



# Long working weeks and health

A research memoir



**Harald Hannerz**

Doctoral dissertation, August 2023

© Harald Hannerz

Cover pictures by  
William Max Hannerz (front) and Alexander Storm Hannerz (back)

The Faculty of Health and Medical Sciences at the University of Copenhagen has accepted this dissertation for public defence for the doctoral degree in medicine.

Copenhagen, 6 March 2023.

Bente Merete Stallknecht  
Head of Faculty

The public defence will take place Monday 21.8.2023 at 13.30 in room number 7-0-34 at CSS, Øster Farimagsgade 5, 1353 Copenhagen K.

*For my grandchildren*

# Contents

Preface .....	3
List of acronyms and abbreviations .....	4
<i>1. Introduction</i> .....	6
<i>2. Protocol-based examination of health and safety in relation to weekly working hours among full-time employees in the general population of Denmark</i> .....	9
2.1 Methods.....	12
2.1.1. Data sources.....	12
2.1.2. Exposure .....	13
2.1.3. Clinical endpoints .....	14
2.1.4. Covariates.....	16
2.1.5. Strategies to ensure that family wise error rates are $\leq 0.05$ .....	17
2.1.6. Statistical methods .....	17
2.1.7. Feasibility studies.....	17
2.1.8. Sensitivity analyses .....	18
2.1.9. Protocol publication dates in relation to dates of establishment of the research data sets.....	22
2.2. Results .....	23
2.3. Discussion .....	26
2.3.1. Main findings .....	26
2.3.2. Methodological considerations .....	27
2.3.3. Generalisability.....	30
2.3.4. Previous research.....	30
2.3.4. Conclusions .....	36
<i>3. Methodological perspectives</i> .....	37
3.1. Adverse side effects of statistical significance testing.....	38
3.1.1. Re. multiple testing .....	39
3.1.2. Re. post-hoc testing .....	40
3.1.3. Re. insufficient statistical power .....	42
3.1.4. Another side effect .....	44
3.2. Initiatives to improve quality and credibility of statistical analyses and reports .....	45
3.2.1. Initiatives by journal editors .....	47
3.2.2. Local initiatives by the statistical society at NRCWE.....	50
3.2.3. Personal initiatives .....	54
3.2.4. Discussion.....	56

3.3. Statistical power calculations as a means of reducing the risk of being fooled by publication bias and faulty statistical significance declarations .....	61
3.3.1. A systematic review on work-related psychosocial factors and ischaemic heart disease .	62
3.3.2. Another example of power calculations in a literature overview .....	64
3.3.3. Discussion.....	65
4. <i>Concluding remarks and recommendations</i> .....	67
<i>English summary</i> .....	69
<i>Dansk resumé</i> .....	70
<i>References</i> .....	71

## **Preface**

The summarising chapters of the present dissertation were written by me as a leisure-time activity from July 2017 to July 2020. The attached research papers, which I co-authored as a research scientist at the National Research Centre for the Working Environment (NRCWE) in Denmark, were published in the time period 2014 – 2020.

The research associated with the papers was supported by external financing, partly through a grant from the German Federal Institute for Occupational Safety and Health, partly through a grant from the Velliv Association and partly through two grants from the Danish Work Environment Research Foundation.

Permissions to reproduce the papers have been granted by each of the concerned copyright owners.

Ole Melkevik, Pia Dukholm and Sannie Vester Thorsen are hereby thanked for having contributed with pictures and drawings to the summarising chapters of the dissertation.

Damien Hannerz and Helle Soll-Johanning are hereby thanked for helping me to translate the summary of the dissertation into Danish.

Each of the following researchers is hereby thanked for having contributed as a co-author to one or more of the attached articles: Martin Lindhardt Nielsen, Karen Albertsen, Kim Dalhoff, Hermann Burr, Ute Latza, Ann Dyreborg Larsen, Anne Helene Garde, Jan Hyld Pejtersen, Helle Soll-Johanning, Simone Visbjerg Møller, Jens Peter Bonde, Johnni Hansen, Johnny Dyreborg, Henrik Albert Kolstad, Alba Fishta, Nanna Hurwitz Eller, Åse Marie Hansen, Hans Bay and Sannie Vester Thorsen.

The Danish Health Data Authority, The Danish National Institute for Health Data and Disease Control, NRCWE, and Statistics Denmark. Are hereby thanked for having contributed with data to one or more of the covered research projects.

Some of the data had been obtained through questionnaire-based surveys, and in that respect I would like to extend my gratitude to all survey participants who responded to the questionnaires.

I would also like to extend my gratitude to all the editorial staff and peer-reviewers that were involved in the publication of the attached research papers.

Finally, I would like to extend my gratitude to NRCWE for having provided me with an effective interdisciplinary research environment, with a multitude of inspiring and competent co-workers, throughout the above mentioned time periods.

Harald Hannerz  
Copenhagen July 2020

## List of acronyms and abbreviations

ATC:	Anatomical Therapeutic Chemical
ARRIVE:	Animals in Research: Reporting In Vivo Experiments
BMI:	Body mass index
BMJ:	British Medical Journal
CDC:	Centers for Disease Control and Prevention
Cf:	Confer
CI:	Confidence interval
CL:	Confidence limit
CMS:	The Copenhagen Male Study
CONSORT:	CONsolidated Standards of Reporting Trials
COPE:	The Committee on Publication Ethics
COPSOQ:	The Copenhagen Psychosocial Questionnaire
CVD:	Cardiovascular disease
DANES:	The Danish National Working Environment Survey
DWECS:	The Danish Work Environment Cohort Study
E.g.:	Exempli gratia
EQUATOR:	Enhancing the QUALity and Transparency Of health Research
ESeC:	The European Socioeconomic Classification
Et al.:	Et alia
EU:	European Union
EUWTD:	European Working Time Directive
H:	Hours
Harking	Hypothesizing After the Results are Known
ICD:	International classification of diseases
ICMJE:	The International Committee of Journal Editors
I.e.:	Id est
IHD:	Ischaemic heart disease
IHS:	Iskæmisk hjertesygdom
IPD:	Individual participant data
LTPA:	Leisure time physical activity
MOOSE:	Meta-analysis Of Observational Studies in Epidemiology
NHST:	Null Hypothesis Significance Tests
NIMS:	The Northern Ireland Mortality Study
NRCWE:	The National Research Centre for the Working Environment
OR:	Odds ratio
PRISMA:	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QRP:	Questionable research practices
RCT:	Randomised controlled trial
Re:	Regarding
RR:	Rate ratio
SES:	Socioeconomic status
SOCIO:	Statistics Denmark's socioeconomic classification
STARD:	Standards for Reporting Diagnostic accuracy studies



StdErr: Standard error  
STROBE: STrengthening the Reporting of OBservational studies in Epidemiology  
TREND: Transparent Reporting of Evaluations with Nonrandomized Designs  
USA: United States of America  
UK: United Kingdom  
Vs: Versus  
WHO: World Health Organisation  
WWH: Weekly working hours

# 1. Introduction

A first part of the present dissertation (chapter 2) synthesizes results and statistical conclusions from a series of cohort studies aimed at estimating prospective associations between weekly working hours and ill health among full-time employees in the general population of Denmark. A second part (chapter 3) elaborates on the methodological principles on which the research was based. The concluding remarks and recommendations of the thesis are given in chapter 4.

The dissertation is structured as a compilation thesis i.e. a collection of published research papers, with a summarising set of introductory chapters. The papers are divided into two sections, one for the papers that are addressed in chapter 2 and one for the papers that are addressed in chapter 3. The following papers are included:

## ***Section A: Protocol-based examination of health and safety in relation to weekly working hours among employees in the general population of Denmark***

Psychotropic medicine study      Paper I  
**Hannerz H**, Albertsen K. Long working hours and subsequent use of psychotropic medicine: a study protocol. JMIR Res Protoc. 2014 Sep 19;3(3):e51. doi: 10.2196/resprot.3301.

Paper II  
**Hannerz H**, Albertsen K. Long working hours and use of psychotropic medicine: a follow-up study with register linkage. Scand J Work Environ Health. 2016;42(2):153-61.

Paper III  
**Hannerz H**, Albertsen K, Nielsen ML, Garde AH. Prospective Associations Between Working Time Arrangements and Psychiatric Treatment in Denmark: Protocol for a Cohort Study. JMIR Res Protoc. 2020;9(6):e18236. Published 2020 Jun 15. doi:10.2196/18236

Paper IV  
**Hannerz H**, Albertsen K, Nielsen ML, Garde AH. Long working hours and psychiatric treatment: A Danish follow-up study. Scand J Work Environ Health. 2020 Nov 17:3936. doi: 10.5271/sjweh.3936. Epub ahead of print.

Ischaemic Heart Disease study      Paper V (feasibility study)  
**Hannerz H**, Dalhoff K, Burr H, Latza U. Correlation between relative rates of hospital treatment or death due to ischaemic heart disease (IHD) and of IHD-related medication among socio-occupational and economic activities groups in Denmark, 1996-2005. Int J Occup Med Environ Health. 2014 Aug;27(4):536-46.

Paper VI

**Hannerz H**, Larsen AD, Garde AH. Working Time Arrangements as Potential Risk Factors for Ischemic Heart Disease among Workers in Denmark: A Study Protocol. JMIR Res Protoc. 2016 Jun 22;5(2):e130.

Paper VII

**Hannerz H**, Larsen AD, Garde AH. Long weekly working hours and ischaemic heart disease: a follow-up study among 145 861 randomly selected workers in Denmark. BMJ Open. 2018 Jun 15;8(6):e019807. doi: 10.1136/bmjopen-2017-019807.

Stroke study

Paper VIII

**Hannerz H**, Albertsen K, Burr H, Nielsen ML, Garde AH, Larsen AD, Pejtersen JH. (2017): The association between long working hours and stroke in the general workforce of Denmark – a study protocol. figshare.  
<https://doi.org/10.6084/m9.figshare.4684951.v1>.

Paper IX

**Hannerz H**, Albertsen K, Burr H, Nielsen ML, Garde AH, Larsen AD, Pejtersen JH. Long working hours and stroke among employees in the general workforce of Denmark. Scand J Public Health. 2018 May;46(3):368-374.

All-cause mortality study

Paper X

**Hannerz H**, Soll-Johanning H. (2017): General mortality in relation to the EU Working Time Directive: a Danish study protocol. figshare.  
<https://doi.org/10.6084/m9.figshare.5297062.v1>

Paper XI

**Hannerz H**, Soll-Johanning H. Working hours and all-cause mortality in relation to the EU Working Time Directive: a Danish cohort study. Eur J Public Health. 2018 Oct 1;28(5):810-814.

Injury study

Paper XII

Larsen AD, **Hannerz H**, Møller SV, Dyreborg J, Bonde JP, Hansen J, Kolstad HA, Hansen ÅM, Garde AH. (2016): Study protocol for examining long working hours and night work as risk factors for injuries. figshare.  
<https://doi.org/10.6084/m9.figshare.3408220.v1>

Paper XIII

Larsen AD, **Hannerz H**, Møller SV, Dyreborg J, Bonde JP, Hansen J, Kolstad HA, Hansen ÅM, Garde AH. Night work, long work weeks, and risk of accidental injuries. A register-based study. Scand J Work Environ Health. 2017 Nov 1;43(6):578-586.

***Section B: Statistical power calculations as a means of reducing the risk of being fooled by random variation***

Paper XIV

**Hannerz H**, Thorsen SV, Bay H. Re: Sainani K. Interpreting "null" results. *PMR*. 2018 May;10(5):562-563.

Paper XV

Pejtersen JH, Burr H, **Hannerz H**, Fishta A, Hurwitz NE. Update on work-related psychosocial factors and the development of ischemic heart disease: a systematic review. *Cardiol Rev*. 2015;23(2):94-8.

## ***2. Protocol-based examination of health and safety in relation to weekly working hours among full-time employees in the general population of Denmark***



Figure 2.1. Drawing by Sannie Vester Thorsen at the National Research Centre for the Working Environment, Denmark, 2014

*“Everyone has the right to rest and leisure, including reasonable limitation of working hours and periodic holidays with pay “.* [United Nations. Universal Declaration of Human Rights, Article 24]

Long working hours have often been associated with short sleep [cf. Hayashi et al, 1996; Kageyama et al., 2001; Sasaki et al., 1999] and fatigue due to insufficient recovery between work shifts [cf. Hayashi et al, 1996, Sasaki et al., 1999; Iwasaki et al., 1998; Proctor et al., 1996; Spurgeon et al., 1997]. It has, moreover been shown that short sleep and fatigue are risk factors for psychiatric disorders [Glozier et al., 2010; Breslau et al., 1996; Chang et al., 1997; Szklo-Coxe et al., 2010; Harvey et al., 2009; Huibers et al., 2007; Skapinakis et al., 2004], accidental injuries [Uehli et al., 2014], hypertension [Wang et al., 2015], ischaemic heart disease [Cappuccio et al., 2011] and stroke [Koo et al., 2018].

Long working hours are, however, all else being equal, also associated with an increased income and thereby a decreased risk or intensity of financial strain, which is a very important stressor among contemporary employees [cf. Sinclair and Cheung, 2016]. It has been shown that financial strain is associated with an increased risk of psychiatric disorders [Weich and Lewis, 1998; Dunn et al., 2008; Sareen et al., 2011], hypertension [Steptoe et al., 2005] and acute myocardial infarction [Rosengren et al., 2004]. It has also been shown that there is a very clear and important inverse association between income and risks of psychiatric disorders [Sareen et al., 2011], stroke [Toivanen, 2011] and IHD [Andersen et al., 2003]; the higher the income the lower the risk. There are, in other words, both advantages and disadvantages associated with long working hours, which suggests a U-shaped relation, where the risk of ill health decreases with weekly working hours up to a point where they become too long to allow sufficient sleep and recovery.

Another advantage of long working hours is that they usually contribute positively to a nation’s gross domestic product, and thereby create more governmental funds that may be spent on health care and other health promoting initiatives [Swift, 2011]. Such indirect effects of long working hours are however beyond the scope of the present work.

Long working hours are recognised as a risk factor for death due to mental and cardiovascular disorders by the governments of Japan, Taiwan and South Korea [cf. Ke, 2012; Eguchi et al., 2016; Lin et al., 2017]. They are also recognised as occupational safety and health hazards by the European Parliament and the Council of the European Union. The issue is addressed in the EU Working Time Directive, which, among other things, stipulates that member states shall take the measures necessary to ensure that the average working time for each 7-day period, including overtime, does not exceed 48 hours. The purpose of the directive, which has been in force since 1993, with slight amendments in the year 2000 and 2003, is to protect the safety and health of workers [EU, 2003].

The cut-point at 48 hours is, however, arbitrary and when the directive was enacted, it had not been established that a threshold at 48 working hours a week would be low enough to protect against excess rates of ill health from long workweeks.

The aim of the present work was to examine mortality and morbidity rates among fulltime employees in the general workforce of Denmark, as a function of their usual weekly working hours.

The following outcome measures were regarded: Redeemed prescription for psychotropic drugs, psychiatric hospital treatment due to mood, anxiety or stress-related disease, redeemed prescription for antihypertensive drugs, hospital treatment or death due to ischaemic heart disease, hospital treatment or death due to stroke, all-cause mortality and accidental injuries.

In all analyses except the one dealing with stroke morbidity, the participants' usual weekly working hours were treated as a categorical variable, with 32 – 40 hours (the reference) to represent normal weekly working hours, 41 – 48 hours to represent overtime work that lies within the limits of the working time directive, and 49 – 100 hours to represent overtime work beyond the threshold of the directive. Age, gender and socio-economic status were taken into account, and a series of sub-analyses were conducted to test if hypothetical relationships between weekly working hours and the examined outcome measures depended on gender and socio-economic status.

As mentioned in Paper VII, the above working hour categorisation “facilitates interpretation of the results in relation to the EU Working Time Directive in a Danish context, in accordance with the following arguments:

- If a rate ratio is statistically significantly high among workers with 41–48 working hours a week, then it might be of practical importance, since it suggests that the 48 hours threshold of the EU Working Time Directive may need to be lowered to protect against [adverse health effects] from long working hours.
- If a rate ratio is statistically significantly low among workers with more than 48 working hours a week then it might be of practical importance, since it suggests that the 48 hours threshold of the EU Working Time Directive either is unnecessary or unnecessarily low (when it comes to protecting employees against [adverse health effects] from long working hours).
- If a rate ratio is statistically significantly high among workers with more than 48 working hours a week but not among employees with 41–48 working hours a week, then the results do not indicate any need to change the threshold of the Working Time Directive. The elevated rate ratio may, however, be of practical importance from a public health perspective, since it identifies a group of people who might be in need of health promotion.”

All analyses were preceded by the publication of a detailed study protocol in which all hypotheses and statistical models were completely defined before we looked at any relation between the exposure data and the outcome measures of the study.

The studies on the association between working hours and psychotropic drugs are described in Paper I (study protocol), Paper II (results), Paper III (study protocol), Paper IV (results). The study on the association with psychiatric hospital treatment due to mood, anxiety or stress-related disease is described in Paper III (study protocol) and Paper IV (results). The study on the association with antihypertensive drug usage and ischaemic heart disease is described in Paper V (feasibility study), Paper VI (study protocol) and Paper VII (results). The study on the association with stroke is described in Paper VIII (study protocol) and Paper IX (results). The study on the association with all-cause mortality is described in Paper X (study protocol) and Paper XI (results). The study on the association with accidental injuries is described in Paper XII (study protocol) and Paper XIII (results).

The statistical analysis plans for the first study on psychotropic drugs as well as the studies on ischaemic heart disease, antihypertensive drugs, and accidental injuries were published before the exposure data of the studies had been linked to any health register data. These analyses were, in other words, completely blinded.

The statistical analysis plan for the second study on psychotropic drugs as well as the studies on psychiatric hospital treatment, all-cause mortality and stroke were published before we had looked at any relation between the concerned exposure data and psychotropic drugs, psychiatric hospital treatment, all-cause mortality or stroke. They were, however, written and published after we had looked at the relation between the concerned exposure data and ischaemic heart disease, antihypertensive drugs and accidental injuries.

A synopsis of the studies will be given in the present chapter.

## **2.1 Methods**

The investigations were designed as prospective cohort studies. The target population was defined as employees in the general working population of Denmark. Information on working time arrangements came from questionnaire data obtained in a series of randomly selected samples of the target population that were drawn in the time period 1995 - 2013. The follow-ups were accomplished through national registers. The linkage between the questionnaire data and the register data were based on the participants' personal identification numbers. The participants were censored if they died or emigrated or the study-period ended before they reached the particular end-point of the study. Participants who had experienced the end-point of the study during the calendar year preceding the start of the follow-up were excluded from the analysis, in all studies except the one dealing with accidental injuries. The average follow-up was 1.8 years for accidental injuries, 3.9 years for psychotropic drugs, 4.8 years for psychiatric hospital treatment, 6.7 years for antihypertensive drugs and 7.7 years for ischaemic heart disease, stroke and all-cause mortality.

### **2.1.1. Data sources**

The questionnaire data for the first study on the association between working hours and psychotropic drug usage were obtained from "the Copenhagen Psychosocial Questionnaire (COPSOQ) study sample of 2004 [Pejtersen et al., 2010], the Danish National Working Environment Survey (DANES) of 2008 [Thorsen et al., 2013], and the Danish Work Environment Cohort Study (DWECS) of 1995 [Burr et al., 2003], 2000 [Burr et al., 2003], 2005 [Feveile et al., 2007], and 2010 [Bach et al, 2011]." Further details on these surveys are given in Paper I.

The questionnaire data for the remaining studies were obtained from the Danish Labour Force Surveys of 1999–2013, which, in accordance with an EU-directive, were based on random samples of 15 to 74-year-old people in the Danish population. New samples were drawn each quarter of the years and each of the drawn individuals were invited to participate in up to four interview rounds over a period of 1.5 years [Statistics Denmark, 2019]. Further details on these surveys are given in Paper VI.



The following national registers were used: The central person register (for information on dates of birth deaths and migrations) [Pedersen, 2011], the employment classification module (for information on occupation, industry and socio-economic status) [Petersson et al., 2011], the national patient register (for information on dates and diagnoses of hospital contacts) [Lynge et al., 2011], the national cause of death register (for information on death diagnoses) [Helweg-Larsen, 2011], the national prescription register (for information on redeemed prescriptions) [Kildemoes et al., 2011] and the psychiatric central research register. (for information on psychiatric hospital treatments) [Mors et al., 2011]. Further details on these registers are given in Paper III and VI.

### **2.1.2. Exposure**

The exposure variable “weekly working hours” was based on the sum of the hours worked in the primary and secondary jobs and, as previously mentioned, categorised into 32 – 40, 41 – 48 and > 48 working hours a week. The only exception from this rule was the stroke study (Paper VIII and IX), in which the working hours were categorised into 35 – 40, 41 – 48, 49 – 54 and > 54 hours a week. The reason for the departure from the rule was that the stroke study aimed at replicating a previous study [Kivimaki et al., 2015a], in which the latter categorisation was used.

The working hours questions that were used in the first study of psychotropic drug usage are described as follows in Paper I: “The COPSOQ, the DANES, and the DWECS surveys contain person-based information on weekly working hours, calculated by adding the hours worked in secondary jobs to the ones in a primary job. The wording of the questions differs, however, slightly between the various questionnaires. The DWECS questionnaires of 1995, 2000, and 2005 ask for weekly working hours in current jobs or (if the person is momentarily out of work) in the last held job. DWECS 2010 asks for current weekly working hours without further specification. COPSOQ and DANES ask for average working hours during the one-year period preceding the time of the interview. The COPSOQ questionnaire only allowed participants to report between 0 and 99 working hours per week, while the other questionnaires allowed an unlimited number of hours. Another peculiarity of the COPSOQ questionnaire is that it uses a single question to ask for the combined number of hours worked in primary and secondary jobs, while the other questionnaires use one question for the number of hours worked in the primary job and another one for the hours worked in secondary jobs.” The exact wordings of the used questions are given in the supplemental appendices of Paper I.

The working hour questions that were used in the remaining studies are described as follows in Paper VII “The labour force surveys gather person-based information on WWH (weekly working hours), calculated by adding the hours worked in secondary jobs to the ones worked in a primary job. The participants are asked first how many hours they usually work and then how many hours they worked during the reference week (a predetermined work week, which occurred 1–4 weeks prior to the interview). ... The questions used to gather this information have changed slightly with time. Before 2001 there was no mention of whether meal breaks should be counted as working hours. During 2001–2006 all participants were instructed to exclude meal breaks when they counted their work hours. As of 2007 the time used for meal breaks were to be counted if the person got paid while eating and excluded otherwise. From 2001, the average of the actual WWH during the 4 weeks preceding the interview was used as substitute for usual WWH among participant whose

working hours varied a lot. In the time-period 2001–2006, weeks in which the participant was absent due to for example, holidays, vacation or sick leave were to be disregarded when this average was calculated, but from 2007 all of the 4 weeks were to be included in the calculation.”

### **2.1.3. Clinical endpoints**

The following clinical endpoints were regarded as primary outcomes: Redeemed prescription for psychotropic drugs; psychiatric hospital treatment due to mood, anxiety or stress-related disease, redeemed prescription for antihypertensive drugs; hospital treatment or death due to ischaemic heart disease; hospital treatment or death due to stroke; hospital treatment or death due to accidental injuries; death from any cause.

The endpoint “redeemed prescriptions for psychotropic drugs” was met if and when the participant redeemed a prescription for a drug in the ATC-code category N05 (psycholeptica) or N06 (psychoanaleptica). “The psycholeptic category contains antipsychotics, anxiolytics, hypnotics, and sedatives, while the psychoanaleptic category contains antidepressants, psychostimulants, and antidementia drugs” [Paper I and II].

The endpoint “psychiatric hospital treatment due to mood, anxiety or stress-related disease” was met if and when the participant received hospital treatment with a principal diagnosis in the ICD-10 code category: F30 manic episode, F31 bipolar affective disorder, F32 depressive episode, F33 recurrent depressive disorder, F34 persistent mood [affective] disorders, F38 other mood [affective] disorders, F39 unspecified mood [affective] disorder, F40 phobic anxiety disorders, F41 other anxiety disorders, and F43 reaction to severe stress, and adjustment disorders. [Paper III and IV].

The endpoint “redeemed prescriptions for antihypertensive drugs” was met if and when the participant redeemed a prescription for a drug in the ATC-code category C02 antihypertensives, C03 diuretics, C07 alpha- and beta-blockers, C08 calcium channel blockers or C09 angiotensin-converting enzyme inhibitors and angiotensin-II antagonists [Paper VI and VII].

The endpoint “hospital treatment or death due to ischaemic heart disease” was met if and when the participant received hospital treatment or died with a principal diagnosis or cause of death in the ICD-10 code category: I20 angina pectoris, I21 acute myocardial infarction, I22 subsequent myocardial infarction, I23 certain current complications following acute myocardial infarction, I24 other acute ischaemic heart disease or I25 chronic ischaemic heart disease [Paper VI and VII].

The endpoint “hospital treatment or death due to stroke” was met if and when the participant received hospital treatment or died with a principal diagnosis or cause of death in the ICD-10 code category: I60 subarachnoid haemorrhage; I61 intracerebral haemorrhage; I63 cerebral infarction; I64 stroke, not specified as haemorrhage or infarction [Paper VIII and IX].

The endpoint “death from any cause” was met if and when the participant died regardless of the cause of death [Paper X and XI].

The endpoint “hospital treatment or death due to accidental injuries” [Paper XII and XIII] was met if and when the participant received hospital treatment or died with one of the following ICD-10 codes as a principal diagnosis or cause of death, respectively:

#### Hospital diagnoses

- S00-S09 Injuries to the head
- S10-S19 Injuries to the neck
- S20-S29 Injuries to the thorax
- S30-S39 Injuries to the abdomen, lower back, lumbar spine and pelvis
- S40-S49 Injuries to the shoulder and upper arm
- S50-S59 Injuries to the elbow and forearm
- S60-S69 Injuries to the wrist and hand
- S70-S79 Injuries to the hip and thigh
- S80-S89 Injuries to the knee and lower leg
- S90-S99 Injuries to the ankle and foot
- T00-T07 Injuries involving multiple body regions
- T08-T14 Injuries to unspecified part of trunk, limb or body region
- T15-T19 Effects of foreign body entering through natural orifice
- T20-T32 Burns and corrosions
- T33-T35 Frostbite
- T36-T50 Poisoning by drugs, medicaments and biological substances
- T51-T65 Toxic effects of substances chiefly nonmedicinal as to source
- T66-T78 Other and unspecified effects of external causes
- T79-T79 Certain early complications of trauma
- T90-T98 Sequelae of injuries, of poisoning and of other consequences of external causes

#### Death causes

- V01-V99 Transport accidents
- W00-X59 Other external causes of accidental injury
- X85-Y09 Assault
- Y10-Y34 Event of undetermined intent
- Y35-Y36 Legal intervention and operations of war
- Y85-Y89 Sequelae of external causes of morbidity and mortality
- Y90-Y98 Supplementary factors related to causes of morbidity and mortality classified elsewhere

The following endpoints were regarded in exploratory sub-analyses: Redeemed prescription for anxiolytics (ATC-code N05B); redeemed prescription for hypnotics and sedatives (ATC-code N05C); redeemed prescription for antidepressants (ATC-code N06A); hospital treatment or death due to ischaemic stroke (ICD-10: I63); hospital treatment or death due to haemorrhagic stroke (ICD10: I60 and I61); hospital treatment or death due to acute myocardial infarction (ICD-10: I21).

#### **2.1.4. Covariates**

Age, gender and socioeconomic status (SES) were included as covariates in all of the studies.

In the studies of accidental injuries, psychotropic drug use and psychiatric hospital treatment, the age variable was fixed at the start of the follow-up. In the other studies it was treated as a dynamic (time-varying) variable. In the study of accidental injuries, it was categorised into 20–24, 25–29, 30–39, 40–49 and 50–59 years. In the other studies, it were categorised into 10-year classes.

In the studies on psychotropic drug usage and psychiatric hospital treatment, SES was coded into “legislators, senior officials and managers; professionals; technicians and associate professionals; workers in occupations that require skills at a basic level; workers in elementary occupations; gainfully occupied people with an unknown occupation” in accordance with SOCIO (Statistics Denmark’s socioeconomic classification) [Statistics Denmark, 1997]. In the other studies, it was coded into “low; intermediate; high; unknown” in accordance with ESeC (The European Socioeconomic Classification) [cf. Harrison and Rose, 2006].

Calendar time (2000–2004; 2005–2009; 2010–2014) and time passed since start of follow-up (0–4 years; 5–9 years; >=10 years) were included as dynamic variables in the studies of IHD, antihypertensive drug usage, stroke and all-cause mortality.

In the study of accidental injuries, the second study on psychotropic drug use and the study of psychiatric hospital treatment, we include calendar year of interview as a covariate. It was categorised as (1999–2000; 2001–2006; 2007–2013) in the study of accidental injuries and as (2000–2004; 2005–2009; 2010–2013) in the second study on psychotropic drug use and the study of psychiatric hospital treatment.

Night work (‘Yes, regularly’ or ‘Yes, occasionally’ vs. ‘No’) was used as a covariate in the second study on psychotropic drug use and the study of psychiatric hospital treatment as well as in the studies on IHD, antihypertensive drug usage, accidental injuries and all-cause mortality.

Shift work (fixed night shifts or rotational shift work schedules versus other) was used as a covariate in the study on psychotropic drug usage.

Sample (DWECS 1995; DWECS 2000; COPSOQ 2004; DWECS 2005; DANES 2008; DWECS 2010) was used as a covariate in the study on psychotropic drug usage.

Employment in the healthcare industry (Yes vs No) was used as a covariate in the study on IHD and hypertensive drug usage.

Industrial sector (agriculture and fishing; manufacturing; construction; wholesale and retail trade; transport and storage; accommodation and food service; human health and social work; other; unknown) was included as a covariate in the study on accidental injuries.

### **2.1.5. Strategies to ensure that family wise error rates are $\leq 0.05$**

In the stroke study [Paper VIII and IX] there was only one pre-specified hypothesis to be tested, namely, that the rates of overall stroke increase with weekly working hours among full-time employees in the general working population of Denmark. Hence, there was no need to adjust for multiple comparisons in that study.

In the study of accidental injuries, the study of psychiatric hospital treatment and the second study of psychotropic drug usage, the multiplicity of comparisons was handled by testing individual hypotheses at the significance level 0.01. In the first study of psychotropic drug usage and the study of all-cause mortality, it was handled by nested hypothesis testing.

In the study on IHD and antihypertensive drug usage, we regarded relative rates of antihypertensive drug usage as a proxy measure for relative rates of IHD and performed multiple tests on each of these outcome measures. To ascertain that the overall significance level would be less than or equal to 0.05, we applied a Bonferroni correction to adjust for the multiplicity of outcome measures and a nested hypothesis testing procedure to adjust for the multiplicity of tests that were performed on each of two outcome measures.

### **2.1.6. Statistical methods**

Parameters were estimated by use of Poisson regression and hypotheses were tested by use of likelihood ratio tests. The analyses were implemented in the GENMOD procedure of SAS version 9.4.

### **2.1.7. Feasibility studies**

#### *2.1.7.1. Transitions between working hour categories*

Before we commenced with any of the studies on the relationship between long working hours and health, we wanted to know if the exposure to long working hours was stable enough to allow a long follow-up period on the basis of working hours at baseline. To shed light on this issue, we performed a feasibility study [Paper I] in which we cross-tabulated the reported weekly working hours in the year 2000 with the ones in the year 2005 among respondents in DWECs, who were 30 years or older and worked 32 or more hours a week in both waves. We categorised the reported weekly working hours into standard (32 – 40) and long (> 40 hours). We found that 68% of the workers with long working hours and 83% with standard working hours in 2005 belonged to the same category five years earlier, while 68% of the workers with long hours and 84% of the workers with standard hours in 2000 belonged to the same category five years later. Cohen's kappa for agreement between the hours worked in 2000 and the hours worked in 2005 was estimated at 0.52 (95% CI 0.48-0.55) - a moderate agreement according to Landis and Koch (1977). On the basis of this feasibility study, we concluded that the exposure was stable enough to make the studies worth the while.

#### *2.1.7.2. Statistics on smoking, overweight and obesity*

The labour force surveys do not provide any information on smoking habits and body mass index (BMI), which are known to be important predictors of, inter alia, IHD [Møller 2013; Hannerz and Holtermann, 2014]. We therefore wanted to know to what degree we could expect our results to be influenced by differences in smoking habits and BMI, before we commenced with the study on the relationship between long working hours and IHD. To shed light on this issue, we used data from

DWECS 2010 to estimate the age and gender standardised prevalence of smoking, moderate overweight ( $25 \leq \text{BMI} < 30$ ) and obesity ( $\text{BMI} \geq 30$ ) among employees in each of the following categories: 32 – 40, 41 – 48 and  $> 48$  working hours a week [Paper VI]. No discernible signs of any association between weekly working hours and smoking or BMI were found. Hence, we concluded that associations between working hours and health in the labour force surveys were unlikely to be distorted by differences in the prevalence of smoking, overweight or obesity.

#### *2.1.7.3. Antihypertensive drug usage*

In the study on the relationship between long working hours and IHD, we included redeemed prescriptions on antihypertensive drugs as an auxiliary outcome measure to increase the statistical power of the analyses. A statically significant association with hospital treatment or death due to IHD would afford direct statistical evidence of an association between long working hours and IHD, while the results obtained for antihypertensive drug usage would afford “indirect statistical evidence of an association with IHD if they (i) were statistically significant and (ii) showed a similar pattern to the results obtained for hospital treatment or death due to IHD” [Paper VI].

Before we were allowed to use rate ratios for antihypertensive drug usage as a proxy for rate ratios of IHD among occupational groups in Denmark, we needed to establish that the correlation between the two outcome measures was large enough to justify such a decision. We therefore conducted a feasibility study [Paper V] in which rate ratios of IHD related medication among socio-occupational and industrial groups in Denmark were compared with corresponding ratios based on hospital treatment or death due to IHD. The following conclusion was drawn: “Apart from a few caveats, the strong correlations obtained in the present study signify that purchase of a prescription for IHD-related medication is a usable risk indicator for IHD in the working population of Denmark. The usage of medicine data in addition to or instead of the use of death or hospital data in epidemiological studies on work-related IHD risk will bring about a tremendous increase in statistical power” [Paper V].

#### *2.1.7.4. Statistical power calculations*

Each of the studies on the association between long working hours and health was preceded by a power calculation to ascertain that the chance of detecting an effect of practical importance would be at least 80% [cf. Paper I, III, VI, VIII, X and XII].

### **2.1.8. Sensitivity analyses**

The term sensitivity analysis is defined in “A Dictionary of Epidemiology” [Portia et al., 2014] as “a method to determine the robustness of an assessment by examining the extent to which results are affected by changes in methods, models, values of unmeasured variables, or assumptions.”

#### *2.1.8.1. Pre-specified sensitivity analyses*

In the study of psychotropic drug usage [Paper I and II], we performed two sensitivity analysis. One of them explored the effect of excluding workers with poor self-rated mental health at baseline while the other explored the effect of controlling for job insecurity and job satisfaction.

In the study of IHD [Paper VI and VII], we performed a sensitivity analysis in which we only included participants who belonged to the same category according to their usual working hours as they did according to the actual hours worked during the reference week of the interview.

In the study of IHD [Paper VI and VII] and stroke [Paper VIII and IX], we performed a sensitivity analysis in which we excluded all participants who had received hospital treatment for IHD and stroke, respectively, sometime during a five-year period preceding baseline.

In the second study of psychotropic drug usage [Paper III and IV], we performed a sensitivity analysis in which we excluded all participants who received psychiatric hospital treatment or redeemed a prescription for psychotropic drugs sometime during a five-year period prior to the start of follow-up. We, moreover, performed a sensitivity analysis in which estimated rate ratios for psychotropic drug usage were controlled for industrial sector.

In the study of IHD [Paper VI and VII] and all-cause mortality [Paper X and XI], we performed a sensitivity analysis where we stratified the results by calendar period of the interview (1999–2000, 2001–2006, and 2007–2013).

In the second study of psychotropic drug usage [Paper III and IV] as well as in the study of stroke [Paper VIII and IX] and all-cause mortality [Paper X and XI], we performed a sensitivity analysis which only included people who i) had participated in more than one interview, and ii) had not moved more than one step among the ordered working time categories between their first and last interview.

Further details about the rationale and methods of the above sensitivity analyses are given in the papers.

#### *2.1.8.2. Regression with and without survey weights – a post hoc sensitivity analysis*

The results of Paper IV, VII, IX, XI and XIII were based on data obtained in the Danish labour force survey, which are sampled and weighted in accordance with the following principles:

Each quarter of a calendar year, a random sample of 15 – 74 year old people is drawn from the Population Statistics Register. An extra sample of unemployed people is drawn from the Register-based unemployment Statistics (RAM). Approximately 20 % of the combined sample will consist of unemployed people.

Due to researcher protections [forskerbeskyttelse], emigrations, deaths, non-response, random variation and the oversampling of unemployed people, the distribution of the participants by gender, age, unemployment, income, socioeconomic status, education, ethnicity and geographical region will be different from that in the target population. (Researcher protection means that the person, through a registration in the central person register, is protected against inquiries in connection with statistical and scientific studies, which are based on random samples from the central person register.)

To make the results of the survey representative of the target population, a weight is attached to each of the obtained replies. The purpose of the weights is (i) to adjust for the oversampling of unemployed people and (ii) to adjust for other differences between the responding participants and the target population with regard to the distribution of gender, age, unemployment, income, socioeconomic status, education, immigration status and geographical region.

The weights are based on information from the following national registers: The Central Person Register (CPR) · The Population Statistics Register · The Register-based Unemployment Statistics (RAM) · The Education Classification Module (DISCED) · The Register-based Labour Force Statistics (RAS).

The weight of a particular observation denotes the number of persons in the target population that are represented by that observation, at the time of the interview. If we, for example, want to estimate the total number of women in the target population who usually worked more than 40 hours in a particular quarter of a calendar year, we can do so by adding the weights of all observations that fall within this category. [Statistics Denmark, 2019]

The weights make a lot of sense and are crucial in the estimation of the number of people in the target population that falls within a given response category. Without the weights we would, for example, grossly overestimate the number and proportion of unemployed people in the target population and thereby grossly underestimate the actual size of the labour force.

Weights that are designed for a particular purpose may, however, be counterproductive if they are used for another purpose.

The aim of the studies of the present project was to estimate the association between weekly working hours and a variety of health outcomes (after adjustment for a number of preselected covariates) among full-time employees in the Danish labour force.

We had the opportunity to incorporate the weights in our analyses but chose not to, for, inter alia, the following reasons:

1. We believe that the problem with the oversampling of unemployed people, for all practical purposes, was circumvented by our decision (i) to only include full-time employees and (ii) to exclude all participants who were registered in the employment classification module as unemployed or otherwise not economically active during the calendar year preceding the start of the follow-up.
2. The weights would lose their meaning in the pooled data set. If the analyses had been restricted to a single quarter of a single calendar year and the study population had been based on all people who were interviewed in that quarter, then the weighting of the observations could be nicely interpreted as a way to make the distribution of gender, age, unemployment, income, socioeconomic status, education, immigration status and geographical region in the study population approximately equal to what it was in the target population, at that particular time period. Our primary analyses included however only the responses obtained in the first out of four



interview rounds and the demographic mix of the people who participate in their first interview may very well differ significantly from the demographic mix among the people who participate in their second, third or fourth round. Our analyses were, moreover, based on 60 different samples, one for each quarter of the calendar years 1999 – 2013. The weighting procedure has changed several times [Statistics Denmark, 2019] and the sample sizes as well as the demographic mix of the target population have changed with time. An application of the weights to the observations used in the present project would thereby create a data set with a demographic mix that is representative of some hypothetical population, which does not exist.

3. We felt quite confident that bias due to differences in demographic and socioeconomic distributions among the exposed versus unexposed workers of the study would be satisfactorily controlled or minimised by our statistical model. It is, however, possible that an examined association tends to be stronger or weaker among the responders than it is among the non-responders (after adjustment for demographic and socio-economic factors). If this is the case then our estimates would suffer from non-response bias, which we could not control for in our statistical model. From this viewpoint, it did not sound like a good idea to apply survey weights which up-weight observations from sub-populations with a low response rate and down-weight observations from sub-populations with a high response rate.

4. The weighting procedure is poorly described. Information has been given about the overall purpose of the weights and the registers that are currently utilised to obtain them [refs], but there are no descriptions as to how they were created. Without a proper method description, it would be difficult to publish the results in journals which require full methodological transparency.

It should, however, be noted that if a sample is free from non-response bias and the statistical model is able to satisfactorily control for bias from differences in demographic and socioeconomic distributions among the exposed versus unexposed workers then the effects of weight-induced changes in the socioeconomic and demographic distribution of the sample would be nullified by the statistical model. I.e. the difference between the point estimates obtained with and without weighting would be negligible, regardless of what standard population the weights were designed to emulate.

From this viewpoint it would have been interesting to see what the rate ratios presented in Paper IV, VII, IX, XI and XIII would have looked like if we had applied the survey weights in our regression analyses. To shed some light on this issue, I conducted a post-hoc sensitivity analysis, in which rate ratios obtained with survey weights are compared to rate ratios obtained without survey weights. The weighted analyses were performed by use of Cox-regression in the SURVEYPHREG procedure of SAS version 9.4. The non-weighted analyses were performed as described in the method section of the present chapter.

### 2.1.9. Protocol publication dates in relation to dates of establishment of the research data sets

The study protocol for the analyses of the association between long working hours and psychotropic drugs [Paper I] was published 19 September 2014 while the application for permission to link the exposure data to the outcome data (Figure 2.1) was dated at 29 November 2014.

The study protocols for the analyses on IHD [Paper VI] and accidental injuries [Paper XII] were published 22 June 2016 and 1 June 2016, respectively, while the request for the outcome data of the projects (Figure 2.2) was submitted the 24 June 2016.

<b>Forskningservice</b> Opgave nr. 702560	<b>Dato 29.11.2014</b>
<b>Indstilling om godkendelse af udvidelse af projekt</b>	
<b>Autoriseret institution</b> Det Nationale Forskningscenter for Arbejdsmiljø (NFA)	
<b>Projekttitel</b> Lange arbejdstider og brug af psykofarmaka	
<b>Projektbeskrivelse</b> Formålet er at undersøge betydningen af lange arbejdstider for risiko for brug af psykofarmaka blandt lønmodtagere i Danmark. Information om arbejdsforhold findes i form af individbaseret spørgeskemadata der er indsamlet af NFA år 1995, 2000, 2004, 2005 og 2010 samt af Danmarks Statistik på vegne af NFA år 2008. Disse data uploades til opgave nr. 702560 på Danmarks Statistiks forskermaskine og kobles ved hjælp af CPR-nr. til udvalgte eksisterende oplysninger fra lægemiddelsstatistikregisteret samt Erhvervs og Hospital Registeret (EHR). Se uddybet projektbeskrivelse her: <a href="http://www.researchprotocols.org/2014/3/e51/">http://www.researchprotocols.org/2014/3/e51/</a>	
<b>Population</b> Projektet vil behandle oplysning om personer som besvaret arbejdsmiljørelaterede spørgeskemaer fra NFA. <ul style="list-style-type: none"><li>• Ca. 18000 personer som deltog i den nationale arbejdsmiljøkohorte (NAK (på engelsk kaldet DWECS)) som lønmodtager i noget af følgende interviewrunder (årstal): 1995, 2000, 2005 eller 2010.</li><li>• Ca. 6000 lønmodtagere som deltog i det nationale arbejdsmiljøtværnsnit (NAT (på engelsk kaldet DANES)) år 2008</li><li>• Ca. 3500 lønmodtagere som besvarede NFAs tredækkerspørgeskema (på engelsk kaldet COPSOQ) år 2004-05.</li></ul>	
<b>Variabelindhold</b> Fra spørgeskemaerne vil vi bruge oplysninger om arbejdsforhold og selv vurderet mentalt helbred. Disse oplysninger vil vi ved hjælp af cpr-nummeret koble til oplysninger om brug af psykofarmaka fra lægemiddelsstatistikregisteret (ATC-kode + dato for indløsning af recept), samt oplysninger om social status, død og udvandring fra EHR.	
<b>Autoriserede forskere</b> Harald Hannerz, Det Nationale Forskningscenter for Arbejdsmiljø, <a href="mailto:hha@nrcwe.dk">hha@nrcwe.dk</a> , tlf 39165460. (Er allerede autoriseret til opgave nr. 702560)	

Figure 2.1. Application for permission to link the questionnaire data of the study on the association between long working hours and psychotropic medicines to data from national register at Statistics Denmark.

**Fra:** Jesper Møller Pedersen (JMP)  
**Sendt:** 24 June 2016 14:16  
**Til:** Karina Buchwald (KBU@dst.dk)  
**Cc:** Karin Ørum Elwert (KAE@dst.dk); Elsa Bach (EBA); Harald Hannerz (HHA)  
**Emne:** Data til projekt 704291  
**Vedhæftede filer:** Hannerz et al 2016.pdf; Larsen et al 2016.docx

Kære Karina,

Vi ønsker følgende data tilføjet projekt 704291 på Forskermaskinen:

**Fra LPR 1995 – 2014:** pnr; indskrivningsdato; udskrivningsdato; aktionsdiagnose; patienttype

**Fra dødsårsagsregisteret 2000 – 2014:** pnr; dødsdato; primær dødsårsag

**Fra LMDB 1995 – 2014** (Vi er kun interesserede af records hvor ATC-koden begynder på et C (Cardiovaskulære system); pnr; ekspeditionsdato (EKSD); atc-kode (ATC); barns alder (BALD)

**Fra vandrings-registeret 1995 - 2014:** pnr; Hændelsesdato (HÆNDDTO); Kode for ind-/udvandring (INDUD)

**Fra CPR-registeret 1999 - 2014:** pnr; aar; køn; fødselsår; dødsdato

**Fra arbejdsklassifikationsmodulet 1999 – 2013:** pnr; aar; branchekode (for væsentligste beskæftigelse i året); fagkode (for væsentligste beskæftigelse i året)

Go'weekend

Mvh

Jesper Møller

---

**Jesper Møller Pedersen (JMP)**

Fuldmægtig (analytiker), Cand.scient.adm.

Tlf: 39 16 52 89

e-mail: [jmp@arbejdsmiljoforskning.dk](mailto:jmp@arbejdsmiljoforskning.dk)

Figure 2.2. A request for and specification of the data from national registers that were to be linked with the questionnaire data of the Labour Force Survey.

## 2.2. Results

We did not find any statistically significant effects of interaction between weekly working hours and age, gender, socioeconomic status and night-time work, respectively, and we did not find any statistically significant main effects of weekly working hours on the incidence of psychotropic drug usage, mood, anxiety or stress-related disease, ischaemic heart disease, antihypertensive drug usage, accidental injuries or stroke. We found, however, that employees with moderate overtime work (41 – 48 hours a week) had significantly low rates of all-cause mortality ( $P < 0.0001$ ) compared with employees with 32 – 40 working hours a week. The pre-specified sensitivity analyses did not have any important impact on the results.

Table 2.1 gives the number of person years at risk, the number of cases and the rate ratios for each category of working hours, stratified by the outcome measures of the confirmatory analyses of the project. Table 2.2 gives the corresponding numbers stratified by outcome measures that were considered in exploratory analyses. The results of the post-hoc sensitivity analysis, which examined

the effect of survey weights on the estimates obtained on the data from the labour force survey, are given in Table 2.3.

The results of the sub-group analyses and the pre-specified sensitivity analyses of the project are given in the respective papers.

Table 2.1. Endpoint specific rate ratios with 95% confidence interval (CI) as a function of weekly working hours among Danish employees. Results from confirmatory analyses.

Endpoint	Weekly working hours	Person years at risk	Cases	Rate ratio	95% CI
Redeemed prescription for psychotropic drugs *	> 48	10 458	305	1.15	1.02 – 1.30
	41 - 48	17 665	500	1.04	0.94 – 1.15
	32 - 40	70 896	2109	1.00	-
Redeemed prescription for psychotropic drugs **	> 48	32 718	978	1.08	1.01 – 1.15
	41 - 48	57 164	1568	0.94	0.89 – 0.99
	32 - 40	432 094	13 280	1.00	-
Psychiatric hospital treatment due to mood, anxiety or stress-related disease **	> 48	38 628	78	0.96	0.76 - 1.21
	41 - 48	66 186	132	0.90	0.75 - 1.08
	32 - 40	531 859	1270	1.00	-
Hospital treatment or death due to ischaemic heart disease ***	> 48	71 258	284	1.07	0.94 – 1.21
	41 - 48	124 106	380	0.95	0.85 – 1.06
	32 - 40	931 403	2971	1.00	.
Redeemed prescription for antihypertensive drugs ***	> 48	55 221	1350	1.02	0.97 – 1.08
	41 - 48	99 090	2339	0.99	0.95 – 1.04
	32 - 40	680 240	16 959	1.00	.
Hospital treatment or death due to accidental injury ****	> 48	16 175	1484	1.02	0.97 - 1.08
	41 – 48	27 062	2281	0.96	0.91- 1.00
	32 – 40	230 463	19 730	1.00	-
Death from any cause *****	> 48	78 543	228	0.92	0.80 – 1.05
	41 - 48	133 667	275	0.75	0.66 – 0.85
	32 - 40	1 025 789	2871	1.00	-
Hospital treatment or death due to stroke *****	> 48	80 868	130	1.00	0.84 – 1.20
	41 - 48	133 906	189	0.97	0.83 - 1.13
	35 - 40	943 219	1418	1.00	-

\* adjusted for age, gender, socioeconomic status (SES), sample and shift work

\*\* adjusted for age, gender, SES, night shift work and calendar time of the interview

\*\*\* adjusted for age, gender, SES, night-time work, calendar-time, time passed since start of follow-up and health-care work (Yes vs. No)

\*\*\*\* adjusted for age, gender, SES, night-time work, year of interview and industry

\*\*\*\*\* adjusted for age, gender, SES, night-time work, calendar-time and time passed since start of follow-up

\*\*\*\*\* adjusted for age, gender, SES, calendar-time and time passed since start of follow-up

Please note that the reporting in Table 2.1 has been harmonised through the following actions: (i) the results of the working hour categories 49 – 54 and > 54 of the stroke study [Paper IX] have been pooled into > 48 hours a week, and (ii) the 99% confidence intervals of the rate ratios given in the second study on redeemed prescriptions for psychotropic drugs [Paper IV] and the study on injuries [Paper XIII] have been converted into 95% confidence intervals.

Table 2.2. Endpoint specific rate ratios with 95% confidence interval (CI) as a function of weekly working hours among Danish employees. Results from exploratory analyses.

Endpoint	Weekly working hours	Person years at risk	Cases	Rate ratio	95% CI
Redeemed prescription for anxiolytics *	> 48	11 684	113	1.15	0.94 – 1.41
	41 - 48	19 656	152	0.83	0.70 – 0.99
	32 - 40	79 351	849	1.00	-
Redeemed prescription for hypnotics & sedatives *	> 48	11 423	164	1.23	1.04 – 1.46
	41 - 48	19 348	247	1.03	0.90 – 1.19
	32 - 40	79 058	1007	1.00	-
Redeemed prescription for antidepressants *	> 48	11 519	159	0.98	0.83 – 1.16
	41 - 48	19 208	286	1.02	0.89 – 1.16
	32 - 40	77 713	1251	1.00	-
Hospital treatment or death due to acute myocardial infarction **	> 48	72 280	107	0.98	0.80 - 1.20
	41 – 48	125 404	147	0.96	0.80 - 1.14
	32 – 40	942 120	1119	1.00	-
Hospital treatment or death due to haemorrhagic stroke ***	> = 55	40 300	18	1.33	0.82 - 2.15
	49 - 54	41 058	21	1.58	1.01 - 2.46
	41 - 48	134 642	47	1.10	0.81 - 1.50
	35 - 40	948 535	310	1.00	-
Hospital treatment or death due to ischaemic stroke ***	> = 55	40 238	33	0.86	0.61 - 1.22
	49 - 54	40 991	32	0.85	0.60 - 1.22
	41 - 48	134 341	111	1.01	0.83 - 1.23
	35 - 40	946 216	815	1.00	-

\* adjusted for age, gender, socioeconomic status (SES), sample and shift work

\*\* adjusted for age, gender, SES, night-time work, calendar-time, time passed since start of follow-up and health-care work (Yes vs. No)

\*\*\* adjusted for age, gender, SES, calendar-time and time passed since start of follow-up

Table 2.3. Endpoint specific rate ratios with 95% confidence interval (CI) as a function of weekly working hours among Danish employees. Weighted results from a post-hoc sensitivity analysis compared with the unweighted results of the confirmatory analyses.

Endpoint	Weekly working hours	Unweighted Poisson		Weighted Cox	
		RR	95% CI	RR	95% CI
Hospital treatment or death due to ischaemic heart disease *	> 48	1.07	0.94 – 1.21	1.09	0.95 – 1.24
	41 - 48	0.95	0.85 – 1.06	0.93	0.83 – 1.05
	32 - 40	1.00	-	1.00	-
Redeemed prescription for antihypertensive drugs *	> 48	1.02	0.97 – 1.08	1.01	0.95 – 1.08
	41 - 48	0.99	0.95 – 1.04	0.99	0.94 – 1.04
	32 - 40	1.00	-	1.00	-
Redeemed prescription for psychotropic drugs **	> 48	1.08	1.01 – 1.15	1.07	0.99 – 1.15
	41 - 48	0.94	0.89 – 0.99	0.94	0.89 – 1.00
	32 - 40	1.00	-	1.00	-
Hospital treatment or death due to accidental injury ***	> 48	1.02	0.97 – 1.08	1.02	0.97 – 1.09
	41 – 48	0.96	0.91 – 1.00	0.95	0.91 – 1.00
	32 – 40	1.00	-	1.00	-
Death from any cause ****	> 48	0.92	0.80 – 1.05	0.87	0.75 – 1.02
	41 - 48	0.75	0.66 – 0.85	0.75	0.65 – 0.86
	32 - 40	1.00	-	1.00	-
Hospital treatment or death due to stroke *****	> = 55	0.89	0.69 – 1.16	0.93	0.70 – 1.25
	49 - 54	1.10	0.86 – 1.39	1.08	0.83 – 1.40
	41 - 48	0.97	0.83 – 1.13	0.96	0.81 – 1.13
	35 - 40	1.00	-	1.00	-

\* adjusted for age, gender, SES, night-time work, calendar-time, time passed since start of follow-up and health-care work (Yes vs. No)

\*\* adjusted for age, gender, SES, night shift work and calendar time of the interview

\*\*\* adjusted for age, gender, SES, night-time work, year of interview and industry

\*\*\*\* adjusted for age, gender, SES, night-time work, calendar-time and time passed since start of follow-up

\*\*\*\*\* adjusted for age, gender, SES, calendar-time and time passed since start of follow-up

## 2.3. Discussion

### 2.3.1. Main findings

The present work estimated rate ratios for psychotropic drug usage, mood, anxiety and stress-related disease, ischaemic heart disease, antihypertensive drug usage, accidental injuries, stroke and all-cause mortality, respectively, as a function of weekly working hours among full-time employees in the general population of Denmark. Each of the studies was preceded by a power analysis, which ascertained that the chance of detecting an effect of practical importance would be at least 80%. We did not find any statistically significant detrimental effects of long working hours (41 – 48 or > 48 hours a week) on any of the examined outcome measures. We found, however, that moderately

long work weeks (41 – 48 h) were statistically significantly associated with a decreased rate of all-cause mortality.

The exploratory analyses suggested a U-shaped association between weekly working hours and use of anxiolytics, with lowest rates among employees with moderate overtime work. They also suggested that very long hours (> 48) may be associated with increased rates of haemorrhagic stroke and usage of hypnotics and sedatives. These exploratory findings need, however, to be reproduced in an independent data set before they can be regarded as statistically significant.

### **2.3.2. Methodological considerations**

The project was conducted in accordance with the guidance given in three speeches – Speech I (Figure 3.2.4), Speech II (Figure 3.2.5) and Speech III (Figure 3.2.6). The first speech tells of the importance to ascertain that the statistical power of a hypothesis test is large enough to allow publication of the results even if they do not reach statistical significance, and that we should refrain from performing statistical hypothesis tests that are so underpowered that their outcomes only can be published if they are positive. We followed this advice and consequently were able to publish the results obtained in all of our studies on long working hours [Paper II, IV, VII, IX, XI and XIII] even though only one of them contained results that were statistically significant [Paper XI]. The second speech tells of the importance of differentiating between confirmatory analyses and exploratory hypothesis generating analyses. It, moreover, recommends that research protocols are published before studies of a confirmatory nature are commenced. The third speech tells us that in a confirmatory statistical analysis, it is not enough to state the hypothesis before we look at the results; we also need to state exactly how the hypothesis test will be performed. We followed all of these advices. Consequently, all of our statistical hypothesis tests were preceded by a completely specified statistical analysis plan [Paper I, III, VI, VIII, X and XII], which were written and published before we looked at any relation between the exposure and outcome data of the test.

The pre-published study-protocols, the statistical power and the hypothesis testing strategies, which ensured that the probability of a chance finding would be less than 5% in each of the separate studies, are some of the major strengths of the project.

Other major strengths are well summarised by the following text in Paper I: “Since the clinical endpoint of the study is determined through national registers, which cover all residents of Denmark, and we are able to censor for deaths and emigrations, we have eliminated bias from missing follow-up data. The study is further strengthened by its prospective design, the exclusion of prevalent cases and the use of a study population that has been randomly sampled from the target population.”

Another advantage is that the working hour categories were based on the sum of the hours worked in the primary and secondary jobs. If we had only regarded hours worked in the primary job, then 25% of the participants in the labour force surveys with long working hours would have been misclassified as having worked less than they actually did.

Since we are dealing with observational studies, we cannot rule out the possibility that the results have been biased by uncontrolled selection factors.

We had person-based information on night or shift work and since the EU Working Time Directive [EU, 2003] recognises night work as a potential health hazard, we included night or shift work as a covariate in all studies except the one that dealt with the outcome measure stroke. The reason for the exception was that the stroke study was aimed at replicating a previous study [Kivimaki et al., 2015a], which did not include night or shift work as a covariate.

We had also person-based information on age, gender, socioeconomic status and calendar time and since each of these factors have been associated with ischaemic heart disease [Mozaffarian et al., 2015; Tüchsen and Endahl, 1999; Sanchis-Gomar et al., 2016], stroke [Kissela et al., 2012; Appelros et al., 2009; Jakovljevic et al., 2001; Cox et al., 2006], psychotropic drug usage [Pratt et al., 2005; Wittchen et al., 2001; Tjepkema, 2005; Steinhausen and Bisgaard, 2014; Hudson, 2005], all-cause mortality [Hannerz, 1999] and injuries [Do et al., 2013; Haagsma et al., 2016; Hannerz et al., 2007], we included them as covariates in all of the examined associations between working hours and health.

We had, moreover, person-based information about the industry of the participants and since industry is highly associated with injuries [Kines et al., 2007; Pedersen et al., 2010], we added the industrial group of the participant as a covariate in the examination of the association between working hours and injuries. In the study of injuries, we also added a parameter to adjust for the effect of interaction between age and gender [cf. ref til mannhod trials].

We were, however, not able to include any person-based data on life style, body mass index, and working environment exposures.

We know that smoking, overweight and excessive drinking are risk factors for IHD, stroke and premature death [West, 2017; Sturm, 2002; WHO, 2018]. Excessive drinking is, moreover, associated with an increased risk of injury [WHO, 2018]. We also know that leisure time physical activity (LTPA) is associated with a decreased risk of IHD, stroke, premature death and depression [Warburton et al., 2006].

Fortunately, one of our feasibility studies [Paper VI] enabled us to rule out the possibility that our results had been importantly distorted by differences in the prevalence of smoking, moderate overweight and obesity.

A systematic review of the relation between occupational factors and LTPA suggests a negative association between total hours worked and LTPA [Kirk and Rhodes, 2011]. Long working hours have, moreover, been associated with a slightly increased risk of excessive drinking [Virtanen et al., 2015]. A negative association with LTPA and a positive association with excessive drinking would contribute toward an increased risk of ill-health among employees with long hours. It is therefore unlikely that the inability to find an adverse effect of long working hours in the present project was due to a failure to control for LTPA and excessive drinking.



It is possible that employees in a toxic work environment are more reluctant to working long hours, compared with employees in a healthy work environment, and if this is the case then the rate ratios among employees with long working hours may have been biased downwards.

In the present project, we were not able to control for specific occupational exposures. We were however able to perform a sensitivity analysis, which showed that the rate ratios in the study on psychotropic drug usage were virtually unaffected when we controlled for job satisfaction and job insecurity [Paper II].

Since job satisfaction is strongly associated with the work environment [Nübling et al., 2006] as well as with mental and physical health [Faragher et al., 2005] and it still did not influence the estimated rate ratios for psychotropic drug usage, we believe that the potential bias from differences in work environmental exposures in the present project (after control for age, sex and SES) is small. We can, however, not rule out the possibility that uncontrolled differences in working environmental exposures may have biased rate ratios among employees with long working hours slightly downward.

The studies were further limited by their lack of person-based data on income and sleeping habits: “As previously mentioned, one of the main theoretical reasons for a detrimental effect of a long work week is that it has been linked to short sleep, while one of the main theoretical reasons for a beneficial effect is its association with an increased income. Hence, it would have been of interest to study effect modification by income and sleeping habits” [Paper VII].

Regarding the exposure data (weekly working hours), we had two concerns. Firstly, they did not contain any information on the duration of the exposure and thereby limited the scope of the studies into examining the effect of the participants’ usual weekly working hours at baseline. Secondly, they were self-reported and thereby open to recall bias.

Misclassifications of working hours may result in a conservative bias towards the null. It is, however, also possible that they result in a bias away from null. E.g.: If workers perceive their work as more arduous when they are in a poor health condition than they do when they are in good physical or mental health then it is possible that they also perceive (recall) their working hours as longer when they are in a poor health condition than they do when they are in good health, and if this is the case it would bias the results towards the hypothesis of an increased risk among workers with long hours. To shed some light on this issue, we performed a sensitivity analysis [Paper VI and VII] in which we only included participants whose “usual working hours” were close to their “actual working hours during the reference week of the interview” (which were less likely to be distorted by recall bias). In the primary analysis, the rate ratio for IHD among workers with more than 48 working hours a week was estimated at 1.07 (95% CI: 0.94 – 1.21). In the sensitivity analysis it was estimated at 1.00 (95% CI: 0.86 – 1.18), which suggests that the failure to find an adverse effect of long working hours was not due to recall bias.

When we use codes in administrative health care registers as outcome measures there is always a possibility of detection and referral or prescription bias. There is, moreover, a possibility of bias due to non-differential misclassification of diagnoses. This shortcoming applies to all of the studies

except the one that dealt with all-cause mortality. Hence, it was interesting to note that the rate ratios among employees with long working hours were lower for all-cause mortality than they were for the life-shortening diseases IHD, stroke and mental disorders, which suggests that the failure to find an adverse effect of long working hours on the incidence of IHD, stroke or mental disorders is unlikely to be explained by diagnostic errors or by detection, referral or prescription bias.

The studies were further weakened by low response rates at the baseline interviews, which ranged between 48 and 80 percent. This particular drawback was discussed in Paper II by the following text: “It has been shown that the response rates to public health questionnaires in Denmark tend to be especially low among young men, unmarried people, people with a low educational level and people with an ethnic background other than Danish [Christensen et al., 2012; 2014]. It is possible that the response rates as well as the reasons for nonresponse in the present study differ between the exposed and unexposed workers. Long working hours imply, for example, less time to answer questionnaires. We believe, however, that any such bias was mitigated by our decision to control for age, gender and SES. We also believe that it was further mitigated by our decision to focus on relative rather than absolute rates and by our decision to exclude prevalent cases.”

Last but not least, we need to consider the possibility of a healthy worker effect. Employees with good health may be more prone to work overtime than employees with poor health, which would result in a bias towards decreased rates of ill-health among employees with long working hours. There is, however, also a possibility of an unhealthy worker effect, especially among employees with very long working hours, where many may suffer from workaholism, which has been associated with obsessive compulsive disorders, ADHD, anxiety, and depression [Andreassen et al., 2016].

### **2.3.3. Generalisability**

There may be nations in which the findings of the present work do not hold good. Concerns with regard to the generalisability (external validity) of our findings are well expressed as follows in Paper IX: “The finding pertains to the general working population of Denmark – a country with generous sick-leave benefits, relatively strong work environment legislations, free medical care, five to six weeks of paid vacation, a full-time working week of 37 h and a wage level making it realistic for most people to live an ordinary life with the income from a single 37-hour job. The results from the present study may not hold good in nations with a less regulated work environment and where long working hours for more people are necessary to survive or to keep their job, and where access to health care is more costly ...”

### **2.3.4. Previous research**

The present section will focus firstly on comparable studies which targets employees in Denmark and secondly on comparable studies which i) targets employees in Europe and ii) are large enough to convey meaningful information. Further discussions of previous research are given in Papers I, II, IV, VI, VII, IX, XI and XIII.

Apart from the present work, there are four projects, which have produced a total of 15 relevant prospective analyses that are large enough to convey meaningful information about the association between weekly working hours and morbidity or mortality among workers in Europe. One of the

projects is based on a cohort from Denmark, another is based on a cohort from Northern Ireland, a third one is based on a series of meta-analyses on a large variety of cohorts from Europe, Australia and USA and a fourth one is based on an Italian cohort.

#### *2.3.4.1. The Copenhagen Male Study*

The Danish project, named the Copenhagen Male Study (CMS), was established in 1970, with a primary purpose “to elucidate the roles of physical activity and physical fitness as predictors of ischaemic heart disease (IHD)” [Gyntelberg et al., 2004]. All 40 – 59-year-old male employees at 14 companies in the urban area of Copenhagen were invited to participate in the baseline examination 1970 -1971, which collected data on work, lifestyle and health through a series of clinical tests and questionnaires. In total 5249 employees agreed to participate, which gave a response rate of 87%. The cohort has recently been followed-up in national registers to study rates of deaths [Holtermann et al., 2010] and dementia [Nabe-Nielsen et al., 2017] as a function of self-reported weekly working hours at baseline. Estimated rate ratios, adjusted for age [Holtermann et al.] and age, time since exposure measurement, calendar year, shift work and socio-economic status [Nabe-Nielsen et al.] are given in Table 2.4. The long follow-up periods (30 years in the study by Holtermann et al. and 22 - 44 years in the study by Nabe-Nielsen et al.) ascertained that the statistical power was sufficiently high to detect a clinically important effect on all-cause mortality and dementia, respectively. The studies were further strengthened by the high response rates at baseline. A drawback of the CMS is that it only included males in a selected set of companies, which limits the generalisability of the results.

#### *2.3.4.2. The Northern Ireland Mortality Study*

The Northern Irish project, named the Northern Ireland Mortality Study (NIMS), is based on a record-linkage between the 2001 census returns for the whole enumerated population of Northern Ireland and registered deaths up to 2009 [O’Reilly et al., 2012]. The census contained information on inter alia age, sex, socio-economic status, marital status and weekly working hours, defined as the average number of work hours per week in the participant’s main job during the 4 weeks period prior to the census. “The Census placed a legal obligation on every household in which someone was usually resident on Census Day, and on every person who was a usual resident of a communal establishment, to complete a Census form” [The Northern Ireland Statistics and Research Agency, 2002], which resulted in a response rate of 95 percent [The Office for National Statistics, 2004]. The NIMS data were used by O’Reilly and Rosato (2013) to examine the relationship between weekly working hours and mortality among full-time workers in the general population of Northern Ireland. Their analyses were stratified by sex and the data set was large enough to provide estimates with an acceptable precision for all-cause mortality among the men. A first model was adjusted for age and marital status, a second model added adjustment for socio-economic status, a third model added adjustment for “dependent children and caregiving” and a fourth model added adjustment for “limiting long-term illness and general health”. The rate ratios in the first model, decreased monotonically with working hours to 0.88 (95% CI: 0.80 – 0.96) for > 54 vs 35 – 40 hours a week. This association disappeared, however, after adjustment for socio-economic status (Table 2.4) and was not further influenced by the added control for “dependent children and caregiving” and “limiting long-term illness and general health”.

Compared with our mortality study [Paper XI], the NIMS study had several advantages – (i) it was larger (ii) it had a higher response rate, (iii) it had a higher proportion of workers with very long hours, which enabled acceptably precise estimates for both of the categories 49–54 and > 54 hours a week, and (iv) it had access to data on “limiting long-term illness and general health” and could thereby rule out the healthy worker effect as a possible explanation for not finding a generally increased mortality among people with long working hours. The drawbacks of the NIMS study were (i) that it only regarded the hours worked in the participant’s main job, and (ii) that it only regarded the hours worked during the 4-week period immediately preceding the census [O’Reilly and Rosato, 2013].

It should also be noted that workers in UK have the right to opt out of the 48-hour rule of the EU Working Time Directive, i.e. they can choose to work more than 48 hours a week on average in their primary job, but they cannot be forced to do so [O’Reilly and Rosato, 2013], and that workers in Denmark do not have this option.

#### *2.3.4.3. The IPD-Work Consortium*

The third project, named the Individual-Participant-Data Meta-analysis in Working Populations (IPD-Work) Consortium, is a collaborative effort among researchers in many different countries, who each contribute with one or more samples of individual participant data to be harmonised, analysed and subsequently combined in a meta-analysis. “The aim of the consortium is to estimate reliably the associations of work-related psychosocial factors with chronic diseases, disability, and mortality” [Kivimäki et al., 2015b]. The consortium was established in 2008 and originally included samples of 17 cohorts from Finland, Sweden, Denmark, the Netherlands, Belgium, France, Germany, and UK. The project has thereafter been expanded. In 2015 it comprised 26 cohorts from Europe, USA, and Australia [Kivimäki et al., 2015b]. The major strengths of the IPD-Work consortium are (i) that its combined data material often is large enough to afford meaningful estimates of the strength of associations between occupational exposures and ill-health, and (ii) that it endorses the principle that exposures, outcome measures and hypotheses should be defined prior to the commencement of analyses [Kivimäki et al., 2015b].

The IPD-Work consortium has, so far, examined the prospective association between weekly working hours and coronary heart disease [Kivimäki et al., 2015a], stroke [Kivimäki et al., 2015a], atrial fibrillation [Kivimäki et al., 2017], cancer [Heikkilä et al., 2016], type 2 diabetes [Kivimäki et al., 2015c] and depressive symptoms [Virtanen et al., 2018]. Main effect estimates of the studies, adjusted for age, gender and socio-economic status, are given in Table 2.4.

The advantages of the above mentioned IPD-Work studies, compared with the present work, were (i) that the diversity of the included samples (with regard to time-periods, nations and types of population - some were company based, some were population based ...) [cf. Madsen et al., 2014] enabled the researchers to explore heterogeneity of effects and thereby shed some light on the generalisability of their findings, and (ii) that it had a higher proportion of workers with very long hours, which often enabled acceptably precise estimates for both of the categories 49–54 and > 54 hours a week. The main drawback of the studies is that all of the cases of depressive symptoms, 61 % of the coronary heart disease cases, 52% of the diabetes cases, and 30% of the stroke cases were based on self-reports, which resulted in a tremendous loss to follow-up. This drawback was,

however, not present in the cancer study [Heikkila et al., 2016], where all outcomes were ascertained through national cancer, hospitalisation and death registers.

The value of the studies was, furthermore, slightly weakened by (i) the missing information on whether the working hours in the various cohorts were based on the hours worked in the main job only or on the sum of the hours worked in the main and secondary jobs, and (ii) the missing rate ratios for the categories 41 – 48 and 49 – 54 working hours a week, in the studies of type 2 diabetes [Kivimäki et al., 2015c] and depressive symptoms [Virtanen et al., 2018].

#### *2.3.4.4. The 2011 Italian census*

The 2011 Italian census included a form with questions on sociodemographic and occupational factors, of which one asked about the participants' usual weekly working hours. This particular form was sent to "all the households in the municipalities with less than 20,000 residents and to a sample (one-third) of the households in municipalities with 20,000 residents or more and in all provincial capitals" [Alicandro et al., 2020]. The collected data were used by Alicandro et al. (2020) to examine the relationship between weekly working hours and mortality from cardiovascular disease (CVD) among full-time workers in the general population of Italy. Their analyses were stratified by sex and their data set was large enough to provide estimates with an acceptable precision for cardiovascular as well as non-cardiovascular mortality. The study was, moreover, large enough to provide quite precise subanalyses on ischaemic heart disease and stroke mortality among the men.

The conclusion of the study was that it "does not support the hypothesis of an association between long working hours and increased rates of CVD mortality among active Italian men. It supports, however, the hypothesis that moderately long working hours (41–54 h a week) are associated with a slight decrease in CVD mortality among men."

The mortality study by Alicandro et al. encompassed approximately 12 million individuals and 85 000 deaths and is thereby the largest study ever on the relationship between long working hours and mortality. It had, moreover, the advantage of being based on a survey in which participation was mandatory, which resulted in an extraordinary high response rate of 99%.

#### *2.3.4.5. Summary of previous results*

Estimated risk ratios from the CMS, NIMS and IPD studies are given in Table 2.4 while the estimates of the Italian census-based mortality study are given in Table 2.5.

Table 2.4. Endpoint specific risk (odds or rate) ratios with 95% confidence interval (CI) as a function of weekly working hours (WWH) among a variety of published cohort studies

Paper	Endpoint	Sex	WWH	Cases	RR/OR	95% CI
Nabe-Nielsen et al., 2017*	Dementia	M	> 45	118	0.97	0.79 – 1.19
			< = 45	516	1.00	–
Holtermann et al., 2010*	Death from any cause	M	> 45	444	0.91	0.79 – 1.05
			41 – 45	1886	1.07	0.95 – 1.20
			36 – 40	345	1.00	–
O’Reilly and Rosato, 2013**	Death from any cause	M	> 54	566	0.96	0.87 – 1.05
			49 – 54	418	1.01	0.91 – 1.12
			41 – 48	663	0.96	0.88 – 1.04
			35 – 40	2800	1.00	–
Kivimäki et al., 2015a***	Coronary heart disease	M/F	> 54	132	1.08	0.94 – 1.23
			49 – 54	117	1.07	0.92 – 1.24
			41 – 48	241	1.02	0.91 – 1.15
			35 – 40	774	1.00	–
	Stroke	M/F	> 54	347	1.33	1.11 – 1.61
			49 – 54	281	1.27	1.03 – 1.56
			41 – 48	460	1.10	0.94 – 1.28
			35 – 40	1393	1.00	–
Kivimäki et al., 2017***	Atrial fibrillation	M/F	> 54	91	1.42	1.13 – 1.80
			49 – 54	88	1.17	0.93 – 1.49
			41 – 48	195	1.02	0.85 – 1.23
			35 – 40	586	1.00	–
Heikkila et al., 2016***	Any incident cancer	M/F	> 54	269	0.93	0.81 – 1.06
			49 – 54	320	1.07	0.94 – 1.21
			41 – 48	681	0.97	0.87 – 1.07
			35 – 40	1820	1.00	–
Kivimäki et al., 2015c***	Type 2 diabetes	M/F	> 54	NR	1.05	0.87 – 1.25
			35 – 40	NR	1.00	–
Virtanen et al., 2018****	Depressive symptoms	M/F	> 54	NR	1.11	1.00 – 1.22
			35 – 40	NR	1.00	–

\* Copenhagen Male Study; \*\* Northern Ireland Mortality Study; \*\*\* IPD-Work Consortium: Cohorts from Europe, Australia and North America; \*\*\*\* IPD-Work Consortium: Cohorts from Europe; M = Males; F = Females; M/F = Males and females; NR = Not reported

Table 2.5. Cause specific mortality rate ratios (RR) with 95% confidence interval (CI) in the period 2012 – 2016 as a function of weekly working hours (WWH) among participants in the 2011 Italian census [Alicandro et al., 2020]

Cause of death	WWH	Men			Women		
		Cases	RR*	95% CI	Cases	RR*	95% CI
All cardiovascular diseases	> 54	967	0.95	0.89 – 1.02	88	1.19	0.95–1.49
	49 – 54	1142	0.90	0.85 – 0.96	88	0.95	0.76–1.18
	41 – 48	2250	0.93	0.89 – 0.98	221	0.96	0.83–1.11
	35 – 40	10 903	1.00	–	1547	1.00	–
Ischaemic heart diseases	> 54	477	0.95	0.86 – 1.05	29	1.18	0.79 – 1.76
	49 – 54	568	0.91	0.83 – 1.00	27	0.92	0.61 – 1.37
	41 – 48	1063	0.91	0.85 – 0.97	60	0.86	0.65 – 1.14
	35 – 40	5287	1.00	–	443	1.00	–
Cerebrovascular diseases	> 54	127	0.95	0.79 – 1.15	22	0.98	0.62 – 1.53
	49 – 54	152	0.93	0.78 – 1.10	28	0.99	0.67 – 1.46
	41 – 48	334	1.05	0.93 – 1.18	72	1.02	0.79 – 1.32
	35 – 40	1439	1.00	–	477	1.00	–
Non-cardiovascular diseases	> 54	3233	0.95	0.92 – 0.99	589	1.02	0.94 – 1.11
	49 – 54	4155	0.97	0.94 – 1.01	722	0.98	0.91 – 1.06
	41 – 48	8024	0.98	0.96 – 1.01	1659	0.95	0.90 – 1.00
	35 – 40	37 360	1.00	–	11 889	1.00	–

\*adjusted for age category, marital status, number of dependent children, geographic area, education and occupation

Among workers with moderately long working hours (41 – 48 hours a week), we note that all of the estimated rate or odds ratios were close to unity in Table 2.4 as well as in Table 2.5. We also note that the rate ratios for CVD and IHD mortality, among men with 41 – 48 working hours a week in the Italian census-based mortality study, were statistically significantly lower than unity. In the IPD-Work studies, there was a tendency towards increased rate ratios of overall stroke and atrial fibrillation among the participants with more than 48 working hours a week. The increased rates of overall stroke could, however, not be reproduced in the present work [Paper IX], nor in the Italian census-based mortality study, and the findings of the study on atrial fibrillation has, as far as I know, never been properly tested outside of the IPD-Work consortium.

Another interesting observation was that the estimated rate ratios for all-cause mortality among the workers in the category with the longest working hours were lower than they were among the workers with standard working hours, with, coincidentally, an upper limit of the 95% confidence interval at 1.05 in all of the three all-cause mortality studies [Paper XI; Holtermann et al., 2010; O’Reilly and Rosato, 2013].

A review article by Wagstaff and Sigstad (2011) cites a handful of studies on the association between working hours and occupational injuries and concludes that long work-shifts are associated with an increased risk. We have, however, not found any study that is comparable to the injury study of the present work [Paper XIII], which looked at the risk of injury from a public health perspective without

regard of whether it happened at work or leisure. The two outcomes measures (occupational injury versus any injury) are not comparable, since long hours at work imply less time to be injured outside of work.

#### **2.3.4. Conclusions**

We note that the upper limits of the 95% confidence intervals of the rate ratios among employees with 41 – 48 working hours a week lie either within or below the “no association” region [cf. Monson, 1990] in all of the confirmatory analyses of the present work (Table 2.1). Moreover, all comparable rate ratios among people in this working hour category (Table 2.4 and 2.5) lied within the no association region and none of them was statistically significantly high. Hence, for all practical purposes, we may conclude that overtime work that lies within the limits of the EU working time directive is not associated with increased rates of psychotropic drug usage, psychiatric hospital treatment due to mood, anxiety or stress-related disease, antihypertensive drug usage, overall accidental injuries, ischaemic heart disease, overall stroke or all-cause mortality among employees in the general working population of Denmark.

We also note that all rate ratios of the present confirmatory analyses among employees with more than 48 working hours a week lie within the no association region, and that none of them is statistically significant. Hence, the findings of our studies do not support the notion that long weekly working hours constitute a public health problem in the general population of Denmark.



### ***3. Methodological perspectives***

During 27 years of involvement in public health related statistical inference I have encountered some interesting problems, which have been solved or mitigated through some coping strategies that I have decided to share in the present dissertation.

In the beginning of my carrier, I was not a very critical reader and I tended to trust research findings that were presented in peer-reviewed journals, i.e. in articles that had been written by scholars and scrutinised by presumed experts before they were accepted for publication.

Later on, I came to realise that the study of peer-reviewed public health literature actually is associated with a tremendous risk of being fooled by faulty statistical significance declarations and bias from non-publication and selective reporting of results.

I also came to realise that the value of a statistical analysis would be considerably enhanced if the researchers could document that their study was free from faulty statistical significance declarations and bias from selective reporting.

Consequently, I have adopted some methodological principles and coping strategies aimed at i) reducing the risk of being fooled by others' research reports; and ii) increasing the credibility of one's own research.

In the present chapter, I will address the methodological principles and coping strategies that governed the research presented in chapter 2. I will also share my thoughts and reflections, on them as well as the problems they were designed to solve or mitigate. The problem formulation is given in subchapter 3.1. The coping strategies are addressed in subchapter 3.2 – 3.3.

### 3.1. Adverse side effects of statistical significance testing



Figure 3.1.1. Collage by Pia Dukholm, text by the statistical society at the National Research Centre for the Work Environment, Denmark, 2008

Statistical significance testing is a time-honoured method to reduce the risk of being fooled by random variation [Fisher, 1935]. How to do it:

1. State the hypothesis
2. Set the significance level
3. Design the experiment
4. Establish a testing strategy and statistical significance criteria which (in accordance with laws of probability and mathematical statistics) ensures that the probability of a type 1 error is less than or equal to the predefined significance level
5. Ascertain that the statistical power will be sufficient to detect effects of practical importance
6. Collect the data
7. Perform the test

Sometimes the test only involves a single comparison e.g. between exposed and unexposed people in the study population; other times it involves several comparisons performed in different ways and/or in different subsamples of the study population. When a test is based on a single comparison it may be deemed statistically significant at the significance level  $\alpha$  if an  $(1 - \alpha)\%$  confidence interval of the estimated contrast does not contain the expected value of the null-hypothesis (usually 0 for a difference and 1 for a ratio). This strategy can, however, not be used when the test involves multiple comparisons. In such situations, we need a test strategy that takes the multiplicity of comparisons into account. Sometimes it is possible to find a test which processes all the concerned variables simultaneously; other times we would need to divide the analysis into a series of subordinate hypothesis tests with individual significance levels that are adjusted in a way that ensures that the overall significance level (family wise error rate) is less than or equal to  $\alpha$  [cf. Neyman and Pearson, 1933; Dunn, 1961; Holm, 1979; Benjamini and Hochberg, 1995].

Properly performed statistical significance tests reduces the risk of being fooled by coincidences and thereby play an important role in the evaluation of hypothetical associations between environmental factors and health. The introduction of statistical significance tests in public health research is, however, also associated with some unfortunate side effects. Firstly, it has put us at risk of being fooled by faulty statistical significance declarations. Secondly, since results are more likely to be published if they are statistically significant [Ingre, 2017], it has put us at increased risk of being fooled by bias from non-publication and selective reporting of results.

I believe that a substantial part of faulty statistical significance declarations is due to statistical misconceptions. In the present subchapter, I will list and thoroughly debunk some of these misconceptions.

### **3.1.1. Re. multiple testing**

I have noticed that some epidemiologists hold that an observed 95% confidence interval which does not contain the expected value of the null-hypothesis always denotes statistical significance at the significance level 0.05 regardless of the context in which it was observed, and that as long as the significance is determined by use of confidence intervals there is no need to adjust for multiple comparisons. One of the reviewers of Paper VII even went so far as to propose that adjustment for

multiple comparisons was an unethical practice which should be banned from publication in medical, public health, or occupational health science journals [See supplemental material to Paper VII]. In response to this comment I wrote the following text:

“Let’s say that Mr X suspects that the three dice that Mr Y is using are loaded to favour the outcome six. To check this, he borrows the dice and rolls them repeatedly, looking for a triple six. It takes him 200 attempts to obtain a triple 6. He wants to check if this is statistically significant at a significance level of 0.01. If he chooses to abide by the rules set out by reviewer 1 (do not take the multiplicity of trials into account) then he will arrive at conclusion A (see below). If he, on the other hand, chooses to base his test on logics and the laws of probability theory he will arrive at conclusion B.

A. He did observe a triple 6 and the probability of such an event in a single trial with three fair dice equals  $1/216 < 0.01$ . Hence, the result of his experiment supports the hypothesis that the dice are loaded to favour the outcome six.

B. It took him 200 trials to obtain a triple 6, and the probability of observing at least one triple 6 in 200 trials with three fair dice equals 0.6. Hence, the result of his experiment does not support the hypothesis that the dice are loaded to favour the outcome six.

According to the rules set out by reviewer 1, A is correct and B is incorrect. We do not agree with this notion. The reason for our disagreement lies in the logical contradiction, we do not feel that it would be right to say that the probability of observing a triple 6 in the context of the experiment was less than 0.01 when the laws of probability theory dictates that it was equal to 0.6. We hold that probability theoretic operations should be based on the rules and definitions of probability theory in the same way as we hold that arithmetic operations should be based on the rules and definitions of arithmetic. Hence, we reject the notion that A is correct and B is incorrect in the same way as we would reject the notion that  $2 + 2 = 5$ .

Of less importance, but still of interest is that the STROBE explanatory article [Vandenbroucke et al., 2007] states that authors should take multiplicity of analyses into account when they interpret their results.”

The words “in the same way as we would reject the notion that  $2 + 2 = 5$ ” were, however, deleted before we uploaded the response.

### **3.1.2. Re. post-hoc testing**

As suggested by Figure 3.1.1., a faulty statistical significance declaration may also arise from a failure to differentiate between estimates that are based on a pre-specified hypothesis and estimates that have been obtained in response to a hypothesis that was generated by the data, which can easily be explained by the following probability theoretic arguments.

Let’s say that an urn contains 100 balls, that each of them is marked with a unique integer number between 1 and 100, and that they are equally likely to be drawn at random. One ball is drawn randomly from the urn. If the number is greater than 80 then it will be announced, otherwise the

ball will be put back in the urn, the balls will be remixed and a new draw will be made. What is the probability that an announced number will be greater than 95?

Solution: Let A be the event that the number on the ball is greater than 95 and let B be the event that it is greater than 80. Since all balls are equally likely to be drawn at random, we note that  $P(A) = 0.05$ , that  $P(B) = 0.2$  and, since A is a proper subset of B, that  $P(A \cap B) = P(A)$ . From elementary probability theory [Kolmogorov, 1933], we know that the probability of A given B is given by the equation

$$P(A|B) = P(A \cap B)/P(B) \quad (3.1.1)$$

Hence, the probability that an announced number will be greater than 95 is equal to 0.25. We also note that the expected average of drawn numbers  $E[X] = (1 + 100)/2 = 50.5$ , while the expected average of announced numbers  $E[X|B] = (81 + 100)/2 = 90.5$ .

In the above example, there was a world of a difference between the unconditional probability  $P(A)$  and the conditional probability of A given B. The very same reasoning can be applied to show that there may be a world of difference between “associations that have been estimated to test a hypothesis that was specified before the researcher looked at the data” and “associations that have been estimated in response to a hypothesis that was generated by the data”. Let’s say that a researcher is browsing some descriptive tables with outcome-specific averages among individuals who have been grouped in relation to a variety of exposures. Whenever an observed difference makes sense and moreover seems large enough to generate the hypothesis that there is a significant association between the exposure and outcome in question, it will be calculated and presented with a 95% confidence interval. What is the probability that such a confidence interval (by chance alone) does not contain the null-value?

Solution: Let A be the event that a 95% confidence interval (by chance alone) does not contain the null-value. Let B be the event that the difference in the outcome among exposed and unexposed individuals (by chance alone) makes sense to the researcher and moreover is large enough to generate the hypothesis that there is a significant association between the exposure and the outcome. Here we note that the probability of the event A is equal to 0.05 if the hypothesis is pre-specified i.e. if the decision to test is independent of the data. We also note that the probability of A given B (according to Equation 3.1.1) may be much higher than 0.05 and that it theoretically can take on any value between 0.05 and 1.

The research behaviour described above is not unusual. As a practicing statistician, I have on numerous occasions been approached by people (including peer-reviewers) who have observed things in data that they want me to check for statistical significance, and in a survey among more than 2000 academic psychologists at major US universities, it was estimated that 54% of the participants had “in a paper, reported an unexpected finding as having been predicted from the start” [John et al., 2012].

Here, it should be noted that it is not necessarily wrong to estimate associations with a confidence interval in response to a hypothesis that was generated by the data. It is, however, as indicated by

equation 3.1.1, very important to differentiate between estimates with a pre-specified hypothesis and estimates that have been obtained in response to a hypothesis that was generated by the data. A failure to recognise that an estimate was obtained post hoc may fool people into believing that something is statistically significant when it isn't. It may also fool people into believing that the estimate is free from researcher selection bias.

The importance to differentiate between pre-specified and post hoc estimation is acknowledged in the guidelines on good publication practice by the Committee on Publication Ethics (COPE), which states that "Failure to disclose that the analysis was post hoc is unacceptable" [COPE, 2000]. It is also acknowledged in some widely endorsed reporting guidelines, namely, the CONSORT (CONsolidated Standards of Reporting Trials) 2010 guideline, which states that "Analyses that were pre-specified in the trial protocol are much more reliable than those suggested by the data, and therefore authors should report which analyses were pre-specified" [Moher et al., 2010], and the STROBE (STrengthening the Reporting of OBServational studies in Epidemiology) guideline, which states that "When a study is reported, authors should tell readers whether particular analyses were suggested by data inspection [...] Readers need to know which subgroup analyses were planned in advance, and which arose while analysing the data." [Vandenbroucke et al., 2007].

There are, however, some researchers in the field who do not understand that the need to differentiate between pre-specified and post-hoc tests is based on a mathematical fact (cf. Eq. 3.1.1), which is not open for discussion. The existence of such misunderstandings is evidenced by the following misguided statements, which I found in some debate articles in the high impact journal EPIDEMIOLOGY:

"One thread in the arguments favoring registration of observational studies is that conclusions from data that bear on a "prespecified hypothesis" are stronger than if the hypothesis was not specified before seeing the data. Although seemingly intuitive, logicians have a hard time proving this; the explanation a scientist arrives at might be exactly the same, whether thought of before or after seeing the data, so what's the difference?" [Vandenbroucke, 2010]

"Furthermore, the timing of a research hypothesis (whether before or after data collection) is irrelevant to its validity." [The Editors, 2010]

### **3.1.3. Re. insufficient statistical power**

Another source of faulty statistical significance declarations lies in a failure to ascertain a statistical power that is large enough to detect effects of practical importance, before the test is performed.

Most researchers understand that a null-finding from a severely underpowered test neither can confirm nor reject the existence of a clinically important effect and that it therefore is unlikely to impart any meaningful information. It is however not generally understood that the result of such a test is inconclusive also when the P-value of the test is less than its significance level; and it is often believed that a P-value that is less than the significance level of the test proves that the statistical power of the test was sufficient.

This misconception is commented on in Paper XIV, where I and my colleagues Sannie Vester Thorsen and Hans Bay point out that *“statistical power may be an issue even if the null hypothesis is rejected by the authors. If the hypothesis is rejected due to a test statistic result that is unlikely to occur (e.g. less than 5% chance) under the null-hypothesis, but the power of the test was unacceptably low (e.g. only a 6% chance of detecting a true positive effect under any realistic alternative hypothesis) then all we know is that we have obtained a test result that is unlikely to occur both under the null and the alternative hypothesis. A low probability of occurrence under the null-hypothesis is therefore not a sufficient criterion for statistical significance.”*

If, on the other hand, the statistical power of the above test had been acceptably high (e.g. > 80% chance of detecting a realistically chosen and clinically important effect) then we would have obtained a test result that we know is much more likely to occur under the alternative hypothesis than it is under the null-hypothesis, and due to the striking difference between the significance level and the statistical power of the test ( $\leq 5\%$  vs.  $> 80\%$ ) we would be able to rule in favour of the alternative hypothesis. Statistical significance can, in other words, only be declared if the test result is associated with a low probability of occurrence under the null-hypothesis in combination with a high probability of occurrence under the alternative hypothesis. It is therefore necessary to take the statistical power into account when we interpret statistical tests, regardless of their associated P-values. And it goes without saying that this principle also holds good when the test is based on a confidence interval: i.e. we need to take the statistical power of the test into account regardless of whether or not its associated confidence interval contains the expected value of the null-hypothesis.

Another insightful comment on the importance of differentiating between low-powered and high-powered statistical tests when interpreting research findings was given as follows by Ingre (2017):

*“The high inconsistency and poor credibility that is expected from positive findings in small low-powered studies is just another way of expressing the limited value of information that low-powered statistical tests have in science, regardless of their observed “statistical significance” from NHST [Null Hypothesis Significance Tests]. Researchers should be aware whether they are interpreting a high-powered test or a low powered test; when positive findings from low-powered tests are interpreted, they should be considered speculative and susceptible to type-1 errors; and any observed “significant” association should be assumed to be inflated from its true size. High-powered tests can be trusted to a higher degree and thus, can be used to argue support for an observed association when positive, or the absence of an association when negative.”*

In a perfect world, all test results would be useful, also the ones from severely underpowered tests since they could serve as input in Meta-analyses. Tests that are underpowered to a degree where no meaningful information would be imparted by a null finding, are, however, not only associated with large statistical errors but also with publication and within-study selection bias [cf. Scargle, 2000; Ingre, 2017; Ferguson and Heene, 2012; Yarkoni, 2009; Speech I]. Hence, unless we are dealing with a set of underpowered tests that we know are free from selective reporting and publication bias, it might be a good idea to ignore results from underpowered tests whenever we summarise the existing evidence of a hypothetical association in the research literature. Such a strategy would be good for two reasons. Firstly, it would reduce the risk of being fooled by faulty statistical significance

declarations and bias from non-publication and selective reporting of results. Secondly, it would reduce time wasted on meaningless information.

#### **3.1.4. Another side effect**

An unfortunate consequence of faulty statistical significance declarations, and their ensuing crying wolf effects, is an increased risk that results of properly performed statistical significance tests will be met with disbelief, and that important information thereby will be ignored.

*“... when evaluating published findings in the current academic environment, we should assume a priori that p-hacking and publication bias is likely to be present. This means that we cannot accept published findings at face value, unless there is explicit evidence indicating that the research was protected from these biases.”* [Michael Ingre, 2017].

To counteract the aforementioned side effects, I have contributed to the development of some strategies to i) decrease the risk of being fooled by faulty statistical significance declarations and bias from non-publication and selective reporting of results in the research literature; and ii) increase the readers' certainty that my own research is free from faulty statistical significance declarations and bias from selective reporting of results. These strategies will be delineated in subchapter 3.3 and 3.2, respectively.



### 3.2. Initiatives to improve quality and credibility of statistical analyses and reports

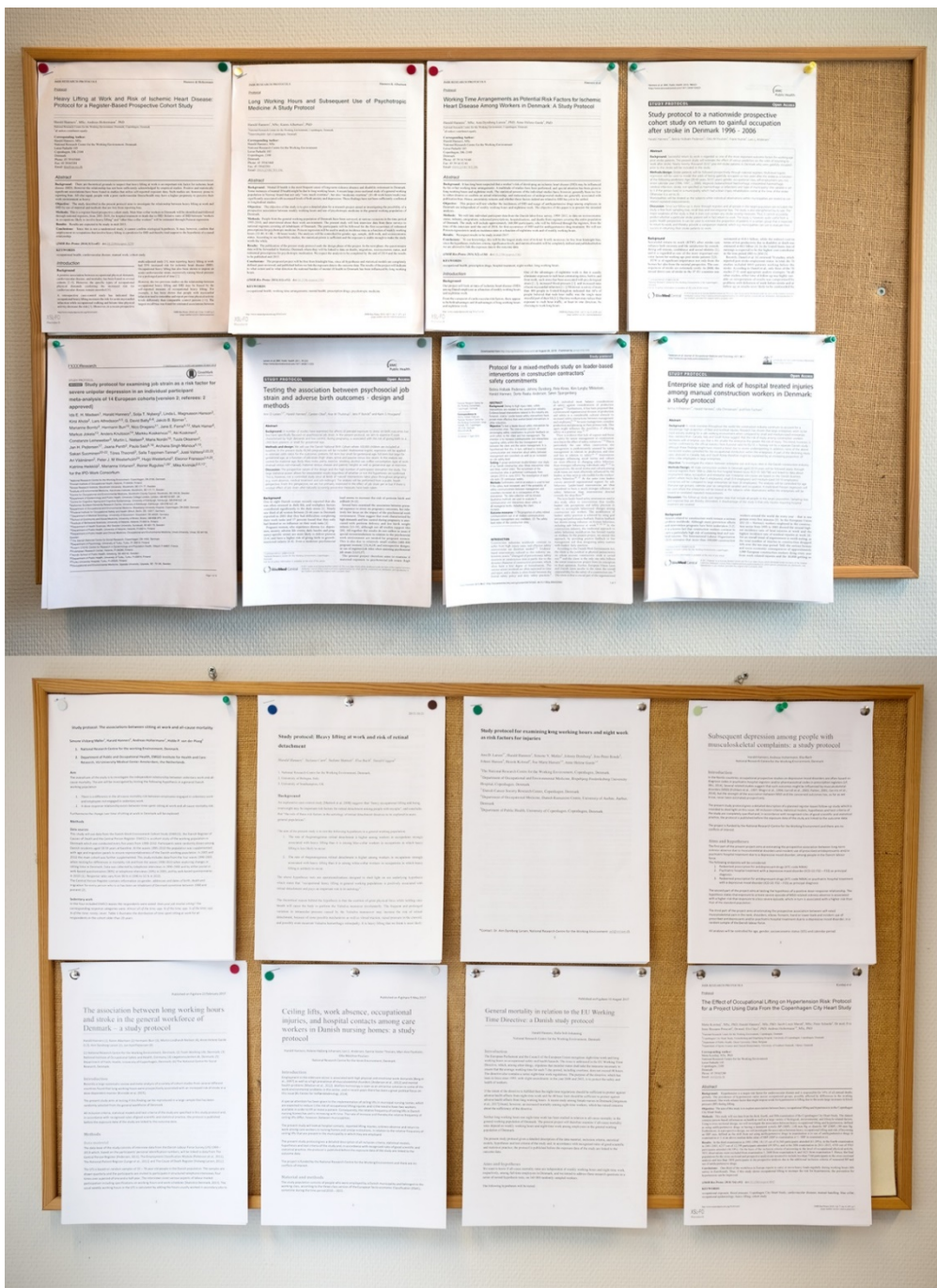


Figure 3.2.1. Pre-published study protocols with H. Hannerz as co-author; photographed by Ole Melkevik, 2018.

Literature on scientific misconduct usually distinguishes between “outright frauds” and “questionable research practices (QRP)” [cf. Fanelli, 2009; John et al., 2012]. In relation to Figure 3.1.1, the first item on the list “Manufacture data that fit your hypothesis” would be classified as “fraud” while the remaining items would be classified as QRPs. Outright fraud appears to be a rare

event, with self-admission rates at 2.0% according to a meta-analysis by Fanelli (2009) and at 1.7% according to a study among 2000 research psychologists, by John et al. (2012). QRPs, on the other hand, are quite common, as illustrated by the following quotation: “Combining three different estimation methods, we found that the percentage of respondents who have engaged in questionable practices was surprisingly high. This finding suggests that some questionable practices may constitute the prevailing research norm” [John et al., 2012].

It has been suggested that a single observational epidemiologic study can be taken seriously only if i) it is very large, ii) there is a very strong association between disease and risk factor (the lower limit of a 95% confidence interval of a risk ratio is greater than three) and iii) there is a highly plausible biological mechanism [cf. Taubes, 1995]. An adoption of such criteria would definitely decrease the risk of being fooled by faulty statistical significance declarations, and personally I would not mind if someone would deem such a finding statistically significant (unlikely to occur by chance), even in the presence of data dredging and harking (Hypothesizing After the Results are Known). It is, however, recognised that an association can have a major public health impact even if it is weak (e.g. with a relative risk at 1.2 or 1.3) [cf. Wynder, 1996; Doll, 1996]. It has, moreover, been proposed that the more conspicuous determinants of non-infectious diseases already have been found [Taubes, 1995] and that from now, “weak associations, like moderate effects of drugs, are the best that we can hope to find in most of our research” [Doll, 1996]. The recommendation to ignore weak associations has therefore been rejected by most researchers in the field, and the continued search for weak associations in combination with a widespread occurrence of questionable research practises and seriously under-powered studies has obviously led to a litany of faulty statistical significance declarations and non-reproducible research findings [cf. Mayes et al., 1989; Young and Alan, 2011]. Consequently, there has been some criticism, which is well illustrated by the below quotations:

“The news about health risks comes thick and fast these days, and it seems almost constitutionally contradictory.” [Taubes, 1995]

“Health-conscious Americans increasingly find themselves beset by contradictory advice. No sooner do they learn the results of one research study than they hear of one with the opposite message.” [Angel and Kassirer, 1994]

“We are fast becoming a nuisance to society. People don’t take us seriously anymore, and when they do take us seriously, we may unintentionally do more harm than good.” [Dimitrios Trichopoulos, head of the epidemiology department at the Harvard School of Public Health (cf. Taubes, 1995)].

“The simple expedient of closing down most university departments of epidemiology could both extinguish this endlessly fertile source of anxiety mongering while simultaneously releasing funds for serious research.” [Le Fanu, 1999]

“Because the BMJ and other major weekly medical journals have cornered the market in splashing data dredged, biased, and confounded associations across the media through their press releases, the profile of quality journals is reduced, much to the chagrin of their editors.” [Smith and Ebrahim, 2002]

The public health research community has undoubtedly contributed to remarkable public health achievements all over the globe [cf. Lash, 2010; CDC, 1999; CDC, 2011a; CDC, 2011b] but, as noted above, there is still room for improvement. In the present subchapter I will address some initiatives to improve the quality and credibility of health and safety related research papers. I will start with initiatives that have been taken by journal editors, then I will deal with some local initiatives that I contributed to as a member of the statistical society at my work place and finally I will present my own personal initiative.

### **3.2.1. Initiatives by journal editors**

To evaluate the reliability and validity of a reported statistical analysis, we need to be able to differentiate between pre-specified and post hoc analyses. We also need to know the pre-specified status of the pre-specified analyses (e.g. confirmatory hypothesis test, exploratory analysis or sensitivity analysis), and to what extent the set of presented analyses and methods differs from that which was planned before the researchers looked at any relation between the exposure and outcome data of the study.

Unfortunately, many health research papers do not contain the information necessary to determine i) whether or not an analysis had been pre-specified and ii) whether or not the results of the study had been selectively reported from a larger set of statistical tests.

Several initiatives aimed at facilitating efforts to evaluate the quality of research papers in peer-reviewed journals have been taken by journal editors. I will not give an exhaustive account of all such efforts, but I will briefly tell of three widely endorsed initiatives, namely, the endorsement of reporting guidelines, trial registrations and registered reports, and how these efforts are expected to reduce the risk that readers will be fooled by faulty statistical significance declarations and selective reporting of results.

#### *3.2.1.1. Reporting guidelines*

Reporting guidelines are tools intended to be used by authors, peer-reviewers and journal editors to ascertain that health research is reported in a way that makes it possible for an educated reader to evaluate the validity and generalisability of the findings. They typically come with a checklist of items that need to be reported, together with an explanatory text that provides a motivation for each of the included items [cf. Stevens et al., 2014; Altman and Simera, 2016].

During the last decades, numerous reporting guidelines, covering different types of health research, have been developed and many of these have been widely endorsed by journal Editors, e.g. CONSORT (Consolidated Standards of Reporting Trials) [Begg et al., 1996; Altman et al., 2001; Moher et al., 2010], PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [Moher et al., 2009], MOOSE (Meta-analysis Of Observational Studies in Epidemiology) [Stroup et al., 2000], TREND (Transparent Reporting of Evaluations with Nonrandomized Designs) [Des Jarlais et al., 2004], STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) [Vandenbroucke et al., 2007], STARD (Standards for Reporting Diagnostic accuracy studies) [Bossuyt et al., 2015] and ARRIVE (Animals in Research: Reporting In Vivo Experiments) [Kilkenny et al., 2010]. Information on these and many other reporting guidelines can be found on the website of the EQUATOR (Enhancing the QUALity and Transparency Of health Research) Network [<http://www.equator-network.org/>],

which has been set up to advance high quality reporting of health research studies [Simera et al. 2010]. In September 2015, the EQUATOR database contained 282 guidelines [Altman and Simera, 2016].

If the reporting of observational studies and randomised controlled trials would follow the spirit of STROBE and CONSORT, respectively, then the readers would be able to differentiate between post hoc and pre-specified analyses. They would, moreover, be able to judge whether or not multiplicity of analyses had been dealt with satisfactorily. It is therefore reasonable to expect that the risk of being fooled by harking, selective outcome reporting and faulty statistical significance declarations would be lower if the research is published in a journal which endorses the guidelines than it would if the research is published in a journal without such guidelines. It is, however, not guaranteed that a paper abides by the guidelines, even if they are endorsed by the journal. It has, for example, been proposed that “a reporting guideline might encourage some authors to report fictitiously the information suggested by the guidance rather than what was actually done” [Schulz et al., 2010].

In a Cochrane review, which included 50 evaluations of RCT reports from a total 16,604 trials, it was determined that the reporting of trials tended to be more complete in CONSORT-endorsing journals compared to non-endorsing journals [Turner et al., 2012], with statistical significance ( $p < .01$ ) for 5 out of 27 evaluated items. It was, nonetheless, concluded that trial reporting often was sub-optimal also among journals that endorsed the CONSORT statement. The effects of the endorsement of a series of other reporting guidelines, among them the STROBE statement, were reviewed and evaluated by Stevens et al. (2014). Stevens et al. did not find any statistically significant difference between papers in STROBE-endorsing and non-endorsing journals nor between papers in STROBE-endorsing journals before and after the endorsement. Their overall conclusion was that the evidence for an association between the completeness of reporting of health research and the endorsement of the studied reporting guidelines was insufficient.

We can, in other words, not trust that the endorsement of reporting guidelines, by itself, will eliminate the risk of being fooled by faulty statistical significance declarations and selective reporting of results.

### *3.2.1.2. Trial registrations*

It is known that selective reporting of trials and measured outcomes within trials tends to contaminate the evidence needed to adjudicate effects (and side-effects) of medical treatments [cf. Chan et al., 2004; Williamson et al., 2005; Von Elm et al., 2008; Smyth et al., 2008; Song et al., 2010; Dwan et al., 2013]. To counter these tendencies the international committee of journal editors (ICMJE) published an editorial in 2004 in which it was announced that, in order to be considered for publication in any of the ICMJE journals ( $n=11$ ), “any research project that prospectively assigns human subjects to intervention and comparison groups to study the cause-and-effect relationship between a medical intervention and a health outcome” would need to be registered in a public trials registry before the onset of patient enrolment [De Angelis et al., 2004]. It was, moreover, stated that “an acceptable registry must include at minimum the following information: a unique identifying number, a statement of the intervention (or interventions) and comparison (or comparisons) studied, a statement of the study hypothesis, definitions of the primary and secondary outcome measures, eligibility criteria, key trial dates (registration date, anticipated or actual start date,

anticipated or actual date of last follow-up, planned or actual date of closure to data entry, and date trial data considered complete), target number of subjects, funding source, and contact information for the principal investigator.” [De Angelis et al, 2004]

Since then, plenty of other journals and organisations have endorsed the ICMJE guidelines [De Angelis et al, 2005; Krleza-Jerić et al, 2005; Hooft et al., 2014] and thereby committed themselves to enforce trial registrations in the same way as the ICMJE journals do. In January 2011, the guidelines were endorsed by 695 journals [Hoft et al., 2014]. On 26 July 2018 the number had increased to 4539 [<http://icmje.org/journals-following-the-icmje-recommendations/>].

The ICMJE guidelines does not mandate observational studies to be registered, but some of the ICMJE-endorsing journals, e.g. The Lancet (2010) and British Medical Journal [Loder, 2010], have announced that they recommend investigators to register their observational studies in a publicly accessible trial registry before the data analysis begins. Moreover, in response to the question “What is your journal’s policy regarding registration of observational studies?” which was part of a survey among Editors of ICMJE-endorsing journals in 2011, 15 out of 153 respondents stated that they required registration, 55 stated that they recommended registration while 83 stated that registration was unnecessary [Hoft et al., 2014].

A proper trial registration will enable journal editors to detect discrepancies between pre-specified and reported outcome measures and thereby reduce their risk of publishing selection biased results. It is, however, not guaranteed that an RCT-report will be free from outcome reporting bias just because it is published in a journal which endorses the ICMJE-guidance. In a recent survey [Hoft et al., 2014] 82% of the editors of ICMJE-endorsing journals answered “no” to the question “For submitted manuscripts, does your journal cross-check the reported data in the manuscript against the prospectively registered data?” while another survey [Mathieu et al., 2013] reported that “only one-third of the peer reviewers surveyed examined registered trial information and reported any discrepancies to journal editors.” Authors have, in other words, plenty of opportunities to get a selectively reported RCT-study accepted for publication.

Fortunately, a publicly accessible trial registry is open to everyone, which gives the readers of the journal in which the trial is reported an opportunity to reduce their risk of being fooled by selective outcome reporting.

### *3.2.1.3. Registered reports*

In 2013, the journal Cortex introduced a new publishing option for confirmatory statistical analyses, which would be called a registered report [Chambers, 2013]. A registered report mandates that the rationale, aims, hypotheses, design and statistical analysis plan are completely defined and peer-reviewed before the researchers collect the data. It also requires that the statistical power to detect an important effect is estimated to be at least 90%. If the protocol is accepted then, given that the researchers abide by their protocol, the ensuing paper will, in principle, be accepted for publication.

Trial registrations, ordinarily, do not require the statistical analysis plan to be pre-specified and completely defined before the data are collected. Hence, the researchers can test their hypothesis in several different ways and selectively report the method which provided the “best” result. The main

advantage of a registered report vs a trial registration is that it closes this loophole. Another advantage is that it reduces the risk of publication bias, since the peer-reviewers and editors are mandated to make their decisions about publication before the results are available. A third advantage is that it mandates the statistical power to be sufficiently large to allow meaningful judgements regardless of the outcome of the test.

By now (31 July 2018), the registered report policy has been adopted by a total of 98 journals [<https://cos.io/rr/#journals>].

### **3.2.2. Local initiatives by the statistical society at NRCWE**

During the last twenty years I have been employed at the Danish National Research Centre for the Working Environment (NRCWE), a governmental research institution, which during the period of my employment has had, on average, approximately 100 researchers and four statisticians on its staff. The vast majority of the research has been quantitative, i.e. evaluated by use of statistical analysis. Until 2011, we had a statistical society at the institute, which consisted of statisticians as well as some methodologists with a special interest in statistical analysis, but with roots in other disciplines (e.g. civil engineering, epidemiology or demography). In the society we met at irregular intervals to discuss methodological challenges, problems and solutions that we had come across in our everyday work as in-house statistical consultants.

As mentioned in the introduction to this subchapter, there had been a lot of scientific media attention on the remarkably high rate of false positive observational epidemiological studies. In 2007, the topic of false positive findings was discussed in one of our society meetings. As statisticians we couldn't help but feeling some responsibility for the quality of statistical analyses that were performed at the institute, even in projects we were not directly involved in. We therefore wanted to make sure that all researchers at NRCWE understood the importance of differentiating between a pre-specified hypothesis test and a post hoc exploratory analysis and how a failure to do so would inflate the probability of false positive results. We also wanted to make sure that they understood how a failure to adjust for multiple testing would lead to an increased probability of false positive results. Last but not least, we wanted to make sure that they understood how selective publication of studies and selective reporting of outcomes within studies may have dire consequences on the validity of the overall evidence on a given subject. To forward these goals, we launched an in-house information campaign in 2008, which consisted of i) a series of lectures, ii) a series of video-speeches and iii) an educational poster.

#### *3.2.2.1. The lectures*

The lectures ( $n = 3$ ) were held by Ole Olsen, a statistician at NRCWE 2002 – 2011, with a previous employment at the Nordic Cochrane Centre 1995 – 2001. The first lecture was entitled “False positive findings in epidemiological studies” and the audience consisted of members of the psychosocial research network at NRCWE. Here, Ole explained the potential problems associated with harking and multiple testing, from a probability theoretic perspective. He also presented empirical results from, inter alia, a study of false positive findings in occupational cancer epidemiology by Swaen et al. (2001), in which the following was concluded: “The strongest factor associated with the false positive or true positive study outcome was if the study had a specific a

priori hypothesis. Fishing expeditions had an over threefold odds ratio of being false positive.” The second and third lecture dealt with publication bias in clinical trials and occupational health research, respectively. The lectures were very popular and the meeting rooms were typically packed with an attentive and enthusiastic audience.

### *3.2.2.2. The video speeches*

During my summer holiday 2008, I drafted three speeches intended to be delivered by some prominent senior research psychologists at NRCWE. The first speech explained why it is important to secure a sufficient statistical power before performing a statistical hypothesis test and how a failure to do so can be expected to bias the research literature away from unity/zero. The second speech dealt with the importance of differentiating between hypothesis generation and hypothesis testing and how a confusion of these two concepts may lead to faulty statistical significance declarations. The third speech told us that, when we do a confirmatory statistical significance test, it is not enough that we state the hypothesis before we look at the results. We also need to state exactly how the hypothesis test will be performed. It explained why this is important and how a failure to comply with this principle may lead to faulty statistical significance declarations.

The drafts of the speeches were read and commented on by members of the statistical society. With the guidance and assistance of Helene Feveile (statistician at NRCWE 1995 – 2010) I thereafter fine-tuned and timed revised versions of the speeches to make sure that they would not be too long and boring for our intended audience (current and future graduate students and researchers at the institute).

The next step in the process was to enlist the participation of the intended speakers-- Karina Nielsen [Speech I], Karen Albertsen [Speech II] and Annie Høgh [Speech III]. They all agreed to participate. The videos were recorded and uploaded to the website of NRCWE. The speeches were moreover published in the in-house magazine of NRCWE.

The respective speeches ended with the following critical statements:

- “It is surely a great criticism of our profession that we do not refrain from performing statistical hypothesis tests that are so underpowered that their outcomes only can be published if they are positive.” [Speech I (Figure 3.2.4)]
- “It is surely a great criticism of our profession that we do not publish research protocols before commencing studies of a confirmatory nature.” [Speech II (Figure 3.2.5)]
- “It is surely a weakness of the editorial process in most of our peer-reviewed journals that we demand that the test should be performed and the results should be submitted before we begin the review process.” [Speech III (Figure 3.2.6)].

Here I note that the reforms that were called for in the video speeches are exactly the same as the ones that were introduced by Chambers (2013) in his registered reports initiative. I will therefore end this section with a thank you to Professor Chambers as well as all other journal editors who are endorsing the concept.

### 3.2.2.3. The educational poster

The educational poster is shown in Figure 3.1.1. It was produced at NRCWE in 2009 with text by the statistical society and photo shopping by Pia Dukholm. The poster likens the statistical testing of a hypothesis to the placing of a bet in roulette. The first box of the poster lists five theoretical ways of cheating in roulette. Each of the methods of cheating in roulette has an analogous method of cheating in confirmatory statistical analysis, and these are listed in the second box.

The initial plan was to make one large poster to be hanged at a strategic place of the institute, where it would serve as a constant reminder of the rules that we need to abide by whenever we perform a confirmatory statistical hypothesis test. Consequently, we applied for permission to print the image in the form of a large poster. The application was, however, denied and we had to settle with a smaller poster in size A3 (297 × 420 mm), which we posted at a bulletin board in one of the corridors of the institute.

**The roulette survey**

**1. How long have you been involved in research activities?**

- *Never* \_\_\_\_\_
- *<3 years* \_\_\_\_\_
- *3-9 years* \_\_\_\_\_
- *>= 10 years* \_\_\_\_\_

If 'never' then end.

**2. Have you assisted in or performed any statistical hypothesis test?**

- *Yes* \_\_\_\_\_
- *No* \_\_\_\_\_
- *Don't know* \_\_\_\_\_

If 'no' or 'do not know' then end.

[Please read the 'poster' on the back page before answering the following questions.]

**3. Have you ever witnessed any of the behaviours listed in the lower box of the poster?**

- *Yes* \_\_\_\_\_
- *No* \_\_\_\_\_
- *No comment* \_\_\_\_\_

Figure 3.2.2. The questionnaire

To increase the awareness of the existence of the poster and the wisdom it imparts, we initiated a mini-survey with a questionnaire presented on a single sheet of paper that had an image of the poster on one side and three survey questions (see Figure 3.2.2.) on the other. The questionnaires were handed to research staff at NRCWE (n = 93) and data were obtained from December 2009 to



March 2010. One researcher refused to participate and one did not return with an answer. The remaining invitees replied to the questionnaire. To ascertain anonymity the responders were asked to put their response into a ballot box, at which time they were checked off as responders. They were thereafter asked if they wanted to receive an A3 sized copy of the poster. Consequently, we had to hand out a total of 75 poster copies.

In total, 91 persons responded to the questionnaire whereof one answered 'Never' to the first screening question "How long have you been involved in research activities?" and 20 answered 'No' to the second screening question "Have you assisted in or performed any statistical hypothesis test?". The remaining participants (n = 70) went on to the third and final question "Have you ever witnessed any of the behaviours listed in the lower box of the poster?" The distribution of the answers to the last question is given in Table 3.2.1.

Table 3.2.1. "Have you ever witnessed any of the behaviours listed in the lower box of the poster?"

Research experience	Yes	No	No comment	"Not sure"	Total
< 3 years	5	7	0	0	12
3 – 9 years	12	9	1	0	22
>= 10 years	29	5	0	2	36
<b>Total</b>	46	21	1	2	70

We were glad to see that only about half of the researchers with less than 10 year of experience had ever witnessed any of the research behaviours listed in the poster. We were also glad to see so many positive reactions to the survey as well as the poster.

In a real life casino it is very difficult (almost impossible) to cheat in roulette. The investment of the casino owners is secured not only by surveillance cameras and security guards but also by the transparency of the game itself. Everything happens in plain view in front of the players, the croupier and whatever audience is present at the table. We note, however, that all of the "five ways of cheating in roulette" would be possible if the gamblers were allowed to place their bets behind a closed curtain and wait until the ball landed, before they unveiled the curtain to reveal their bets and the number the ball landed on. We also note that such conditions would be analogous to the way that most observational health studies were conducted prior to the introduction of trial registrations and registered reports.

In that regard, I hope that our national research funding agencies will adopt some of the security thinking of the casinos and thereby realise that they can reduce the risk of being cheated by research behaviours 2 – 5 of Figure 3.1.1 simply by adding, as a condition for the grant, that any confirmatory statistical analysis should be preceded by a trial registration and pre-publication of a completely specified statistical analysis plan.

### 3.2.3. Personal initiatives

Towards the end of the 2000s, after having gone through some literature on questionable research practices, I had come to realise that the credibility and thereby the value of a confirmatory statistical analysis would be considerably enhanced if the researchers could document that their study is free from faulty statistical significance declarations and bias from selective reporting. I had also come to realise that a confirmatory statistical analysis without such documentation runs a considerable risk of having its statistical conclusions dismissed as probable hindsight rationalisations.

These realisations prompted me to start looking for ways to make the pre-specified statistical analysis plan available to the readers whenever the results of a confirmatory statistical analysis were published, so as to enable them to see exactly what was planned before the researchers looked at any relation between the exposure and outcome data of the study.

At that time I found out that some public health related journals encouraged researchers to submit their study protocols for peer-review and possible publication, and a few years later I found out that it was possible to publish research protocols for free at Figshare.com.

Consequently, I have co-authored a series of study protocols of which some have been published in peer-reviewed journals [Hannerz et al., 2010, 2014a, 2014b, 2016a; Korshøj et al., 2018; Larsen et al., 2011; Madsen et al., 2014; Pedersen et al., 2010, 2011] and some have been published on Figshare.com [Hannerz et al., 2013, 2016b, 2017a, 2017b, 2018; Larsen et al., 2016; Møller et al., 2014].

With time I have come to regard the publication of non-confidential confirmatory statistical analysis plans not only as an option but also as a duty, especially if the project is funded by tax money, and during the last five years I have felt obliged to abide by a self-developed code of conduct, which reads:

If I am in charge of a tax payer funded statistical significance test then, whenever it is possible to do so, I have a moral obligation to ascertain

- i. that the statistical analysis plan is completely specified and documented, before I or anyone else in the research team looks at any relation between the exposure and outcome data of the test
- ii. that the documented analysis plan (exactly as written before the researchers looked at any relation between the exposure and outcome data of the test) will be available to the readers whenever the results of the test are published

I do not regard this code as a self-imposed burden. Au contraire, I see it as a tremendous aid and stress reliever, especially if the analysis plan is published before the analysis phase of the project begins. Firstly, it will make the co-authors realise that they cannot wait until the results are known before they give their opinions about the design. Secondly, it will block pressures from co-authors, peer-reviewers and editors to change the narrative of the statistical analysis plan after the results are known. Thirdly, it may decrease the risk that a boss will shut down the project after the results are known, in order to hide politically incorrect or otherwise disappointing findings. Fourthly, it

reduces the risk that valid statistical conclusions will be dismissed as probable hindsight rationalisations.

There is, however, one drawback associated with the publication of study protocols that needs to be mentioned, namely the risk of being accused of plagiarism when methodological details of the study protocol are repeated in the paper in which the results of the study are published. This happened to me in connection with the publication of Paper VII where I received an e-mail from the Editorial office of BMJ Open, which contained the following text:

*“The reviewer(s) have recommended publication. However, our editorial checks have identified a high level of text overlap in your manuscript. We require that you make some revisions to the text of your article in order to remove text overlap with previously published articles. We noted that in several places your article contains whole sentences/paragraphs of text overlap as indicated in the attached report.*

*Please be aware that copying extracts from previous publications is not acceptable. As a member of Committee on Publication Ethics (COPE), BMJ Open takes seriously all suspected cases of plagiarism. In line with the COPE guidelines for cases involving plagiarism, we require that provide an explanation as well as a revised version of your manuscript. Please revise the text to remove overlap.”*

I responded to the plagiarism allegation with the following text:

*“EXPLANATION FOR THE 26% OVERLAP WITH [WWW.RESEARCHPROTOCOLS.ORG](http://WWW.RESEARCHPROTOCOLS.ORG)  
A pre-published study protocol, in which the aims, hypotheses, inclusion criteria, statistical significance criteria and statistical methods are completely defined before the exposure data of the project are linked to the outcome data, is valuable because it allows the readers to evaluate the credibility of the study when the results are published. By comparing the method section in the study protocol with the one in the paper in which the results are given, they are able identify protocol violations and thereby reduce their risk of being fooled by faulty statistical significance declarations and bias from selective reporting of results.*

*Such a comparison would obviously be easier if the aims, hypotheses and methods were worded and described in the same way in the result publication as they were in the study protocol than it would if they were worded and described in a completely different way (no overlaps).*

*We therefore wanted the method description of our paper to be as similar as possible to the method description of our study protocol, and tried to accomplish this through the following strategy:*

*1. We published our study protocol in a paper with the following copyright statement:*

*“Copyright*

*©Harald Hannerz, Ann Dyreborg Larsen, Anne Helene Garde. Originally published in JMIR Research Protocols (<http://www.researchprotocols.org>), 22.06.2016.*

*This is an open-access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Research Protocols, is properly cited. The complete bibliographic information, a link to the original publication on <http://www.researchprotocols.org>, as well as this copyright and license information must be included.” [<http://www.researchprotocols.org/2016/2/e130/>]*

2. We initiated the method section of our manuscript with the following paragraph:

*The statistical analyses were governed by a study protocol [14] which was written, peer-reviewed, and published before we linked the exposure data to the outcome data. The protocol defined all hypothesis and statistical methods for two studies. One study concerned the association between weekly working hours and risk of IHD or antihypertensive drug usage (reported here) while the other concerned the association between night-time work and risk of IHD or antihypertensive drug usage (results to be reported elsewhere). The present method section will repeat methodological details of the study protocol which pertain to the study on weekly working hours.*

3. We included the complete bibliographic information of the study protocol in the reference list of our manuscript and inserted an additional link to it in the abstract.

*We failed, however, to notice that we needed to include the above copyright and license information in the manuscript. We apologise for this mistake.”*

I thereafter uploaded a revised version of the manuscript in which I i) included the copyright statement of the study protocol and ii) added quotation marks around all text passages that were repeated from it. The manuscript was thereby accepted without removal of the text overlap between the manuscript and the previously published study protocol.

The moral to this story is that text from the method section of a study protocol can be repeated when the results of the study are published if i) the authors owns the copyright of the protocol and ii) all quoted text passages are surrounded by quotation marks.

Another way of circumventing the self-plagiarism problem, developed and practiced by Karin Sørig Hougaard at NRCWE, is to refrain from publishing the protocol and instead keeping it in a safe, where it is put after it has been dated and signed by all involved co-authors.

#### **3.2.4. Discussion**

It should be noted that the publication of a pre-specified study protocol does not hinder researchers from performing post-hoc exploratory analyses nor does it hinder them from changing the design of the study after they have looked at the results. It merely inhibits them from misreporting their post-hoc exploratory analyses as confirmatory hypothesis tests.

It should also be noted that a recommendation to publish the statistical analysis plan of a confirmatory statistical hypothesis test in no way discourages nor disparages the worth of properly reported exploratory analyses.



Figure 3.2.3. Karin Sørig Hougaard; photographed by Ole Melkevik, 2018.

### Statistical power and publication bias

My name is Karina Nielsen. I am a senior psychologist at the Danish National Research Centre for the Work Environment. Today I am going to talk to you about the importance of securing a sufficient statistical power before performing a statistical hypothesis test and how a failure to do so can be expected to bias the research literature away from unity/zero.

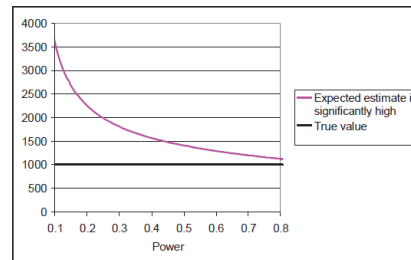
Consider the following situation: A questionnaire has been used on a representative sample of people from a certain population. From the questionnaire we have, among a lot of other things, information about height and monthly pay. One of our hypotheses is that women who are taller than 170 cm are better paid than women who are shorter than 160 cm and we want to estimate the difference in the average monthly income between these groups. The estimation and the hypothesis testing will take place simultaneously. If the lower boundary of a 95% confidence interval is greater than zero then we will accept the hypothesis.

Before performing the test, we calculate the power. Let's say we believe that the true difference is 1000 Dkr per month (which means that the tall women earns approximately 5% more than the short ones) and that our calculations show that the power of finding such a difference in our material is 0.2. Since there is a one-to-one relation between the power of the test and the width of its associated confidence interval we can also determine that should the estimate be what we expect it to be then the estimated 95% confidence interval would go from -750 to 2750. [Dr. Nielsen draws the confidence interval on the white board.] It is clear to us that this interval is so wide that the estimate would be considered meaningless and that we wouldn't be able to publish such a result.

At this point we can choose not to test the hypothesis. We can also reason that we most probably would be able to publish the result if the estimated difference were significantly high and choose the following strategy: We test the hypothesis. If the difference is significantly high then we submit it for publication otherwise we pretend that we didn't test it.

This is, however, a seriously biased estimation strategy. If the power of the test was 20% and the true income difference was 1000 Dkr. then the least possible significantly high estimate of income difference would be 1750 Dkr. [Dr. Nielsen draws a parallel displacement of the earlier drawn confidence interval. (See photo to the right).] We would, in other words, only be able to publish the result if the estimated difference was at least 75% larger than the true difference.

[Dr. Nielsen shows and comments on the graph below] This is a graph, which exemplifies the publication bias that will arise from the strategy to publish only if significantly high, as a function of the power of the test. The upper line gives the expected estimate given that it is significantly high, while the lower line gives the true value. The gap between the two lines indicates the publication bias and, as you can see, the lower the power the greater the bias. The graph shows that in or example, where the power was 20% and the true income difference was 1000 Dkr, the expected estimated income difference, given that it is significantly high, is about 2200 Dkr. —more than twice as large as the true one.



This is something one should be aware of not only as a researcher but also as a reader of research literature. When we encounter a statistically significant estimate we have to ask ourselves if that estimate would have been published if it wasn't statistically significant.

It is also important to be aware of this mechanism when we formulate inclusion criteria for meta-analyses.

It is surely a great criticism of our profession that we do not refrain from performing statistical hypothesis tests that are so underpowered that their outcomes only can be published if they are positive.



The above text was written by Harald Hanerz in collaboration with members of the statistical society and the psychosocial competence forum at the Danish National Research Centre for the Work Environment, 2008.

Figure 3.2.4. Speech I: Nielsen KM (speaker), Hanerz H (speechwriter). Statistical power and publication bias

## Use and misuse of statistical tests

My name is Karen Albertsen. I am a senior research psychologist at the Danish National Research Centre for the Work Environment. Today I am going to talk to you about the importance of differentiating between hypothesis generation and hypothesis testing and how a confusion of these two concepts may lead to false P-values and falsified research reports.

As an example, let's say that we want to explore the possibility that people with extra sensory perception (ESP) exist, and we want to screen for a suitable candidate who we can subject to a variety of ESP tests. Consider the following experiment: An interviewer stands in a street corner with a portable computer, which, at the click on a button, will generate a random integer between 0 and 999. When a person walks by he/she is asked if he/she wants to participate in the study. If the person accepts, a random number is generated and the participant, who is not allowed to look at the computer screen, is asked to tell the interviewer which number he/she believes it is. Before doing so he/she is told that the number is an integer between 0 and 999.

This process is continued until we find a person who answers the question correctly. Let's say that the 100<sup>th</sup> person, whose name is Florence, is the first one who answers correctly. Now we are done with our exploratory study and we have generated the hypothesis that Florence is a psychic.

At this point we can choose to do a confirmatory study where we test the hypothesis that was generated in the exploratory study. We can also choose to pretend that the exploratory analysis was a confirmatory analysis. In other words, we pretend that we had hypothesised that Florence was a psychic already before she answered the question. If we do this we can

go ahead and publish a paper and since the chance that a mere guess would provide a correct answer to our question is 1/1000 we can report a P-value of 0.001 (which is generally considered to be quite significant). This P-value is, however, false. In this context, the true probability of a correct answer equals 1 minus the probability of 100 consecutive incorrect answers [ $1 - 0.999^{100} = 0.095$ ], which is almost 100 times greater than the reported P-value.

Can we ever know whether something published as a confirmatory analysis actually was a confirmatory analysis?

Well, sometimes we can. If a research protocol exists, which predates the data collection, and if the protocol was adhered to then we can truly know that we are dealing with a confirmatory analysis.

It is surely a great criticism of our profession that we do not publish research protocols before commencing studies of a confirmatory nature.



Dr. Karen Albertsen

The above text was written by Harald Hamerz in collaboration with members of the statistical society and the psychosocial competence forum at the Danish National Research Centre for the Work Environment, 2008.

Figure 3.2.5. Speech II: Albertsen K (speaker), Hannerz H (speechwriter). Use and misuse of statistical tests

### More on the dos and don'ts of statistical testing

My name is Annie Høgh. I am a senior research psychologist at the Danish National Research Centre for the Working Environment.

When I supervise graduate students I always tell them that, when they do a confirmatory analysis, it is not enough that they state the hypothesis before they look at the results. They also have to state exactly how the hypothesis test will be performed. Now I will explain why this is important and how a failure to comply with this principle may lead to false P-values and falsified research reports.

Consider the following example: We want to test the hypothesis that there is an association between social isolation at work and backache. To our disposal we have a data set obtained through a questionnaire used on a representative sample of people from a certain population.

Before we perform the test we make the following decisions:

- We will use logistic regression to test the hypothesis.
- The significance level will be set to 0.05
- The outcome will be based on the question 'Do you currently suffer from backache?', which could be answered with either 'Yes' or 'No'.
- The explanatory variable will be based on the question 'Is it possible for you to talk with colleagues when you are working?' which could be answered with one of the following reply categories: 'Almost all the time', 'Approx. ¾ of the time', 'Approx. ½ of the time', 'Approx. ¼ of the time', 'Seldom or Never'.
- The explanatory variable will be dichotomised.

Since there are five reply categories there are four possible cut-points to choose between when we do the dichotomisation. Our strategy is to try all dichotomisations and use the one which gives the 'best' result. Let's say that one of the dichotomisations renders a P-value of 0.048. Let's also say that we have forgotten that the cut-point of the dichotomisation was not decided upon until after we had looked at the result. Since P is below 0.05, we can now report that we found a statistically significant association between social isolation at work and backache.

This P-value is, however, false. If there is no association between the examined variables, then, for

any given dichotomisation, the probability that a P-value will be less than or equal to 0.05 is 0.05, but the probability that at least one of the four different dichotomisations would yield a P-value that is less than 0.05 is much higher than that. I recently performed a Monte Carlo simulation, which is a test of repeated random sampling to estimate parameters and probabilities, where I found that this probability was equal to 0.16. In other words, the true P-value should be more than three times higher than the falsely reported one.

The moral to this story is that the ideal situation with regard to a confirmatory statistical analysis would be that

1. The statistical model is completely defined before the test is performed.
2. If the work is to be subject to peer-review then the peer-reviewing of the methodology is done before the test is performed.
3. When the results of the test have been obtained, the statistical model is not to be changed.

Unfortunately the above sequence and principles are seldom adhered to. The following sequence is, however, quite common:

1. The test is performed
2. The work is submitted for peer-review
3. The reviewer (after having looked at the result) suggests changes to the statistical model
4. The Editor demands that the statistical model is changed in accordance with reviewer's comments
5. ...

It is surely a weakness of the editorial process in most of our peer-reviewed journals that we demand that the test should be performed and the results should be submitted before we begin the review process.



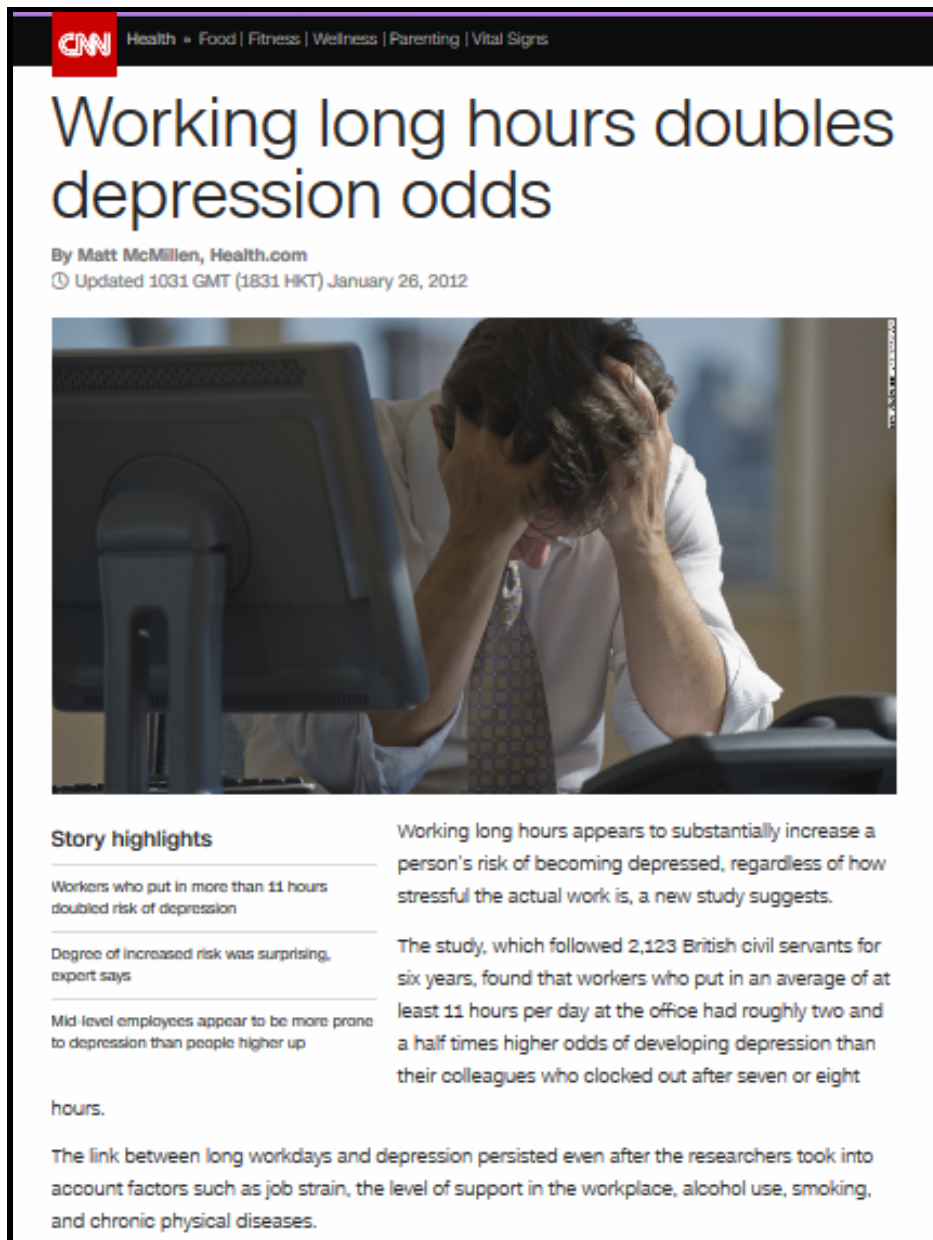
Dr. Annie Høgh

The above text was written by Harald Hannerz in collaboration with members of the statistical society and the psychosocial competence forum at the Danish National Research Centre for the Work Environment, 2008.

Figure 3.2.6. Speech III: Høgh A (speaker), Hannerz H (speechwriter). More on the dos and don'ts of statistical testing.



### 3.3. Statistical power calculations as a means of reducing the risk of being fooled by publication bias and faulty statistical significance declarations



The image is a screenshot of a CNN news article. At the top left is the CNN logo. To its right, a navigation bar lists categories: Health, Food, Fitness, Wellness, Parenting, and Vital Signs. The main headline reads "Working long hours doubles depression odds" in a large, bold, black font. Below the headline, it says "By Matt McMillen, Health.com" and "Updated 1031 GMT (1831 HKT) January 26, 2012". A photograph shows a man in a white shirt and tie sitting at a desk, looking stressed with his hands on his head. Below the photo, the article text is organized into sections. A "Story highlights" section on the left contains three bullet points: "Workers who put in more than 11 hours doubled risk of depression", "Degree of increased risk was surprising, expert says", and "Mid-level employees appear to be more prone to depression than people higher up". The main text on the right states: "Working long hours appears to substantially increase a person's risk of becoming depressed, regardless of how stressful the actual work is, a new study suggests. The study, which followed 2,123 British civil servants for six years, found that workers who put in an average of at least 11 hours per day at the office had roughly two and a half times higher odds of developing depression than their colleagues who clocked out after seven or eight hours. The link between long workdays and depression persisted even after the researchers took into account factors such as job strain, the level of support in the workplace, alcohol use, smoking, and chronic physical diseases."

Figure 3.3.1. Screen dump [<https://edition.cnn.com/2012/01/25/health/working-overtime-doubles-depression/index.html>]

Recently, I and my colleagues Hermann Burr and Jan Hyld Pejtersen devised a very simple procedure to deal with underpowered tests in a literature review (see Paper XV). It goes like this: 1. State an effect that you believe is realistic and of practical importance. 2. State a criterion for acceptable power. 3. Calculate the statistical powers of the tests of interest. 4. Exclude the underpowered tests from your literature overview.

There is nothing new with setting a minimum sample size as a criterion for inclusion in a literature review. This was, for example, done by Albertsen et al. (2006) who required that a study should comprise at least 100 persons to be eligible for inclusion in their systematic review of the impact of work environment on smoking cessation, relapse and amount smoked. An inclusion criterion that is based on formal power calculations had, however, as far as I know, never been seen before. A summary of the systematic review (Paper XV) in which the power-based inclusion criterion was introduced is given below.

### **3.3.1. A systematic review on work-related psychosocial factors and ischaemic heart disease**

#### *3.3.1.1. Objective*

The aim of the study was to update a previous systematic review [Eller et al., 2009] on work-related psychosocial factors and the development of ischaemic heart disease (IHD).

#### *3.3.1.2. Inclusion criteria*

Our updated review had the following four inclusion criteria of which the first three were used in the original review:

1. Study: a prospective or case-control study if exposure was not self-reported (prognostic studies excluded)
2. Outcome: definite IHD determined externally
3. Exposure: psychosocial factors at work (excluding shift work, trauma, violence or accidents, and social capital)
4. Statistical power: acceptable to detect a 20% increased risk of IHD.

The original review comprised 33 papers. The search date of the updated review was set at April 28, 2013. Eleven new papers which met the inclusion criteria 1–3 were found. Hence, a total of 44 papers were to be evaluated regarding inclusion criteria 4. Nine of the papers concerned prospective studies with aggregated data [Reed et al., 1989; Alfredsson et al., 1985; Johnson et al., 1989; Alterman et al., 1994; Steenland et al., 1997; Andersen et al., 2004; Eaker et al., 2004; Vahtera et al., 2004; Bonde et al., 2009]; 28 of the papers concerned prospective studies with person-based exposure [Kivimäki et al., 2012; Suadicani et al., 1993; Haan, 1988; Theorell and Floderus-Myrhed, 1977; Netterstrøm and Juel, 1988; Siegrist et al., 1992; Lynch et al., 1997; Kivimäki et al., 2002; Lee et al., 2002; Matthews and Gump, 2002; Lee et al., 2004; De Bacquer et al., 2005; Kivimäki et al., 2005; Elovainio et al., 2006; Kornitzer et al., 2006; Kuper et al., 2006; Netterstrøm et al., 2006; André-Petersson et al., 2007; Chandola et al., 2008; Kivimäki et al., 2008; Väänänen et al., 2008; Nyberg et al., 2009; Allesøe et al., 2010; Holtermann et al., 2010; Netterstrøm et al., 2010; Virtanen et al., 2010; Kivimäki et al., 2011; Slopen et al., 2012]; and 7 of the papers concerned case-control studies [Alfredsson et al., 1982; Iversen et al., 1989; Johnson et al., 1996; Ferrie et al., 1998; Hammar et al., 1998; Theorell et al., 1998; Sokejima and Kagamimori, 1998].

#### *3.3.1.3. Power Calculations*

In observational cohort studies, Monson (1990) recommends epidemiologists to interpret rate ratios in the open interval 0.9 to 1.2 as 'no association' to allow for possible effects of selection bias, misclassifications and uncontrolled confounding. We recognised that death or hospital treatment due to IHD is a serious endpoint and that IHD is the leading cause of death worldwide [Lozano et al.,

2012]. Effects of individual occupational risk factors may therefore be regarded as clinically important even if the rate ratio is less than 1.2 [cf. Ha et al., 2011]. However, in keeping with Monson’s recommendation, we did not want our critical value for clinical significance to be less than 1.2. We therefore considered a 20% increase in risk of IHD due to work environmental exposures to be an important effect, and we wanted to know the power to detect such an effect in the 44 individual papers found. In our opinion, a 95% power is desirable and an 80% power is acceptable. For each paper, we calculated the power to detect a rate ratio of 1.2 through the following procedure:

When available, the confidence interval of a published rate ratio was used to estimate the *Stderr* of its logarithm by means of the equation

$$Stderr = \frac{\log(UpperCL) - \log(LowerCL)}{2\Phi^{-1}(1-\alpha)} \quad (3.3.1)$$

where UpperCL is the upper confidence limit and LowerCL is the lower confidence limit of a 100(1 – 2α)% confidence interval. The power to detect a rate ratio of 1.2 as a function of *Stderr* and α was thereafter approximated by the equation

$$Power = \Phi\left(\frac{\log(1.2)}{Stderr} - \Phi^{-1}(1 - \alpha)\right) \quad (3.3.2)$$

where Φ is the standard normal distribution function. For papers, which gave the expected number of cases but no confidence interval, we replaced *Stderr* in Equation 2 with

$$E[Stderr] = \sqrt{\frac{1}{e_1} + \frac{1}{e_2}} \quad (3.3.3)$$

where  $e_1$  is the expected number of cases among the exposed and  $e_2$  is the expected number of cases in the comparison group. Three of the papers contained neither of the above [Suadicani et al., 1993; Haan, 1988; Theorell and Floderus-Myrhed, 1977] but it was obvious from the extraordinary low number of cases that the power would be < 0.1.

In some cases, the expected direction of the relationship between compared groups was reversed: for example, the rate ratio between workers with low vs high control was expected to be > 1, but the rate ratio of high vs low control was expected to be <1. In such cases, a rate ratio was deemed clinically important if it was ≤ 1/1.2. Because the standard normal probability density function is symmetric around zero, the power to detect a rate ratio of 1/1.2 would be equal to the power to detect a rate ratio of 1.2. The equations are based on the central limit theorem and Gauss’ propagation of error formulas. The derivation of the power formula is given by Bickel and Doksum (1977).

#### 3.3.1.4. Results

We were able to calculate the statistical power for 169 out of 170 significance tests in the 44 papers. Thirty-six tests were found among the prospective studies with aggregated data, 111 were found among the prospective studies with self-reported data and 22 were found among the case-control

studies. The statistical power to detect a rate ratio at 1.2 ranged from 0.04 to 0.99. The median power was 0.11. Only 10 tests had an acceptable power ( $\geq 80\%$ ). Four of the acceptably powered tests were performed in a meta-analysis by Kivimäki et al. (2012). The remaining 6 tests were performed in a case-control study by Hammar et al (1998) using a job exposure matrix.

In the meta-analysis by Kivimäki et al. (2012), the risk ratios for coronary heart disease were estimated at 1.04 (95% CI: 0.92 to 1.17) and 0.86 (0.79 to 0.96) for 2 standard deviations increase in job demands and job control, respectively, after adjustment for age and gender. The risk ratio for coronary heart disease among workers with vs without job strain was estimated at 1.23 (1.10 to 1.37) after adjustment for age and gender, and at 1.17 (1.05 to 1.31) after adjustment for age, gender and socioeconomic status.

In the case-control study by Hammar et al. (1998), the rate ratios of myocardial infarction among men in occupations with low vs. high decision latitude in Sweden were estimated at 1.37 (95% CI: 1.25 to 1.50) and 1.12 (1.05 to 1.19) for the age categories 30 - 54 and 55 - 64 years, respectively. The corresponding rate ratios for high vs low demands were estimated at 0.93 (0.84 to 1.02) and 0.95 (0.89 to 1.01) and the corresponding rate ratios for low vs high social support were estimated at 1.28 (1.17 to 1.41) and 1.10 (1.04 to 1.17).

Comment: Regarding the study by Hammar et al. (1998), I noticed the following errors in Table 1 of Paper XV: Men in the age category 30 - 54 years have been miscoded as “men without further specification” while men in the age category 55 – 64 years have been miscoded as “women without further specification”.

#### *3.3.1.5. Concluding remarks*

It was concluded that the studies aimed at examining associations between the psychosocial work environment and IHD often have been too small to detect important effects, and that there is a need for considering statistical power in the planning of such studies.

#### **3.3.2. Another example of power calculations in a literature overview**

In connection with Paper II, I and my colleague Karen Albertsen made a literature overview of the existing evidence of a prospective association between long working hours and mental health problems. We found a total of 44 disjoint risk ratios for various type of mental health problems among workers with long vs standard working hours - 2 from our own study and 42 from 12 other research papers. For each of the concerned tests, we wanted to know the statistical power to detect an effect of long working hours. The effect sizes of interest were firstly a risk ratio of 1.2, which is classified as a weak association according to Monson's guide to strength of association [Monson, 1990], and secondly a risk ratio of 1.5, which is classified as a moderate association. Only 3 and 10 of the 44 tests had an acceptable power to detect a weak and moderate effect, respectively.

The risk ratios and statistical powers are given in the supplementary appendix of paper II while a plot of the risk ratios against the statistical power to detect a weak effect is given in Figure 3.3.2. The figure suggests that risk ratios in tests with acceptable power tend to be close to unity (no association according to Monson). It also suggests that the news about a 2.5 time increase in depression among workers with long hours (Figure 3.3.1) may have been a false alarm, caused by a

severely underpowered test in combination with a misconception that an unacceptably low power only is a problem if the P-value of the test is greater than 0.05.

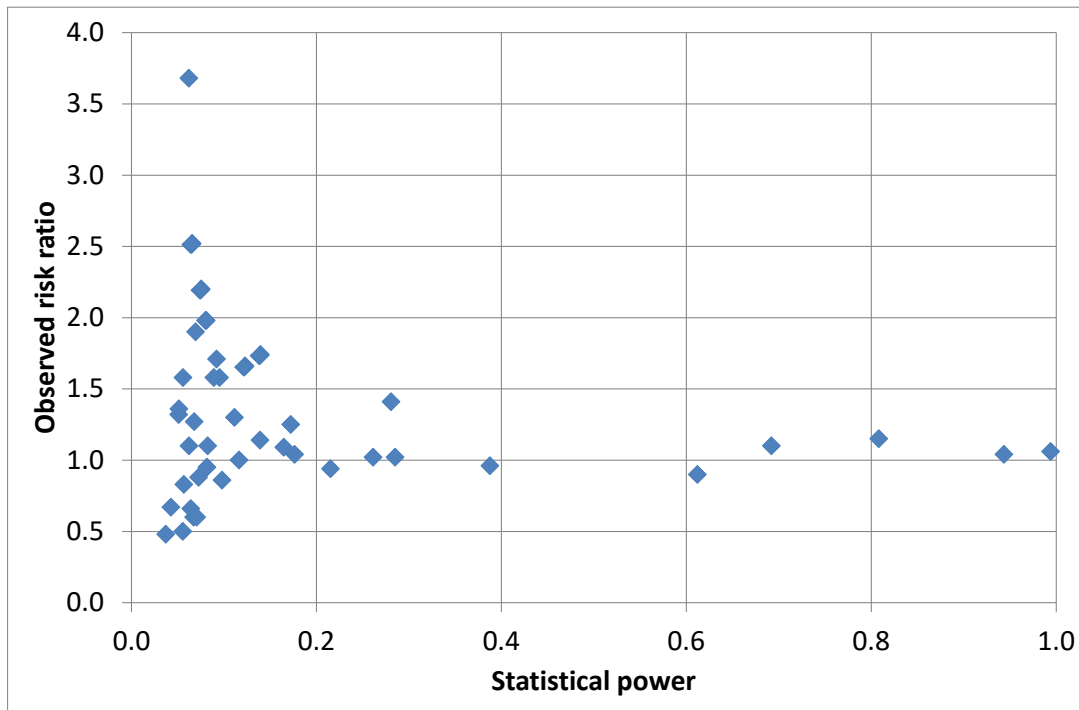


Figure 3.3.2. Observed risk ratio as a function of the statistical power to detect a hypothetical 20% increased risk of mental ill health among workers with long versus standard working hours.

### 3.3.3. Discussion

For the sake of simplicity, I referred to estimated risk ratios as “tests” in the above text. Given the standard error and significance level of the “test”, I thereafter calculated the probability that the lower limit of a confidence interval of the risk ratio would be higher than unity if the true rate ratio equalled 1.2, and subsequently referred to this as the statistical power to detect a positive association if the true rate ratio equalled 1.2. The reason for calling them tests is that they often are interpreted as such in literature overviews on the existing evidence of hypothetical associations. It should, however, be noted that some estimates may have been obtained for exploratory purposes without any a priori hypothesis attached to them. It should also be noted that a bona fide statistical significance test needs to take the context of the experiment into account, which is not always the case with tests that are based on confidence intervals. The procedure can therefore be viewed as a quick and dirty but still quite effective way of reducing the risk of being misled, dispersed or confused by estimates that are too uncertain to impart any meaningful information.

There are, however, some additional drawbacks and limitations of the method that need to be mentioned. Firstly, it requires two arbitrary decisions, namely, i) how strong the studied association should be in order to be of practical importance, and ii) how large the statistical power should be in order to be acceptable. Researchers have often read a substantial portion of the available literature on a subject before they start their literature overview. It is therefore possible that their decisions about the above cut-points may be influenced by their a priori expectations of the association they intend to examine. Secondly, although studies with a high statistical power are less prone to

publication bias than studies with a low statistical power, there is still no guarantee that a literature review which only includes studies with a sufficient statistical power will be free from publication bias [cf. Ingre, 2017]. In other words, the method can reduce random variation related bias but it cannot eliminate it. If we want to accomplish a literature review that is completely free from bias due to selective reporting of results then we would need to restrict our inclusion to papers which come with a pre-published and completely defined statistical analysis plan.

The usefulness as well as the necessity of having access to study protocols in order to rule out within-study selection bias was well formulated in a debate article by Bracken (2011), where he summarised the results of a study by Chan et al. (2004): *“In a seminal paper, Chan et al compared protocols for randomized trials with the final trial reports to document significant bias in outcome reporting. In 34% of reports, the protocol primary outcome was published as a secondary outcome; in 26%, the protocol primary outcome was not reported at all; in 19%, the protocol secondary outcomes were published as primary; and in 17%, the published primary outcome was not mentioned in the protocol. Overall, 62% of published trials showed discrepancies between the protocol and the published primary outcome. Perhaps not surprisingly, statistically significant outcomes in the originally declared protocols were 2 to 3 times more likely to be fully reported in the trial report than nonsignificant outcomes. Of particular relevance to observational epidemiology, statistically significant outcomes concerning harm (which can often be studied only by observational epidemiology) and defined in the protocol were 4 to 5 times more likely to be reported than nonsignificant outcomes. Interestingly, 86% of authors denied the existence of unreported outcomes despite clear evidence to the contrary.”*

## ***4. Concluding remarks and recommendations***

To protect workers against overwork-related safety and health problems, the EU Working Time Directive [EU, 2003] stipulates that Member States shall take the measures necessary to ensure that:

- a) “every worker is entitled to a minimum daily rest period of 11 consecutive hours per 24-hour period” (Article 3)
- b) “where the working day is longer than six hours, every worker is entitled to a rest break.” (Article 4)
- c) “per each seven-day period, every worker is entitled to a minimum uninterrupted rest period of 24 hours plus the 11 hours' daily rest referred to in Article 3” (Article 5)
- d) “in keeping with the need to protect the safety and health of workers: (a) the period of weekly working time is limited by means of laws, regulations or administrative provisions or by collective agreements or agreements between the two sides of industry; (b) the average working time for each seven-day period, including overtime, does not exceed 48 hours.” (Article 6)
- e) “every worker is entitled to paid annual leave of at least four weeks in accordance with the conditions for entitlement to, and granting of, such leave laid down by national legislation and/or practice.” (Article 7)

In accordance with Article 22 of the directive (see below), member states may opt out of Article 6. They may, however, not opt out of Article 3, 4, 5 and 7.

Article 22: “A Member State shall have the option not to apply Article 6, while respecting the general principles of the protection of the safety and health of workers, and provided it takes the necessary measures to ensure that:

- a) no employer requires a worker to work more than 48 hours over a seven-day period, calculated as an average for the reference period referred to in Article 16(b), unless he has first obtained the worker's agreement to perform such work;
- b) no worker is subjected to any detriment by his employer because he is not willing to give his agreement to perform such work;
- c) the employer keeps up-to-date records of all workers who carry out such work;
- d) the records are placed at the disposal of the competent authorities, which may, for reasons connected with the safety and/or health of workers, prohibit or restrict the possibility of exceeding the maximum weekly working hours;
- e) the employer provides the competent authorities at their request with information on cases in which agreement has been given by workers to perform work exceeding 48 hours over a period of seven days, calculated as an average for the reference period referred to in Article 16(b).”

At the onset of the work presented in chapter 2, it had not been established whether or not an upper threshold at 48 hours a week is low enough to protect against excess morbidity and mortality among employees who habitually work more than 40 hours a week. A primary objective of the

present work was to settle this question. A series of cohort studies were conducted (cf. Paper I – Paper XIII) and the results of the studies enabled me to establish that the 48-hour limit generally affords more than ample protection against excess rates of overwork-related safety and health problems among employees in the general population of Denmark.

The results thereby imply that the EUWTD without application of the opt-out clause provides sufficient protection against overwork-related safety and health problems. The data sets that were used in the present work were, however, not large enough to test if the EUWTD would provide sufficient protection when Article 22 is applied. I therefore propose that a natural next step in this line of research would be to acquire a data set that is large enough to test if Article 22 may be applied without compromising the health and safety of workers.

An opt-out in accordance with Article 22, without further modifications, would make it possible for an employee to work for the same employer for up to 67 hours a week. The limit at 67 hours is implied by the resting rules of  $7 \times 11 + 24$  hours a week. From this perspective, it is of interest to test the hypothesis that rates of health and safety problems tend to be higher among employees with approximately 67 non-compulsory working hours a week than they are among employees with standard full-time working hours. If we can reject this hypothesis in sufficiently powered statistical tests that control for age, gender and socioeconomic status, then we have provided evidence in support of the notion that the opt-out clause (Article 22) may be applied without compromising the safety and health of workers. If the tested hypothesis is confirmed, then we have provided evidence in support of the notion that Article 22 may need to be modified in order to protect the safety and health of workers. For example through the addition of the following condition: “f) the average working time for each seven-day period, including overtime, may exceed 48 hours, but it may not exceed x hours” (where x is to be replaced by a number between 48 and 67).



# *English summary*

The present dissertation covers a series of studies about health and safety in relation to weekly working hours among full-time employees in the general population of Denmark [Paper I – XIII].

The studies aimed at estimating rate ratios for psychotropic drug usage, ischaemic heart disease, antihypertensive drug usage, accidental injuries, stroke, all-cause mortality and psychiatric hospital treatment due to mood, anxiety or stress-related disease, respectively, as a function of weekly working hours (32 – 40; 41 – 48; > 48 hours a week). Each of the studies was preceded by a statistical power analysis, which ascertained that the chance of detecting an effect of practical importance would be at least 80%. Each study was, moreover, preceded by a completely specified statistical analysis plan that was written and published before I or anyone else in the research team were allowed to look at any relation between the exposure and outcome data of the study.

The studies did not show any statistically significant effects of interaction between weekly working hours and age, gender, socioeconomic status or night-time work, and they did not show any statistically significant main effects of weekly working hours on the incidence of psychotropic drug usage, ischaemic heart disease, antihypertensive drug usage, accidental injuries, stroke or psychiatric hospital treatment due to mood, anxiety or stress-related disease. They showed, however, that employees with moderate overtime work (41 – 48 working hours a week) had significantly low rates of all-cause mortality, compared with employees with 32 – 40 working hours a week ( $P < 0.0001$ ), after adjustment for age, gender, calendar year, night-time work and socioeconomic status.

In conclusion, the findings of our studies do not support the notion that long weekly working hours constitute a public health problem in Denmark. Moreover, they imply that the 48-hour limit of the EU-working time directive generally affords more than ample protection against excess rates of overwork-related safety and health problems among employees in the general population of Denmark.

# *Dansk resumé*

Foreliggende afhandling dækker en række undersøgelser om sammenhæng mellem ugentlig arbejdstid og helbred blandt fuldtidsansatte i den generelle danske befolkning [Papir I - XIII].

The studies aimed at estimating rate ratios for psychotropic drug usage, ischaemic heart disease, antihypertensive drug usage, accidental injuries, stroke, all-cause mortality and psychiatric hospital treatment due to mood, anxiety or stress-related disease, respectively, as a function of weekly working hours (32 – 40; 41 – 48; > 48 hours a week). Each of the studies was preceded by a statistical power analysis, which ascertained that the chance of detecting an effect of practical importance would be at least 80%. Each study was, moreover, preceded by a completely specified statistical analysis plan that was written and published before I or anyone else in the research team were allowed to look at any relation between the exposure and outcome data of the study.

Undersøgelserne havde til formål at estimere rate ratioer for hhv. brug af psykofarmaka, psykiatrisk hospitalbehandling på grund af affektive, angst eller stressrelaterede sindslidelser, iskæmisk hjertesygdom, brug af antihypertensiv medicin, ulykke, slagtilfælde og død som en funktion af den ugentlige arbejdstid (32 - 40; 41 - 48; > 48 timer om ugen). I hver af undersøgelserne blev der foretaget en statistisk styrkeberegning, som fastslog, at chancen for at opdage en klinisk signifikant effekt ville være mindst 80 %. For hvert studie blev en fuldstændig specificeret statistisk analyseplan skrevet og offentliggjort, inden nogen i forskergruppen fik lov til at se på forhold mellem studiets eksponerings- og udfaldsdata.

Undersøgelserne viste ingen statistisk signifikante effekter af interaktion mellem ugentlig arbejdstid og hhv. alder, køn, socioøkonomisk status og nat- eller skifteholdsarbejde. De viste heller ingen statistisk signifikante hovedeffekter af ugentlig arbejdstid på incidens af hhv. brug af psykofarmaka, psykiatrisk hospitalsbehandling, iskæmisk hjertesygdom, brug af antihypertensiv medicin, ulykke og slagtilfælde. De viste imidlertid, at medarbejdere med moderat overarbejde (41-48 arbejdstimer om ugen) havde betydelig lavere risiko for død end medarbejdere med 32-40 arbejdstimer om ugen ( $P < 0,0001$ ) efter justering for alder, køn, kalenderår, natarbejde og socioøkonomisk status.

Det konkluderes, at resultaterne af vores undersøgelser ikke understøtter forestillingen om, at lang ugentlig arbejdstid udgør et folkesundhedsmæssigt problem i Danmark. Desuden konkluderes, at 48-timersgrænsen i EU-arbejdstidsdirektivet giver rigelig beskyttelse mod overarbejdsrelaterede sikkerhed og sundhedsmæssige problemer blandt ansatte i Danmark.

# References

Albertsen K, Borg V, Oldenburg B. A systematic review of the impact of work environment on smoking cessation, relapse and amount smoked. *Prev Med* 2006;43(4):291-305.

Alfredsson L, Karasek R, Theorell T. Myocardial infarction risk and psychosocial work environment: an analysis of the male Swedish working force. *Soc Sci Med*. 1982;16:463–467.

Alfredsson L, Spetz CL, Theorell T. Type of occupation and near-future hospitalization for myocardial infarction and some other diagnoses. *Int J Epidemiol*. 1985;14:378–388.

Alicandro G, Bertuccio P, Sebastiani G, La Vecchia C, Frova L. Long working hours and cardiovascular mortality: a census-based cohort study. *Int J Public Health*. 2020;65(3):257-266.

Allesøe K, Hundrup YA, Thomsen JF, Osler M. Psychosocial work environment and risk of ischaemic heart disease in women: the Danish Nurse Cohort Study. *Occup Environ Med*. 2010 May;67(5):318-22.

Alterman T, Shekelle RB, Vernon SW, Burau KD. Decision latitude, psychologic demand, job strain, and coronary heart disease in the Western Electric Study. *Am J Epidemiol*. 1994 Mar 15;139(6):620-7.

Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gøtzsche PC, Lang T; CONSORT GROUP (Consolidated Standards of Reporting Trials). The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med*. 2001 Apr 17;134(8):663-94.

Altman DG, Simera I. A history of the evolution of guidelines for reporting medical research: the long road to the EQUATOR Network. *J R Soc Med*. 2016 Feb;109(2):67-77.

Andersen I, Burr H, Kristensen TS, Gamborg M, Osler M, Prescott E, Diderichsen F. Do factors in the psychosocial work environment mediate the effect of socioeconomic position on the risk of myocardial infarction? Study from the Copenhagen Centre for Prospective Population Studies. *Occup Environ Med*. 2004 Nov;61(11):886-92.

Andersen I, Osler M, Petersen L, Grønbaek M, Prescott E. Income and risk of ischaemic heart disease in men and women in a Nordic welfare country. *Int J Epidemiol* 2003 Jun;32(3):367-374.

Andreassen CS, Griffiths MD, Sinha R, Hetland J, Pallesen S. The Relationships between Workaholism and Symptoms of Psychiatric Disorders: A Large-Scale Cross-Sectional Study. *PLoS One*. 2016 May 18;11(5):e0152978. doi: 10.1371/journal.pone.0152978.

André-Petersson L, Engström G, Hedblad B, Janzon L, Rosvall M. Social support at work and the risk of myocardial infarction and stroke in women and men. *Soc Sci Med*. 2007 Feb;64(4):830-41.

Angell M, Kassirer JP. Clinical research--what should the public believe? *N Engl J Med.* 1994 Jul 21;331(3):189-90.

Appelros P, Stegmayr B, Terént A. Sex differences in stroke epidemiology: a systematic review. *Stroke.* 2009 Apr;40(4):1082-90.

Bach E, Andersen LL, Bjørner JB, Borg V, Clausen T, Flyvholm M, Hansen ÅM, Garde AH, Holtermann A, Jørgensen MB, Kines P, Lund SP, Nielsen K, Rugulies R, Sørensen OH, Thorsen SV. *Arbejds miljø og helbred i Danmark 2010. Resumé og resultater.* Copenhagen: National Research Centre for the Working Environment; 2011.

Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA.* 1996 Aug 28;276(8):637-9.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* 57: 289-300.

Bickel PJ, Doksum KA. (1977). *Mathematical statistics — basic ideas and selected topics.* New Jersey: Prentice Hall.

Bonde JP, Munch-Hansen T, Agerbo E, Suadicani P, Wieclaw J, Westergaard-Nielsen N. Job strain and ischemic heart disease: a prospective study using a new approach for exposure assessment. *J Occup Environ Med.* 2009 Jun;51(6):732-8.

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, de Vet HCW, Kressel HY, Rifai N, Golub RM, Altman DG, Hooft L, Korevaar DA, Cohen JF, For the STARD Group. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. *BMJ.* 2015;351:h5527.

Bracken MB. Preregistration of epidemiology protocols: a commentary in support. *Epidemiology.* 2011 Mar;22(2):135-7.

Breslau N, Roth T, Rosenthal L, Andreski P. Sleep disturbance and psychiatric disorders: A longitudinal epidemiological study of young adults. *Biol Psychiatry* 1996 Mar 15;39(6):411-418.

Burr H, Bjørner JB, Kristensen TS, Tüchsen F, Bach E. Trends in the Danish work environment in 1990-2000 and their associations with labor-force changes. *Scand J Work Environ Health* 2003 Aug;29(4):270-279.

Cappuccio FP, Cooper D, D'Elia L, Strazzullo P, Miller MA. Sleep duration predicts cardiovascular outcomes: a systematic review and meta-analysis of prospective studies. *Eur Heart J* 2011 Jun;32(12):1484-1492.

Centers for Disease Control and Prevention (CDC). Ten great public health achievements--United States, 1900-1999. *MMWR Morb Mortal Wkly Rep.* 1999 Apr 2;48(12):241-3.

Centers for Disease Control and Prevention (CDC). Ten great public health achievements--United States, 2001-2010. *MMWR Morb Mortal Wkly Rep.* 2011 May 20;60(19):619-23.

Centers for Disease Control and Prevention (CDC). Ten great public health achievements--worldwide, 2001-2010. *MMWR Morb Mortal Wkly Rep.* 2011 Jun 24;60(24):814-8.

Chambers CD. Registered Reports: a new publishing initiative at Cortex. *Cortex* 49, 609–610 (2013).

Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA.* 2004 May 26;291(20):2457-65.

Chandola T, Britton A, Brunner E, Hemingway H, Malik M, Kumari M, Badrick E, Kivimaki M, Marmot M. Work stress and coronary heart disease: what are the mechanisms? *Eur Heart J.* 2008 Mar;29(5):640-8.

Chang PP, Ford DE, Mead LA, Cooper-Patrick L, Klag MJ. Insomnia in young men and subsequent depression. The Johns Hopkins Precursors Study. *Am J Epidemiol* 1997 Jul 15;146(2):105-114.

Committee on Publication Ethics. Committee on Publication Ethics: the COPE report 1999. Guidelines on good publication practice. *Occup Environ Med.* 2000 Aug;57(8):506-9.

Christensen AI, Ekholm O, Glümer C, Andreasen AH, Hvidberg MF, Kristensen PL, Larsen FB, Ortiz B, Juel K. The Danish National Health Survey 2010. Study design and respondent characteristics. *Scand J Public Health.* 2012 Jun;40(4):391-7.

Christensen AI, Ekholm O, Glümer C, Juel K: Effect of survey mode on response patterns: comparison of face-to-face and self-administered modes in health surveys. *Eur J Public Health.* 2014;24(2):327–32.

Cox AM, McKeivitt C, Rudd AG, Wolfe CD. Socioeconomic status and stroke. *Lancet Neurol.* 2006 Feb;5(2):181-8.

De Angelis C, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, Kotzin S, Laine C, Marusic A, Overbeke AJ, Schroeder TV, Sox HC, Van Der Weyden MB; International Committee of Medical Journal Editors. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *N Engl J Med.* 2004 Sep 16;351(12):1250-1.

De Angelis CD, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, Kotzin S, Laine C, Marusic A, Overbeke AJ, Schroeder TV, Sox HC, Van Der Weyden MB; International Committee of Medical Journal Editors. Is this clinical trial fully registered?--A statement from the International Committee of Medical Journal Editors. *N Engl J Med.* 2005 Jun 9;352(23):2436-8.

De Bacquer D, Pelfrene E, Clays E, Mak R, Moreau M, de Smet P, Kornitzer M, De Backer G. Perceived job stress and incidence of coronary events: 3-year follow-up of the Belgian Job Stress Project cohort. *Am J Epidemiol*. 2005 Mar 1;161(5):434-41.

Des Jarlais DC, Lyles C, Crepaz N. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. *Am J Public Health*. 2004;94(3):361-6.

Do MT, Fréchette M, McFaul S, Denning B, Ruta M, Thompson W. Injuries in the North--analysis of 20 years of surveillance data collected by the Canadian Hospitals Injury Reporting and Prevention Program. *Int J Circumpolar Health*. 2013 Sep 20;72. doi: 10.3402/ijch.v72i0.21090.

Doll R. Weak Associations in Epidemiology : Importance, Detection, and Interpretation. *J Epidemiol*,1996 ;6: S11-S20.

Dunn N, Inskip H, Kendrick T, Oestmann A, Barnett J, Godfrey K, Cooper C. Does perceived financial strain predict depression among young women? Longitudinal findings from the Southampton Women's Survey. *Ment Health Fam Med*. 2008 Mar;5(1):15-21.

Dunn OJ. Multiple comparisons among means. *J Am Stat Assoc* 1961; 56: 52–64.

Dwan K, Gamble C, Williamson PR, Kirkham JJ; Reporting Bias Group. Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. *PLoS One*. 2013 Jul 5;8(7):e66844. doi: 10.1371/journal.pone.0066844.

Eaker ED, Sullivan LM, Kelly-Hayes M, D'Agostino RB Sr, Benjamin EJ. Does job strain increase the risk for coronary heart disease or death in men and women? The Framingham Offspring Study. *Am J Epidemiol*. 2004 May 15;159(10):950-8.

Eller NH, Netterstrøm B, Gyntelberg F, Kristensen TS, Nielsen F, Steptoe A, Theorell T. Work-related psychosocial factors and the development of ischemic heart disease: a systematic review. *Cardiol Rev*. 2009 Mar-Apr;17(2):83-97.

Elovainio M, Leino-Arjas P, Vahtera J, Kivimäki M. Justice at work and cardiovascular mortality: a prospective cohort study. *J Psychosom Res*. 2006 Aug;61(2):271-4.

Eguchi H, Wada K, Smith DR. Recognition, Compensation, and Prevention of Karoshi, or Death due to Overwork. *J Occup Environ Med*. 2016 Aug;58(8):e313-4.

European Parliament, Council of the European Union. Directive 2003/88/EC of The European Parliament and of The Council of 4 November 2003 concerning certain aspects of the organisation of working time. *Official J Eur Union* 2003:9–19.

Fanelli D. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS One*. 2009 May 29;4(5):e5738. doi:10.1371/journal.pone.0005738.

Faragher EB, Cass M, Cooper CL. The relationship between job satisfaction and health: A meta-analysis. *Occup Environ Med* 2005 Feb;62(2):105-112.

Ferguson CJ, Heene M. (2012). A Vast Graveyard of Undead Theories: Publication Bias and Psychological Science's Aversion to the Null. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 7(6), 555–561.

Ferrie JE, Shipley MJ, Marmot MG, Stansfeld S, Davey Smith G. The health effects of major organisational change and job insecurity. *Soc Sci Med*. 1998 Jan;46(2):243-54.

Feveile H, Mikkelsen KL, Hannerz H, Olsen O. Quantifying inequality in health in the absence of a natural reference group. *Sci Total Environ*. 2006;367:112-22.

Feveile H, Olsen O, Burr H. Danish work environment cohort study 2005: from idea to sampling design. *Statistics in transition-new series*. 2007;8(3):441-458.

Fisher RA (1935): *The Design of Experiments* (8th edition). New York: Hafner Publishing Company;1971.

Glozier N, Martiniuk A, Patton G, Ivers R, Li Q, Hickie I, Senserrick T, Woodward M, Norton R, Stevenson M. Short sleep duration in prevalent and persistent psychological distress in young adults: the DRIVE study. *Sleep*. 2010 Sep;33(9):1139-45.

Gyntelberg F, Hein HO, Suadicani P. [The Copenhagen Male Study]. *Ugeskr Laeger*. 2004 Apr 5;166(15-16):1444-8. Review. Danish.

Ha J, Kim SG, Paek D, Park J. The Magnitude of Mortality from Ischemic Heart Disease Attributed to Occupational Factors in Korea - Attributable Fraction Estimation Using Meta-analysis. *Saf Health Work*. 2011 Mar;2(1):70-82.

Haan MN. Job strain and ischaemic heart disease: an epidemiologic study of metal workers. *Ann Clin Res*. 1988;20:143–145.

Haagsma JA, Graetz N, Bolliger I, Naghavi M, Higashi H, Mullany EC, Abera SF, Abraham JP, Adofo K, Alsharif U, Ameh EA, Ammar W, Antonio CA, Barrero LH, Bekele T, Bose D, Brazinova A, Catalá-López F, Dandona L, Dandona R, Dargan PI, De Leo D, Degenhardt L, Derrett S, Dharmaratne SD, Driscoll TR, Duan L, Petrovich Ermakov S, Farzadfar F, Feigin VL, Franklin RC, Gabbe B, Gosselin RA, Hafezi-Nejad N, Hamadeh RR, Hajar M, Hu G, Jayaraman SP, Jiang G, Khader YS, Khan EA, Krishnaswami S, Kulkarni C, Lecky FE, Leung R, Lunevicius R, Lyons RA, Majdan M, Mason-Jones AJ, Matzopoulos R, Meaney PA, Mekonnen W, Miller TR, Mock CN, Norman RE, Orozco R, Polinder S, Pourmalek F, Rahimi-Movaghar V, Refaat A, Rojas-Rueda D, Roy N, Schwebel DC, Shaheen A, Shahrzad S, Skirbekk V, Sørreide K, Soshnikov S, Stein DJ, Sykes BL, Tabb KM, Temesgen AM, Tenkorang EY, Theadom AM,

Tran BX, Vasankari TJ, Vavilala MS, Vlassov VV, Woldeyohannes SM, Yip P, Yonemoto N, Younis MZ, Yu C, Murray CJ, Vos T. The global burden of injury: incidence, mortality, disability-adjusted life years and time trends from the Global Burden of Disease study 2013. *Inj Prev*. 2016 Feb;22(1):3-18.

Hammar N, Alfredsson L, Johnson JV. Job strain, social support at work, and incidence of myocardial infarction. *Occup Environ Med*. 1998;55: 548–553.

Hannerz H. (1999). *Methodology and applications of a new law of mortality*. Lund: Department of Statistics. University of Lund, Sweden.

Hannerz H. Presentation and derivation of a five-parameter survival function intended to model mortality in modern female populations. *Scandinavian Actuarial Journal* 2001a;101:176-187.

Hannerz H. Manhood trials and the law of mortality. *Demographic Research* 2001b;4:185-202.

Hannerz H. An extension of relational methods in mortality estimation. *Demographic Research* 2001c;4:337-367.

Hannerz H, Albertsen K. Long working hours and subsequent use of psychotropic medicine: a study protocol. *JMIR Res Protoc*. 2014a Sep 19;3(3):e51. doi: 10.2196/resprot.3301.

Hannerz H, Albertsen K, Burr H, Nielsen ML, Garde AH, Larsen AD, Pejtersen JH. (2017b, February 23). The association between long working hours and stroke in the general workforce of Denmark – a study protocol. figshare. doi:10.6084/m9.figshare.4684951.v1

Hannerz H, Borgå P. Mortality among persons with a history as psychiatric inpatients with functional psychosis. *Soc Psychiatry and Psychiatr Epidemiol* 2000;35:380-387.

Hannerz H, Borgå P, Borritz M. Life expectancies for individuals with psychiatric diagnoses. *Public Health* 2001a;115:328-337.

Hannerz H, Curti S, Mattioli S, Bach E, Coggon D. (2013, October 22). Study protocol: Heavy lifting at work and risk of retinal detachment. figshare. doi:10.6084/m9.figshare.829535.v1

Hannerz H, Holtermann A. Heavy lifting at work and risk of ischemic heart disease: protocol for a register-based prospective cohort study. *JMIR Res Protoc*. 2014b Aug 20;3(3):e45. doi: 10.2196/resprot.3270.

Hannerz H, Holtermann A, Bach E. (2016b, December 22). Subsequent depression among people with musculoskeletal complaints: a study protocol. figshare. doi:10.6084/m9.figshare.4491335.v1

Hannerz H, Johansen HH, Andersen L, Thorsen SV, Flyvholm MA, Poulsen OM. (2018, July 11). Ceiling lifts, work absence, occupational injuries, and hospital contacts among care workers in Danish nursing homes: a study protocol. 2nd version (revisions in red). figshare. doi:10.6084/m9.figshare.6803291.v1



Hannerz H, Larsen AD, Garde AH. Working Time Arrangements as Potential Risk Factors for Ischemic Heart Disease Among Workers in Denmark: A Study Protocol. *JMIR Res Protoc*. 2016a Jun 22;5(2):e130. doi: 10.2196/resprot.5563.

Hannerz H, Mikkelsen KL, Nielsen ML, Tüchsen F, Spangenberg S. Social inequalities in injury occurrence and in disability retirement attributable to injuries: a 5 year follow-up study of a 2.1 million gainfully employed people. *BMC Public Health*. 2007 Aug 23;7:215.

Hannerz H, Nielsen ML. Life expectancies among survivors of acute cerebrovascular disease. *Stroke* 2001b;32:1739-1744.

Hannerz H, Pedersen BH, Poulsen OM, Humle F, Andersen LL. Study protocol to a nationwide prospective cohort study on return to gainful occupation after stroke in Denmark 1996 - 2006. *BMC Public Health*. 2010 Oct 19;10:623. doi: 10.1186/1471-2458-10-623.

Hannerz H, Tüchsen F, Pedersen BH, Dyreborg J, Rugulies R, Albertsen K. Work-relatedness of mood disorders in Denmark. *Scand J Work Environ Health*. 2009;35(4):294-300.

Hannerz H, Tüchsen F, Spangenberg S, Albertsen K. Industrial differences in disability retirement rates in Denmark 1996 - 2000. *IJOMEH*, 2004;17:465-471.

Hannerz H, Soll-Johanning H. (2017a, August 10). General mortality in relation to the EU Working Time Directive: a Danish study protocol. figshare. doi:10.6084/m9.figshare.5297062.v1

Harrison, E and Rose, D. The European Socio-economic Classification (ESeC) User Guide. Institute for Social and Economic Research, University of Essex, Colchester, UK; 2006.

Harvey SB, Wessely S, Kuh D, Hotopf M. The relationship between fatigue and psychiatric disorders: Evidence for the concept of neurasthenia. *J Psychosom Res* 2009 May;66(5):445-454.

Hayashi T, Kobayashi Y, Yamaoka K, Yano E. Effect of overtime work on 24-hour ambulatory blood pressure. *J Occup Environ Med* 1996 Oct;38(10):1007-1011.

Heikkila K, Nyberg ST, Madsen IE, de Vroome E, Alfredsson L, Bjorner JJ, Borritz M, Burr H, Erbel R, Ferrie JE, Fransson EI, Geuskens GA, Hooftman WE, Houtman IL, Jöckel KH, Knutsson A, Koskenvuo M, Lunau T, Nielsen ML, Nordin M, Oksanen T, Pejtersen JH, Pentti J, Shipley MJ, Steptoe A, Suominen SB, Theorell T, Vahtera J, Westerholm PJ, Westerlund H, Dragano N, Rugulies R, Kawachi I, Batty GD, Singh-Manoux A, Virtanen M, Kivimäki M; IPD-Work Consortium. Long working hours and cancer risk: a multi-cohort study. *Br J Cancer*. 2016 Mar 29;114(7):813-8.

Helweg-Larsen K. The Danish Register of Causes of Death. *Scand J Public Health* 2011 Jul;39(7 Suppl):26-29.

Hooft L, Korevaar DA, Molenaar N, Bossuyt PM, Scholten RJ. Endorsement of ICMJE's Clinical Trial Registration Policy: a survey among journal editors. *Neth J Med*. 2014 Sep;72(7):349-55.

Holm, S. (1979). "A simple sequentially rejective multiple test procedure". *Scandinavian Journal of Statistics*. 6 (2): 65–70.

Holtermann A, Mortensen OS, Burr H, et al. Long work hours and physical fitness: 30-year risk of ischaemic heart disease and all-cause mortality among middle-aged Caucasian men. *Heart*. 2010;96:1638–1644.

Hudson CG. Socioeconomic status and mental illness: Tests of the social causation and selection hypotheses. *Am J Orthopsychiatry* 2005 Jan;75(1):3-18.

Huibers MJ, Leone SS, van Amelsvoort LG, Kant I, Knottnerus JA. Associations of fatigue and depression among fatigued employees over time: A 4-year follow-up study. *J Psychosom Res* 2007 Aug;63(2):137-142.

Ingre, M. (2017). P-hacking in academic research: a critical review of the job strain model and of the association between night work and breast cancer in women. Stockholm: University, Faculty of Social Sciences, Department of Psychology.

Iversen L, Sabroe S, Damsgaard MT. Hospital admissions before and after shipyard closure. *BMJ*. 1989;299:1073–1076.

Iwasaki K, Sasaki T, Oka T, Hisanaga N. Effect of working hours on biological functions related to cardiovascular system among salesmen in a machinery manufacturing company. *Ind Health* 1998 Oct;36(4):361-367.

Jakovljević D, Sarti C, Sivenius J, Torppa J, Mähönen M, Immonen-Räihä P, Kaarsalo E, Alhainen K, Kuulasmaa K, Tuomilehto J, Puska P, Salomaa V. Socioeconomic status and ischemic stroke: The FINMONICA Stroke Register. *Stroke*. 2001 Jul;32(7):1492-8.

John LK, Loewenstein G, Prelec D. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol Sci*. 2012 May 1;23(5):524-32.

Johnson JV, Hall EM, Theorell T. Combined effects of job strain and social isolation on cardiovascular disease morbidity and mortality in a random sample of the Swedish male working population. *Scand J Work Environ Health*. 1989;15:271–279.

Johnson JV, Stewart W, Hall EM, Fredlund P, Theorell T. Long-term psychosocial work environment and cardiovascular mortality among Swedish men. *Am J Public Health*. 1996 Mar;86(3):324-31.

Kageyama T, Nishikido N, Kobayashi T, Kawagoe H. Estimated sleep debt and work stress in Japanese white-collar workers. *Psychiatry Clin Neurosci* 2001 Jun;55(3):217-219.

Ke DS. Overwork, stroke, and karoshi-death from overwork. *Acta Neurol Taiwan*. 2012 Jun;21(2):54-9.

Kines P, Hannerz H, Mikkelsen KL, Tüchsen F. Industrial sectors with high risk of women's hospital-treated injuries. *Am J Ind Med*. 2007 Jan;50(1):13-21.

Kildemoes HW, Sørensen H, Hallas J. The Danish National Prescription Registry. *Scand J Public Health* 2011 Jul;39(7 Suppl):38-41.

Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG (2010) Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research. *PLoS Biol* 8(6): e1000412.

Kirk MA, Rhodes RE. Occupation correlates of adults' participation in leisure-time physical activity: a systematic review. *Am J Prev Med*. 2011 Apr;40(4):476-85.

Kissela BM, Khoury JC, Alwell K, Moomaw CJ, Woo D, Adeoye O, Flaherty ML, Khatri P, Ferioli S, De Los Rios La Rosa F, Broderick JP, Kleindorfer DO. Age at stroke: temporal trends in stroke incidence in a large, biracial population. *Neurology*. 2012 Oct 23;79(17):1781-7.

Kivimäki M, Batty GD, Hamer M, Ferrie JE, Vahtera J, Virtanen M, Marmot MG, Singh-Manoux A, Shipley MJ. Using additional information on working hours to predict coronary heart disease: a cohort study. *Ann Intern Med*. 2011 Apr 5;154(7):457-63.

Kivimäki M, Ferrie JE, Brunner E, Head J, Shipley MJ, Vahtera J, Marmot MG. Justice at work and reduced risk of coronary heart disease among employees: the Whitehall II Study. *Arch Intern Med*. 2005 Oct 24;165(19):2245-51.

Kivimäki M, Jokela M, Nyberg ST, Singh-Manoux A, Fransson EI, Alfredsson L, Bjorner JB, Borritz M, Burr H, Casini A, Clays E, De Bacquer D, Dragano N, Erbel R, Geuskens GA, Hamer M, Hooftman WE, Houtman IL, Jöckel KH, Kittel F, Knutsson A, Koskenvuo M, Lunau T, Madsen IE, Nielsen ML, Nordin M, Oksanen T, Pejtersen JH, Pentti J, Rugulies R, Salo P, Shipley MJ, Siegrist J, Steptoe A, Suominen SB, Theorell T, Vahtera J, Westerholm PJ, Westerlund H, O'Reilly D, Kumari M, Batty GD, Ferrie JE, Virtanen M; IPD-Work Consortium. Long working hours and risk of coronary heart disease and stroke: a systematic review and meta-analysis of published and unpublished data for 603,838 individuals. *Lancet*. 2015a Oct 31;386(10005):1739-46.

Kivimäki M, Leino-Arjas P, Luukkonen R, Riihimäki H, Vahtera J, Kirjonen J. Work stress and risk of cardiovascular mortality: prospective cohort study of industrial employees. *BMJ*. 2002 Oct 19;325(7369):857-860.

Kivimäki M, Nyberg ST, Batty GD, Fransson EI, Heikkilä K, Alfredsson L, Bjorner JB, Borritz M, Burr H, Casini A, Clays E, De Bacquer D, Dragano N, Ferrie JE, Geuskens GA, Goldberg M, Hamer M, Hooftman WE, Houtman IL, Joensuu M, Jokela M, Kittel F, Knutsson A, Koskenvuo M, Koskinen A, Kouvonen A, Kumari M, Madsen IE, Marmot MG, Nielsen ML, Nordin M, Oksanen T, Pentti J, Rugulies R, Salo P, Siegrist J, Singh-Manoux A, Suominen SB, Väänänen A, Vahtera J, Virtanen M,

Westerholm PJ, Westerlund H, Zins M, Steptoe A, Theorell T; IPD-Work Consortium. Job strain as a risk factor for coronary heart disease: a collaborative meta-analysis of individual participant data. *Lancet*. 2012 Oct 27;380(9852):1491-7.

Kivimäki M, Nyberg ST, Batty GD, Kawachi I, Jokela M, Alfredsson L, Bjorner JB, Borritz M, Burr H, Dragano N, Fransson EI, Heikkilä K, Knutsson A, Koskenvuo M, Kumari M, Madsen IEH, Nielsen ML, Nordin M, Oksanen T, Pejtersen JH, Pentti J, Rugulies R, Salo P, Shipley MJ, Suominen S, Theorell T, Vahtera J, Westerholm P, Westerlund H, Steptoe A, Singh-Manoux A, Hamer M, Ferrie JE, Virtanen M, Tabak AG; IPD-Work consortium. Long working hours as a risk factor for atrial fibrillation: a multi-cohort study. *Eur Heart J*. 2017 Sep 7;38(34):2621-2628.

Kivimäki M, Singh-Manoux A, Virtanen M, Ferrie JE, Batty GD, Rugulies R. IPD-Work consortium: pre-defined meta-analyses of individual-participant data strengthen evidence base for a link between psychosocial factors and health. *Scand J Work Environ Health*. 2015b May 1;41(3):312-21.

Kivimäki M, Theorell T, Westerlund H, Vahtera J, Alfredsson L. Job strain and ischaemic disease: does the inclusion of older employees in the cohort dilute the association? The WOLF Stockholm Study. *J Epidemiol Community Health*. 2008 Apr;62(4):372-4.

Kivimäki M, Virtanen M, Kawachi I, Nyberg ST, Alfredsson L, Batty GD, Bjorner JB, Borritz M, Brunner EJ, Burr H, Dragano N, Ferrie JE, Fransson EI, Hamer M, Heikkilä K, Knutsson A, Koskenvuo M, Madsen IEH, Nielsen ML, Nordin M, Oksanen T, Pejtersen JH, Pentti J, Rugulies R, Salo P, Siegrist J, Steptoe A, Suominen S, Theorell T, Vahtera J, Westerholm PJM, Westerlund H, Singh-Manoux A, Jokela M. Long working hours, socioeconomic status, and the risk of incident type 2 diabetes: a meta-analysis of published and unpublished data from 222 120 individuals. *Lancet Diabetes Endocrinol*. 2015c Jan;3(1):27-34.

Kolmogorov AN (1950). *Foundations of the theory of probability*. New York: Chelsea Publishing Company.

Koo DL, Nam H, Thomas RJ, Yun CH. Sleep Disturbances as a Risk Factor for Stroke. *J Stroke*. 2018 Jan;20(1):12-32. doi: 10.5853/jos.2017.02887.

Kornitzer M, deSmet P, Sans S, Dramaix M, Boulenguez C, DeBacker G, Ferrario M, Houtman I, Isacson SO, Ostergren PO, Peres I, Pelfrene E, Romon M, Rosengren A, Cesana G, Wilhelmsen L. Job stress and major coronary events: results from the Job Stress, Absenteeism and Coronary Heart Disease in Europe study. *Eur J Cardiovasc Prev Rehabil*. 2006 Oct;13(5):695-704.

Korshøj M, Hannerz H, Marott JL, Schnohr P, Prescott EIB, Clays E, Holtermann A. The Effect of Occupational Lifting on Hypertension Risk: Protocol for a Project Using Data From the Copenhagen City Heart Study. *JMIR Res Protoc*. 2018 Apr 27;7(4):e93. doi: 10.2196/resprot.9692

Krleza-Jerić K, Chan AW, Dickersin K, Sim I, Grimshaw J, Gluud C. Principles for international registration of protocol information and results from human trials of health related interventions:

Ottawa statement (part 1). *BMJ*. 2005 Apr 23;330(7497):956-8. Review. Erratum in: *BMJ*. 2005 May 28;330(7502):1258.

Kuper H, Adami HO, Theorell T, Weiderpass E. Psychosocial determinants of coronary heart disease in middle-aged women: a prospective study in Sweden. *Am J Epidemiol*. 2006 Aug 15;164(4):349-57.

Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977 Mar;33(1):159-174.

Larsen AD, Hannerz H, Møller SV, Dyreborg J, Bonde JP, Hansen J, Garde AH. (2016, June 1). Study protocol for examining long working hours and night work as risk factors for injuries. figshare. doi:10.6084/m9.figshare.3408220.v1

Larsen AD, Hannerz H, Obel C, Thulstrup AM, Bonde JP, Hougaard KS. Testing the association between psychosocial job strain and adverse birth outcomes—design and methods. *BMC Public Health*. 2011 Apr 21;11:255. doi: 10.1186/1471-2458-11-255.

Lash TL . Preregistration of study proposals is unlikely to improve the yield from our sciences, but other strategies might. *Epidemiology*. 2010;21:612–613.

Lee S, Colditz G, Berkman L, et al. A prospective study of job strain and coronary heart disease in US women. *Int J Epidemiol*. 2002;31:1147–53.

Lee S, Colditz GA, Berkman LF, et al. Prospective study of job insecurity and coronary heart disease in US women. *Ann Epidemiol*. 2004;14:24–30.

Le Fanu J. *The Rise and Fall of Modern Medicine*. New York: Carrol and Graf; 1999.

Lin RT, Lin CK, Christiani DC, Kawachi I, Cheng Y, Verguet S, Jong S. The impact of the introduction of new recognition criteria for overwork-related cardiovascular and cerebrovascular diseases: a cross-country comparison. *Sci Rep*. 2017 Mar 13;7(1):167. doi: 10.1038/s41598-017-00198-5. Erratum in: *Sci Rep*. 2018 Mar 12;8(1):4654.

Loder E, Groves T, Macauley D. Registration of observational studies. *BMJ*. 2010 Feb 18;340:c950. doi: 10.1136/bmj.c950.

Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, Abraham J, Adair T, Aggarwal R, Ahn SY, Alvarado M, Anderson HR, Anderson LM, Andrews KG, Atkinson C, Baddour LM, Barker-Collo S, Bartels DH, Bell ML, Benjamin EJ, Bennett D, Bhalla K, Bikbov B, Bin Abdulhak A, Birbeck G, Blyth F, Bolliger I, Boufous S, Bucello C, Burch M, Burney P, Carapetis J, Chen H, Chou D, Chugh SS, Coffeng LE, Colan SD, Colquhoun S, Colson KE, Condon J, Connor MD, Cooper LT, Corriere M, Cortinovis M, de Vaccaro KC, Couser W, Cowie BC, Criqui MH, Cross M, Dabhadkar KC, Dahodwala N, De Leo D, Degenhardt L, Delossantos A, Denenberg J, Des Jarlais DC, Dharmaratne SD, Dorsey ER, Driscoll T, Duber H, Ebel B, Erwin PJ, Espindola P, Ezzati M, Feigin V, Flaxman AD, Forouzanfar MH, Fowkes FG, Franklin R, Fransen M, Freeman MK, Gabriel SE, Gakidou E, Gaspari F, Gillum RF, Gonzalez-Medina D,

Halasa YA, Haring D, Harrison JE, Havmoeller R, Hay RJ, Hoen B, Hotez PJ, Hoy D, Jacobsen KH, James SL, Jasrasaria R, Jayaraman S, Johns N, Karthikeyan G, Kassebaum N, Keren A, Khoo JP, Knowlton LM, Kobusingye O, Koranteng A, Krishnamurthi R, Lipnick M, Lipshultz SE, Ohno SL, Mabweijano J, MacIntyre MF, Mallinger L, March L, Marks GB, Marks R, Matsumori A, Matzopoulos R, Mayosi BM, McAnulty JH, McDermott MM, McGrath J, Mensah GA, Merriman TR, Michaud C, Miller M, Miller TR, Mock C, Mocumbi AO, Mokdad AA, Moran A, Mulholland K, Nair MN, Naldi L, Narayan KM, Nasser K, Norman P, O'Donnell M, Omer SB, Ortblad K, Osborne R, Ozgediz D, Pahari B, Pandian JD, Rivero AP, Padilla RP, Perez-Ruiz F, Perico N, Phillips D, Pierce K, Pope CA 3rd, Porrini E, Pourmalek F, Raju M, Ranganathan D, Rehm JT, Rein DB, Remuzzi G, Rivara FP, Roberts T, De León FR, Rosenfeld LC, Rushton L, Sacco RL, Salomon JA, Sampson U, Sanman E, Schwebel DC, Segui-Gomez M, Shepard DS, Singh D, Singleton J, Sliwa K, Smith E, Steer A, Taylor JA, Thomas B, Tleyjeh IM, Towbin JA, Truelsen T, Undurraga EA, Venketasubramanian N, Vijayakumar L, Vos T, Wagner GR, Wang M, Wang W, Watt K, Weinstock MA, Weintraub R, Wilkinson JD, Woolf AD, Wulf S, Yeh PH, Yip P, Zabetian A, Zheng ZJ, Lopez AD, Murray CJ, AlMazroa MA, Memish ZA. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*. 2012 Dec 15;380(9859):2095-128.

Lynch J, Krause N, Kaplan GA, Tuomilehto J, Salonen JT. Workplace conditions, socioeconomic status, and the risk of mortality and acute myocardial infarction: the Kuopio Ischemic Heart Disease Risk Factor Study. *Am J Public Health*. 1997 Apr;87(4):617-22.

Lynge E, Sandegaard JL, Rebolj M. The Danish National Patient Register. *Scand J Public Health* 2011 Jul;39(7 Suppl):30-33.

Madsen IE, Hannerz H, Nyberg ST, Magnusson Hanson LL, Ahola K, Alfredsson L, Batty GD, Bjorner JB, Borritz M, Burr H, Dragano N, Ferrie JE, Hamer M, Jokela M, Knutsson A, Koskenvuo M, Koskinen A, Leineweber C, Nielsen ML, Nordin M, Oksanen T, Pejtersen JH, Pentti J, Salo P, Singh-Manoux A, Suominen S, Theorell T, Toppinen-Tanner S, Vahtera J, Väänänen A, Westerholm PJ, Westerlund H, Fransson E, Heikkilä K, Virtanen M, Rugulies R, Kivimäki M; IPD-Work Consortium. Study protocol for examining job strain as a risk factor for severe unipolar depression in an individual participant meta-analysis of 14 European cohorts. Version 2. *F1000Res*. 2013 Nov 5 [revised 2014 Jan 1];2:233. doi: 10.12688/f1000research.2-233.v2. eCollection 2013.

Mathieu S, Chan AW, Ravaud P. Use of trial register information during the peer review process. *PLoS One*. 2013;8:e59910.

Matthews KA, Gump BB. Chronic work stress and marital dissolution increase risk of posttrial mortality in men from the Multiple Risk Factor Intervention Trial. *Arch Intern Med*. 2002;162:309-315.

Mayes LC, Horwitz RI, Feinstein AR. A collection of 56 topics with contradictory results in case-control research. *Int J Epidemiol* 1989; 3:725-27.

Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010; 340:c869.

Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement: *PLoS Med* 2009;6(7):e1000097.

Monson R. *Occupational Epidemiology*, 2nd edition. Boca Raton, Florida: CRC Press Inc., 1990.

Mors O, Perto GP, Mortensen PB. The Danish Psychiatric Central Research Register. *Scand J Public Health* 2011 Jul;39(7 Suppl):54-57.

Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, American Heart Association Statistics Committee Stroke Statistics Subcommittee. Heart disease and stroke statistics--2015 update: a report from the American Heart Association. *Circulation* 2015 Jan 27;131(4):e29-322.

Møller S. [A Follow-up Study of Sedentary Work and Ischemic Heart Disease]. Copenhagen: University of Copenhagen; 2013. URL: [https://figshare.com/articles/New\\_draft\\_item/3187753](https://figshare.com/articles/New_draft_item/3187753).

Møller SV, Hannerz H, Holtermann A, van der Ploeg HP. (2014, April 3). Study protocol: The associations between sitting at work and all-cause mortality. figshare. doi:10.6084/m9.figshare.980714.v1

Nabe-Nielsen K, Garde AH, Ishtiaq-Ahmed K, Gyntelberg F, Mortensen EL, Phung TKT, Rod NH, Waldemar G, Westendorp RG, Hansen ÅM. Shift work, long working hours, and later risk of dementia: A long-term follow-up of the Copenhagen Male Study. *Scand J Work Environ Health*. 2017 Nov 1;43(6):569-577.

Netterstrøm B, Juel K. Impact of work-related and psychosocial factors on the development of ischemic heart disease among urban bus drivers in Denmark. *Scand J Work Environ Health*. 1988;14:231-238.

Netterstrøm B, Kristensen TS, Jensen G, Schnor P. Is the demand-control model still a useful tool to assess work-related psychosocial risk for ischemic heart disease? Results from 14 year follow up in the Copenhagen City Heart study. *Int J Occup Med Environ Health*. 2010;23(3):217-24.

Netterstrøm B, Kristensen TS, Sjørl A. Psychological job demands increase the risk of ischaemic heart disease: a 14-year cohort study of employed Danish men. *Eur J Cardiovasc Prev Rehabil*. 2006;13:414-420.

Neyman J, Pearson ES. (1933). "On the Problem of the Most Efficient Tests of Statistical Hypotheses". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 231 (694-706): 289-337.

Nübling M, Stößel U, Hasselhorn HM, Michaelis M, Hofmann F. Measuring psychological stress and strain at work - Evaluation of the COPSQ questionnaire in Germany. *Psychosoc Med* 2006;3:Doc05.

Nyberg A, Alfredsson L, Theorell T, Westerlund H, Vahtera J, Kivimäki M. Managerial leadership and ischaemic heart disease among employees: the Swedish WOLF study. *Occup Environ Med.* 2009 Jan;66(1):51-5.

O'Reilly D, Rosato M. Worked to death? A census-based longitudinal study of the relationship between the numbers of hours spent working and mortality risk. *Int J Epidemiol* 2013;42(6):1820–30.

O'Reilly D, Rosato M, Catney G, Johnston F, Brolly M. Cohort description: the Northern Ireland Longitudinal Study (NILS). *Int J Epidemiol.* 2012 Jun;41(3):634-41.

Pedersen BH, Dyreborg J, Kines P, Mikkelsen KL, Hannerz H, Andersen DR, Spangenberg S. Protocol for a mixed-methods study on leader-based interventions in construction contractors' safety commitments. *Inj Prev.* 2010 Jun;16(3):e2. doi: 10.1136/ip.2009.025403.

Pedersen BH, Hannerz H, Tüchsen F, Mikkelsen KL, Dyreborg J. Industry and injury related hospital contacts: a follow-up study of injuries among working men in Denmark. *J Occup Health.* 2010;52(3):147-54.

Pedersen BH, Hannerz H, Christensen U, Tüchsen F. Enterprise size and risk of hospital treated injuries among manual construction workers in Denmark: a study protocol. *J Occup Med Toxicol.* 2011 Apr 21;6:11. doi: 10.1186/1745-6673-6-11.

Pedersen CB. The Danish Civil Registration System. *Scand J Public Health* 2011 Jul;39(7 Suppl):22-25.

Pejtersen JH, Kristensen TS, Borg V, Bjorner JB. The second version of the Copenhagen psychosocial questionnaire. *Scand J Public Health* 2010 Feb;38(3 Suppl):8-24

Petersson F, Baadsgaard M, Thygesen LC. Danish registers on personal labour market affiliation. *Scand J Public Health* 2011 Jul;39(7 Suppl):95-98.

Porta M, Greenland S, Hernan M, dos Santos Silva I, Last M, editors. *A dictionary of epidemiology.* 6th ed. New York: Oxford University Press, 2014.

Pratt LA, Brody DJ, Gu Q. NCHS Data Brief No. 76. Antidepressant use in persons aged 12 and over: United States, 2005-2008. Hyattsville (MD): National Center for Health Statistics (US); 2005. URL: <http://www.cdc.gov/nchs/data/databriefs/db76.htm> [accessed 2018-10-18]

Proctor SP, White RF, Robins TG, Echeverria D, Rocskay AZ. Effect of overtime work on cognitive function in automotive workers. *Scand J Work Environ Health* 1996 Apr;22(2):124-132.



Reed DM, LaCroix AZ, Karasek RA, Miller D, MacLean CA. Occupational strain and the incidence of coronary heart disease. *Am J Epidemiol.* 1989 Mar;129(3):495-502.

Rosengren A, Hawken S, Ounpuu S, Sliwa K, Zubaid M, Almahmeed WA, et al. Association of psychosocial risk factors with risk of acute myocardial infarction in 11119 cases and 13648 controls from 52 countries (the INTERHEART study): case-control study. *Lancet* 2004;364(9438):953-962.

Sanchis-Gomar F, Perez-Quilis C, Leischik R, Lucia A. Epidemiology of coronary heart disease and acute coronary syndrome. *Ann Transl Med.* 2016 Jul;4(13):256. doi: 10.21037/atm.2016.06.33.

Sareen J, Afifi TO, McMillan KA, Asmundson GJ. Relationship between household income and mental disorders: Findings from a population-based longitudinal study. *Arch Gen Psychiatry* 2011 Apr;68(4):419-427.

Sasaki T, Iwasaki K, Oka T, Hisanaga N, Ueda T, Takada Y, Fujiki Y. Effect of working hours on cardiovascular-autonomic nervous functions in engineers in an electronics manufacturing company. *Ind Health.* 1999 Jan;37(1):55-61.

Scargle JD. Publication bias: the “file-drawer” problem in scientific inference. *J Sci Explor.* 2000; 14:91–106.

Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ.* 2010 Mar 23;340:c332. doi: 10.1136/bmj.c332.

Siegrist J, Peter R, Motz W, Strauer BE. The role of hypertension, left-ventricular hypertrophy and psychosocial risks in cardiovascular-disease—prospective evidence from blue-collar men. *Eur Heart J.* 1992; 13:89–95.

Simera I, Moher D, Hirst A, Hoey J, Schulz KF, Altman DG. Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. *BMC Med.* 2010 Apr 26;8:24.

Sinclair RR, Cheung JH. Money Matters: Recommendations for Financial Stress Research in Occupational Health Psychology. *Stress Health.* 2016 Aug;32(3):181-93. doi: 10.1002/smi.2688.

Skapinakis P, Lewis G, Mavreas V. Temporal relations between unexplained fatigue and depression: Longitudinal data from an international study in primary care. *Psychosom Med* 2004;66(3):330-335.

Slopen N, Glynn RJ, Buring JE, Lewis TT, Williams DR, Albert MA. Job strain, job insecurity, and incident cardiovascular disease in the Women's Health Study: results from a 10-year prospective study. *PLoS One.* 2012;7(7):e40512.

Smith GD, Ebrahim S. Data dredging, bias, or confounding. *BMJ.* 2002 Dec 21;325(7378):1437-8.

Smyth RM, Kirkham JJ, Jacoby A, Altman DG, Gamble C, Williamson PR. Frequency and reasons for outcome reporting bias in clinical trials: interviews with trialists. *BMJ*. 2011 Jan 6;342:c7153. doi: 10.1136/bmj.c7153.

Sokejima S, Kagamimori S. Working hours as a risk factor for acute myocardial infarction in Japan: case-control study. *BMJ*. 1998;317:775-780.

Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, Hing C, Kwok CS, Pang C, Harvey I. Dissemination and publication of research findings: an updated review of related biases. *Health Technol Assess*. 2010 Feb;14(8):iii, ix-xi, 1-193. doi: 10.3310/hta14080.

Spurgeon A, Harrington JM, Cooper CL. Health and safety problems associated with long working hours: a review of the current position. *Occup Environ Med*. 1997 Jun;54(6):367-75.

Statistics Denmark. SOCIO Danmarks Statistiks Socioøkonomiske Klassifikation. Copenhagen: Statistics Denmark; 1997.

Statistics Denmark. Arbejdskraftundersøgelsen. 2019. URL: <http://www.dst.dk/da/statistik/dokumentation/statistikdokumentation/arbejdskraftundersogelsen> [accessed 2019-02-23]

Steenland K, Johnson J, Nowlin S. A follow-up study of job strain and heart disease among males in the NHANES1 population. *Am J Ind Med*. 1997;31:256-260.

Steinhausen HC, Bisgaard C (2014) Nationwide time trends in dispensed prescriptions of psychotropic medication for children and adolescents in Denmark. *Acta Psychiatr Scand* 129:221-231

Stephoe A, Brydon L, Kunz-Ebrecht S. Changes in financial strain over three years, ambulatory blood pressure, and cortisol responses to awakening. *Psychosom Med* 2005;67(2):281-287.

Stevens A, Shamseer L, Weinstein E, Yazdi F, Turner L, Thielman J, Altman DG, Hirst A, Hoey J, Palepu A, Schulz KF, Moher D. Relation of completeness of reporting of health research to journals' endorsement of reporting guidelines: systematic review. *BMJ*. 2014 Jun 25;348:g3804.

Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting: meta-analysis of observational studies in epidemiology (MOOSE) group. *JAMA*. 2000;283(15):2008-12.

Sturm R. The effects of obesity, smoking, and drinking on medical problems and costs. *Health Aff (Millwood)*. 2002 Mar-Apr;21(2):245-53.

Suadicani P, Hein HO, Gyntelberg F. Are social inequalities as associated with the risk of ischaemic heart disease a result of psychosocial working conditions? *Atherosclerosis*. 1993;101:165-175.

Swaen GG, Teggeler O, van Amelsvoort LG. False positive outcomes and design characteristics in occupational cancer epidemiology studies. *Int J Epidemiol*. 2001 Oct;30(5):948-54.

Swift R. The relationship between health and GDP in OECD countries in the very long run. *Health Econ*. 2011 Mar;20(3):306-22.

Szklo-Coxe M, Young T, Peppard PE, Finn LA, Benca RM. Prospective associations of insomnia markers and symptoms with depression. *Am J Epidemiol* 2010 Mar 15;171(6):709-720.

Taubes G. Epidemiology faces its limits. *Science*. 1995;269:164–169.

The editors. The registration of observational studies—when metaphors go bad. *Epidemiology*. 2010;21:607–609.

The Lancet. Should protocols for observational research be registered? *Lancet*. 2010 Jan 30;375(9712):348.

The Northern Ireland Statistics and Research Agency. Northern Ireland Census 2001 Population Report and Mid-Year Estimates. A National Statistics Publication. Norwich: Her Majesty's Stationery Office; 2002. Available at <https://www.nisra.gov.uk/publications/2001-census-population-report-and-mid-year-estimates>

The Office for National Statistics. Census 2001. Definitions. A National Statistics publication. London: Her Majesty's Stationery Office; 2004. Available at [https://census.ukdataservice.ac.uk/media/51185/2001\\_defs\\_intro.pdf](https://census.ukdataservice.ac.uk/media/51185/2001_defs_intro.pdf)

Theorell T, Floderus-Myrhed B. 'Workload' and risk of myocardial infarction—a prospective psychosocial analysis. *Int J Epidemiol*. 1977;6:17–21.

Theorell T, Tsutsumi A, Hallquist J, Reuterwall C, Hogstedt C, Fredlund P, Emlund N, Johnson JV. Decision latitude, job strain, and myocardial infarction: a study of working men in Stockholm. The SHEEP Study Group. Stockholm Heart epidemiology Program. *Am J Public Health*. 1998 Mar;88(3):382-8.

Thorsen SV, Rugulies R, Hjarsbech PU, Bjorner JB. The predictive value of mental health for long-term sickness absence: The major depression inventory (MDI) and the mental health inventory (MHI-5) compared. *BMC Med Res Methodol* 2013;13:115 [FREE Full text] [doi: 10.1186/1471-2288-13-115]

Tjepkema M. Insomnia. *Health Rep* 2005 Nov;17(1):9-25.

Toivanen S. Income differences in stroke mortality: a 12-year follow-up study of the Swedish working population. *Scand J Public Health* 2011;39:797–804.

Tüchsen F, Endahl LA. Increasing inequality in ischaemic heart disease morbidity among employed men in Denmark 1981-1993: the need for a new preventive policy. *Int J Epidemiol* 1999 Aug;28(4):640-644.

Turner L, Shamseer L, Altman DG, Schulz KF, Moher D. Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. *Syst Rev*. 2012 Nov 29;1:60.

Uehli K, Mehta AJ, Miedinger D, Hug K, Schindler C, Holsboer-Trachsler E, Leuppi JD, Künzli N. Sleep problems and work injuries: a systematic review and meta-analysis. *Sleep Med Rev*. 2014;18(1):61-73.

Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M; STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *PLoS Med*. 2007 Oct 16;4(10):e297.

Vandenbroucke JP. Pre-registration of epidemiologic studies: an illfounded mix of ideas. *Epidemiology*. 2010;21:619-620.

Vahtera J, Kivimäki M, Pentti J, Linna A, Virtanen M, Virtanen P, Ferrie JE. Organisational downsizing, sickness absence, and mortality: 10-town prospective cohort study. *BMJ*. 2004 Mar 6;328(7439):555.

Virtanen M, Ferrie JE, Singh-Manoux A, Shipley MJ, Vahtera J, Marmot MG, Kivimäki M. Overtime work and incident coronary heart disease: the Whitehall II prospective cohort study. *Eur Heart J*. 2010 Jul;31(14):1737-44.

Virtanen M, Jokela M, Madsen IE, Magnusson Hanson LL, Lallukka T, Nyberg ST, Alfredsson L, Batty GD, Bjorner JB, Borritz M, Burr H, Dragano N, Erbel R, Ferrie JE, Heikkilä K, Knutsson A, Koskenvuo M, Lahelma E, Nielsen ML, Oksanen T, Pejtersen JH, Pentti J, Rahkonen O, Rugulies R, Salo P, Schupp J, Shipley MJ, Siegrist J, Singh-Manoux A, Suominen SB, Theorell T, Vahtera J, Wagner GG, Wang JL, Yiengprugsawan V, Westerlund H, Kivimäki M. Long working hours and depressive symptoms: systematic review and meta-analysis of published studies and unpublished individual participant data. *Scand J Work Environ Health*. 2018 May 1;44(3):239-250.

Virtanen M, Jokela M, Nyberg ST, Madsen IE, Lallukka T, Ahola K, Alfredsson L, Batty GD, Bjorner JB, Borritz M, Burr H, Casini A, Clays E, De Bacquer D, Dragano N, Erbel R, Ferrie JE, Fransson EI, Hamer M, Heikkilä K, Jöckel KH, Kittel F, Knutsson A, Koskenvuo M, Ladwig KH, Lunau T, Nielsen ML, Nordin M, Oksanen T, Pejtersen JH, Pentti J, Rugulies R, Salo P, Schupp J, Siegrist J, Singh-Manoux A, Steptoe A, Suominen SB, Theorell T, Vahtera J, Wagner GG, Westerholm PJ, Westerlund H, Kivimäki M. Long working hours and alcohol use: systematic review and meta-analysis of published studies and unpublished individual participant data. *BMJ*. 2015 Jan 13;350:g7772. doi: 10.1136/bmj.g7772.

von Elm E, Röllin A, Blümle A, Huwiler K, Witschi M, Egger M. Publication and non-publication of clinical trials: longitudinal study of applications submitted to a research ethics committee. *Swiss Med Wkly*. 2008 Apr 5;138(13-14):197-203.

Väänänen A, Koskinen A, Joensuu M, Kivimäki M, Vahtera J, Kouvonen A, Jäppinen P. Lack of predictability at work and risk of acute myocardial infarction: an 18-year prospective study of industrial employees. *Am J Public Health*. 2008 Dec;98(12):2264-71.

Wagstaff AS, Sigstad Lie JA. Shift and night work and long working hours--a systematic review of safety implications. *Scand J Work Environ Health*. 2011;37(3):173-85.

Wang Y, Mei H, Jiang YR, Sun WQ, Song YJ, Liu SJ, Jiang F. Relationship between Duration of Sleep and Hypertension in Adults: A Meta-Analysis. *J Clin Sleep Med*. 2015 Sep 15;11(9):1047-56.

Warburton DE, Nicol CW, Bredin SS. Health benefits of physical activity: the evidence. *CMAJ*. 2006 Mar 14;174(6):801-9.

Weich S, Lewis G. Poverty, unemployment, and common mental disorders: Population based cohort study. *BMJ* 1998 Jul 11;317(7151):115-119.

West R. Tobacco smoking: Health impact, prevalence, correlates and interventions. *Psychol Health*. 2017 Aug;32(8):1018-1036.

WHO . Global status report on alcohol and health 2018. Geneva: World Health Organization; 2018. Available at:

[http://www.who.int/substance\\_abuse/publications/global\\_alcohol\\_report/gsr\\_2018/en/](http://www.who.int/substance_abuse/publications/global_alcohol_report/gsr_2018/en/)

Williamson PR, Gamble C, Altman DG, Hutton JL. Outcome selection bias in meta-analysis. *Stat Methods Med Res*. 2005 Oct;14(5):515-24.

Wittchen HU, Hoyer J. Generalized anxiety disorder: Nature and course. *J Clin Psychiatry* 2001;62 Suppl 11:15-19.

Wynder EL. Invited commentary: response to Science article, "Epidemiology faces its limits". *Am J Epidemiol*. 1996 Apr 15;143(8):747-9.

Yarkoni, T. (2009). Big Correlations in Little Studies: Inflated fMRI Correlations Reflect Low Statistical Power-Commentary on Vul et al. (2009). *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 4(3), 294-298.

Young SS, Alan K. "Deming, data and observational studies. A process out of control and needing fixing." *Significance* 8 (2011): 116-120.





ISBN 978-87-94336-90-1



Printed by:  
Campus Print - University of Copenhagen