognitive and externalizing problem outcomes (if anything the effect was in the different direction for low academic achievement).

Table 4. Test of carry-over and sibling contagion effects.

	Effect estimates (standard error)		**Interaction effect;**
Outcome	**SDP_1 on OUT_2**	**SDP_2 on OUT_1**	**p-value**
Birth-weight	-152.5 (1.7)	-148.4 (1.9)	0.110
Preterm birth	0.32 (0.02)	0.13 (0.02)	<0.001
Born small for gestational age	0.86 (0.01)	0.72 (0.01)	<0.001
Low academic achievement	1.00 (0.02)	1.08 (0.02)	<0.001
General cognitive ability	-0.59 (0.03)	-0.64 (0.04)	0.282
Crime	0.67 (0.02)	0.72 (0.02)	0.098
Violent crime	1.15 (0.05)	1.12 (0.05)	0.624
Drug misuse	0.69 (0.03)	0.80 (0.05)	0.064

5.4 STUDY IV – THE CO-DEVELOPMENT OF ADHD AND EXTERNALIZING BEHAVIOR

Background. Although it is well established that ADHD and externalizing traits, such as conduct disorder and oppositional defiant disorder, are comorbid, the developmental trajectories are not clearly understood. In prison populations it has been estimated that ADHD is present in between 24% and 45% of the inmates,[80] compared to a world-wide adult prevalence of about 4%.[81] Both ADHD-like traits and externalizing traits have been shown to have substantial heritable components, explaining 60-90% and around 50%, of ADHD-like and externalizing traits, respectively.[82-85]

Aim. We wanted to investigate how ADHD and externalizing behavior co-developed from childhood to adulthood; is one trait predicting the other? We also wanted to make use of multiple raters, and be able to assess genetic and environmental sources of covariance between ADHD and externalizing behavior.

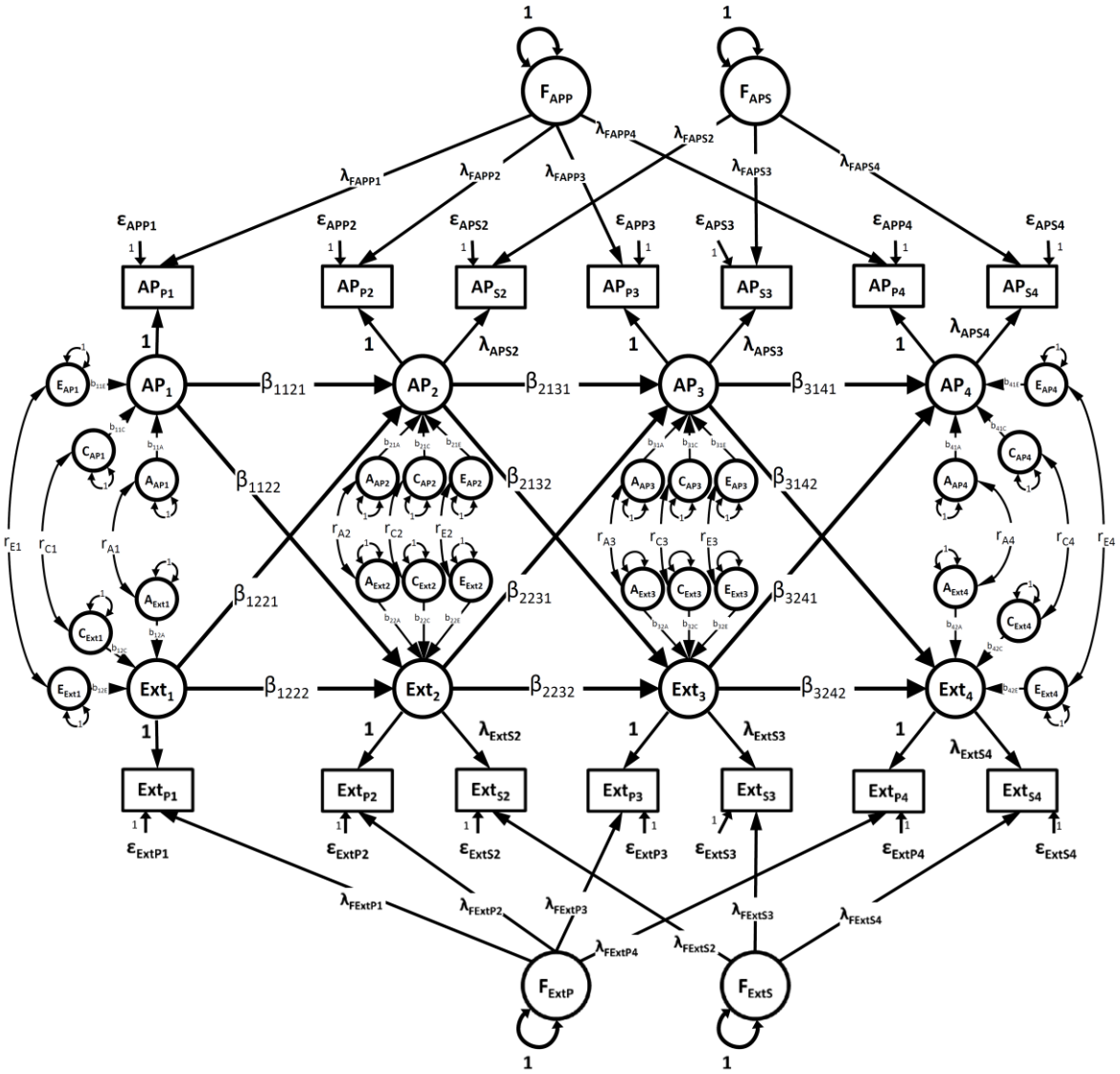
Methods. The study used data from the TCHAD-study where twins and their parents have been approached to answer questionnaires at ages 8-9, 13-14, 16-17 and 19-20.[27] The Child Behavior Checklist (CBCL) was used for parents to evaluate ADHD and externalizing behavior, and analogues were used for self-rating. In line with previous literature the scales Attention Problems (AP; an ADHD sensitive scale) and Externalizing behavior (Ext; including both an aggressive and a delinquent subscale) were used as scales for ADHD-like and externalizing traits.[86,87]

To be able to answer the questions posed in our aims we developed a novel model for the available longitudinal data with multiple raters on multiple traits. The cross-lagged model[88] was extended to cover more time-points, and combined with a model for the measurements which factored the raters together.[31]

Both parent and self-rating were available at all time-points but the first; the model estimated a shared view of each trait from the two raters (Figure 13). By combining rater effects, shared over time-points, and random errors for each rater at each time-point, we aimed at using as much as possible of the information and to reduce rater and time-specific biases. The cross-lagged structural model (Figure 13) was used to discern developmental patterns. At each time-point, in this model, each of the constructs (AP and Ext) was regressed on both constructs the previous time-point. This allows for separating effects which may be explained by variation at earlier time-points (stable effects) from variation which is new at the present time-point (innovative effects). We separated the covariance matrix at each time-point into stable and innovative effects, and we further decomposed this into A, C and E sources.

Results. The results are summarized in Figures 14 and 15. No reduced model fitting was performed; the results are from the full, unreduced, model. In Figure 14 the standardized regression parameters (cross-age stability [within trait] and cross-lagged [between traits]) are presented. All cross-age stability paths were large (0.52-0.88) and differed significantly from 0. Only two cross-lagged paths were significantly different from 0; from Ext at age 8-9 to AP at age 13-14 and from AP at age 16-17 to Ext at age

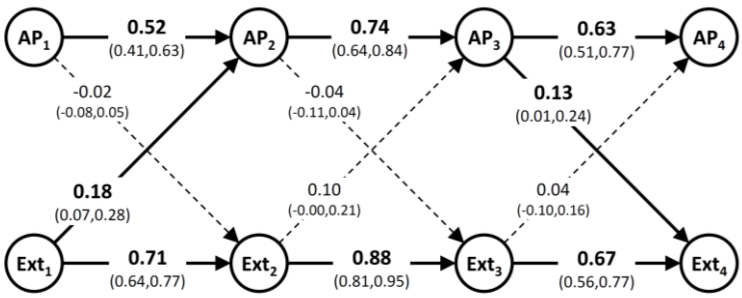
Figure 13. Path diagram representing the model.



Note: Path diagram within an individual, for estimation twin-pairs are used. Sub-index 1 to 4 indicates time-point, and sub-index P and S indicate parent and twin, respectively. F_{APP} represents rater effects of parent on AP, F_{APS} rater effects of twin on AP, F_{ExtP} and F_{ExtS} rater effects of parent and twin, respectively, on Ext. ε represents rater-specific errors.

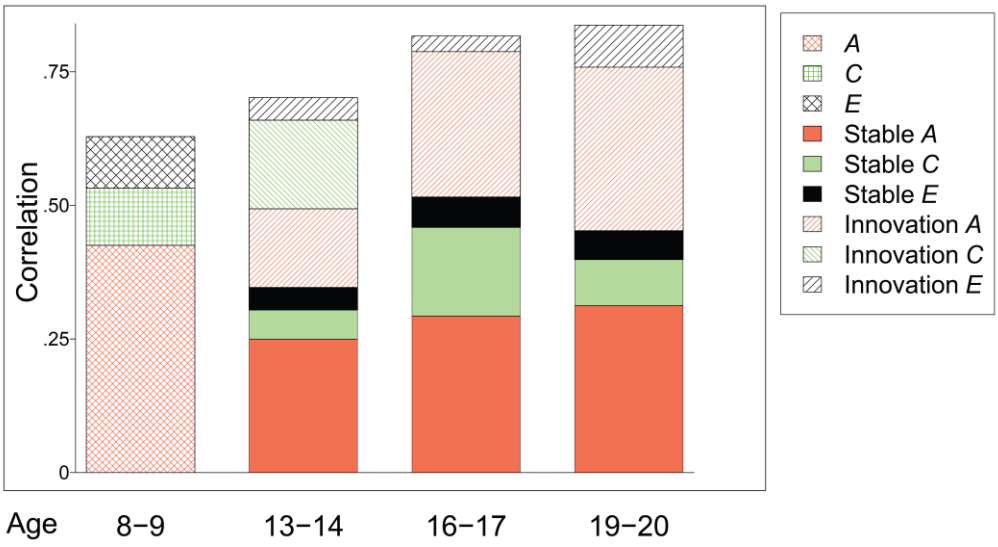
19-20. In Figure 15 the correlation and relative contributions of stability and innovation A, C and E are presented. The phenotypic correlation (i.e., the correlation between the latent AP and Ext constructs) at age 8-9 was relatively high, 0.63, and increased significantly the subsequent time-points to 0.70,

Figure 14. Results for the structural model.



0.82 and 0.84 (Likelihood ratio test for equal correlation; χ^2=36.30, df=4, p-value<0.001).

Figure 15. Correlation between AP and Ext and relative contribution of stability and innovation A, C and E.



Discussion. Many previous studies has focused on the effect of childhood ADHD on later externalizing behaviors, and not considered the possibility of the reverse direction. The only study that we are aware of which investigated whether externalizing behavior predicted ADHD found that it did, while ADHD did not predict externalizing behavior (in mid-childhood to adolescent children, and in both cases adjusting for concomitant associations).[89] Interestingly, we were able to reproduce this result where we found externalizing behavior in mid-childhood to predict ADHD-like traits in early adolescence, but not the other way around. The observed associations of childhood ADHD and late adolescent/adult externalizing problems are, however, not contradicted by our results since we saw that ADHD-like traits in mid to late adolescence predicted externalizing behavior in young adulthood.

The source of covariation between ADHD and Ext was explained in almost equal proportions by innovation and stable effects throughout development. This indicates that developmental changes in etiologic factors are the rule, rather than the exception. We also found that the correlation between the two traits increased over time.

Methodological considerations. The main reason for using this cross-lagged model was that it fitted well with the research question at hand; the development of co-occurrence of ADHD-like and externalizing traits. The model, however, puts constraints on how the traits are allowed to co-vary in genetic and environmental sources between time-points. Specifically it is more constrictive than a "Cholesky" model, where A, C and E components more freely are allowed to co-vary across time-points.

The “shared-view” approach, which was used to try and capture less biased factors representing ADHD-like and externalizing traits, focus the analysis on a fraction of the available variation. By focusing on this specific fraction of variation, we aim at getting closer to inferences representing the “true nature” of the associations.

6 GENERAL DISCUSSION

In this thesis I have examined risk factors for the development of cognitive and externalizing problems. This has been done using within-family and quantitative genetic designs. The methods used have been adaptations of existing methods, as well as novel methods developed within the work with this thesis.

6.1 GENERAL FINDINGS

Within-family designs identified substantial familial confounding; i.e., for the associations between paternal age and ever committing a violent crime and between SDP and long-term cognitive and externalizing problems. The results are in conflict with some researcher's views regarding causality of the associations (e.g., Slotkin),[78] and instead congruent with non-causal interpretations. This demonstrates that simply looking at associations, whether crude or adjusted for measured covariates, is not always sufficient. Hence, family designs are important for making less biased inferences and should be used more often in epidemiology to provide information on the mechanisms behind found associations.

In our attempts to understand the reasons for the familial confounding, I applied quantitative genetic models to the associations of SDP and a variety of cognitive and externalizing problem outcomes; the results showed that the familial confounding was primarily due to substantial genetic overlaps. Again, within-family designs combined with quantitative genetic designs seem to be very informative of mechanisms behind associations in epidemiology, and other associations not investigated in this thesis would probably benefit from such analyses. Genetic effects were also important for the co-development between ADHD and externalizing behavior, regardless if sources stable over time or innovative at specific time-points during development were considered.

6.2 METHODOLOGICAL DEVELOPMENT

To be able to perform analyses to respond to my research questions I had to develop new methodological approaches.

In study I developed a novel method to include a truncated Poisson outcome in a within-sibling design using GEE; this made it possible to investigate the important aspect of repeated violent offending, where data suggested a within-family effect of paternal age on repeated violence.

Within-family analyses have been performed using Swedish data before, data has been assembled in different ways, and different methodological approaches have been used. For example, in a study of advancing paternal age and bipolar disorder the family history of bipolar disorder were adjusted for;[52] a study of SDP and school performance used permutations of exposure and outcome within sibling pairs;[72] a study on co-occurrence of schizophrenia and bipolar disorder used quantitative genetic models.[90] A great deal of my thesis work included to develop a program package in which both data

generation as well as analyses can be done more efficiently, which will be beneficial for future research. This package was developed during the work with study II and III, and used in study III. The data managing program is written in the SAS software (SAS Institute, Cary NC), and allows for identification of extended families using the Swedish total-population registries. The analytic program is written for the OpenMx package[91,92] in software R,[93] and performs quantitative genetic analyses. My goal in doing this was to simplify analyses on extended family pedigrees for researchers, and make use of the family data available in Sweden. Hopefully future use of this package will make both analyses and interpretation of family studies (particularly in the Swedish setting) easier and more streamlined. In fact, others have already started using this program.[94]

Using this program, in study III, I developed a parameterization of intergenerational associations in the quantitative genetic model for extended families. This model provided insights in genetic and environmental contributions to the familial confounding of associations.

Finally, in study IV I extended the cross-lagged model[88] and combined it with an existing measurement model[31] to be able to study a developmental model over four time-points using multiple raters. This model provide considerable advantages, since models using multiple raters better handle rater bias,[31,86,87] and no such model has up to now been presented.

6.3 METHODOLOGICAL CONSIDERATIONS

In section 5 methodological considerations specific to each study has been discussed, below follows some general methodological considerations.

Within-family analyses. All study designs have limitations, as does all statistical analyses, and within-family design and analyses are no exceptions. The within-parameter is not always easy to interpret, as has been discussed. It has been shown that correlation between both unmeasured confounders and measurement error in the exposure may bias the estimate.[37] Within-sibling analyses assume no carry-over and sibling contagion effects (see D'Onofrio et al.;[2] which extensively discuss assumptions and limitations in different within-family designs). Clearly, no single study is enough to prove causal (or lack of causal) effects for a specific association. Instead, to understand mechanisms for associations several designs (including animal studies) are needed, especially if it is not possible to perform a randomized controlled trial. In study III I used several types of analyses to assess the validity of some of the assumptions in the within-sibling design, and all these analyses suggested that the associations were due to familial confounding.

Within-family estimates may be less biased when there are unmeasured confounders shared between relatives. However, it is inherent in the design that explicit tests whether unmeasured confounders are present cannot be made. It would require measurements of the confounding variables, and the absence of measurements is the reason for doing the analyses in the first place.

Utilizing data on relatives and performing within-relative comparisons have the potential to enhance knowledge in many areas and do it for many associations. It is, however, important to be wary about making strong statements about causality of within-relative estimates. Throughout this thesis I have tried to express this causative approach using phrases such as "congruent with causal inference".

Equal environments assumption. In contrast to genetic similarity between relatives, where there is a theoretical reason for assumed correlations, there is no theoretical value for similarity between relatives of environmental effects on traits. Therefore, rather than assuming anything about shared effect of environment, the estimable correlated environmental effect is split into C and E in twin studies.[38,39] The C-parameter is interpreted as environment making twins alike in the trait under study. It is important to note that C does not mean any environment shared, it means environments *making the specific traits similar* between twins, and further, similar to the same amount in MZ and DZ twins; this is the equal environments assumption. Whether the equal environments assumption is valid or not is an empirical question specific to each studied trait (or association between traits), and several tests of the assumption have been made, both on univariate, e.g. for psychiatric traits,[95] and multivariate, e.g. for aggressive traits,[96] data. These tests generally support the equal environments assumption.

Dominance deviance and assortative mating. In the quantitative genetic analyses performed in the current thesis neither dominance genetics nor assortative mating has been considered. Exclusion of both these sources of (co)variance partly follows from an attempt to make the models simple to grasp for the average potential future user. Further, I was not specifically interested in estimating any of these parameters in included studies. However, the model may be extended to include both sources if need be. Although not included, it is unlikely that the exclusion has biased the results to a large extent. Firstly, dominance deviation from additive genetic influence is likely to not influence the overall variation due to genetics greatly; exhaustive simulations found that even the classical twin design tended to estimate variation due to genetics relatively close to simulation values.[97] Secondly, a study on violent criminal convictions, using the same data sources as in this thesis, investigating assortative mating found that level of assumed assortment did not have large effects on the other estimates (heritability estimates ranging from 46% to 59% between the two most extreme scenarios).[98]

Extending shared environments to other relatives and over generations. When there is no empirical knowledge regarding environmental inheritance it is hard to decide by which pathways the, possibly intergenerational, associations may be assumed to work. In study III different potential environmental sources of (co)variance are assumed, modeling the association between relatives in studied traits. The modeled environmental sources of (co)variance may be viewed as an elongation of the twin studies' C to other type of relatives and environments. Approaches of intergenerational environmental associations have been suggested previously; the social homogamy model of Rao et al.[50] and its extension in the social homogamy and cultural inheritance model by Eaves et al.[51] being two examples. However, the data used in the current

thesis is different than what Rao et al. and Eaves et al. had in mind, particularly since different traits in the parental and offspring generations are modeled. In the *ACMPE*-model the *C*, *M*, *P* and *E* represents environments shared or not shared between relatives of different relation. Whether the assumptions regarding environmental similarities/differences are valid remains to be tested.

No interaction between genetic variation and environments. One of the simplest ways to conceptualize genetic variation-by-environment interaction in quantitative genetic analyses is that the (co)variance due to genetics is different in different environments.[39] An example in current thesis is if maternal SDP would be less heritable in, e.g., high socioeconomic strata than in low, perhaps due to higher education and/or more stigmatization among mothers with high socioeconomic status, making the genetic liability to smoke while pregnant less important for the variability of the trait. In this case it would be informative to stratify the quantitative genetic estimation on socioeconomic level and estimate separate (co)variance components, as has been done for, e.g., heritability of intelligence in different socioeconomic levels.[99] No such attempts have been made in this thesis, mainly because no prior hypotheses regarding genetic variation-by-environment interactions have been identified. Failing to acknowledge such differences makes the estimates of genetic variation averages over different environments, thus yielding population-averaged inferences.

6.4 ETHICAL CONSIDERATIONS

According to Swedish law research on human subjects may only be carried out if the risks study participants are exposed to are counterbalanced by the scientific value of the research.[100] Observational studies, in contrast to experimental/interventional studies, do not intervene on the study participants. Therefore observational studies generally have lower risk of inflicting harm on the included subjects. However, in the studies in this thesis sensitive information on study subjects is used, this may be thought to harm the personal integrity of participants. To avoid identifiability of study participants the linkage of national registries in study I-III and the twin cohort in study IV were de-identified; the civic registration numbers were replaced by random identifiers. Since no interventions on study subjects have been made, and the results of research in this thesis may guide future research and prevention/intervention efforts, I believe that the risks for study subjects in the research in this thesis are outweighed by its scientific value. Furthermore, in study IV informed consent has been collected, and all studies have been approved by the regional ethical committee.

One of the results presented in this thesis is that some of the associations were due to familial confounding, particularly familial confounding of genetic origin. Genetics is probably not well understood by the average person, and may often be interpreted in a deterministic manner. In contrast, environmental exposures may be interpreted as more modifiable. To avoid such misunderstandings it is important to try to communicate that effect of genes, especially on complex and multi-factorial traits such as cognitive and externalizing problems, are definitely modifiable.[39] Moreover, they are not deterministic in the sense that an individual is deemed by her genes to behave in certain ways. In study III an attempt to communicate this in the discussion section has been

made. Even more important, I believe that this point should be particularly stressed when communicating with a more general population, e.g. via mainstream media.

Many fields of science suffers from under-reporting of negative, or unwanted, findings,[101] resulting in a biased view of, e.g., risk factors for certain outcomes. Many of the findings in this thesis can be viewed as "negative", i.e. reported null effects, hence the current work does not add to such under-reporting.

7 CONCLUSIONS

Novel methodological approaches were developed to be able to respond to specific aims in each study; A GEE hurdle Poisson model was implemented to separate effects on committing one offense from repeated offending. A method of identifying extended families, and analyze the data intergenerationally was developed to systematize within-family analyses and estimate genetic and environmental sources of (co)variance. The cross-lagged model was extended to include more time-points and multiple raters. These methodological advances helped to produce novel insights:

Study I. Advancing paternal age seems to cause higher violent offending rates among violent criminal re-offenders, supporting the *de novo* mutations hypothesis.

Study II and **III**. The associations between maternal SDP and pregnancy related outcomes were congruent with causal inferences, while associations with long-term cognitive and externalizing problem outcomes were congruent with non-causal inferences. The associations between SDP and cognitive and externalizing problem outcomes were mainly explained by genetic sources.

Study IV. ADHD and externalizing behaviors were, to a large extent, co-occurring. Externalizing behavior predicted ADHD early in development and ADHD predicted externalizing behavior late in development. Even though new variability was introduced throughout development the correlation between the two traits increased over time, and was largely explained by genetic sources.

8 ACKNOWLEDGEMENTS

My PhD-student time has been a wonderful journey. Learning to become an independent researcher, and realizing the multitude of talents required to be just that, has been possible because of all the individuals I have met having those attributes. I have done this, and met these people, both at the department of Medical Epidemiology and Biostatistics and at the Swedish Prison and Probation Services, between which I've divided my PhD-student time.

I would like to thank a whole bunch of people:

Paul Lichtenstein, my main supervisor. You have supported me through my many excursions to unknown territory, knowing I'd come back on track eventually. You have guided me in the, often messy, translational field of "behavioral epidemiology", where social sciences, epidemiology and behavioral genetics meet. Foremost, while doing so you have taught me how to do science.

Niklas Långström, my co-supervisor. Both as a colleague and as a boss you have helped me straighten out any question mark I've had regarding psychiatry. You've also helped me see that the social aspects of research are just as valuable as any old regression model.

Yudi Pawitan, my co-supervisor. My PhD-studies started with me re-learning the basis of statistics in your excellent course. From there on I've kept a foot in mathematics, and I've had your help in doing so.

Brian D'Onofrio, my co-supervisor overseas. Even though we're separated by the Atlantic Ocean you've been a terrific help in learning how to be a researcher. Questioning assumptions, methods and results, and forcing me to do the same, has improved all my studies.

Although supervision might have been important for my scientific development, friendship of colleagues has meant at least as much. **Thomas Frisell**, your professional friendship has taught me to question everything, and your friendship outside MEB has always added a little something extra to everyday life. **Iffat Rahman**, your intelligent and down-to-earth views on research have always kept me alert. But above all you have been, and are, a good friend, always with an ear to lend and an opinion on all that matters. **Therese Ljung**, without you the days at MEB would've been boring, thanks for always wanting to spend a moment contemplating all and everything. **Emma Frans**, thanks for all the scientific, and not-so-scientific, discussions, I think we've learnt a lot asking and answering questions together.

I've enjoyed working together with so many interesting and intelligent persons during this time; **Henrik Larsson**, **Sebastian Lundström**, **Amber Beckley**, **Zheng Chang**, **Arvid Sjölander**, **Hasse Walum**, **Alexander Viktorin**, **Amir Sariaslan**, **Martin Cederlöf**, **Jie Song**, and **Sven Sandin**. Thank you! I look forward to continued collaborations.

All my friends and colleagues here at MEB, past and present, senior and junior, remembered and forgotten, especially **Sara Christensen**, **Lisa Möller**, **Adina Feldman**, **Maria Sandberg**, **Stephanie Bonn**, **Martin Fransson**, **Anne Örtqvist**, **Patrik Magnusson**, **Erik Petterson**, **Shuyang Yao**, **Lennart Martinsson**, **Anastasia Iliadou Nyman**, **Gunilla Sonnebring**, **Camilla Ahlqvist**, **Monica Sagerstål**, **Camilla Palm**, **Christina Norrby**, **Isabelle Kizling**, **Marie Jansson**, **Karin Dellenvall**, **Tong Gong**, **Miriam Elfström**, **Carolyn Cesta**, **Linda Halldner Henriksson**, **Amy Leval**, **Camilla Gard**, **Robert Karlsson**, **Sandra Eloranta**, **Therese Andersson**, **Caroline Weibull**, **Anna Johansson**, **Robert Szulkin**, **Henrik Olsson**, **Hatef Darabi**, **Fang Fang**, **Yasmina Molero-Samuelson**, **Anna Kähler**, **Victoria Johansson**, **Simon Kyaga**, **Nancy Pedersen**, **Paul Dickman**, **Alex Ploner**, and **Christina Hultman**; You've all made my PhD-student time a joy.

All my friends at kriminalvården, **Lena Lundholm**, **Mats Forsman**, **Andrea Mogues**, **Jenny Yourstone**, **Ulrika Haggård**, **Elisabeth Nordén**, **Petra Buxrud**, **Gustav Tallving**, **Diana Lidman**, **Axel Fors**, **Bengt Holmgren**, and **Ingrid Freij**; Working with you have always been fun and inspiring.

Henrik Grönberg, head of the department during most of my PhD-student time. Thanks for seeing to that MEB had a wonderful atmosphere, and that it has been an excellent working environment.

But not all my friends are found in academia, to all of you - a big thanks for being you! Especially **Fredrik**, **Madeleine**, **Mia**, **David**, **Anna-Karin**, **Elin**, and **Jimmy**, if it weren't for you I wouldn't be who I am today.

To my (extended) family: **Roine**, **Lucas**, **Emma**, **Camilla**, **Christoffer**, **Lasse, Linnea**, **Karin**, **Åke**, **Stig**, **Bjarne**, **Inga-Lill**, and **Gunnar**; thank you for all support during these years.

The most important person in my strife for knowledge has always been my mother, **Ann-Christin**. Mum, you taught me that I can do whatever I want! Thank you for being there for me and my family.

My wonderful daughter **Hilda**, you don't know it yet, but you're what keep my feet on the ground.

My love **Stina**, you mean the world to me! Thank you for all the love and support during this time, especially at the excruciating end.

9 REFERENCES

1. Rothman KJ, Greenland S, Lash TL. *Modern epidemiology.* 3. ed. Philadelphia: Lippincott Williams & Wilkins; 2008.

2. D'Onofrio BM, Lahey BB, Turkheimer E, Lichtenstein P. Critical need for family-based, quasi-experimental designs in integrating genetic and social science research. *American journal of public health.* Oct 2013;103 Suppl 1:S46-55.

3. D'Onofrio BM, Singh AL, Iliadou A, et al. Familial confounding of the association between maternal smoking during pregnancy and offspring criminality: a population-based study in Sweden. *Archives of general psychiatry.* May 2010;67(5):529-538.

4. Begg MD, Parides MK. Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Statistics in medicine.* Aug 30 2003;22(16):2591-2602.

5. Krug EG, Mercy JA, Dahlberg LL, Zwi AB. The world report on violence and health. *Lancet.* Oct 5 2002;360(9339):1083-1088.

6. Krug EG, World Health Organization. *World report on violence and health.* Geneva: World Health Organization; 2002.

7. Regier DA, Kuhl EA, Kupfer DJ. The DSM-5: Classification and criteria changes. *World psychiatry : official journal of the World Psychiatric Association.* Jun 2013;12(2):92-98.

8. Krueger RF, Hicks BM, Patrick CJ, Carlson SR, Iacono WG, McGue M. Etiologic connections among substance dependence, antisocial behavior, and personality: modeling the externalizing spectrum. *Journal of abnormal psychology.* Aug 2002;111(3):411-424.

9. Krueger RF, Markon KE, Patrick CJ, Benning SD, Kramer MD. Linking antisocial behavior, substance use, and personality: an integrative quantitative model of the adult externalizing spectrum. *Journal of abnormal psychology.* Nov 2007;116(4):645-666.

10. World Health Organization. *Disease control priorities related to mental, neurological, developmental and substance abuse disorders.* Geneva: World Health Organization; 2006.

11. Maulik PK, Mascarenhas MN, Mathers CD, Dua T, Saxena S. Prevalence of intellectual disability: a meta-analysis of population-based studies. *Research in developmental disabilities.* Mar-Apr 2011;32(2):419-436.

12. Moffitt TE. The new look of behavioral genetics in developmental psychopathology: gene-environment interplay in antisocial behaviors. *Psychological bulletin.* Jul 2005;131(4):533-554.

13. Frisell T, Lichtenstein P, Langstrom N. Violent crime runs in families: a total population study of 12.5 million individuals. *Psychological medicine.* Jan 2011;41(1):97-105.

14. Mick E, Faraone SV. Family and genetic association studies of bipolar disorder in children. *Child and adolescent psychiatric clinics of North America.* Apr 2009;18(2):441-453, x.

15. Svensson AC, Lichtenstein P, Sandin S, Oberg S, Sullivan PF, Hultman CM. Familial aggregation of schizophrenia: the moderating effect of age at onset, parental immigration, paternal age and season of birth. *Scandinavian journal of public health.* Feb 2012;40(1):43-50.

16. Prescott CA, Kendler KS. Age at first drink and risk for alcoholism: a noncausal association. *Alcoholism, clinical and experimental research.* Jan 1999;23(1):101-107.

17. Oberg S, Cnattingius S, Sandin S, Lichtenstein P, Iliadou AN. Birth weight predicts risk of cardiovascular disease within dizygotic but not monozygotic twin pairs: a large population-based co-twin-control study. *Circulation.* Jun 21 2011;123(24):2792-2798.

18. Knopik VS, Heath AC, Jacob T, et al. Maternal alcohol use disorder and offspring ADHD: disentangling genetic and environmental effects using a children-of-twins design. *Psychological medicine.* Oct 2006;36(10):1461-1471.

19. Statistics Sweden. Multi-generation register 2012. A description of contents and quality. Statistiska centralbyrån; 2013: http://www.scb.se/statistik/_publikationer/BE9999_2012A01_BR_BE96BR1302.pdf.

20. Truett KR, Eaves LJ, Walters EE, et al. A model system for analysis of family resemblance in extended kinships of twins. *Behavior genetics.* Jan 1994;24(1):35-49.

21. Keller MC, Medland SE, Duncan LE, et al. Modeling extended twin family data I: description of the Cascade model. *Twin research and human genetics : the official journal of the International Society for Twin Studies.* Feb 2009;12(1):8-18.

22. Hatemi PK, Hibbing JR, Medland SE, et al. Not by Twins Alone: Using the Extended Family Design to Investigate Genetic Influence on Political Beliefs. *Am J Polit Sci.* Jul 2010;54(3):798-814.

23. D'Onofrio BM, Turkheimer EN, Eaves LJ, et al. The role of the Children of Twins design in elucidating causal relations between parent characteristics and child outcomes. *J Child Psychol Psyc.* Nov 2003;44(8):1130-1144.

24. Coyne CA, Langstrom N, Rickert ME, Lichtenstein P, D'Onofrio BM. Maternal age at first birth and offspring criminality: using the children of twins design to test causal hypotheses. *Development and psychopathology.* Feb 2013;25(1):17-35.

25. Sverige. Riksskatteverket. *Population registration in Sweden.* Stockholm: Swedish Tax Authority [Riksskatteverket]; 2000.

26. Magnusson PK, Almqvist C, Rahman I, et al. The Swedish Twin Registry: establishment of a biobank and other recent developments. *Twin research and human genetics : the official journal of the International Society for Twin Studies.* Feb 2013;16(1):317-329.

27. Lichtenstein P, Tuvblad C, Larsson H, Carlstrom E. The Swedish Twin study of CHild and Adolescent Development: the TCHAD-study. *Twin research and human genetics : the official journal of the International Society for Twin Studies.* Feb 2007;10(1):67-73.

28. Centre for Epidemiology. The Swedish Medical Birth Register -A summary of content and quality. The national board of health and welfare 2003: http://www.socialstyrelsen.se/Lists/Artikelkatalog/Attachments/10655/2003-112-3_20031123.pdf.

29. Carlsen B. *Cognitive abilities – aspects of structure, process and measurement*, Göteborgs universitet; 2000.

30. Sweden S. Utbildningsregistret. Statistics Sweden; 2012: http://www.scb.se/Statistik/UF/UF0506/_dokument/UF0506_DO_2012.pdf.

31. Kendler KS, Gardner CO, Lichtenstein P. A developmental twin study of symptoms of anxiety and depression: evidence for genetic innovation and attenuation. *Psychological medicine.* Nov 2008;38(11):1567-1575.

32. Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *American journal of epidemiology.* Jan 15 2002;155(2):176-184.

33. Pawitan Y. *In all likelihood : statistical modelling and inference using likelihood.* Oxford: Clarendon; 2001.

34. Neuhaus JM, McCulloch CE. Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *J Roy Stat Soc B.* 2006;68:859-872.

35. Wright S. The Method of Path Coefficients. 1934/09 1934:161-215.

36. McGue M, Osler M, Christensen K. Causal Inference and Observational Research: The Utility of Twins. *Perspectives on psychological science : a journal of the Association for Psychological Science.* Oct 11 2010;5(5):546-556.

37. Frisell T, Oberg S, Kuja-Halkola R, Sjolander A. Sibling comparison designs: bias from non-shared confounders and measurement error. *Epidemiology.* Sep 2012;23(5):713-720.

38. Neale MC, Maes HH. Methodology for Genetic Studies of Twins and Families. Dordrecht, The Netherlands: Kluwer Academic Publishers B.V.; 2002: http://ibgwww.colorado.edu/workshop2002/cdrom/HTML/BOOK/book2002c.pdf.

39. Plomin R. *Behavioral genetics.* 6 ed. New York: Worth Publishers; 2013.

40. Falconer DS, Mackay TFC. *Introduction to quantitative genetics.* 4. ed. Harlow: Longman; 1996.

41. Fisher RA. The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Philosophical Transactions of the Royal Society of Edinburgh.* 1918;52:399-433.

42. Wright S. Systems of Mating. I. the Biometric Relations between Parent and Offspring. *Genetics.* Mar 1921;6(2):111-123.

43. Crick F. Central dogma of molecular biology. *Nature.* Aug 8 1970;227(5258):561-563.

44. Pearson H. Genetics: what is a gene? *Nature.* May 25 2006;441(7092):398-401.

45. Consortium EP, Bernstein BE, Birney E, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature.* Sep 6 2012;489(7414):57-74.

46. Bollen KA. *Structural equations with latent variables.* New York: Wiley; 1989.

47. Sewell G. *Computational methods of linear algebra.* 2. ed. Hoboken: Wiley; 2005.

48. Loehlin JC. The Cholesky approach: A cautionary note. *Behavior genetics.* Jan 1996;26(1):65-69.

49. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era--concepts and misconceptions. *Nature reviews. Genetics.* Apr 2008;9(4):255-266.

50. Rao DC, Morton NE, Yee S. Analysis of family resemblance. II. A linear model for familial correlation. *American journal of human genetics.* May 1974;26(3):331-359.

51. Eaves LJ, Fulker DW, Heath AC. The effects of social homogamy and cultural inheritance on the covariances of twins and their parents: a LISREL model. *Behavior genetics.* Jan 1989;19(1):113-122.

52. Frans EM, Sandin S, Reichenberg A, Lichtenstein P, Langstrom N, Hultman CM. Advancing paternal age and bipolar disorder. *Archives of general psychiatry.* Sep 2008;65(9):1034-1040.

53. Malaspina D, Harlap S, Fennig S, et al. Advancing paternal age and the risk of schizophrenia. *Archives of general psychiatry.* Apr 2001;58(4):361-367.

54. Reichenberg A, Gross R, Weiser M, et al. Advancing paternal age and autism. *Archives of general psychiatry.* Sep 2006;63(9):1026-1032.

55. Fazel S, Gulati G, Linsell L, Geddes JR, Grann M. Schizophrenia and violence: systematic review and meta-analysis. *PLoS medicine.* Aug 2009;6(8):e1000120.

56. Fazel S, Lichtenstein P, Grann M, Goodwin GM, Langstrom N. Bipolar disorder and violent crime: new evidence from population-based longitudinal studies and systematic review. *Archives of general psychiatry.* Sep 2010;67(9):931-938.

57. Crow JF. Spontaneous mutation in man. *Mutation research.* Jul 1999;437(1):5-9.

58. Moffitt TE. Adolescence-limited and life-course-persistent antisocial behavior: a developmental taxonomy. *Psychological review.* Oct 1993;100(4):674-701.

59. Min YY, Agresti A. Random effect models for repeated measures of zero-inflated count data. *Stat Model.* Mar 2005;5(1):1-19.

60. Hinshaw SP, Lahey BB, Hart EL. Issues of Taxonomy and Comorbidity in the Development of Conduct Disorder. *Development and psychopathology.* Win-Spr 1993;5(1-2):31-49.

61. Patterson GR, DeBaryshe BD, Ramsey E. A developmental perspective on antisocial behavior. *Am Psychol.* Feb 1989;44(2):329-335.

62. Andres RL, Day MC. Perinatal complications associated with maternal tobacco use. *Seminars in neonatology : SN.* Aug 2000;5(3):231-241.

63. Cnattingius S. The epidemiology of smoking during pregnancy: smoking prevalence, maternal characteristics, and pregnancy outcomes. *Nicotine & tobacco research : official journal of the Society for Research on Nicotine and Tobacco.* Apr 2004;6 Suppl 2:S125-140.

64. Cnattingius S, Haglund B. Decreasing smoking prevalence during pregnancy in Sweden: the effect on small-for-gestational-age births. *American journal of public health.* Mar 1997;87(3):410-413.

65. D'Onofrio BM, Singh AL, Iliadou A, et al. A quasi-experimental study of maternal smoking during pregnancy and offspring academic achievement. *Child development.* Jan-Feb 2010;81(1):80-100.

66. D'Onofrio BM, Van Hulle CA, Waldman ID, et al. Smoking during pregnancy and offspring externalizing problems: an exploration of genetic and environmental confounds. *Development and psychopathology.* Winter 2008;20(1):139-164.

67. Ernst M, Moolchan ET, Robinson ML. Behavioral and neural consequences of prenatal exposure to nicotine. *Journal of the American Academy of Child and Adolescent Psychiatry.* Jun 2001;40(6):630-641.

68. Haglund B, Cnattingius S. Cigarette smoking as a risk factor for sudden infant death syndrome: a population-based study. *American journal of public health.* Jan 1990;80(1):29-32.

69. Huijbregts SC, Seguin JR, Zoccolillo M, Boivin M, Tremblay RE. Associations of maternal prenatal smoking with early childhood physical aggression, hyperactivity-impulsivity, and their co-occurrence. *Journal of abnormal child psychology.* Apr 2007;35(2):203-215.

70. Kafouri S, Leonard G, Perron M, et al. Maternal cigarette smoking during pregnancy and cognitive performance in adolescence. *International journal of epidemiology.* Feb 2009;38(1):158-172.

71. Knopik VS. Maternal smoking during pregnancy and child outcomes: real or spurious effect? *Developmental neuropsychology.* 2009;34(1):1-36.

72. Lambe M, Hultman C, Torrang A, Maccabe J, Cnattingius S. Maternal smoking during pregnancy and school performance at age 15. *Epidemiology.* Sep 2006;17(5):524-530.

73. Lundberg F, Cnattingius S, D'Onofrio B, et al. Maternal smoking during pregnancy and intellectual performance in young adult Swedish male offspring. *Paediatric and perinatal epidemiology.* Jan 2010;24(1):79-87.

74. Vaglenova J, Birru S, Pandiella NM, Breese CR. An assessment of the long-term developmental and behavioral teratogenicity of prenatal nicotine exposure. *Behavioural brain research.* Apr 2 2004;150(1-2):159-170.

75. Muthén LK, Muthèn BO. *Mplus User's Guide.* Los Angeles, CA: Muthén & Muthén; 1998-2207.

76. Thapar A, Rice F, Hay D, et al. Prenatal smoking might not cause attention-deficit/hyperactivity disorder: evidence from a novel design. *Biological psychiatry.* Oct 15 2009;66(8):722-727.

77. Gaysina D, Fergusson DM, Leve LD, et al. Maternal smoking during pregnancy and offspring conduct problems: evidence from 3 independent genetically sensitive research designs. *JAMA psychiatry.* Sep 2013;70(9):956-963.

78. Slotkin TA. Maternal smoking and conduct disorder in the offspring. *JAMA psychiatry.* Sep 2013;70(9):901-902.

79. Abbott LC, Winzer-Serhan UH. Smoking during pregnancy: lessons learned from epidemiological studies and experimental studies using animal models. *Critical reviews in toxicology.* Apr 2012;42(4):279-303.

80. Young S, Thome J. ADHD and offenders. *The world journal of biological psychiatry : the official journal of the World Federation of Societies of Biological Psychiatry.* Sep 2011;12 Suppl 1:124-128.

81. Wilens E, Faraone SV, Biederman J. Attention-deficit/hyperactivity disorder in adults. *Jama-J Am Med Assoc.* Aug 4 2004;292(5):619-623.

82. Burt SA. Are there meaningful etiological differences within antisocial behavior? Results of a meta-analysis. *Clinical psychology review.* Mar 2009;29(2):163-178.

83. Burt SA. Rethinking environmental contributions to child and adolescent psychopathology: a meta-analysis of shared environmental influences. *Psychological bulletin.* Jul 2009;135(4):608-637.

84. Faraone SV, Perlis RH, Doyle AE, et al. Molecular genetics of attention-deficit/hyperactivity disorder. *Biological psychiatry.* Jun 1 2005;57(11):1313-1323.

85. Rhee SH, Waldman ID. Genetic and environmental influences on antisocial behavior: a meta-analysis of twin and adoption studies. *Psychol Bull.* May 2002;128(3):490-529.

86. Chang Z, Lichtenstein P, Asherson PJ, Larsson H. Developmental twin study of attention problems: high heritabilities throughout development. *JAMA psychiatry.* Mar 2013;70(3):311-318.

87. Wichers M, Gardner C, Maes HH, Lichtenstein P, Larsson H, Kendler KS. Genetic innovation and stability in externalizing problem behavior across development: a multi-informant twin study. *Behavior genetics.* May 2013;43(3):191-201.

88. Burt SA, McGue M, Krueger RF, Iacono WG. How are parent-child conflict and childhood externalizing symptoms related over time? Results from a genetically informative cross-lagged study. *Development and psychopathology.* Winter 2005;17(1):145-165.

89. Lahey BB, Loeber R, Burke J, Rathouz PJ, McBurnett K. Waxing and waning in concert: dynamic comorbidity of conduct disorder with other disruptive and emotional problems over 7 years among clinic-referred boys. *Journal of abnormal psychology.* Nov 2002;111(4):556-567.

90. Lichtenstein P, Yip BH, Bjork C, et al. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet.* Jan 17 2009;373(9659):234-239.

91. Boker S, Neale M, Maes H, et al. OpenMx: An Open Source Extended Structural Equation Modeling Framework. *Psychometrika.* Apr 2011;76(2):306-317.

92. Boker SM, Neale MC, Maes HH, et al. OpenMx User Guide, Release 1.3;. 2012.

93. *R: A language and environment for statistical computing* [computer program]. Vienna, Austria: R Foundation for Statistical Computing; 2013.

94. Chang Z, Lichtenstein P, D'Onofrio BM, Kuja-Halkola R, Sjolander A, Larsson H. Maternal age at childbirth and risk for ADHD in offspring: A genetically informative study. Manuscript.

95. Kendler KS, Neale MC, Kessler RC, Heath AC, Eaves LJ. A test of the equal-environment assumption in twin studies of psychiatric illness. *Behavior genetics.* Jan 1993;23(1):21-27.

96. Derks EM, Dolan CV, Boomsma DI. A test of the equal environment assumption (EEA) in multivariate twin studies. *Twin research and human genetics : the official journal of the International Society for Twin Studies.* Jun 2006;9(3):403-411.

97. Keller MC, Medland SE, Duncan LE. Are extended twin family designs worth the trouble? A comparison of the bias, precision, and accuracy of parameters estimated in four twin family models. *Behavior genetics.* May 2010;40(3):377-393.

98. Frisell T, Pawitan Y, Langstrom N, Lichtenstein P. Heritability, assortative mating and gender differences in violent crime: results from a total population sample using twin, adoption, and sibling models. *Behavior genetics.* Jan 2012;42(1):3-18.

99. Turkheimer E, Haley A, Waldron M, D'Onofrio B, Gottesman, II. Socioeconomic status modifies heritability of IQ in young children. *Psychological science.* Nov 2003;14(6):623-628.

100. Swedish Riksdag. Lag (2003:460) om etikprövning av forskning som avser människor. 2003.

101. Dickersin K, Chalmers I. Recognizing, investigating and dealing with incomplete and biased reporting of clinical research: from Francis Bacon to the WHO. *Journal of the Royal Society of Medicine.* Dec 2011;104(12):532-538.