The effect of exposure to long working hours on depression: A systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury

Reiner Rugulies, Kathrine Sørensen, Cristina Di Tecco, Michela Bonafede, Bruna Rondinone, Seoyeon Ahn, Emiko Ando, Jose Luis Ayuso-Mateos, Maria Cabello, Alexis Descatha, et al.

▶ **To cite this version:**

Reiner Rugulies, Kathrine Sørensen, Cristina Di Tecco, Michela Bonafede, Bruna Rondinone, et al.. The effect of exposure to long working hours on depression: A systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environment International, 2021, 155, pp.106629. 10.1016/j.envint.2021.106629 . hal-03282577

HAL Id: hal-03282577
https://hal.science/hal-03282577

Submitted on 25 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Environment International 155 (2021) 106629

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Systematic review

The effect of exposure to long working hours on depression: A systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury

Reiner Rugulies [a,b,c,*], Kathrine Sørensen [a], Cristina Di Tecco [d], Michela Bonafede [d], Bruna M. Rondinone [d], Seoyeon Ahn [e], Emiko Ando [f], Jose Luis Ayuso-Mateos [g,h,i], Maria Cabello [g,h], Alexis Descatha [j,k,l], Nico Dragano [m], Quentin Durand-Moreau [n], Hisashi Eguchi [o,p], Junling Gao [q], Lode Godderis [r,s], Jaeyoung Kim [t], Jian Li [u], Ida E.H. Madsen [a], Daniela V. Pachito [v], Grace Sembajwe [w,x], Johannes Siegrist [y], Kanami Tsuno [z], Yuka Ujita [aa], JianLi Wang [ab], Amy Zadow [ac], Sergio Iavicoli [d], Frank Pega [ad]

[a] National Research Centre for the Working Environment, Copenhagen, Denmark
[b] Department of Public Health, University of Copenhagen, Copenhagen, Denmark
[c] Department of Psychology, University of Copenhagen, Copenhagen, Denmark
[d] Inail, Department of Occupational and Environmental Medicine, Epidemiology and Hygiene, Monte Porzio Catone (Rome), Italy
[e] National Pension Research Institute, Jeonju-si, Republic of Korea
[f] National Cancer Center, Tokyo, Japan
[g] Department of Psychiatry, Universidad Autonoma de Madrid, Madrid, Spain
[h] Instituto de Salud Carlos III, Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Spain
[i] Instituto de Investigación Sanitaria Princesa (IIS-Princesa), Madrid, Spain
[j] Univ Angers, CHU Angers, Univ Rennes, Inserm, EHESP, Irset (Institut de recherche en santé, environnement et travail) - UMR_S 1085, F-49000 Angers, France
[k] AP-HP (Paris Hospital), Occupational Health Unit, Poincaré University Hospital, Garches, France
[l] Inserm Versailles St-Quentin Univ – Paris Saclay Univ (UVSQ), UMS 011, UMR-S 1168, Villejuif, France
[m] Institute of Medical Sociology, Medical Faculty, University of Düsseldorf, Düsseldorf, Germany
[n] Division of Preventive Medicine, Department of Medicine, University of Alberta, Edmonton, Canada
[o] Department of Mental Health, Institute of Industrial Ecological Sciences, University of Occupational and Environmental Health, Kitakyushu, Japan
[p] Department of Public Health, Kitasato University School of Medicine, Sagamihara, Knagawa, Japan
[q] School of Public Health, Fudan University, Shanghai, People's Republic of China
[r] Centre for Environment and Health, KU Leuven, Leuven, Belgium
[s] KIR Department (Knowledge, Information & Research), IDEWE, External Service for Prevention and Protection at Work, Leuven, Belgium
[t] Department of Preventive Medicine, College of Medicine, Keimyung University, Daegu, Republic of Korea
[u] Department of Environmental Health Sciences, Fielding School of Public Health, School of Nursing, University of California, Los Angeles, United States
[v] Hospital Sírio-Libanês, Sao Paulo, Brazil
[w] Department of Occupational Medicine Epidemiology and Prevention, Zucker School of Medicine at Hofstra University, Feinstein Institutes for Medical Research, Northwell Health, New York, United States
[x] Department of Environmental Occupational and Geospatial Sciences, CUNY Institute for Implementation Science in Public Health, CUNY Graduate School of Public Health and Health Policy, New York, United States
[y] Life Science Centre, University of Düsseldorf, Düsseldorf, Germany
[z] School of Health Innovation, Kanagawa University of Human Services, Japan
[aa] Labour Administration, Labour Inspection and Occupational Safety and Health Branch, International Labour Organization, Geneva, Switzerland
[ab] Institute of Mental Health Research, University of Ottawa, Canada
[ac] University of South Australia, Adelaide, Australia
[ad] Department of Environment, Climate Change and Health, World Health Organization, Geneva, Switzerland

* Corresponding author at: National Research Centre for the Working Environment, Lersø Parkallé 105, DK-2100 Copenhagen, Denmark.
E-mail addresses: rer@nfa.dk (R. Rugulies), ksn@nfa.dk (K. Sørensen), c.ditecco@inail.it (C. Di Tecco), m.bonafede@inail.it (M. Bonafede), b.rondinone@inail.it (B.M. Rondinone), ahnseoyeon@nps.or.kr (S. Ahn), andoemiko-tky@umin.ac.jp (E. Ando), joseluis.ayuso@uam.es (J.L. Ayuso-Mateos), maria.cabello@uam.es (M. Cabello), alexis.descatha@inserm.fr (A. Descatha), Dragano@med.uni-duesseldorf.de (N. Dragano), durandmo@ualberta.ca (Q. Durand-Moreau), eguchi@med.uoeh-u.ac.jp (H. Eguchi), jlgao@fudan.edu.cn (J. Gao), lode.godderis@kuleuven.be (L. Godderis), jaeykim@dsmc.or.kr (J. Kim), jianli2019@ucla.edu (J. Li), ihm@nfa.dk (I.E.H. Madsen), pachito@uol.com.br (D.V. Pachito), Grace.Sembajwe@sph.cuny.edu (G. Sembajwe), johannes.siegrist@med.uni-duesseldorf.de (J. Siegrist), ktsuno-tky@umin.ac.jp (K. Tsuno), ujita@ilo.org (Y. Ujita), JianLi.Wang@theroyal.ca (J. Wang), Amy.zadow@unisa.edu.au (A. Zadow), s.iavicoli@inail.it (S. Iavicoli), pega@who.int (F. Pega).

https://doi.org/10.1016/j.envint.2021.106629

Available online 15 June 2021
0160-4120/© 2021 World Health Organization and International Labour Organization; licensee Elsevier. This is an open access article under the CC BY IGO license (http://creativecommons.org/licenses/by/3.0/igo/).

ARTICLE INFO

Handling Editor: Paul Whaley

Keywords:
Global burden of disease
Occupational health
Long working hours
Depression
Systematic review
Meta-analysis

ABSTRACT

Background: The World Health Organization (WHO) and the International Labour Organization (ILO) are developing the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury (WHO/ILO Joint Estimates), supported by a large number of individual experts. Evidence from previous reviews suggests that exposure to long working hours may cause depression. In this article, we present a systematic review and meta-analysis of parameters for estimating (if feasible) the number of deaths and disability-adjusted life years from depression that are attributable to exposure to long working hours, for the development of the WHO/ILO Joint Estimates.

Objectives: We aimed to systematically review and meta-analyse estimates of the effect of exposure to long working hours (three categories: 41–48, 49–54 and ≥55 h/week), compared with exposure to standard working hours (35–40 h/week), on depression (three outcomes: prevalence, incidence and mortality).

Data sources: We developed and published a protocol, applying the Navigation Guide as an organizing systematic review framework where feasible. We searched electronic academic databases for potentially relevant records from published and unpublished studies, including the WHO International Clinical Trial Registers Platform, Medline, PubMed, EMBASE, Web of Science, CISDOC and PsycInfo. We also searched grey literature databases, Internet search engines and organizational websites; hand-searched reference lists of previous systematic reviews; and consulted additional experts.

Study eligibility and criteria: We included working-age (≥15 years) workers in the formal and informal economy in any WHO and/or ILO Member State but excluded children (aged <15 years) and unpaid domestic workers. We included randomized controlled trials, cohort studies, case-control studies and other non-randomized intervention studies with an estimate of the effect of exposure to long working hours (41–48, 49–54 and ≥55 h/week), compared with exposure to standard working hours (35–40 h/week), on depression (prevalence, incidence and/or mortality).

Study appraisal and synthesis methods: At least two review authors independently screened titles and abstracts against the eligibility criteria at a first stage and full texts of potentially eligible records at a second stage, followed by extraction of data from qualifying studies. Missing data were requested from principal study authors. We combined odds ratios using random-effects meta-analysis. Two or more review authors assessed the risk of bias, quality of evidence and strength of evidence, using Navigation Guide and GRADE tools and approaches adapted to this project.

Results: Twenty-two studies (all cohort studies) met the inclusion criteria, comprising a total of 109,906 participants (51,324 females) in 32 countries (as one study included multiple countries) in three WHO regions (Americas, Europe and Western Pacific). The exposure was measured using self-reports in all studies, and the outcome was assessed with a clinical diagnostic interview (four studies), interview questions about diagnosis and treatment of depression (three studies) or a validated self-administered rating scale (15 studies). The outcome was defined as incident depression in all 22 studies, with first time incident depression in 21 studies and recurrence of depression in one study. We did not identify any study on prevalence of depression or on mortality from depression. For the body of evidence for the outcome incident depression, we had serious concerns for risk of bias due to selection because of incomplete outcome data (most studies assessed depression only twice, at baseline and at a later follow-up measurement, and likely have missed cases of depression that occurred after baseline but were in remission at the time of the follow-up measurement) and due to missing information on life-time prevalence of depression before baseline measurement.

Compared with working 35–40 h/week, we are uncertain about the effect on acquiring (or incidence of) depression of working 41–48 h/week (pooled odds ratio (OR) 1.05, 95% confidence interval (CI) 0.86 to 1.29, 8 studies, 49,392 participants, I^2 46%, low quality of evidence); 49–54 h/week (OR 1.06, 95% CI 0.93 to 1.21, 8 studies, 49,392 participants, I^2 40%, low quality of evidence); and ≥ 55 h/week (OR 1.08, 95% CI 0.94 to 1.24, 17 studies, 91,142 participants, I^2 46%, low quality of evidence).

Subgroup analyses found no evidence for statistically significant ($P < 0.05$) differences by WHO region, sex, age group and socioeconomic status. Sensitivity analyses found no statistically significant differences by outcome measurement (clinical diagnostic interview [gold standard] versus other measures) and risk of bias ("high"/"probably high" ratings in any domain versus "low"/"probably low" in all domains).

Conclusions: We judged the existing bodies of evidence from human data as "inadequate evidence for harmfulness" for all three exposure categories, 41–48, 48–54 and ≥55 h/week, for depression prevalence, incidence and mortality; the available evidence is insufficient to assess effects of the exposure. Producing estimates of the burden of depression attributable to exposure to long working appears not evidence-based at this point. Instead, studies examining the association between long working hours and risk of depression are needed that address the limitations of the current evidence.

1. Background

`The World Health Organization (WHO) and the International Labour Organization (ILO) are producing the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury (WHO/ILO Joint Estimates) (Ryder, 2017). The organizations are estimating the numbers of deaths and disability-adjusted life years (DALYs) that are attributable to selected occupational risk factors. The WHO/ILO Joint Estimates are based on already existing WHO and ILO methodologies for estimating the burden of disease for selected occupational risk factors (Ezzati et al., 2004; International Labour Organization, 1999, 2014; Prüss-Üstün et al., 2017). They expand these existing methodologies with estimation of the burden of several prioritized additional pairs of occupational risk factors and health outcomes. For this purpose, population attributable fractions (Murray et al., 2004) – the proportional reduction in burden from the health outcome achieved by a reduction of exposure to the risk factor to zero – are being calculated for each additional risk factor-outcome pair, and these fractions are being applied to the total disease burden envelopes for the health outcome from the WHO Global Health Estimates for the years 2000–2016 (World Health Organization, 2019).

The WHO/ILO Joint Estimates may include estimates of the burden of depression attributable to exposure to long working hours, if feasible, as one additional risk factor-outcome pair whose global burden of disease has not previously been estimated. To select parameters with the best and least biased evidence for our estimation models, we conducted a systematic review and meta-analysis of studies on the relationship between exposure to long working hours and depression according to our protocol (Rugulies et al., 2019), and we present its results in the current paper. A systematic review of studies estimating the prevalence of exposure to long working hours is ongoing (Sembajwe et al., forthcoming) and applies novel systematic review methods (Pega et al., 2020b). WHO and ILO have conducted or are conducting several other systematic reviews and meta-analyses on other additional risk factor-outcome pairs (Descatha et al., 2018; Descatha et al., 2020; Godderis et al., 2018; Hulshof et al., 2019; Hulshof et al., 2021a; Hulshof et al., 2021b; Li et al., 2018; Li et al., 2020; Mandrioli et al., 2018; Pachito et al., 2020; Paulo et al., 2019; Pega et al., 2020a; Teixeira et al., 2021a; Teixeira et al., 2019; Teixeira et al., 2021b; Tenkate et al., 2019). To our knowledge, these are the first systematic reviews and meta-analyses (with a pre-published protocol) conducted specifically for an occupational burden of disease study. An editorial provides an overview of this series of systemic reviews and meta-analyses from the WHO/ILO Joint Estimates and outlines its scientific, methodological, policy, editorial and other innovations (Pega et al., 2021a). The WHO/ILO joint estimation methodology and the WHO/ILO Joint Estimates are separate from these systematic reviews, and they are described in more detail and reported elsewhere. WHO/ILO Joint Estimates have just been published of the global, regional and national burdens of ischemic heart disease and stroke attributable to exposure to long working hours for 194 countries for the years 2000, 2010 and 2016 (Pega et al., 2021b).

1.1. Rationale

To consider the feasibility of estimating the burden of depression attributable to exposure to long working hours, and to ensure that potential estimates of burden of depression are reported in adherence with the *Guidelines for Accurate and Transparent Health Estimates Reporting* (GATHER) (Stevens et al., 2016), WHO and ILO require a systematic review of studies on the prevalence of relevant levels of exposure to long working hours (Sembajwe et al., forthcoming), as well as a systematic review and meta-analysis with estimates of the relative effect of exposure to long work hours on depression prevalence, incidence and mortality, compared with the theoretical minimum risk exposure level (the systematic review presented here). The theoretical minimum risk exposure level is the exposure level that would result in the lowest possible population risk, even if it is not feasible to attain this exposure level in practice (Murray et al., 2004). These data and effect estimates should be tailored to serve as parameters for estimating the burden of depression from exposure to long work hours in the WHO/ILO Joint Estimates.

We are aware of at least three prior systematic reviews on the effect of long working hours on depression published since 2015. First, Theorell et al., reported, based on six cohort studies of high or moderate quality, that there was a prospective association of long working hours (denoted as "long working weeks" by the authors) with risk of onset of depressive symptoms (Theorell et al., 2015). Using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach (Morgan et al., 2016), they assessed the evidence as "limited" for women and "very limited" for men. The authors refrained from upgrading the evidence level for long working hours, because they found the estimates of the association between long working hours and depression neither consistent, nor large enough, for qualifying for an upgrade; they also did not conduct a meta-analysis of the included effect estimates. Second, Watanabe et al., examined overtime work and risk of onset of depressive disorders and identified seven cohort studies (Watanabe et al., 2016). The meta-analysis conducted in this systematic review showed an increased, but not statistically significant, association between working ≥50 h/week and risk of depressive disorders (relative risk (RR) 1.24, 95% CI 0.88 to 1.75). Third, Virtanen et al., included in their meta-analysis ten published cohort studies and 18 unpublished cohort studies with individual-participant data, yielding 31 study-specific estimates (as three studies of the published studies had provided estimates stratified by sex) (Virtanen et al., 2018). The outcome was named "depressive symptoms" and included both measures of clinical depression and depressive symptoms and of psychological distress. The overall pooled estimate for the association of long working hours with risk of onset of "depressive symptoms" was an odds ratio (OR) of 1.14 (95% CI 1.03 to 1.25). The association was stronger in studies from Asian countries (OR 1.50, 95% CI 1.13 to 2.01), weaker in European studies (OR 1.11, 95% CI 1.00 to 1.22), and absent in North American studies (OR 0.95, 95% CI 0.70 to 1.29) and in the one study from the Western Pacific (Australia only) (OR 0.95, 95% CI 0.70 to 1.29). When stratified by "depression" (including clinical depression and depressive symptoms) versus psychological distress, the pooled ORs were 1.09 (0.94 to 1.26) and 1.18 (1.06 to 1.32) for clinical depression/depressive symptoms and psychological distress, respectively. Meta-regressions did not show any statistically significant ($P = 0.05$) differences in the estimates for clinical depression, depressive symptoms and psychological distress. In summary, the previous systematic review and meta-analytic evidence appears to be inconclusive with regards to the outcome of depression. To our knowledge, prior systematic reviews did not have a pre-published protocol and/or missed other essential aspects of a systematic review. Our systematic review is fully compliant with latest systematic review methods (including use of a pre-published protocol: Rugulies et al., 2019) and expands the scope of the existing systematic review evidence by covering evidence from studies published up to 28 November 2019.

Our systematic review covers workers in the formal and in the informal economy. The informal economy is defined as "all economic activities by workers and economic units that are – in law or in practice – not covered or insufficiently covered by formal arrangements" (104th International Labour Conference 2015). It does not comprise "illicit activities, in particular the provision of services or the production, sale, possession or use of goods forbidden by law, including the illicit production and trafficking of drugs, the illicit manufacturing of and trafficking in firearms, trafficking in persons and money laundering, as defined in the relevant international treaties" (104th International, 2015). Work in the informal economy may lead to different exposures and exposure effects than work in the formal economy does. Consequently, formality of work (informal versus formal economy) may modify the effect of long working hours on depression. Therefore, we consider in the systematic review the formality of the economy reported in included studies.

1.2. Description of the risk factor

Burden of disease estimation requires unambiguous definition of the risk factor, risk factor levels and the theoretical minimum risk exposure level. Long working hours are defined as working hours exceeding standard working hours, i.e. any working hours of >40 h/week (Table 1). Based on results from earlier studies on long working hours and health endpoints (Kivimäki et al., 2015; Virtanen et al., 2015), the preferred four exposure level categories for our systematic review are 35–40, 41–48, 49–54 and ≥55 h/week (Table 1).

The theoretical minimum risk exposure is standard working hours defined as 35–40 h/week (Table 1). We acknowledge that it is possible that the theoretical minimum risk exposure might be lower than standard working hours, but we have to exclude working hours < 35 h/week, because studies indicate that some persons working less than standard hours do so because of existing health problems or family care obligations (Kivimäki et al., 2015; Sokejima and Kagamimori, 1998). In other words, persons working less than standard hours might belong to a

Table 1
Definitions of the risk factor, risk factor levels and the minimum risk exposure level.

	Definition
Risk factor	Exposure to long working hours (including those spent in secondary jobs), defined as working hours >40 h/week, i.e. working hours exceeding standard working hours (35–40 h/week).
Risk factor levels	Four levels: 35–40 h/week. 41–48 h/week. 49–54 h/week. ≥55 h/week.
Theoretical minimum risk exposure level	Standard working hours, defined as working hours of 35–40 h/week.

Source: (Rugulies et al., 2019).

health-selected group or a group concerned with family care and therefore cannot serve as comparators. Consequently, if a study used as the reference group persons working less than standard hours or a combination of persons working standard hours and persons working less than standard hours, it was excluded from the systematic review and meta-analysis. The category 35–40 h/week is the reference group used in many large studies and in the most recent systematic review by Virtanen et al. (Virtanen et al., 2018).

1.3. Definition of the outcome

The WHO Global Health Estimates group outcomes into standard burden of disease categories (World Health Organization, 2017), based on standard codes from the *International Statistical Classification of Diseases and Related Health Problems 10th Revision* (ICD-10) (World Health Organization, 2015). The relevant WHO Global Health Estimates category for this systematic review is "II.E.1 Major depressive disorders" (World Health Organization, 2017). In line with the WHO Global Health Estimates, we define the health outcome covered in this systematic review as depression, defined as conditions with ICD-10 codes F32, F33 and F34.1 (Table 2). Table 2 presents for each disease or health problem included in the WHO Global Health Estimates category the inclusion in this systematic review, showing that this review covers all the relevant categories.

1.4. How the risk factor may impact the outcome

Fig. 1 presents the logic model for our systematic review of the causal relationship between exposure to long working hours and depression, taken from our protocol (Rugulies et al., 2019). This logic model is an a priori, process-orientated one (Rehfuess et al., 2018) that seeks to capture the complexity of the risk factor–outcome causal relationship (Anderson et al., 2011).

Based on knowledge of previous research on long working hours and depression, we assume that the effect of long working hours on risk of depression may be mediated via (a) disturbance of work/life balance; (b) exhaustion; (c) emotional distress; (d) health-related behaviors, such as lack of physical activity, high alcohol consumption and reduced sleeping hours; and (e) psycho-physiological changes, such as activation of the hypothalamic–pituitaryadrenal (HPA) axis, inflammation processes, circadian disruptions, and sleep impairment (Baglioni et al., 2011; Bannai and Tamakoshi, 2014; Bergs et al., 2018; Boden and Fergusson, 2011; Fujimura et al., 2014; Gold, 2015; Kronfeld-Schor and Einat, 2012; McEwen, 2004; 2012; Pariante and Lightman, 2008; Pittenger and Duman, 2008; Virtanen et al., 2009; Virtanen et al., 2015).

Table 2
ICD-10 codes and disease and health problems covered by the WHO Global Health Estimates category "II.E.1 Major depressive disorders" and their inclusion in the systematic review.

ICD-10 code	Disease or health problem	Included in this review
F32	Depressive episode	Yes
F33	Recurrent depressive episode	Yes
F34.1	Dysthymia	Yes

Source: (Rugulies et al., 2019).

As possible confounders, we consider age, sex and socioeconomic position (also denoted as socioeconomic status, SES); we assume that these variables may impact both exposure to long working hours and risk of depression. It is well established that women and persons with low SES have a higher risk of depression than men and persons with high SES (Kessler et al., 2003; Lorant et al., 2003; Wittchen and Jacobi, 2005). With regard to age, some studies indicate that the 12-month prevalence of depression is modestly higher in young adulthood than in middle adulthood (Kessler et al., 2003; Wittchen and Jacobi, 2005), although birth cohort effects may also play a role, with a higher prevalence of depression in more recent birth cohorts (Kessler et al., 2003). Age, sex and SES may also be related to number of working hours, although the direction of these relationships may depend on other variables and contextual factors (Bannai et al., 2016; Larsen et al., 2017; Lee et al., 2016; O'Reilly and Rosato, 2013; Organisation for Economic Co-operation and Development (OECD) 2018; Wirtz et al., 2012); thus, it appears reasonable to regard these three variables as potential confounders for the association of exposure to long working hours with depression. We addressed this possible confounding by including only studies in the meta-analysis that had adjusted or stratified for age, sex and SES or had been based on study samples that were homogenous with regard to these criteria (e.g., included only men or only individuals from the same occupational group).

It is possible that age, sex and SES are not only confounders, but also effect modifiers for the association of long working hours and depression, and we consequently conducted meta-analyses stratified by age, sex and SES. We further considered as effect modifiers country, industrial sector, occupation and formality of the economy and also conducted meta-analyses stratified by these variables, if data allowed this.

Fig. 1 also considers the macro and meso-level contexts that may impact the prevalence of exposure to long working hours, the effect of long working hours exposure on depression, or both (Commission of Social Determinants on Health, 2008; Dahlgreen and Whitehead, 2006; Martikainen et al., 2002; Rugulies et al., 2004).

2. Objectives

To systematically review and meta-analyse evidence on the effect of exposure to long working hours (three categories: 41–48, 49–54 and ≥55 h/week) on depression prevalence, incidence and mortality among workers of working age, compared with the minimum risk exposure level (standard working hours: 35–40 h/week).

3. Methods

3.1. Developed protocol

We applied the Navigation Guide (Woodruff and Sutton, 2014) methodology for systematic reviews in environmental and occupational health as our guiding methodological framework wherever feasible. The guide applies established systematic review methods from clinical medicine, including standard Cochrane methods for systematic reviews of interventions, to the field of environmental and occupational health to ensure systematic and rigorous evidence synthesis on environmental and occupational risk factors that reduces bias and maximizes transparency (Woodruff and Sutton, 2014). The need for further methodological development and refinement of the relatively novel Navigation Guide has been acknowledged (Woodruff and Sutton, 2014). From the perspective of the Navigation Guide framework, all steps were conducted (i.e., steps 1–6 in Fig. 1 in Woodruff and Sutton, 2014) for the

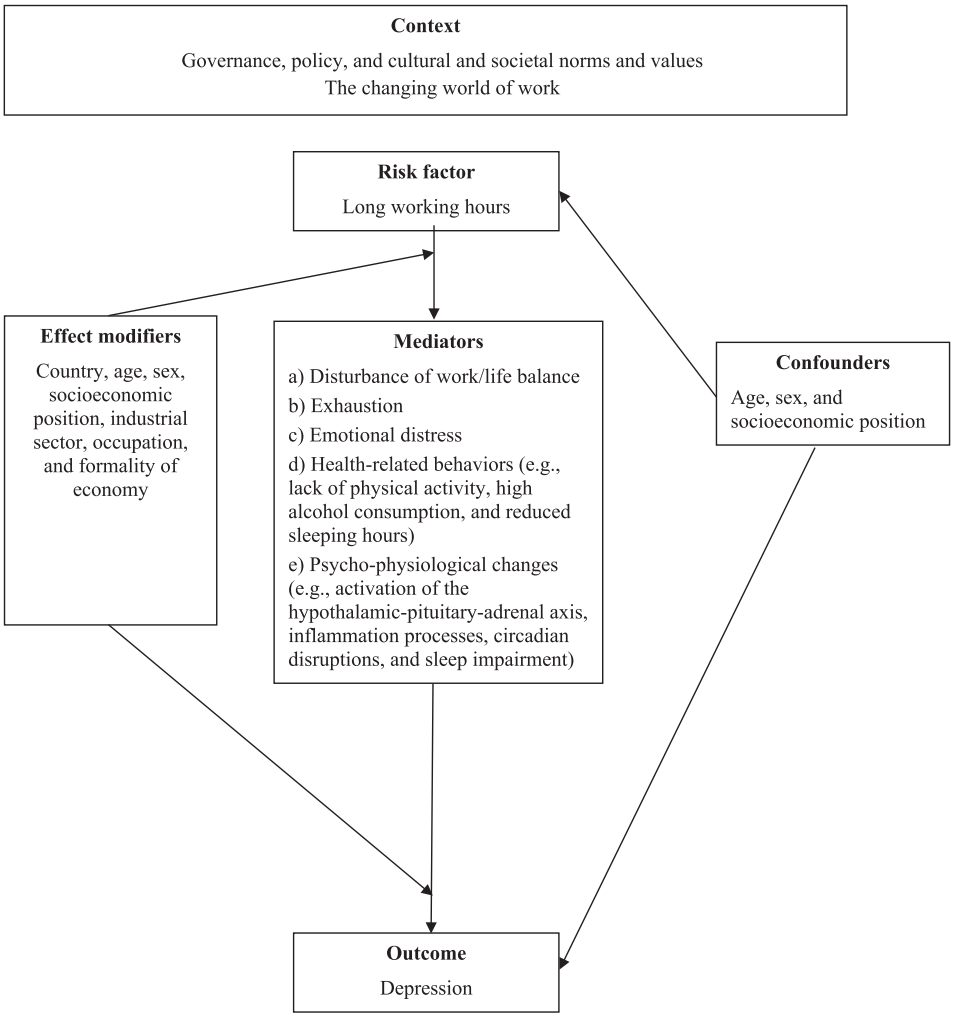

Fig. 1. Logic model of the possible causal relationship between exposure to long working hours and depression.

stream on human data and none of the steps for the stream on non-human data, although we narratively synthesized the mechanistic evidence from non-human data that we were aware of (Section 1.4).

We registered the protocol in PROSPERO under CRD42018085729. The protocol adheres to the preferred reporting items for systematic review and meta-analysis protocols statement (PRISMA-P) (Moher et al., 2015; Shamseer et al., 2015), with the abstract adhering to the reporting items for systematic reviews in journal and conference abstracts (PRISMA-A) (Beller et al., 2013). Any modification of the methods stated in the protocol was registered in PROSPERO and reported in the systematic review itself (*Section 8*). Our systematic review is reported according to the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses Statement* (PRISMA) (Liberati et al., 2009). Our reporting of the parameters for estimating the burden of depression attributable to exposure to long working hours in the systematic review adheres with the requirements of the GATHER guidelines (Stevens et al., 2016), because the WHO/ILO Joint Estimates that may be produced consecutive to the systematic review must also adhere to these reporting guidelines.

3.2. Searched literature

3.2.1. Electronic academic databases

We searched the seven following electronic academic databases:

1. WHO International Clinical Trials Registry Platform (to 18 July 2018).
2. Ovid MEDLINE with Daily Update (1 January 1946 to 11 July 2018).
3. PubMed (1 January 1946 to 11 July 2018).
4. EMBASE (1 January 1947 to 11 July 2018).
5. Web of Science (1 January 1945 to 11 July 2018).
6. CISDOC archived database (1 January 1901 to 17 July 2018).
7. PsycINFO (1 January 1880 to 12 July 2018).

The Ovid MEDLINE search strategy was presented in the protocol (Rugulies et al., 2019). The full search strategies for all databases were revised by a research librarian and are presented in Appendix 1 in the Supplementary data. We performed searches in electronic databases operated in the English language using a search strategy in the English language. When we neared completion of the review, we conducted a top-up search of the MEDLINE database on 28 November 2019 to capture the most recent publications (e.g., publications ahead of print). Deviations from the proposed search strategy and the actual search strategy are documented in Section 8.

3.2.2. Electronic grey literature databases

We also searched the two following grey literature databases.

- OpenGrey (http://www.opengrey.eu/) (up to 25 July 2018)

- Grey Literature Report (http://greylit.org/) (searched on 5 August 2018 but last update of Grey Literature Report database was in January 2017)

The full search strategies for the two grey literature databases are presented in Appendix 1 in the Supplementary data.

3.2.3. Internet search engines

We also searched the Google (www.google.com/) and Google Scholar (www.google.com/scholar/) Internet search engines on 8 October 2018 using the search terms "work*" AND (depression OR depressive OR antidepressant). We screened the first 100 hits for potentially relevant records, as was previously done in Cochrane Reviews (Pega et al., 2015; Pega et al., 2017).

3.2.4. Organizational websites

The websites of the six following international organizations and national government departments were searched between September and November 2018 using the keywords "Behavioural symptoms", "Affective symptoms", "Mood", "Depression", "Depressive disorders", "Dysthymia", "Adjustment disorders", "Antidepress":

1. International Labour Organization (www.ilo.org).
2. World Health Organization (www.who.int).
3. Eurostat (http://www.ec.europa.eu/eurostat/web/main/home).
4. United States National Institute of Occupational Safety and Health (NIOSH) of the United States of America, using the NIOSH data and statistics gateway (https://www.cdc.gov/niosh/data/).
5. Finnish Institute of Occupational Health (https://www.ttl.fi/en/).
6. International Commission of Occupational Health Scientific Committee on Work Organization and Psychosocial Factors (http://www.icohweb.org/site/scientific-committee-detail.asp?sc=33).

3.2.5. Hand-searching and expert consultation

We hand-searched for potentially eligible studies in:

- Reference lists of previous systematic reviews.
- Reference lists of all included trials register records.
- Study records published over the past 24 months in the three peer-reviewed academic journals with the largest number of included studies.
- Study records that have cited the included studies (identified in Web of Science citation database).
- Collections of the review authors.

Additional experts were contacted with a list of included studies, with the request to identify potentially eligible additional studies.

3.3. Selected studies

Study selection was carried out with the Covidence software (Veritas Health Innovation). All records identified in the search were downloaded, and duplicates were identified and deleted. Afterwards, two review authors independently and in duplicate screened titles and abstracts (step 1) and then full texts (step 2) of potentially relevant records. A third review author resolved any disagreements between the two review authors. If a study record identified in the literature search was authored by a review author assigned to study selection or if an assigned review author was involved in the study, the record was re-assigned to another review author for study selection. We present the study selection in a flow chart, as per PRISMA guidelines (Liberati et al., 2009).

3.4. Eligibility criteria

The population, exposure, comparator and ouctome (PECO) criteria (Morgan et al., 2018) are described below. Our protocol paper provides a complete, but briefer overview of the PECO criteria (see Rugulies et al., 2019 in Appendix A).

3.4.1. Types of populations

We included studies of the working-age population (≥15 years) in the formal and informal economy. Studies of children (aged <15 years) and unpaid domestic workers were excluded. Participants residing in any Member State of WHO and/or ILO and any industrial setting or occupation were included. Exposure to long working hours may potentially have further population reach (e.g., across generations for workers of reproductive age), and we acknowledge that the scope of our systematic review does not capture these populations and impacts on them.

3.4.2. Types of exposures

We included studies that defined exposure to long working hours in accordance with our standard definition (Table 1). We prioritized measures of the total number of hours worked, including in both of: main and secondary jobs, self-employment and salaried employment, whether in the informal or the formal economy. We included studies with objective (e.g., by means of time recording technology) or subjective measurements of long working hours, including studies that used measurements by experts (e.g. scientists with subject matter expertise) and self-reports by the worker, workplace administrator or manager. If a study presented both objective and subjective measurements, then we prioritized objective ones. Studies with measures from any data source, including registry data, were included. For studies that reported exposure levels differing from our standard levels (Table 1), we converted the reported levels to the standard levels if possible and reported analyses on these alternate exposure levels if impossible.

3.4.3. Types of comparators

The included comparators were participants exposed to the theoretical minimum risk exposure level: worked 35–40 h/week (Table 1). We excluded all other comparators.

3.4.4. Types of outcomes

This systematic review included three outcomes:

1. Has depression (or, in other words, depression prevalence).
2. Acquired depression (depression incidence).
3. Died from depression (depression mortality).

We included studies that defined depression in accordance with our standard definition (Table 2). Other affective disorders (e.g., bipolar disorders) were excluded. We did not expect documented ICD-10 diagnostic codes in most studies examining exposure to long working hours and its effect on depression, but expected that depression was assessed with other methods.

The following measurements of depression were regarded as eligible, as described in the protocol (Rugulies et al., 2019):

i. Psychiatric diagnostic interview.
ii. Diagnosis by a physician, psychologist or other qualified health professional.
iii. Hospital admission or discharge record.
iv. Administrative data (e.g., disability pensioning with the diagnosis of depression).
v. Register data of treatment for depression with one or both of antidepressant medication and psychotherapy (this measurement was only included if there was documentation that the treatment was for depression specifically, and not for other disorders).
vi. Self-administered rating scale for depression that was previously validated against a clinical measure of depression and that dichotomized respondents into cases versus non-cases (e.g., Center of Epidemiological Studies Depression Scale (CES-D)

(Radloff 1977) or Major Depression Inventory (MDI) (Bech et al., 2001)) or other validated self-administered rating scale.
vii. Medically certified cause of death.

Because the endpoint of our study was binary, studies exclusively reporting depression as a continuous variable (e.g., level of depressive symptoms) were excluded, as were all other measurements.

"More objective" measurements (e.g., diagnostic interview, administrative data or register data) and "more subjective" measurements (e.g., self-reported doctor-diagnosed depression or self-administered rating scale) were eligible. If a study presented both a "more objective" measurement and a "more subjective" measurement for the same outcome, then we prioritized the "more objective" one.

3.4.5. Types of studies

We included studies that investigated the effect of exposure to long working hours on depression for any study years and capturing any period of years. Eligible study designs were randomized controlled trials (including parallel-group, cluster, cross-over and factorial trials), cohort studies (both prospective and retrospective), case-control studies and other non-randomized intervention studies (including quasi-randomized controlled trials, controlled before-after studies and interrupted time series studies). We included a broader set of observational study designs than is commonly included, because a recent augmented Cochrane Review of complex interventions identified valuable additional studies using such a broader set of designs (Arditi et al., 2016). As we have an interest in quantifying risk and not in qualitative assessment of hazard (Barroga and Kojima, 2013), we excluded all other study designs (e.g., uncontrolled before-and-after, cross-sectional, qualitative, modelling, case and non-original studies).

Records published in any year and any language were included. Again, the search was conducted using English language terms, so that records published in any language that present essential information (i.e. title and abstract) in English were included. If a record was written in a language other than those spoken by the authors of this review or those of other reviews (Descatha et al., 2018; Descatha et al., 2020; Godderis et al., 2018; Hulshof et al., 2019; Hulshof et al., 2021a; Hulshof et al., 2021b; Li et al., 2018; Li et al., 2020; Mandrioli et al., 2018; Pachito et al., 2020; Paulo et al., 2019; Pega et al., 2020a; Teixeira et al., 2021a; Teixeira et al., 2019; Teixeira et al., 2021b; Tenkate et al., 2019) in the series (i.e., Arabic, Bulgarian, Chinese, Danish, Dutch, English, French, Finnish, German, Hungarian, Italian, Japanese, Norwegian, Portuguese, Russian, Spanish, Swedish and Thai), then the record was translated into English. Published and unpublished studies were included. Studies conducted using unethical practices were excluded (e.g., studies that deliberately exposed humans to a known risk factor to human health).

3.4.6. Types of effect measures

We included measures of the relative effect of a relevant level of exposure to long working hours on the risk of depression (prevalence, incidence and mortality), compared with the theoretical minimum risk exposure level. We included relative effect measures such as RRs, ORs or hazard ratios; however, all studies we identified reported ORs. Measures of absolute effects (e.g., mean differences in risks or odds) were converted into relative effect measures, but if conversion was impossible, they were excluded. We had aimed to convert OR into RR, however, this was for most studies not possible, as we were unable to extract information on the absolute risk of depression in the non-exposed. We therefore reported ORs and combined these effect estimates in the meta-analyses.

If a study presented estimates for the effect from two or more alternative models that had been adjusted for different variables, then we systematically prioritized the estimate from the model that provided information on the relevant confounders or mediators (at least the core variables defined in Fig. 1: age, sex and SES). We prioritized estimates from models adjusted for more potential confounders over those from models adjusted for fewer. For example, if a study presents estimates from a crude, unadjusted model (Model A), a model adjusted for one potential confounder (e.g. age; Model B) and a model adjusted for two potential confounders (e.g., age and sex; Model C), then we prioritized the estimate from Model C. We prioritized estimates from models unadjusted for mediators over those from models that adjusted for mediators, because adjustment for mediators will introduce bias (Greenland et al., 2016; Greenland and Pearce, 2015). For example, if Model A had been adjusted for two confounders, and Model B had been adjusted for the same two confounders and a potential mediator, then we chose the estimate from Model A. We prioritized estimates from models that can adjust for time-varying confounders that are at the same time also mediators, such as marginal structural models (Pega et al., 2016), over estimates from models that can only adjust for time-invariant confounders, such as fixed-effects models (Gunasekara et al., 2014), and over estimates from models that can adjust for neither time-varying, nor time-invariant confounding. If a study presents effect estimates from two or more potentially eligible models, then we documented why we prioritized the selected model.

3.5. Extracted data

A standard data extraction form was developed and trialled until data extractors reached convergence and agreement. At least two review authors independently extracted data on study characteristics (including study authors, study year, study country, participants, exposure and outcome), study design (including study type, comparator, epidemiological model(s) used and effect estimate measure) and risk of bias (including source population representation, blinding, exposure assessment, outcome assessment, confounding, incomplete outcome data, selective outcome reporting, conflict of interest and other sources of bias). A third review author resolved conflicts in data extraction. Data were entered into and managed with Excel.

We also extracted data on potential conflict of interest in included studies. For each author and affiliated organization of each included study record, we extracted their financial disclosures and funding sources. We used a modification of a previous method to identify and assess undisclosed financial interest of authors (Forsyth et al., 2014). Where no financial disclosure or conflict of interest statements were available, we searched the name of all authors in other study records gathered for this study and published in the prior 36 months and in other publicly available declarations of interests (Drazen et al., 2010a; Drazen et al., 2010b).

3.6. Requested missing data

We requested missing data from the principal study author by email or phone, using the contact details provided in the principal study record. If we did not receive a positive response from the study author, we sent follow-up emails twice, at two and four weeks. We present a description of missing data; the study author from whom the data were requested; the dates of requests sent; the date on which data were received (if any); and a summary of the responses provided by the study authors (Appendix 2 in the Supplementary data). If we did not receive some or all of the requested missing data, we nevertheless retained the study in the systematic review as long as it fulfilled our eligibility criteria.

3.7. Assessed risk of bias

Standard risk of bias tools do not exist for systematic reviews for hazard identification or those for risk assessment in occupational and environmental health. The five such tools developed specifically for occupational and environmental health are for either or both hazard identification and risk assessment, and they differ substantially in the types of studies (randomized, observational and/or simulation studies)

and data they seek to assess (e.g., human, animal and/or *in vitro*) (Rooney et al., 2016). However, all five tools, including the Navigation Guide one (Lam et al., 2016c), assess risk of bias in human studies similarly (Rooney et al., 2016).

Consistent with using the Navigation Guide as our organizing framework, we used its risk of bias tool, which builds on the standard risk of bias assessment methods of Cochrane (Higgins et al., 2011) and the US Agency for Healthcare Research and Quality (Viswanathan et al., 2008). Some further refinements of the Navigation Guide method may be warranted (Goodman et al., 2017), but it has been successfully applied in several completed and ongoing systematic reviews (Johnson et al., 2016; Johnson et al., 2014; Koustas et al., 2014; Lam et al., 2016a; Lam et al., 2014; Lam et al., 2017; Lam et al., 2016b; Vesterinen et al., 2014). In our application of the Navigation Guide method, we drew heavily on one of its latest versions, as presented in the protocol for an ongoing systematic review (Lam et al., 2016c).

We assessed risk of bias on the individual study level and across the body of evidence for each outcome. The nine risk of bias domains included in the Navigation Guide method for human data were: (i) source population representation; (ii) blinding; (iii) exposure assessment; (iv) outcome assessment; (v) confounding; (vi) incomplete outcome data; (vii) selective outcome reporting; (viii) conflict of interest; and (ix) other sources of bias. Risk of bias or confounding ratings for all domains were: "low"; "probably low"; "probably high"; "high"; or "not applicable" (Lam et al., 2016c). To judge the risk of bias in each domain, we applied a priori instructions (Rugulies et al., 2019), which we adapted from an ongoing Navigation Guide systematic review (Lam et al., 2016c) and further described in our protocol (Rugulies et al., 2019). For example, a study was assessed as carrying "low" risk of bias from source population representation, if we judge the source population to be described in sufficient detail (including eligibility criteria, recruitment, enrolment, participation and loss to follow up) and the distribution and characteristics of the study sample to indicate minimal or no risk of selection effects.

All risk of bias assessors jointly trialled the application of the risk of bias criteria until they had synchronized their understanding and application of these criteria. Two or more study authors independently assessed the risk of bias for each study by outcome. Where individual assessments differed, a third author resolved the conflict. For each included study, we reported our risk of bias assessment at the level of the individual study by domain in a standard 'Risk of bias' table (Higgins et al., 2011). For the entire body of evidence, we presented the study-level risk of bias ratings by domains in a 'Risk of bias matrix' (Higgins et al., 2011).

3.8. Synthesised evidence (including conducted meta-analysis)

We conducted separate meta-analyses for all outcomes: Has depression, Acquired depression, and Died from depression. If we found two or more studies with an eligible effect estimate, at least two review authors independently investigated the clinical heterogeneity (Deeks et al., 2011) of the studies in terms of participants (including country, sex, age and occupation or industrial sector), level of risk factor exposure, comparator and outcomes, following our protocol (Rugulies et al., 2019). If we found that effect estimates differed considerably by WHO region, sex and/or age, or a combination of these, then we synthesised evidence for the relevant populations defined by WHO region, sex and/or age, or combination thereof. If we found effect estimates to be clinically homogenous across WHO regions, sex and age groups, we combined studies from all of these populations into one pooled effect estimate that could be applied across all combinations of WHO regions, sexes and age groups in the WHO/ILO Joint Estimates.

If we judged two or more studies for the relevant combination of WHO region, sex and age group, or combination thereof, to be sufficiently clinically homogenous to potentially be combined using quantitative meta-analysis, we tested the statistical heterogeneity of the studies using the I^2 statistic (Figueroa, 2014). If two or more clinically homogenous studies were found to be sufficiently homogenous statistically to be combined in a meta-analysis, we pooled the ORs of the studies in a quantitative meta-analysis, using the inverse variance method with a random effects model to account for cross-study heterogeneity (Figueroa, 2014). The meta-analysis was conducted in RevMan 5.3. We neither quantitatively combined data from studies with different designs (e.g. did not combine cohort studies with case-controls studies), nor did we combine unadjusted with adjusted models. We only combined studies that we judged to have a minimum acceptable level of adjustment for the three core confounders we had identified at protocol stage (Fig. 1, *Section 3.4.5*).

If quantitative synthesis was not feasible (for instance, due to different exposure levels as defined above), we synthesised the study findings narratively and identified the estimates that we judged to be the highest quality evidence available.

3.9. Conducted subgroup and sensitivity analyses

Subgroup analyses were conducted only for the main meta-analysis and comparison of interest (i.e., the meta-analysis for the comparison of worked ≥55 h/week, compared with worked 35–40 h/week). We conducted subgroup analyses by:

- WHO region.
- Sex.
- Age group.
- SES.

We also planned to conduct subgroup analyses by occupation, industrial sector and formality of the economy, but did not find evidence or receive missing data to populate these analyses.

We conducted the following sensitivity analyses:

- Studies judged to be of "high"/"probably high" risk of bias in any domain, compared with those judged as of "low"/"probably low" risk of bias in all domains.
- Studies with documented or approximated ICD-10 diagnostic codes (e.g., as recorded in administrative health records), compared with studies without ICD-10 diagnostic codes (e.g., self-reports).

We planned to also compare studies with "low" or "probably low" risk of bias from conflict of interest with studies with "high" or "probably high" risk of bias in this domain but did not conduct these sensitivity analyses because we rated all included study to have "low"/"probably low" risk of bias from conflict of interest.

3.10. Assessed quality of evidence

We assessed the quality of evidence using a modified version of the Navigation Guide quality of evidence assessment tool (Lam et al., 2016c). The tool is based on the GRADE approach (Schünemann et al., 2011b) adapted specifically to systematic reviews in occupational and environmental health (Woodruff and Sutton 2014). An overview of GRADE and a discussion of its applicability to environmental health research has been presented by Morgan et al. (2016).

At least two review authors assessed the quality of evidence for the entire body of evidence by outcome, with any disagreements resolved by a third review author. We adapted the latest Navigation Guide instructions (Rugulies et al., 2019) for grading the quality of evidence (Lam et al., 2016c) and presented the adapted instructions in our protocol (Rugulies et al., 2019). We downgraded the quality of evidence for the following five GRADE considerations: (i) risk of bias; (ii) inconsistency that cannot be explained (or, in other words, "unexplained heterogeneity"); (iii) indirectness; (iv) imprecision; and (v) publication bias. These considerations were downgrades if they could not be

explained. If our systematic review had included ten or more studies, we generated an Egger's funnel plot to judge concerns regarding publication bias.

We graded the quality of the entire body of evidence by outcome, using the three Navigation Guide standard quality of evidence ratings: "high", "moderate" and "low" (Lam et al., 2016c). Within each of the relevant domains, we rated the concern for the quality of evidence, using the ratings "none", "serious" and "very serious". As per Navigation Guide, we started at "high" for randomized studies and "moderate" for observational studies. Quality was downgraded for no concern by nil levels (0), for a serious concern by one level (-1) and for a very serious concern by two levels (-2). We upgraded the quality of evidence for the following other reasons: (i) large magnitude of effect; (ii) presence of a dose–response relationship; and (iii) the plausibility that potential residual confounding cannot explain the effect. There had to be compelling reasons to downgrade or upgrade. If we had a serious concern for risk of bias in a body of evidence consisting of observational studies (−1 level), but had no other concerns, and had no reasons for upgrading, then we downgraded the quality of evidence by one level from "moderate" to "low".

3.11. Assessed strength of evidence

Our systematic review included studies of human data only and no studies of non-human data. The standard Navigation Guide methodology (Lam et al., 2016c) allows for rating human and non-human animal studies separately, and then combining the strength of evidence for each stream for an overall strength of evidence rating. However, the Navigation Guide also allows for rating one stream of evidence based on the domains described above (i.e., risk of bias, indirectness, inconsistency that cannot be explained, imprecisions, publication bias, large magnitude of effect, dose–response and residual confounding) to arrive at an overall rating of the quality of evidence as 'high', 'moderate' or 'low' (see above and the protocol). The approach of evaluating only the human evidence stream is consistent with the GRADE methodology that has adopted the Bradford Hill considerations (Bradford Hill, 1965; Schünemann et al., 2011a). So, using the method above based on the Navigation Guide incorporates the considerations of Bradford Hill (Table 3).

Table 3
Bradford Hill considerations and their relationship to GRADE and the Navigation Guide for evaluating the overall quality of the evidence for human observational studies.

Bradford Hill	GRADE	Navigation Guide
Strength	Strength of association and imprecision in effect estimate	Strength of association and imprecision in effect estimate
Consistency	Consistency across studies, i.e., across different situations (different researchers)	Consistency across studies, i.e., across different situations (different researchers)
Temporality	Study design, properly designed and conducted observational studies	Study design, properly designed and conducted observational studies
Biological Gradient	Dose response gradient	Dose response gradient
Specificity	Indirectness	Indirectness
Coherence	Indirectness	Indirectness
Experiment	Study design, properly designed and conducted observational studies	Study design, properly designed and conducted observational studies
Analogy	Existing association for critical outcomes leads to not downgrading the quality, indirectness	Existing association for critical outcomes leads to not downgrading the quality, indirectness. Evaluating the overall strength of body of human evidence allows consideration of other compelling attributes of the data that may influence certainty.

Adapted from (Schünemann et al., 2011a) and (Lam et al., 2016c).

There is an additional step that is described in the protocol that integrates the quality of the evidence (the method for assessing it was described above) with other elements, including the direction of the effect, the confidence in the effect and other compelling attributes of the data that may influence our certainty to allow for an overall rating that consists of "sufficient evidence of harmfulness", "limited of harmfulness", "inadequate of harmfulness" and "evidence of lack of harmfulness" based on human evidence. This approach to evaluate only the human evidence has been applied in previous systematic reviews (Lam, 2016c, Lam, 2017) and verified by the United States of America's National Academy of Sciences (National Academies of Sciences, 2017). It also provides two steps that integrate Bradford Hill considerations (i.e., evaluating the quality of the evidence, and then evaluating the overall strength of evidence) (Bradford Hill, 1965). Finally, the GRADE quality of evidence ratings (which are the same as for Navigation Guide) are analogous to the final ratings from Bradford Hill for causality which has been described in Schünemann et al. (2011a) (Table 4).

4. Results

4.1. Study selection

Of the total of 25,550 individual study records identified in our searches, 13 study records reporting results from 22 individual studies fulfilled the eligibility criteria, and these 22 studies were included in the systematic review (Fig. 2). This included ten journal articles providing results from ten studies (Ahn 2018; Berthelsen et al., 2015; Dembe and Yao, 2016; Kato et al., 2014; Kim, 2013; Kim et al., 2016; Shields, 1999; Virtanen et al., 2012; Wang et al., 2012a; Wang et al., 2012b) and one systematic review article (Virtanen et al., 2018) providing results from 18 unpublished studies, of which ten fulfilled the eligibility criteria (Virtanen et al., 2018 – ACL; Virtanen et al., 2018 – DWECS-2000; Virtanen et al., 2018 – ELSA; Virtanen et al., 2018 – HeSSup; Virtanen et al., 2018 – HILDA; Virtanen et al., 2018 – HRS; Virtanen et al., 2018 – NLSY; Virtanen et al., 2018 – PUMA; Virtanen et al., 2018 – SHARE; Virtanen et al., 2018 – SLOSH). The remaining eight studies from the Virtanen et al. (2018) review were excluded, because their outcome was not depression but distress (n = 7) or they assessed depression with an unvalidated rating scale (n = 1). Further, we included results from an unpublished manuscript (Zadow et al., 2019; note: the article is now accepted Zadow et al., 2021) and from an analysis we (the author group) conducted of the National Longitudinal Survey of Youth (NLSY) open access database (NLSY OA Cohort, 2019).

We excluded 397 study records after full-text screening. For the 40 excluded studies that most closely resembled the included studies, the reasons for their exclusion are listed in Appendix 3. In addition, there were three studies that provided insufficient information for determining their eligibility (Laaksonen et al., 2012; Ogasawara et al., 2011; Tokuyama et al., 2003). We tried to contact the study authors for missing data to screen these studies for eligibility but did not receive the requested missing data. These three studies are therefore classified as "studies awaiting classification" in the flow chart (Fig. 2), and their characteristics are briefly described in Appendix 4 in the Supplementary data.

Of the 22 studies included in the review, 17 studies were included in one or more quantitative meta-analyses, whereas five studies were not included in any meta-analysis (Fig. 2). Of these five studies, two studies used a comparator (≤35 h/week (Berthelsen et al., 2015) and ≤50 h/week (Kato et al., 2014)) that was substantially different from our standard comparator (35–40 h/week), and the authors of these studies could not provide us with a re-analysis with a different comparator. Further, we excluded the study by Kim et al. (2016) from the meta-analysis, because it was based on the same study population as the included study by Ahn (2018). We prioritized the study by Ahn et al., because it was more recent and because the author responded to our

Table 4
Interpretation of the GRADE ratings of the overall quality of evidence and the Navigation Guide ratings for strength of evidence evaluation.

GRADE rating for quality of evidence	Interpretation of GRADE rating	Navigation Guide rating for strength of evidence for human evidence	Interpretation of Navigation Guide rating
High	There is high confidence that the true effect lies close to that of the estimate of the effect.	**Sufficient evidence of harmfulness**	A positive relationship is observed between exposure and outcome where chance, bias, and confounding can be ruled out with reasonable confidence. The available evidence includes results from one or more well-designed, well conducted studies, and the conclusion is unlikely to be strongly affected by the results of future studies.
Moderate	There is moderate confidence in the effect estimate: the true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different.	**Limited evidence of harmfulness**	A positive relationship is observed between exposure and outcome where chance, bias, and confounding cannot be ruled out with reasonable confidence. Confidence in the relationship is constrained by such factors as: the number, size, or quality of individual studies, or inconsistency of findings across individual studies. As more information becomes available, the observed effect could change, and this change may be large enough to alter the conclusion.
Low	The panel's confidence in the effect estimate is limited: the true effect may be substantially different from the estimate of the effect	**Inadequate evidence of harmfulness**	The available evidence is insufficient to assess effects of the exposure. Evidence is insufficient because of: the limited number or size of studies, low quality of individual studies, or inconsistency of findings across individual studies. More information may allow an assessment of effects.
Very Low	There is little confidence in the effect estimate: the true effect is likely to be substantially different from the estimate of effect.		

Footnote: Adapted from (Schünemann et al., 2011a) and (Lam et al., 2016c).

request for re-analysis of the data. Finally, we excluded the studies by Dembe and Yao (2016) and Virtanen et al., 2018 – NLSY from the meta-analyses, because these two studies were based on the same study population as the NLSY open access database cohort that we analysed (NLSY OA Cohort, 2019). We prioritized the open access data analysis over the published analyses by Dembe and Yao (2016) and Virtanen et al., 2018 – NLSY because in our analyses we could apply all of the standard exposure categories as described in our protocol (Rugulies et al., 2019).

4.2. *Characteristics of included studies*

The characteristics of the included studies are summarized in Table 5.

4.2.1. *Study type*

All 22 included studies were cohort studies, with 21 being prospective and one being retrospective in design. Of the 22 studies, all examined risk of acquiring depression, with 21 studies examining first onset of depression and one study recurrence of depression among participants who had reported a previous depression but who were free from depression at the time of the baseline assessment (Wang et al., 2012b). The effect estimates reported in all studies were ORs for an eligible category of exposure to long working hours, compared with standard working hours (or a similar comparator). All studies adjusted for our three pre-specified confounders (Fig. 1 and Table 5).

4.2.2. *Population studied*

The 22 included studies captured 109,906 workers (51,324 females and 58,582 males). Twenty-one studies included both female and male workers, whereas one study included male workers only (Kato et al., 2014). The age range was from 17 to 71 years. When studies reported a mean age, this was most often in the 40s (Table 5 for details).

By WHO region, most studies examined populations in the Americas (nine studies of two countries) followed by Europe (eight studies, including seven studies of five countries, plus one study combining data from 28 countries), followed by the Western Pacific (five studies of three countries). The most commonly studied countries were the United States of America (six studies), Canada (three studies), and Australia, Denmark, the Republic of Korea and the United Kingdom of Great Britain and Northern Ireland (two studies each).

Only one study provided a detailed breakdown by occupation and industrial sector (Ahn, 2018). Five studies provided limited information on occupation, industrial sector or both (Kato et al., 2014; Kim, 2013; Virtanen et al., 2012; Wang et al., 2012a), and one study included only one specific occupational group (i.e., nurses; Berthelsen et al., 2015). See Table 5 for details. The other studies did not provide information on industrial sector or occupation.

4.2.3. *Exposure studied*

All studies measured exposure to long working hours by self-report, either by pen-and-paper-survey (four studies), face-to-face-survey (three studies), telephone survey (four studies), an interview (unclear if face-to-face or telephone) (two studies) or did not specify which of these methods were used (nine studies). No included study used non-self-reported measures, such as official or company records of hours worked.

Only two of the 22 included studies (NLSY OA Cohort, 2019; Zadow et al., 2019) provided the standard exposure categories that we had predefined in the protocol. We therefore attempted to contact all study authors asking them to either provide us with the original data or conducting a re-analysis of the data using our standard exposure categories. The author of one study provided us with the original data (Kim, 2013), and the authors of three studies re-analysed the data according to our request (Ahn, 2018; Wang et al., 2012a; Wang et al., 2012b), yielding a total of six studies providing the standard exposure categories. In addition, five published studies (Dembe and Yao, 2016; Kato et al.,

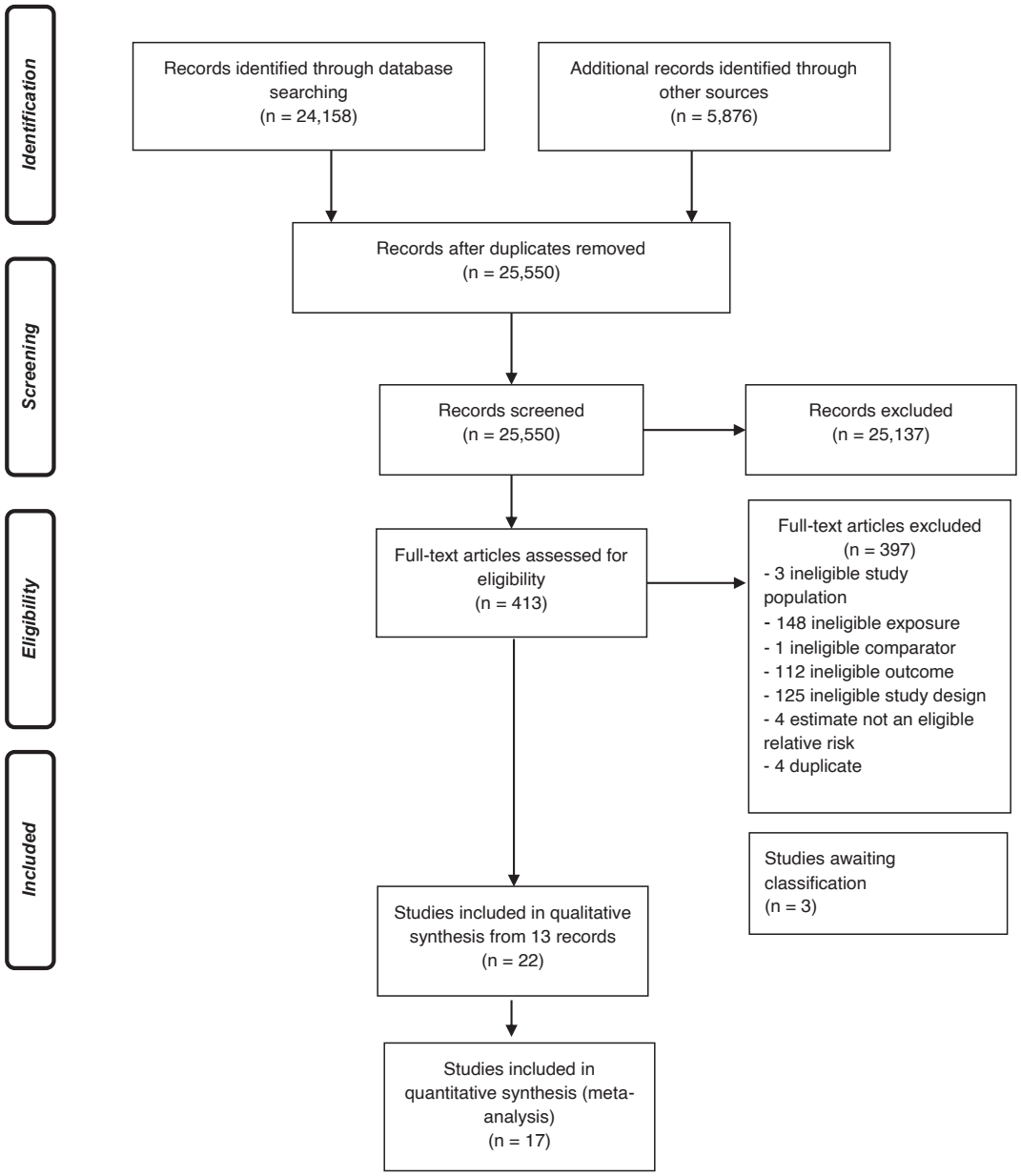

Fig. 2. Flow diagram of study selection.

2014; Kim et al., 2016; Shields, 1999; Virtanen et al., 2012) and all ten unpublished studies from the Virtanen et al. review (Virtanen et al., 2018) reported effect estimates for exposure categories that we regarded as similar enough to our standard exposure categories. One study provided exposure categories (i.e., 35.5–37.5 h/week and > 37.5 h/week) that we judged as non-comparable to our standard exposure categories (Berthelsen et al., 2015), because these exposure categories included values of our standard comparator (35–40 h/week).

4.2.4. Comparator studied

Nineteen studies used the standard comparator of 35–40 h/week, including the four studies that were re-analysed on our request (Ahn, 2018; Kim, 2013; Wang et al., 2012a; Wang et al., 2012b). Of the remaining three studies, one study used a comparator that we judged as similar (30–40 h/week) (Dembe and Yao, 2016). Two studies used a comparator that we judged as non-comparable to the standard comparator, because the comparator either included very low working hours (1–35 h/week) (Berthelsen et al., 2015) that may indicate reduced work ability and existing health problems, or working hours of 1–50 h/week (Kato et al., 2014) that included both very low working hours and working hours that overlapped with our standard exposure categories.

4.2.5. Outcomes studied

Of the three outcomes included in this systematic review, we found no eligible study for the following two outcomes:

- "Has depression" (depression prevalence); and
- "Died from depression" (depression mortality).

Of the 22 studies, all 22 studied the outcome "Acquired depression" (depression incidence), 21 studies capturing the incidence of first onset of depression (first time depression incidence) and one study investigating acquiring depression again (recurrence of depression) (Wang et al., 2012b).

Acquiring depression was measured using:

- A self-administrated rating scale in 15 studies (Ahn 2018; Berthelsen et al., 2015; Kato et al., 2014; Kim et al., 2016; Virtanen et al., 2018 –

Table 5
Characteristics of included studies in the systematic review of long working hours and depression.

(Part I: study population and study type)

Study	Study population							Study type		
Study ID	Total number of study participants	Number of female study participants	Country of study population	Geographic location (specify as 'national' or list regions or sites)	Industrial sector (specify ISIC-4 code provided in worksheet "Industrial sector codes")	Occupation (specify ISCO-08 code provided in worksheet "Occupation code")	Age	Study design	Study period (month of first collection of any data and month of last collection of any data)	Follow-up period (period in months between exposure and outcome)
Ahn 2018 (Ahn, 2018)	**Published:** n = 27,986 observations in logistic regression models **Re-analysis:** n = 7,415 individuals providing n = 26,304 observations	**Original:** n = 13,104 observations (46.8%) **Re-analysis:** n = 3,245 women (43.8%), providing 10,341 observations (39.3%)	Republic of Korea	National	Agriculture, fisheries, mining and quarrying; Manufacturing; Electricity, gas and water supply; Construction; Wholesale and retail trade; Transport, hotels and restaurants; Publishing and communication; Finance, insurance and real estate; R&D and technology service activities; Business activities; Public administration, defense, education, health and social work; Other community, social and personal service activities. (ISIC-code not reported)	Managers; Professionals and technicians; Clerks; Service workers; Skilled agricultural; Craft workers; Machine operators; Labourer; Armed force. (ISCO-code not reported)	**Original:** 25–64 years; Mean 42.86 (SD 0.5) **Re-analysis:** 20–65 + years	Cohort study (prospective)	2007–2013	Between 12 and 84 months
Berthelsen 2015 (Berthelsen et al., 2015)	n = 1,574	n = 1,430 (90.9%)	Norway	National	Nurses	Not reported	<30 years: n = 602; 30–39 years: n = 676; 40–49 years: n = 216; 50–59 years: n = 81; >59 years: n = 7	Cohort study (prospective)	2008–2010	12 months
Dembe 2016 (Dembe and Yao, 2016)	n = 7,492	n = 3,839 (51.2%)	United States of America	National	Not reported	Not reported	Mean age 49.6	Cohort study (prospective)	1979–2010	Up to 384 months
Kato 2014 (Kato et al., 2014)	n = 1,194	n = 0 (0%)	Japan	Local (two factories in Tochigi and Fukushima)	Manufacturing (ISIC-codes not reported)	Blue collar and white-collar workers (ISCO-codes not reported)	18–71 years. Mean: 38.9 (SD 13.4)	Cohort study (prospective)	April 2008 and June 2009	12 months
Kim 2013 (Kim, 2013)	**Original:** n = 35,155 **Re-analysis:** n = 27,975	**Original:** n = 16,911 (48.1%) **Re-analysis:** n = 12,038 (43.0%)	United States of America	National	Not reported	White collar (management, professional sales, office and administration-related occupations; Service); Farming (farming, fishing, and forestry occupations); Blue collar (construction, extraction, and maintenance; and production, transportation and material	**Original:** 18–64 years **Re-analysis:** 17–64 years; Mean: 39.9 (SD: 11.3)	Cohort study (prospective)	2000–2006	6 to18 months (most followed for 18 months)

(*continued on next page*)

Table 5 (*continued*)

(Part I: study population and study type)

Study	Study population							Study type		
Study ID	Total number of study participants	Number of female study participants	Country of study population	Geographic location (specify as 'national' or list regions or sites)	Industrial sector (specify ISIC-4 code provided in worksheet "Industrial sector codes")	Occupation (specify ISCO-08 code provided in worksheet "Occupation code")	Age	Study design	Study period (month of first collection of any data and month of last collection of any data)	Follow-up period (period in months between exposure and outcome)
						moving) (ISCO-codes not reported)				
Kim 2016 (Kim et al., 2016)	n = 2,733 individuals with 6,805 observations	n = not reported; estimated: n = 998 (36.5%) with 2,486 observations	Republic of Korea	National	Not reported	Not reported	20–29 years n = 632; 30–39 years n = 2,111; 40–49 years n = 2,530; 50–59 years n = 1,532	Cohort study (prospective)	2010–2013	12 months
NLSY OA Cohort 2019 (NLSY Oa Cohort, 2019)	n = 4,420	n = 1,884 (42.6%)	United States of America	National	Not reported	Not reported	All participants were 39 to 40 years at baseline	Cohort study (prospective)	1996–2014	120 months
Shields 1999 (Shields, 1999)	n = 3,830	n = 1,649 (43.1%)	Canada	National	Not reported	White-collar (administrative and professional); clerical, sales or service; and blue-collar (ISCO-codes not reported)	25–54 years	Cohort study (prospective)	1994–1997	24 months
Virtanen 2012 (Virtanen et al., 2012)	n = 2,123	n = 497 (23.4%)	United Kingdom of Great Britain and Northern Ireland	Local (Civil servants in London)	Civil Service (ISIC-codes not reported)	Office staff (ISCO-codes not reported)	35–55 years; Mean age: 46.7 years (SD 4.8)	Cohort study (prospective)	Baseline in 1991–1993 and follow-up in 1997–1999.	Mean follow-up: 69,6 months
Virtanen 2018 – ACL (Virtanen et al., 2018)	n = 1,291	n = not reported; estimated: n = 658 (51%)	United States of America	National	Not reported	Not reported	Mean age: 45	Cohort study (prospective)	Year of study entry: 1986	Mean follow-up: 36 months
Virtanen 2018 - DWECS (Virtanen et al., 2018)	n = 3,620	n = not reported; estimated: n = 1,774 (49%)	Denmark	National	Not reported	Not reported	Mean age: 43	Cohort study (retrospective)	Year of study entry: 2000	Mean follow-up: 60 months
Virtanen 2018 – ELSA (Virtanen et al., 2018)	n = 3,220	n = not reported; estimated: n = 1,642 (51%)	United Kingdom of Great Britain and Northern Ireland	National	Not reported	Not reported	Mean age: 55.8	Cohort study (prospective)	Year of study entry: 2002	Mean follow-up: 24 months
Virtanen 2018 - HESSUP (	n = 9,963	n = not reported; estimated: n = 5,579 (56%)	Finland	National	Not reported	Not reported	Mean age: 40	Cohort study (prospective)	Year of study entry: 1998	Mean follow-up: 60 months

(*continued on next page*)

Table 5 (*continued*)

(Part I: study population and study type)

Study	Study population							Study type		
Study ID	Total number of study participants	Number of female study participants	Country of study population	Geographic location (specify as 'national' or list regions or sites)	Industrial sector (specify ISIC-4 code provided in worksheet "Industrial sector codes")	Occupation (specify ISCO-08 code provided in worksheet "Occupation code")	Age	Study design	Study period (month of first collection of any data and month of last collection of any data)	Follow-up period (period in months between exposure and outcome)
Virtanen et al., 2018)										
Virtanen 2018 - HILDA (Virtanen et al., 2018)	n = 5,315	n = not reported; estimated: n = 2,445 (46%)	Australia	National	Not reported	Not reported	Mean age: 39.5	Cohort study (prospective)	Year of study entry: 2005	Mean follow-up: 24 months
Virtanen 2018 – HRS (Virtanen et al., 2018)	n = 7,055	n = not reported; estimated: n = 3,739 (53%)	United States of America	National	Not reported	Not reported	Mean age: 54	Cohort study (prospective)	Year of study entry: 1992	Mean follow-up: 24 months
Virtanen 2018 – NLSY (Virtanen et al., 2018)	n = 5,169	n = not reported; estimated: n = 2,223 (43%)	United States of America	National	Not reported	Not reported	Mean age: 30.9	Cohort study (prospective)	Year of study entry: 1992	Mean follow-up: 24 months
Virtanen 2018 - PUMA (Virtanen et al., 2018)	n = 901	n = not reported; estimated: n = 766 (85%)	Denmark	National	Not reported	Not reported	Mean age: 45	Cohort study (prospective)	Year of study entry: 1999	Mean follow-up: 60 months
Virtanen 2018 - SHARE (Virtanen et al., 2018)	n = 5,302	n = not reported; estimated: n = 2,386 (45%)	27 European countries and Israel	National	Not reported	Not reported	Mean age: 56.4	Cohort study (prospective)	Year of study entry: 2004	Mean follow-up: 24 months
Virtanen 2018 - SLOSH (Virtanen et al., 2018)	n = 5,083	n = not reported; estimated: n = 2,694 (53%)	Sweden	National	Not reported	Not reported	Mean age: 49	Cohort study (prospective)	Year of study entry: 2008	Mean follow-up: 48 months
Wang 2012a (Wang et al., 2012a)	**Original:** n = 2,752 **Re-analysis:** n = 2,752	**Original:** n = not reported; weighted percent = 43.8% **Re-analysis:** n = 1,173 (42.6%)	Canada	Region (Working population in the province of Alberta)	66.9% of the population was selected from the oil and gas industry, the service industry, and the government. (ISIC-codes not reported)	Not reported	**Original:** Mean age: 42.6 (SD 0.21) **Re-analysis:** 25–65 years	Cohort study (prospective)	January 2008-November 2011	12 months

(*continued on next page*)

Table 5 (*continued*)

(Part I: study population and study type)

Study	Study population							Study type		
Study ID	Total number of study participants	Number of female study participants	Country of study population	Geographic location (specify as 'national' or list regions or sites)	Industrial sector (specify ISIC-4 code provided in worksheet "Industrial sector codes")	Occupation (specify ISCO-08 code provided in worksheet "Occupation code")	Age	Study design	Study period (month of first collection of any data and month of last collection of any data)	Follow-up period (period in months between exposure and outcome)
Wang 2012b (Wang et al., 2012b)	**Original:** n = 470 **Re-analysis:** n = 485	**Original:** n = not reported, 58.3% **Re-analysis:** n = 242 (52.8%)	Canada	Region (Working population in the province of Alberta)	Not reported	Not reported	**Original:** Mean age = 43.6 (SE 0.45) **Re-analysis:** 25–65 years	Cohort study (prospective)	January 2008–2009	12 months
Zadow 2019 (Zadow et al., 2019)	**Original:** n = 1,084 **Re-analysis:** Same as original	**Original:** n = 423 (39.0%) **Re-analysis:** Same as original	Australia	Region (Working population in the states of New South Wales, Western Australia, Southern Australia)	Not reported	Not reported	**Original:** 18–75 years; mean age = 47.6 (SD: 10.6) **Re-analysis:** Same as original	Cohort study (prospective)	Not reported	12 months

(Part II: exposure assessment and comparator)

Study	Exposure assessment							Comparator
Study ID	Exposure definition (i.e. how was the exposure defined?)	Unit for which exposure was assessed	Mode of exposure data collection	Exposure assessment methods	Levels/intensity of exposure (specify unit)	Number of study participants in exposed group	Number of study participants in unexposed group	Definition of comparator (define comparator group, including specific level of exposure)
Ahn 2018	Average number of working hours per week	Individual level	Face-to-face survey	Self report	**Original:** <30h/w, 30–40h/w, 41–52h/w, 53–60h/w, ≥ 60h/w **Re-analysis:** 35–40h/w, 41–48h/w, 49–54h/w, ≥55h/w	**Original:** 41–52h/w: 27.62%, 52–60h/w: 12.98%, >61h/w: 9.14% **Re-analysis:** Not reported	**Original:** <30h/w (part-time work): 19.63%, 30–40h/w (reference group): 30.63% **Re-analysis:** Not reported	**Original:** 30–40h/w **Re-analysis:** 35–40h/w
Berthelsen 2015	Working hours per week	Individual level	Pen-and-paper-survey; computer-assisted questionnaire	Self report	≤35h/w, 35.5–37-5h/w, >37.5h/w	35.5–37-5h/w: n= 683, >37.5h/w: n= 309	≤35 h: n= 560	≤35h/w
Dembe 2016	Average work hours per week summed across the entire 32 years study period	Individual level	Interview, type not specified	Self report	30–40h/w, 41–50h/w, 51–60h/w, >60h/w	41–50h/w: n=4,171, 51–60h/w: n=869, >60h/w: n=210	n=2,242	30–40h/w
Kato 2014	Working hours per week	Individual level	Pen-and-paper survey	Self report	≤50h/w, 50.1–60h/w (or 40.1–80h overtime per month),	50.1–60h/w: n=247 >60h/w: n=29	n=918	≤50h/w

(*continued on next page*)

Table 5 (*continued*)

(Part II: exposure assessment and comparator)

Study	**Exposure assessment**							**Comparator**
Study ID	Exposure definition (i.e. how was the exposure defined?)	Unit for which exposure was assessed	Mode of exposure data collection	Exposure assessment methods	Levels/intensity of exposure (specify unit)	Number of study participants in exposed group	Number of study participants in unexposed group	Definition of comparator (define comparator group, including specific level of exposure)
					>60h/w (or >80h overtime per month)			
Kim 2013	Working hours per week	Individual level	Face-to-face survey	Self report	**Original:** ≤40h/w, >40h/w **Re-analysis:** 35–40h/w, 41–48h/w, 49–54h/w, ≥55h/w	**Original:** Unclear **Re-analysis:** 41–48h/w: n=2,864, 49–54h/w: n=2,632, ≥55h/w: n=3,024	**Original:** Unclear **Re-analysis:** n=19,455	**Original:** ≤40h/w **Re-analysis:** 35–40h/w
Kim 2016	Working hours per week	Individual level	Face-to-face survey	Self report	35–40h/w, 41–52h/w, 53–68h/w, >68h/w	41–52h/w: n=2,324 53–68h/w: n=1,405 >68h/w: n=568	n=2,508	35–40h/w
NLSY OA Cohort 2019	Working hours per week	Individual level	Interview, type not specified	Self report	35–40h/w, 41–48h/w, 49–54h/w, ≥55h/w	41–48h/w: n=703, 49–54h/w: n=501, ≥55h/w: n=742	n=2,474	35–40h/w
Shields 1999	Working hours per week	Individual level	Pen-and-paper survey	Self report	35–40h/w, >41h/w	Not reported	Not reported	35–40h/w
Virtanen 2012	Hours of work on an average weekday	Individual level	Not reported, likely interview or questionnaire	Self report	7–8h/day (equivalent to 35–40h/w), 9h/day (equivalent to 45h/w), 10*h*/day (equivalent to 50h/w), 11–12h/day (equivalent to 55–60h/w)	9h/day: n=445, 10*h*/day: n=346, 11–12h/day: n=227	n=1,105	7–8h/day (equivalent to 35–40h/w)
Virtanen 2018 – ACL	Working hours per week	Individual level	Not reported, likely interview or questionnaire	Not reported	35–40h/w, ≥55h/w	Not reported	Not reported	35–40h/w
Virtanen 2018 – DWECS	Working hours per week	Individual level	Telephone survey	Self report	35–40h/w, ≥55h/w	Not reported	Not reported	35–40h/w
Virtanen 2018 – ELSA	Working hours per week	Individual level	Not reported, likely interview or questionnaire	Not reported	35–40h/w, ≥55h/w	Not reported	Not reported	35–40h/w
Virtanen 2018 – HESSUP	Working hours per week	Individual level	Not reported, likely interview or questionnaire	Not reported	35–40h/w, ≥55h/w	Not reported	Not reported	35–40h/w
Virtanen 2018 – HILDA	Working hours per week	Individual level	Not reported, likely interview or questionnaire	Not reported	35–40h/w, ≥55h/w	Not reported	Not reported	35–40h/w
Virtanen 2018 – HRS	Working hours per week	Individual level	Not reported, likely interview or questionnaire	Not reported	35–40h/w, ≥55h/w	Not reported	Not reported	35–40h/w
Virtanen 2018 – NLSY	Working hours per week	Individual level	Not reported, likely interview or questionnaire	Not reported	35–40h/w, ≥55h/w	Not reported	Not reported	35–40h/w

(*continued on next page*)

Table 5 (*continued*)

(Part II: exposure assessment and comparator)

Study	Exposure assessment							Comparator
Study ID	Exposure definition (i.e. how was the exposure defined?)	Unit for which exposure was assessed	Mode of exposure data collection	Exposure assessment methods	Levels/intensity of exposure (specify unit)	Number of study participants in exposed group	Number of study participants in unexposed group	Definition of comparator (define comparator group, including specific level of exposure)
Virtanen 2018 – PUMA	Working hours per week	Individual level	Penn-and-paper survey	Not reported	35–40h/w, ≥55h/w	Not reported	Not reported	35–40h/w
Virtanen 2018 – SHARE	Working hours per week	Individual level	Not reported, likely interview or questionnaire	Not reported	35–40h/w, ≥55h/w	Not reported	Not reported	35–40h/w
Virtanen 2018 – SLOSH	Working hours per week	Individual level	Not reported, likely interview or questionnaire	Not reported	35–40h/w, ≥55h/w	Not reported	Not reported	35–40h/w
Wang 2012a	Average number of working hours per week	Individual level	Telephone survey	Self reported	**Original:** ≤ 35h/w, 35.5–40h/w, ≥40.5h/w **Re-analysis:** 35–40h/w, 41–48h/w, 49–54h/w, ≥55h/w	**Original:** 35.5–40h/w: 43,9%, ≥ 40.5h/w: 38,1% **Re-analysis:** 41–48h/w n=397, 49–54h/w n=296, ≥55h/w n=320	**Original:** ≤ 35h/w: 18.0% **Re-analysis:** n=1,738	**Original:** ≤35h/w **Re-analysis:** 35–40h/w
Wang 2012b	Working hours per week	Individual level	Telephone survey	Self-report	**Original:** ≤ 35h/w, 35.5–40h/w, ≥40.5h/w **Re-analysis:** 35–40h/w, 41–48h/w, 49–54h/w, ≥55h/w	**Original:** 35.5–40h/w: 45.4%, ≥40.5h/w: 33.2% **Re-analysis:** 41–48h/w n=68, 49–54h/w n=63, ≥55h/w n=52	**Original:** 21.4% **Re-analysis:** n=302	**Original:** ≤ 35h/w **Re-analysis:** 35–40h/w
Zadow 2019	Working hours per week	Individual level	Telephone survey	Self-report	**Original:** 35–40h/w, 41–48h/w, 49–54h/w, ≥55h/w **Re-analysis:** Same as original	**Original:** 41–48h/w n=239, 49–54h/w n=147, ≥55h/w n=179 **Re-analysis:** Same as original	**Original:** n=519 **Re-analysis:** Same as original	**Original:** 35–40h/w **Re-analysis:** Same as original

(Part III: outcome assessment and statistical modelling)

Study	Outcome assessment							Statistical modelling					
Study ID	Definition of outcome	Which International Classification of Diseases (ICD) code was reported for the outcome (if any)?	Diagnostic assessment method	Number of cases with outcome of interest in exposed group	Number of non-cases (i.e. without outcome of interest) in exposed group	Number of cases with outcome of interest in unexposed group	Number of non-cases (i.e. without outcome of interest) in unexposed group	Adjusted for confounding by: age	Adjusted for confounding by: sex	Adjusted for confounding by: Socioeconomic status (please specify indicator, e.g. level of education)	Other potential confounders adjusted for (please specify)	Adjustment for any mediators	Treatment effect measure type
Ahn 2018	Depression	Not reported	Self-administered rating scale	**Original:** Not	**Original:** Not	**Original:** Not	**Original:** Not	**Original:** Yes	**Original:** Yes	**Original:** Yes, income	**Original:** Yes, marital	**Original:** No	Odds ratio

(*continued on next page*)

Table 5 (*continued*)

(Part III: outcome assessment and statistical modelling)

Study	Outcome assessment							Statistical modelling					
Study ID	Definition of outcome	Which International Classification of Diseases (ICD) code was reported for the outcome (if any)?	Diagnostic assessment method	Number of cases with outcome of interest in exposed group	Number of non-cases (i.e. without outcome of interest) in exposed group	Number of cases with outcome of interest in unexposed group	Number of non-cases (i.e. without outcome of interest) in unexposed group	Adjusted for confounding by: age	Adjusted for confounding by: sex	Adjusted for confounding by: Socioeconomic status (please specify indicator, e.g. level of education)	Other potential confounders adjusted for (please specify)	Adjustment for any mediators	Treatment effect measure type
			(Center for Epidemiologic Studies Depression, CES-D)	reported **Re-analysis:** Not reported	reported **Re-analysis:** Not reported	reported **Re-analysis:** Not reported	reported **Re-analysis:** Not reported	**Re-analysis:** Yes	**Re-analysis:** Yes	**Re-analysis:** Yes, education, income	status, country (all Republic of Korea) **Re-analysis:** Marital status, industry, occupation, country (all Republic of Korea)	**Re-analysis:** No	
Berthelsen 2015	Depression	Not reported	Self-administered rating scale (Hospital Anxiety and Depression Scale, HADS)	Not reported	Not reported	Not reported	Not reported	Yes	Yes	Yes, occupational grade (all participants were nurses)	Married/ cohabiting, children living at home, baseline level of symptoms of anxiety and depression, country (all Norway)	No	Odds ratio
Dembe 2016	Depression	Not reported	Interview question about self-reported doctor-diagnosed depression	Not reported	Not reported	Not reported	Not reported	Yes	Yes	Yes, education, family income	Race, number of years worked, smoking status, occupation, country (all USA)	No	Odds ratio
Kato 2014	Depression	Not reported	Self-administered rating scale (Center for Epidemiologic Studies Depression, CES-D)	50.1–60h/w: n=25 >60h/w n=8	50.1–60h/w: n=222 >60h/w n=21	n=97	n=821	Yes	Yes, men only	Yes, job grade	Years of experience, shift work, site of work, country (all Japan)	No	Odds ratio
Kim 2013	Depression	ICD-9 codes 296.2 major depression, single episode and 311, depressive disorder, not elsewhere classified	Comprehensive interview to identify episodes of depression by asking about health care utilization, hospital inpatient services, outpatient services and emergency	**Original:** Unclear **Re-analysis:** 41–48h/w n=95; 49–54h/w n=73; ≥55h/w n=84	**Original:** Unclear **Re-analysis:** 41–48h/w n=2,769; 49–54h/w n=2,559; ≥55h/w n=2,940	**Original:** Unclear **Re-analysis:** n=568	**Original:** Unclear **Re-analysis:** n=18,887	**Original:** Yes **Re-analysis:** Yes	**Original:** Unclear **Re-analysis:** Yes	**Original:** Yes, education **Re-analysis**: Yes, education	**Original:** Occupational group, job tenure, work status, smoking, alcohol or substance abuse disorder, exercise, and obesity, added activity	**Original:** Yes, alcohol disorder, exercise, cognitive function impairment **Re-analysis:** No	Odds ratio

(*continued on next page*)

Table 5 (*continued*)

(Part III: outcome assessment and statistical modelling)

Study	Outcome assessment							Statistical modelling					
Study ID	Definition of outcome	Which International Classification of Diseases (ICD) code was reported for the outcome (if any)?	Diagnostic assessment method	Number of cases with outcome of interest in exposed group	Number of non-cases (i.e. without outcome of interest) in exposed group	Number of cases with outcome of interest in unexposed group	Number of non-cases (i.e. without outcome of interest) in unexposed group	Adjusted for confounding by: age	Adjusted for confounding by: sex	Adjusted for confounding by: Socioeconomic status (please specify indicator, e.g. level of education)	Other potential confounders adjusted for (please specify)	Adjustment for any mediators	Treatment effect measure type
			department services. Responses were coded by verbatim interviewers and then coded by professional coders into ICD-9 codes.								limitation because of a chronic medical condition, cognitive function impairment, and comorbidity, self-rated physical and mental health status, country (all USA) **Re-analysis:** Country (all USA)		
Kim 2016	Depression	Not reported	Self-administered rating scale (Center for Epidemiologic Studies Depression, CES-D)	41–52h/w: n=119 53–68h/w: n=74 >68h/w: n=54	41–52h/w: n=2,205 53–68h/w n=1,331 >68h/w n=514	n=101	n=2,407	Yes	Yes	Yes, education, equalized household income	Marital status, country (all Republic of Korea)	No	Odds ratio
NLSY OA Cohort 2019	Depression	Not reported	Interview question about self-reported doctor-diagnosed depression	41–48h/w: n=64; 49–54h/w: n=43; ≥55h/w: n=58	41–48h/w: n=639; 49–54h/w: n=458; ≥55h/w: n=684	n=237	n=2,237	Yes (all 39–40 years old at baseline)	Yes	Yes, income	Country (all United States)	No	Odds ratio
Shields 1999	Depression	Not reported	Composite International Diagnostic Interview (CIDI)	Not reported	Not reported	Not reported	Not reported	Yes	Yes, stratified by sex	Yes, education, household income	White-collar/blue collar, self-employment, shift work, multiple job holdings, marital status, children under age 12 in household, work stress (job strain, job insecurity, supervisor	Yes, work stress	Odds ratio

(*continued on next page*)

Table 5 (*continued*)

(Part III: outcome assessment and statistical modelling)

Study	Outcome assessment							Statistical modelling					
Study ID	Definition of outcome	Which International Classification of Diseases (ICD) code was reported for the outcome (if any)?	Diagnostic assessment method	Number of cases with outcome of interest in exposed group	Number of non-cases (i.e. without outcome of interest) in exposed group	Number of cases with outcome of interest in unexposed group	Number of non-cases (i.e. without outcome of interest) in unexposed group	Adjusted for confounding by: age	Adjusted for confounding by: sex	Adjusted for confounding by: Socioeconomic status (please specify indicator, e.g. level of education)	Other potential confounders adjusted for (please specify)	Adjustment for any mediators	Treatment effect measure type
											support), country (all Canada)		
Virtanen 2012	Depression	None	Composite International Diagnostic Interview, University of Michigan version (UM-CIDI)	9h/day: n=8 10*h*/day: n=10 11–12h/day: n=10	9h/day: n=437 10*h*/day: n=336 11–12h/day: n=217	n=38	n=1,067	Yes	Yes	Yes, occupational grade	Marital status, country (all United Kingdom)	No	Odds ratio
Virtanen 2018 – ACL	Depression	Not reported	Self-administered rating scale (Center for Epidemiologic Studies Depression, CES-D)	Not reported	Not reported	Not reported	Not reported	Yes	Yes	Yes, type not reported	Marital status, country (all USA)	No	Odds ratio
Virtanen 2018 – DWECS	Depression	Not reported	Self-administered rating scale (Mental health inventory, MHI-5)	Not reported	Not reported	Not reported	Not reported	Yes	Yes	Yes, type not reported	Marital status, country (all Denmark)	No	Odds ratio
Virtanen 2018 – ELSA	Depression	Not reported	Self-administered rating scale (Center for Epidemiological Studies Depression Scale, CES-D)	Not reported	Not reported	Not reported	Not reported	Yes	Yes	Yes, type not reported	Marital status, country (all United Kingdom)	No	Odds ratio
Virtanen 2018 - HESSUP	Depression	Not reported	Self-administered rating scale (Beck's Depression Inventory, BDI)	Not reported	Not reported	Not reported	Not reported	Yes	Yes	Yes, type not reported	Marital status, country (all Finland)	No	Odds ratio
Virtanen 2018 – HILDA	Depression	Not reported	Self-administered rating scale (Mental health inventory (MHI-5))	Not reported	Not reported	Not reported	Not reported	Yes	Yes	Yes, type not reported	Marital status, country (all Australia)	No	Odds ratio
Virtanen 2018 – HRS	Depression	Not reported	Not reported, probably self-administered rating scale (Center for Epidemiologic Studies Depression, CES-D)	Not reported	Not reported	Not reported	Not reported	Yes	Yes	Yes, type not reported	Marital status, country (all USA)	No	Odds ratio

(*continued on next page*)

Table 5 (*continued*)

(Part III: outcome assessment and statistical modelling)

Study	Outcome assessment							Statistical modelling					
Study ID	Definition of outcome	Which International Classification of Diseases (ICD) code was reported for the outcome (if any)?	Diagnostic assessment method	Number of cases with outcome of interest in exposed group	Number of non-cases (i.e. without outcome of interest) in exposed group	Number of cases with outcome of interest in unexposed group	Number of non-cases (i.e. without outcome of interest) in unexposed group	Adjusted for confounding by: age	Adjusted for confounding by: sex	Adjusted for confounding by: Socioeconomic status (please specify indicator, e.g. level of education)	Other potential confounders adjusted for (please specify)	Adjustment for any mediators	Treatment effect measure type
Virtanen 2018 – NLSY	Depression	Not reported	Self-administered rating scale (Center for Epidemiologic Studies Depression, CES-D)	Not reported	Not reported	Not reported	Not reported	Yes	Yes	Yes, type not reported	Marital status (all USA)	No	Odds ratio
Virtanen 2018 – PUMA	Depression	Not reported	Self-administered rating scale (Mental health inventory, MHI-5)	Not reported	Not reported	Not reported	Not reported	Yes	Yes	Yes, type not reported	Marital status, country (all Denmark)	No	Odds ratio
Virtanen 2018 – SHARE	Depression	Not reported	Self-administered rating scale (EURO-D)	Not reported	Not reported	Not reported	Not reported	Yes	Yes	Yes, type not reported	Marital status	No	Odds ratio
Virtanen 2018 - SLOSH	Depression	Not reported	Self-administered rating scale, probably Symptom Check List Core Depression scale (SCL-CD6) although this is not completely clear	Not reported	Not reported	Not reported	Not reported	Yes	Yes	Yes, type not reported	Marital status, country (all Sweden)	No	Odds ratio
Wang 2012a	Depression	Not reported	Composite International Diagnosis Interview (CIDI)	**Original:** Not reported **Re-analysis:** 41–48h/w n=5, 49–54h/w n=8, ≥55h/w n=7	**Original:** Not reported **Re-analysis:** 41–48h/w n=313, 49–54h/w n=245, ≥55h/w n=242	**Original:** Not reported **Re-analysis:** n=50	**Original:** Not reported **Re-analysis:** n=1,384	**Original:** Yes Re-analysis: Yes	**Original:** Yes Re-analysis: Yes	**Original:** Yes, education, income, occupational gradient **Re-analysis:** Yes, education, income, occupational gradient	**Original:** Yes, marital status, Part time work, job strain, job insecurity, stress in supervisor support, stress in coworker support, effort-reward imbalance, work to family conflict, family to work conflict, anxiety disorders, country (all Canada) **Re-analysis:** Yes, marital	**Original:** Yes, job strain, job insecurity, stress in supervisor and coworker support, effort-reward imbalance, work to family conflict, family to work conflict **Re-analysis:** No	Odds ratio

(*continued on next page*)

Table 5 (*continued*)

(Part III: outcome assessment and statistical modelling)

Study	**Outcome assessment**							**Statistical modelling**					
Study ID	Definition of outcome	Which International Classification of Diseases (ICD) code was reported for the outcome (if any)?	Diagnostic assessment method	Number of cases with outcome of interest in exposed group	Number of non-cases (i.e. without outcome of interest) in exposed group	Number of cases with outcome of interest in unexposed group	Number of non-cases (i.e. without outcome of interest) in unexposed group	Adjusted for confounding by: age	Adjusted for confounding by: sex	Adjusted for confounding by: Socioeconomic status (please specify indicator, e.g. level of education)	Other potential confounders adjusted for (please specify)	Adjustment for any mediators	Treatment effect measure type
											status, country (all Canada)		
Wang 2012b	Depression	Diagnoses based on DSM-IV criteria (APA, 1994)	Composite International Diagnostic Interview, CIDI)	**Original:** Not reported **Re-analysis**: 41–48h/w n=8, 49–54h/w n=11, ≥55h/w n=5	**Original:** Not reported **Re-analysis**: 41–48h/w n=47, 49–54h/w n=44, ≥55h/w n=34	**Original:** Not reported **Re-analysis:** n=30	**Original:** Not reported **Re-analysis:** n=216	**Original:** Yes **Re-analysis:** Yes	**Original:** Yes **Re-analysis:** Yes	**Original:** No **Re-analysis:** Yes, education, income, occupational gradient	**Original:** Yes, country (all Canada) Re-analysis: Yes, marital status, country (all Canada)	**Original:** No **Re-analysis:** No	Odds ratio
Zadow 2019	Depression	Not reported	Self-administered rating scale (Patient Health Questionnaire 9 (PHQ-9))	**Original:** Not reported Re-analysis: 41–48h/w n=10, 49–54h/w n=5, ≥55h/w n=9	**Original:** Not reported **Re-analysis**: 41–48h/w n=239, 49–54h/w n=147, ≥55h/w n=175	**Original:** Not reported **Re-analysis:** n=13	**Original:** Not reported **Re-analysis:** n=554	**Original:** Yes **Re-analysis:** Yes	**Original:** Yes **Re-analysis:** Yes	**Original:** Yes, income **Re-analysis:** Yes, income	**Original:** Yes, psychosocial safety climate, country (all Australia) **Re-analysis:** Country (all Australia)	**Original:** Yes, psychosocial safety climate **Re-analysis:** No	Odds ratio

ACL; Virtanen et al., 2018 – DWECS; Virtanen et al., 2018 – ELSA; Virtanen et al., 2018 – HeSSup; Virtanen et al., 2018 – HILDA; Virtanen et al., 2018 – HRS; Virtanen et al., 2018 – NLSY; Virtanen et al., 2018 – PUMA; Virtanen et al., 2018 – SHARE; Virtanen et al., 2018 – SLOSH; Zadow et al., 2019):

- 7 studies used the Center of Epidemiologic Studies Depression Scale (CES-D).
- 3 studies used The Mental Health Inventory 5-item scale (MHI-5).
- 1 study used the Beck Depression Inventory (BDI).
- 1 study used the European Depression scale (Euro-D).
- 1 study used the Hospital Anxiety and Depression Scale (HADS).

Navigation Guide (Woodruff and Sutton 2014) risk of bias domain	Study\Navigation Guide risk of bias ratings										
	Ahn (2018)	Berthelsen (2015)	Dembe (2016)	Kato (2014)	Kim, J (2013)	Kim, W (2016)	NLSY OA Cohort (2019)	Shields (1999)	Virtanen (2012)	Virtanen (2018) – ACL	Virtanen (2018) – DWECS2000
1. Are the study groups at risk of not representing their source populations in the manner that might introduce selection bias?	Probably low	Probably high	Probably high	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low
2. Was knowledge of the group assignments inadequately prevented (i.e. blinded or masked) during the study, potentially leading to subjective measurement of either exposure or outcome?	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low
3. Were exposure assessment methods lacking accuracy?	Probably low	Probably high	Probably high	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low
4. Were outcome assessment methods lacking accuracy?	Probably low	Probably low	Probably high	Probably low	Probably low	Probably low	Probably high	Low	Low	Probably low	Probably low
5. Was potential confounding inadequately incorporated?	Low	Low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably high	Probably low	Probably low	Probably low
6. Were incomplete outcome data inadequately addressed?	Probably high	Probably high	Probably high	Probably high	Probably low	Probably high	Probably high	Probably low	Probably high	Probably high	Probably high
7. Does the study appear to have selective outcome reporting?	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low
8. Did the study receive any support from a company, study author, or other entity having a financial interest in any of the exposures studied?	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low
9. Did the study appear to have other problems that could put it at a risk of bias? (Missing information on depressive episodes prior baseline assessment)	Probably high	Probably high	Probably high	Probably high	Probably low	Probably high	Probably high	Probably high	Probably high	Probably high	Probably high

Navigation Guide (Woodruff and Sutton 2014) risk of bias domain	Study\Navigation Guide risk of bias ratings										
	Virtanen (2018) – ELSA	Virtanen (2018) – HESSUP	Virtanen (2018) – HILDA	Virtanen (2018) – HRS	Virtanen (2018) – NLSY	Virtanen (2018) – PUMA	Virtanen (2018) – SHARE	Virtanen (2018) – SLOSH	Wang (2012a)	Wang (2012b)	Zadow (2019)
1. Are the study groups at risk of not representing their source populations in the manner that might introduce selection bias?	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably high	Probably low
2. Was knowledge of the group assignments inadequately prevented (i.e. blinded or masked) during the study, potentially leading to subjective measurement of either exposure or outcome?	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low
3. Were exposure assessment methods lacking accuracy?	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably high	Probably low
4. Were outcome assessment methods lacking accuracy?	Probably low	Probably low	Probably low	Probably high	Probably low	Probably low	Probably low	Probably high	Low	Low	Probably low
5. Was potential confounding inadequately incorporated?	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low	Probably low
6. Were incomplete outcome data inadequately addressed?	Probably high	Probably high	Probably high	Probably high	Probably high	Probably high	Probably high	Probably high	Probably low	Probably low	Probably high
7. Does the study appear to have selective outcome reporting?	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low
8. Did the study receive any support from a company, study author, or other entity having a financial interest in any of the exposures studied?	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low
9. Did the study appear to have other problems that could put it at a risk of bias? (Missing information on depressive episodes prior baseline assessment)	Probably high	Probably high	Probably high	Probably high	Probably high	Probably high	Probably high	Probably high	Probably low	Probably low	Probably high

Fig. 3. Summary of risk of bias, Acquired depression (depression incidence).

  - 1 study used the Symptom Checklist Core Depression 6-item scale (SCL-CD6).
  - 1 study used the Patient Health Questionnaire 9 (PHQ-9).
- A clinical diagnostic interview in four studies (Shields, 1999; Virtanen et al., 2012; Wang et al., 2012a; Wang et al., 2012b).
- A comprehensive interview assessing diagnosed depression and treatment for depression in one study (Kim 2013).
- A question about doctor-diagnosed depression in two studies (Dembe and Yao, 2016; NLSY OA Cohort, 2019).

4.2.6. Statistical modelling

All 22 studies adjusted or stratified for the three key potential confounders (i.e., sex, age and SES), and most studies adjusted for further variables, for example marital status, cohabitation or children living at home. When studies provided multiple statistical models with different adjustments, we aimed for selecting estimates that were adjusted for as many confounders as possible while not being adjusted for potential mediation. See Appendix 5 in the Supplementary data for a description of our rationale for selecting specific estimates from the included studies. For six studies, we conducted re-analyses of the original data, giving us control over the statistical modelling (Ahn, 2018; Kim, 2013; NLSY OA Cohort, 2019; Wang et al., 2012a; Wang et al., 2012b; Zadow et al., 2019).

4.3. Risk of bias at individual study level

We assessed risk of bias based on the information we retrieved from the included study records and additional records on the included studies. For the ten previously unpublished studies presented in the Virtanen et al. (2018) systematic review, risk was assessed based on both information provided by Virtanen et al. (2018) and information from previous publications on the cohorts and datasets analysed by Virtanen et al. (2018). This systematic review only included evidence on one eligible outcome: "Acquired depression" (depression incidence). The risk of bias rating for each domain for all 22 included studies for this outcome are presented in Fig. 3. The justification for each rating for each domain by included study is presented in risk of bias tables in Appendix 6 in the Supplementary data.

4.3.1. Selection bias (not representing source population)

We rated the risk of this selection bias as probably low for 19 studies and as probably high for three studies. We rated the risk of bias as probably low, because the studies comprised large populations of working-age individuals and all studies described their sample criteria extensively, enabling comparisons with the source population. Response rates differed considerably between the studies, however, a low response rate in a cohort study does not necessarily indicate a high risk of bias of the observed association between the exposure and the outcome; it first and foremost means that results cannot necessarily be generalized to the source population (Olsen, 2014; Rothman et al., 2013b), although this has been controversially debated (Ebrahim and Davey Smith, 2013; Rothman et al., 2013a). Under certain circumstances, non-response can introduce bias (Munafó et al., 2018), such as when selection out of the study is differential by exposure status; thus, we cannot rule out bias, but we judged the risk of bias as overall probably low. There are three exception, though: the studies by Berthelsen et al. (2015), Dembe and Yao (2016) and by Wang et al. (2012b). Berthelsen et al. (2015) deliberately did not exclude individuals with depression at baseline, but instead adjusted for baseline depression scores in the analyses. While this adjustment is likely to have reduced bias in the analyses, we are still concerned that individuals with prevalent depression may have been selected into specific working time arrangements. Dembe and Yao (2016) measured exposure as working hours per week, averaged over the whole 32-year observation period. Depression was not assessed at the beginning of the study, but at some point during the study. Thus, several measures of working hours were collected at a time when some individuals might have already had a depression. It is possible that these individuals had been selected into jobs with specific work time arrangements due to their depression, and that these individuals were at increased risk of being diagnosed with depression during follow-up. Consequently, we rated this study as being of probably high risk of bias. Wang et al. (2012b) examined risk of recurrent depression in a cohort of individuals who had a lifetime history of depression but were free of depression at the time of the baseline measurement. It is likely that frequency and severity of past depressive episodes differed in this cohort, and it is possible that individuals with more frequent or more severe previous depressive episodes were both selected into specific work time arrangements and at increased risk of being diagnosed with a recurrent depression during follow-up. Consequently, we rated this study as of probably high risk of bias.

4.3.2. Performance bias

We rated the risk of this bias as probably low for all 22 studies. In experimental studies, such as clinical trials, knowledge about assignment to either intervention or control group may influence behaviour or reporting of participants and can therefore severely bias results. This is of lesser concern in the observational studies included in this review, where participants were not assigned to a treatment or control group but first reported information about exposure and health conditions, then were categorized by researchers into different exposure groups and then were followed-up for assessing incidence of the outcome. We cannot rule out bias, as for example it is possible that individuals sensing that the study might be about working hours and health may over- or underreport numbers of working hours, but we regard the likelihood of this possible risk of bias to be low. None of the studies reported whether study personnel, such as statisticians conducting the analyses, were blinded to the exposure and/or outcome status of the participants. But even if they were unblinded, we doubt that this would have influenced the statistical analyses or reporting.

4.3.3. Detection bias (exposure assessment)

We rated the risk of this bias as probably low for 19 studies, and as probably high risk for three studies. Working hours were recorded in all studies by self-report by the workers; no study used "more objective" measures such as administrative records. We are not concerned about bias here, as we assume that individuals, who are free of depression, are able to estimate their weekly working hours reasonably accurately. This assumption is supported by results from a study by Imai et al. (2016). For three studies, though (Berthelsen et al. (2015), Dembe and Yao (2016) and Wang et al. (2012b)), we rated the risk of this bias as probably high, because it was possible that some participants in these studies either had a current depression (Berthelsen et al., 2015; Dembe and Yao, 2016) or an incomplete recovery from a previous depression (Wang et al., 2012b) when they reported their weekly working hours. This might have introduced bias, as there is evidence that individuals with depression tend to overestimate the adversity and negativity of their environment (Harmer et al., 2009). Berthelsen et al. (2015) deliberately did not exclude individuals with depression at baseline, but instead adjusted for baseline depression scores in the analyses. While this adjustment is likely to have reduced bias in the analyses, we are still concerned with self-reported working hour assessments of individuals with prevalent depression. Dembe and Yao (2016) included measures of working hours over 32 years, including several years before they assessed depression, and therefore did not know the depression status of their participants at the time of some of the exposure assessments (see comments in Section 4.3.1 on selection bias [not representing source population] above). Wang et al. (2012b) examined recurrence of depression among participants with a lifetime history of depression. Although participants with a current depressive episode at baseline were excluded from the analyses, it is possible that incomplete recovery from a previous depressive episode may have affected both the reporting of hours worked at baseline and the risk of depression during follow-up.

4.3.4. Detection bias (outcome assessment)

We rated the risk of this bias as low for four studies, probably low for 14 studies and probably high for four studies. The four studies rated as having low risk of bias (Shields, 1999; Virtanen et al., 2012; Wang et al., 2012a; Wang et al., 2012b) all used the gold standard method for assessing depression in epidemiological studies: a clinical diagnostic interview (Drill et al., 2015). The 14 studies rated as of probably low risk of bias either used a self-administered rating scale that had previously been validated against clinical measures of depression (14 studies) or conducted a comprehensive interview with the participants to identify diagnosis or treatment of depression that was then coded into ICD-9 codes (Kim, 2013). Of the four studies rated as probably high, two used a simple question to assess if the participants ever had a doctor-diagnosed depression (Dembe and Yao, 2016; NLSY OA Cohort, 2019), and we were concerned that several cases of depression were missed with this rather crude method to recall episodes of depression. The two other studies rated as of probably high risk of detection bias were the Virtanen et al., 2018 – HRS study and the Virtanen et al., 2018 – SLOSH study from the systematic review by Virtanen et al. (2018). For the Virtanen et al., 2018 – HRS study, Virtanen et al. (2018) stated that the outcome measure of this study was "depressive symptoms" without naming the instrument. From our knowledge of the study, we assume that the instrument used was the CES-D, but as we cannot be sure and, having indirect evidence only, we rated this study as of probably high risk of bias in this domain. For the Virtanen et al., 2018 – SLOSH study, Virtanen et al. (2018) stated that the outcome measure of this study was "depressive symptoms" measured with the "Symptom Check List, SCL". From our knowledge of this study, we assume that the instrument used was the SCL-CD6, but as we cannot be sure and, having indirect evidence only, we rated this study as having probably high risk of detection bias.

4.3.5. Confounding

We rated the risk of confounding as low for two studies, probably low for 19 studies and probably high for one study. We rated the studies by Ahn (2018) and Berthelsen et al. (2015) as carrying low risk of confounding, because it controlled for all first-tier confounders (i.e., age, sex and SES) and also for all second-tier confounders (i.e., job group, industry and country), either by adjusting, stratifying or study design (e.g., only including employees from a specific job group into the study), as we had prespecified in our protocol (Rugulies et al., 2019). The 19 studies rated as of probably low risk of confounding adjusted for all first-tier confounders but not all second-tier confounders. The one study rated as probably high risk in this domain (Shields, 1999) adjusted for all first-tier and some second-tier confounders, but also further adjusted for a measure of work stress (job strain), which we regard as over-adjustment. One of the components of job strain is high psychological demands (Karasek, 1979), which might conceptually overlap with exposure to long working hours. Further, it seems reasonable to assume that exposure to long working hours can cause the experience that work is highly psychologically demanding, which would make high psychological demands a potential mediator, a step in the causal pathway, for the association between working long hours and risk of depression. Adjusting for a mediator would introduce bias into the analyses; we therefore rated the study by Shields (1999) as having high risk of confounding.

However, if our assumption that exposure to long working hours would cause the work to be highly psychologically demanding was wrong, and instead highly psychologically demanding work would cause long working hours (e.g., because working long hours would be a way to handle high psychological demands at work) and, if further, high psychological demands at work would increase risk of depression, then high psychological demands would be a potential confounder for the association between exposure to long working hours and risk of depression. In this case, our decision to rate the study by Shields (1999) as of high risk of confounding would have been wrong. Further, in this case, our a priori decision to select, if possible, estimates that were not adjusted for high psychological demands or other psychosocial work environment factors, would also have been wrong, and the estimates we selected would have been at risk of being confounded. We were aware of this when we wrote the protocol of the systematic review and made the conscious decision to assume that other psychosocial work environment factors, including psychological demands, are mediators rather than confounders. We acknowledge that this assumption can be contested.

Other psychosocial work environment factors could not only be confounders or mediators but also be effect modifiers, i.e. it is possible that the interaction of exposure to long working hours with other psychosocial work environment factors would cause a stronger or a weaker effect on risk of depression. In the job strain model, it is assumed that high psychological demands in combination with high job control would not be health-hazardous but rather health-beneficial, because this combination would enhance learning and feelings of mastery (Theorell and Karasek, 1996). The combination of high psychological demands with low job control, on the other hand, would inhibit learning, cause strain and ultimately be health-hazardous (Theorell and Karasek, 1996). Empirically, it has indeed been shown that high psychological demands in themselves are not associated with risk of depression, whereas the combination of high psychological demands and low job control is associated with a higher risk of depression (Madsen et al., 2017; Theorell et al., 2015). Similarly, it is conceivable that long working hours in combination with low job control, but also in combination with other psychosocial work environment factors (e.g., low leadership quality or high role conflicts), could have a stronger effect on risk of depression than without these other factors. If this was the case, then our decision not to adjust for other psychosocial work environment factors would remain the correct decision, because adjusting for an effect modifier would give biased results. The appropriate approach would have been to test for interaction, however, this was beyond the scope of this review.

4.3.6. Selection bias (incomplete outcome data)

We rated the risk of this bias as probably low for four studies and probably high for 18 studies. Depression is often, although not always, episodic and self-limiting with a high recurrence rate (Kessler et al., 2003). This constitutes a challenge for epidemiological research. A cohort study measuring depression at baseline and some years later at follow-up, will not be able to identify those episodes of depression that occurred after baseline but were no longer present at the time of the follow-up measurement. This is of particular concern in those studies that measured depression at follow-up with a self-administered rating scale that asked about symptoms of depression during the last one or two weeks only. Consequently, we rated the 15 studies using a self-administered rating scale and the two studies using a simple question about doctor-diagnosed depression (Dembe and Yao, 2016; NLSY OA Cohort, 2019) as of probably high risk of selection bias due to incomplete outcome data. One study, Virtanen et al. (2012), used a clinical diagnostic interview covering depressive episodes during the last 12 months, and this study was also rated as of probably high risk of bias, because the follow-up period was very long (5.8 years). As carrying low risk of this selection bias, we rated Shields (1999), Wang et al. (2012a) and Wang et al. (2012b) that used a clinical diagnostic interview with a relatively short follow-up period of two years (Shields, 1999) and one year (Wang et al., 2012a; Wang et al., 2012b), respectively, and Kim (2013) because in this study doctor-diagnosed depression and treatment for depression was assessed every six months in a comprehensive interview.

4.3.7. Bias due to selective outcome reporting

We rated this bias as of low risk for all 22 studies. None of the included studies had a pre-specified study protocol, but all the studies' outcomes were reported in the results sections of the study record as outlined in the methods sections.

4.3.8. Conflict of interest

We rated all 22 studies as of low risk of this bias, as we did not find

any indication of a conflict of interest. More specifically, the studies:

- Did not receive support from a company or other entity with a financial interest in the study findings;
- Were funded by public research agencies or related organizations that were free from commercial interests in the study findings;
- Were authored only by persons who were not affiliated with companies or other entities with vested interests; and/or;
- Had no conflict of interest declared by study authors.

4.3.9. Other risk of bias

We identified one additional potential risk of bias that applied to 19 of the 22 studies, that is a probably high risk of bias due to lacking information on episodes of depression earlier in the life preceding the time of baseline assessment of depression. While we are confident that 19 of the 22 studies did not include individuals with a prevalent depression at baseline (see Section 4.3.1 on selection bias [not representing source population]), we are not confident that most studies were able to identify depression episodes that occurred during the lifetime before baseline assessment. This is a problem, because if there was a depressive episode before baseline assessment, the outcome of the study would change from "acquired depression for the first time in life" (first time incident depression) to "acquired depression again" (recurrent depression). Whether the association between a risk factor (e.g., exposure to long working hours) and depression is different for first time incidence of depression and recurrent depression is difficult to estimate. This may depend on various factors, such as the severity of the first depression; if the first depression was treated or not; the time span between the first and the recurrent depression; or whether there was only one depressive episode or multiple depression episodes before the measurement of the recurrent depression in the study (Burcusa and Iacono, 2007; Kessler et al., 2003; Moffitt et al., 2010). Thus, that studies were not able to measure lifetime depression prior to baseline assessment does not necessarily mean that estimates were biased, but we judged this as indirect evidence for potential bias and consequently rated the risk of bias as being probably high. As having probably low risk of bias, we rated the studies by Kim et al. (2013) that used comprehensive interviews to assess previous diagnosis and treatment for depression, and Wang et al. (2012a) and Wang et al. (2012b) that assessed lifetime prevalence of depression with a clinical diagnostic interview. Also for these three studies it is possible that previous depression episodes were missed, either because they were not diagnosed or treated (Kim, 2013) or because individuals had forgotten about earlier episodes of depression in the assessment of life-time prevalence (Wang et al., 2012a; Wang et al., 2012b); however, we assessed this risk as probably low. Two studies used a simple question to assess life-time depression before baseline (Dembe and Yao, 2016; NLSY OA Cohort, 2019). Although such a question is superior to not assessing life-time depression at all, we are still concerned that this is not sufficient for measuring lifetime depression; consequently we rated these two studies also as of probably high risk of this other bias.

4.4. Synthesis of results

This systematic review only included evidence on one eligible outcome: "Acquired depression" (depression incidence). We report findings from the three eligible comparisons included in this systematic review.

4.4.1. Comparison: worked 41–48 h/week compared with worked 35–40 h/week

A total of eight studies with 49,392 participants reported effect estimates on the risk of acquiring depression when working 41–48 h/week, compared with working 35–40 h/week. These studies were somewhat heterogeneous in their definition of the exposure category and/or the comparator. Six studies (Ahn, 2018; Kim, 2013; NLSY OA Cohort, 2019; Wang et al., 2012a; Wang et al., 2012b; Zadow et al., 2019) used the standard exposure categories, as defined in our protocol (Rugulies et al., 2019), and two studies (providing three estimates) used exposures categories that were non-standard but judged by us as similar enough for combining them in the meta-analysis (Shields, 1999; Virtanen et al., 2012). The non-standard exposure categories were >40 h/week compared with 35–40 h/week (Shields, 1999) and 45 h/week compared with 35–40 h/week (Virtanen et al., 2012), respectively. Despite these definitional differences in exposure, we judged these studies to be sufficiently homogenous to be combined in one meta-analysis. The eight studies reported nine eligible effect estimates, with one study providing two estimates: one each for women and men. Fig. 4 depicts the forest plot for this meta-analysis of these eight included studies. We retained all nine individual effect estimates in the forest plot (rather than combining the effect estimates for women and men from the same study). Compared with working 35–40 h/week, working 41–48 h/week had similar odds of acquiring a depression when followed up between 1 and 10 years (OR 1.05, 95% CI 0.86 to 1.29, eight studies with nine estimates, 49,392 participants, I^2 46%). When we excluded the one study on recurrent depression (Wang et al., 2012b), the pooled estimate remained virtually unchanged (OR 1.04, 95% CI 0.83 to 1.29, seven studies with eight estimates, 48,997 participants, I^2 52%).

When comparing the estimates from the six studies using the standard exposure categories to the estimates from the two studies using approximated exposure categories, we did not find a statistically significant difference ($p = 0.81$, see Appendix 7 in the Supplementary data).

Fourteen of the 22 studies from this review were not included in this meta-analysis. The ten studies from the Virtanen et al. (2018) review were not included, because they only reported estimates for exposure categories for ≥55 h/week. The study by Berthelsen et al. (2015) was not included, because it used as the comparator all working hours of ≤35 h/week and used as exposures weekly working hours of 35.5–37.5 (OR 1.38, 95% CI 0.92–2.06) and >37.5 (OR 1.04, 95% CI 0.62–1.75), respectively. The study by Kato et al. (2014) was not included, because it used as the comparator all working hours of ≤50 h/week and used as exposures working hours of 50.1–60 h/week (OR 1.14, 95% CI 0.70–1.86) and ≥60 h/week (OR 4.04, 95% CI 1.68–9.75). The study by Dembe and Yao (Dembe and Yao, 2016) was not included, because it was based on the same study population as the NLSY OA Cohort (2019) that was included. Dembe and Yao reported odds ratios of 0.96 (95% CI 0.80–1.15), 1.04 (95% CI 0.79–1.38) and 0.82 (95% CI 0.50–1.35) for 41–50 h/week, 51–60 h/week and >60 h/week, respectively, compared with 30–40 h/week. Kim et al (2016) was not included, because it was based on the same study population as Ahn (2018) that was included. Kim et al. (2016) reported odds ratios of 0.92 (95% CI 0.65–1.30), 1.09 (95% CI 0.82–1.46) and 1.92 (95% CI 1.30–2.85) for 41–52 h/week, 53–68 h/week and >68 h/week, respectively, compared with 35–40 h/week.

4.4.2. Comparison: worked 49–54 h/week compared with worked 35–40 h/week

A total of eight studies with 49,392 participants reported effect estimates on the risk of acquiring depression when working 49–54 h/week, compared with working 35–40 h/week. These studies were also somewhat heterogeneous in their definition of the exposure category and/or the comparator. Six studies (Ahn, 2018; Kim, 2013; NLSY OA Cohort, 2019; Wang et al., 2012a; Wang et al., 2012b; Zadow et al., 2019) used the standard exposure categories, which we had defined in the protocol (Rugulies et al., 2019). Two studies (providing three estimates) used exposures categories that were non-standard but judged (by us) as sufficiently similar to combine them in the meta-analysis (Shields, 1999; Virtanen et al., 2012). The non-standard exposure categories were >40 h/week compared with 35–40 h/week (Shields, 1999) and 50 h/week compared with 35–40 h/week (Virtanen et al., 2012). Despite these definitional differences in exposure, we judged these studies to be

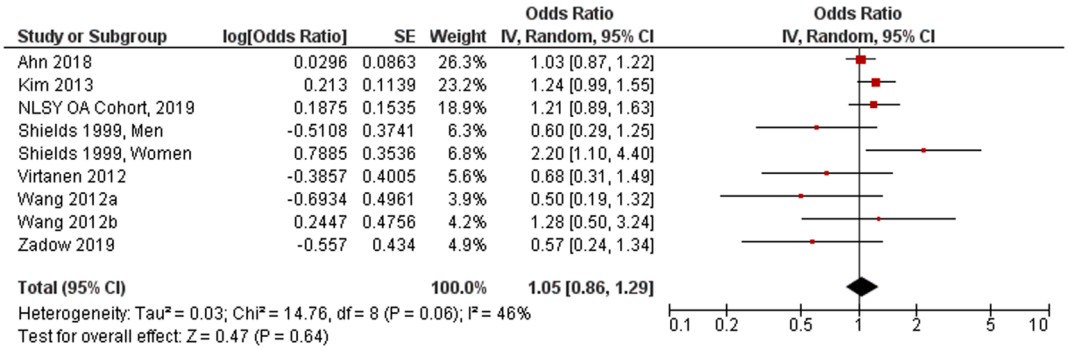

Fig. 4. Main meta-analysis, Acquired depression, worked 41–48 h/week compared with worked 35–40 h/week.

sufficiently homogenous to be combined in one meta-analysis. The eight studies reported nine eligible effect estimates, with one study providing two estimates, namely one each for women and men. Fig. 5 depicts the forest plot for this meta-analysis of the eight included studies. We retained all nine individual effect estimates in the forest plot. Compared with working 35–40 h/week, working 49–54 h/week had similar odds of acquiring a depression when followed up between 1 and 10 years (OR 1.06, 95% CI 0.93 to 1.21, eight studies with nine estimates, 49,392 participants, I^2 40%). When we excluded the one study on recurrent depression (Wang et al., 2012b), the pooled estimate also remained virtually unchanged (OR 1.05, 95% CI 0.92 to 1.19, seven studies with eight estimates, 48,997 participants, I^2 32%).

When comparing the estimates from the six studies using the standard exposure categories with the estimates from the two studies using approximated exposure categories, we did not find a statistically significant difference (p = 0.54; see Appendix 7).

Fourteen of the 22 studies included in this systematic review were not included in this meta-analysis. The ten studies from the review by Virtanen et al. (2018) were not included, because they only reported estimates for exposure categories for ≥55 h/week. Berthelsen et al. (2015) was not included, because this study used as the comparator all working hours of ≤35 h/week and as exposure categories working hours of 35.5–37.5 per week (OR 1.38, 95% CI 0.92–2.06) and >37.5 per week (OR 1.04, 95% CI 0.62–1.75), respectively. Kato et al. (2014) was not included because it used as the comparator all working hours of ≤50 h/week and as exposures working hours of 50.1–60 h/week (OR 1.14, 95% CI 0.70–1.86) and ≥60 h/week (OR 4.04, 95% CI 1.68–9.75). The Dembe and Yao study (Dembe and Yao, 2016) was not included, since it was based on the same study population as the NLSY OA Cohort (2019) that was included. Dembe and Yao reported odds ratios of 0.96 (95% CI 0.80–1.15), 1.04 (95% CI 0.79–1.38) and 0.82 (95% CI 0.50–1.35) for 41–50 h/week, 51–60 h/week and >60 h/week, respectively, when compared with 30–40 h/week. Kim et al. (2016) was not included, because it was based on the same study population as Ahn (2018) that was included. Kim et al. (2016) reported odds ratios of 0.92 (95% CI 0.65–1.30), 1.09 (95% CI 0.82–1.46) and 1.92 (95% CI 1.30–2.85) for 41–52 h/week, 53–68 h/week and >68 h/week, respectively, compared with 35–40 h week.

4.4.3. Comparison: Worked ≥55 h/week compared with worked 35–40 h/week

A total of 17 studies with 91,142 participants reported effect estimates on the risk of acquiring depression when working ≥55 h/week, compared with working 35–40 h/week. These studies were also somewhat heterogeneous in their definition of the exposure category and/or the comparator. Fifteen studies (Ahn, 2018; Kim, 2013; NLSY OA Cohort, 2019; Virtanen et al., 2018 – ACL; Virtanen et al., 2018 – DWECS; Virtanen et al., 2018 – ELSA; Virtanen et al., 2018 – HESSUP; Virtanen et al., 2018 – HILDA; Virtanen et al., 2018 – HRS; Virtanen et al., 2018 – PUMA; Virtanen et al., 2018 – SHARE; Virtanen et al., 2018 – SLOSH; Wang et al., 2012a; Wang et al., 2012b; Zadow et al., 2019) used the standard exposure categories, defined in the protocol (Rugulies et al., 2019). Two studies (with three eligible individual effect estimates) used exposures categories that were non-standard but judged by us as similar enough for combining them in one meta-analysis (Shields, 1999; Virtanen et al., 2012). The non-standard exposure categories were >40 h/week compared with 35–40 h/week (Shields, 1999) and, respectively, 55–60 h/week compared with 35–40 h/week (Virtanen et al., 2012). Despite these definitional differences in exposure, we judged these studies to be sufficiently homogenous to all be combined in the one meta-analysis. The 17 studies reported 18 eligible effect estimates, with one study providing two estimates: one each for women and men. Fig. 6 depicts the forest plot for this meta-analysis of the 17 included studies. We retained all 18 individual effect estimates in the forest plot (rather than combining the effect estimates for women and men from the same study). Compared with working 35–40 h/week, working ≥55 h/week had similar odds of acquiring a depression when followed up between 1 and 10 years (OR 1.08, 95% CI 0.94 to 1.24, 17 studies with 18 estimates, 91,142 participants, I^2 46%). When we excluded the one study on recurrent depression (Wang et al., 2012b), the pooled effect estimate

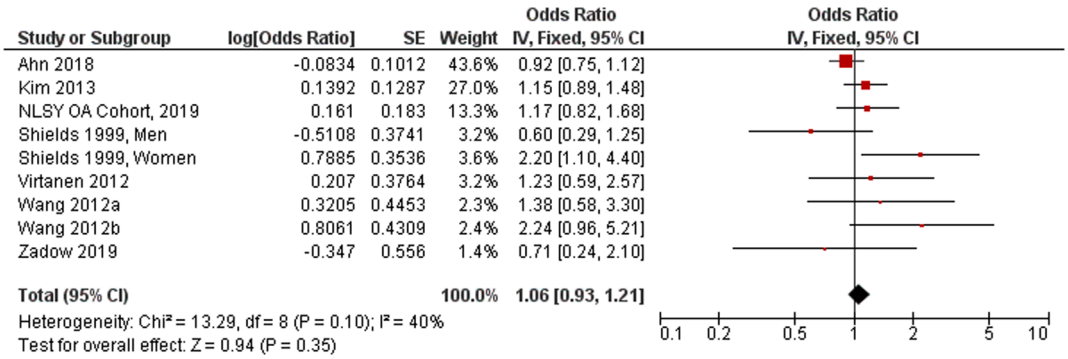

Fig. 5. Main meta-analysis, Acquired depression, worked 49–54 h/week compared with worked 35–40 h/week.

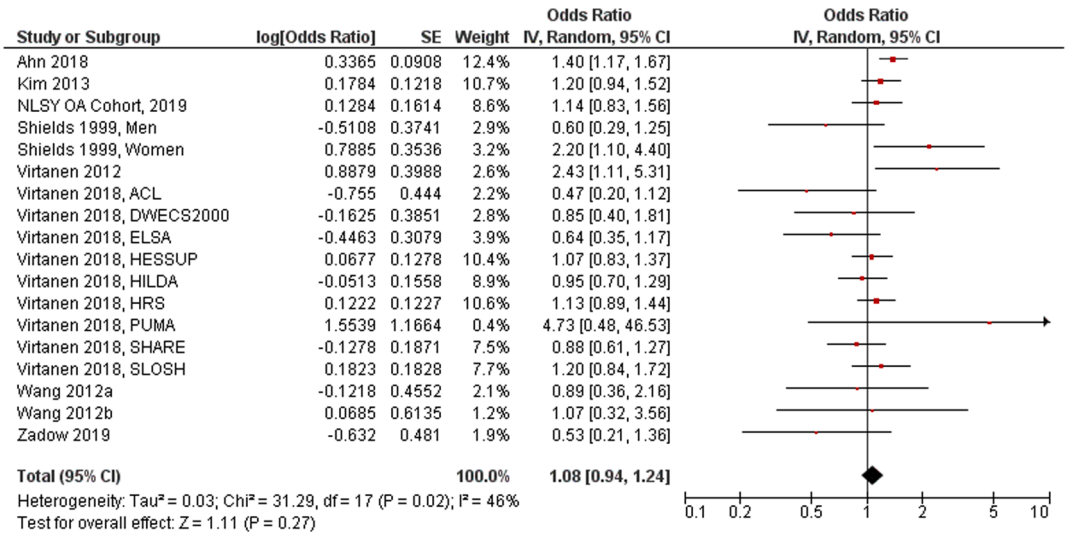

Fig. 6. Main meta-analysis, Acquired depression, worked ≥ 55 h/week compared with worked 35–40 h/week.

remained virtually unchanged (OR 1.08, 95% CI 0.94 to 1.25, 16 studies with 17 estimates, 90,747 participants, I^2 49%).

When comparing the estimates from the 15 studies using the standard exposure categories to the estimates from the two studies using approximated exposure categories, we did not find a statistically significant difference (p = 0.48, see Appendix 7 in the Supplementary data).

Five of the 22 studies included in this systematic review were not included in this meta-analysis. Berthelsen et al. (2015) was not included, because it used as the comparator all working hours of ≤35 h/week, and because it used as exposures weekly working hours of 35.5–37.5 (OR 1.38, 95% CI 0.92–2.06) and >37.5 (OR 1.04, 95% CI 0.62–1.75). Kato et al. (2014) was not included, because it used the comparator of all working hours of ≤50 h/week and the exposures of weekly working hours of 50.1–60 (OR 1.14, 95% CI 0.70–1.86) and ≥60 (OR 4.04, 95% CI 1.68–9.75). Dembe and Yao (Dembe and Yao, 2016) and Virtanen et al., 2018 – NLSY were excluded from the meta-analysis, because they were both based on the same study population as NLSY OA Cohort (2019) that we did include in the meta-analysis. Dembe and Yao reported odds ratios of 0.96 (95% CI 0.80–1.15), 1.04 (95% CI 0.79–1.38) and 0.82 (95% CI 0.50–1.35) for 41–50 h/week, 51–60 h/week and >60 h/week, respectively, when compared with 30–40 h/week. Virtanen et al., 2018 – NLSY reported an odds ratio of 1.06 (95% CI 0.77–1.44) for the comparison of ≥55 h/week versus 35–40 h/week. Kim et al. (2016) was also excluded, since it was based on the same study population as Ahn (2018) that we did include. Kim et al. (2016) reported odds ratios of 0.92 (95% CI 0.65–1.30), 1.09 (95% CI 0.82–1.46) and 1.92 (95% CI 1.30–2.85) for 41–52 h/week, 53–68 h/week and >68 h/week, respectively, compared with 35–40 h week.

Table 6
Summary of results from subgroup analyses on exposure to long working hours (≥55 h/week) and acquiring depression.

Subgroup	Pooled odds ratio (95% confidence interval)
WHO region (n = 17 studies)	
Americas	1.09 (0.89 to1.34)
Europe	1.06 (0.82 to 1.36)
Western Pacific	1.05 (0.70 to 1.58)
p for subgroup differences:	0.97
Sex (n = 6 studies)	
Women	1.15 (0.76 to 1.75)
Men	1.21 (1.01 to 1.44)
p for subgroup differences:	0.85
Age (n = 5 studies)	
15–19 years	Not estimable
20–24 years	1.49 (0.49 to 4.58)
25–29 years	0.84 (0.51 to 1.41)
30–34 years	1.06 (0.67 to 1.65)
35–39 years	1.56 (0.87 to 2.81)
40–44 years	1.15 (0.92 to 1.45)
45–49 years	1.09 (0.48 to 2.46)
50–54 years	1.25 (0.62 to 2.49)
55–59 years	1.41 (0.89 to 2.22)
60–64 years	2.44 (1.41 to 4.22)
≥65 years	1.83 (1.02 to 3.29)
p for subgroup differences	0.31
Socioeconomic status (n = 5 studies)	
Low	1.34 (1.10 to 1.63)
Intermediate	1.14 (0.85 to 1.53)
High	1.27 (0.96 to 1.69)
p for subgroup differences	0.67

4.5. Additional analyses

4.5.1. Subgroup analyses

We conducted subgroup analyses for the comparison between the group that worked ≥55 h/week and the group that worked 35–40 h/week. These analyses include subgrouping by WHO region, sex, age and SES (Table 6). For the subgroup analysis by WHO region, data were available for all studies from the main analysis (n = 17), whereas the subgroup analyses could only be performed for selected studies by sex (n = 6) (Ahn, 2018; Kim, 2013; NLSY OA Cohort, 2019; Shields, 1999; Wang et al., 2012a; Zadow et al., 2019), age (n = 5) (Ahn, 2018; Kim, 2013; Wang et al., 2012a; Zadow et al., 2019) and SES (n = 5) (Ahn, 2018; Kim, 2013; NLSY OA Cohort, 2019; Wang et al., 2012a; Zadow et al., 2019). We could not perform subgroup analyses by occupation, industrial sector or formality of the economy, because we did not identify any studies that provided estimates disaggregated by these subgroups.

None of the subgroup analyses showed a statistically significant difference between the subgroups (all p values ≥0.31). Thus, pooled effect estimates did not differ statistically significantly between WHO regions, women and men, different age groups or different SES groups.

Some of the subgroup analyses yielded estimates that appeared to be considerably higher than the estimate from the main analysis. However, such a comparison has to be made with caution for those subgroup analyses that were conducted with selected studies (i.e., sex, age and SES). For example, the subgroup analysis stratified by sex that was based on six studies yielded OR estimates that were higher for both women (1.15, 95% CI: 0.76–1.75) and men (1.21, 95% CI: 1.01–1.44) than the pooled OR estimate from the main analysis (1.08, 95% CI: 0.94–1.24) that was

based on all 17 studies. The forest plots of all subgroup analyses are presented in Appendix 8 in the Supplementary data.

4.5.2. Sensitivity analyses

We conducted two pre-defined sensitivity analyses (Table 7). One sensitivity analysis compared estimates for studies that had assessed depression with the gold standard method (a clinical diagnostic interview; n = 4) versus studies that used other assessment methods (self-administered rating scales or self-reported doctor diagnosed depression; n = 13). The other sensitivity analysis compared estimates for studies with "low"/"probably low" risk of bias in all RoB domains (n = 2) versus studies with at least one rating of "high"/"probably high" in any RoB domain (n = 15).

The two sensitivity analyses did not show a statistically significant effect. That is, estimates were not statistically significantly different for studies that assessed depression with the gold standard method (clinical diagnostic interview), compared with other methods (p = 0.56), nor for studies with only "low"/"probably low" risk of bias ratings compared to studies with one or more "high"/"probably high" risk of bias ratings (p = 0.52). The forest plots of the sensitivity analyses are presented in Appendix 9 in the Supplementary data.

4.6. Quality of evidence

We now report assessments of the quality of evidence for the entire body of evidence on the outcome "Acquired depression" (depression incidence) for the three eligible comparisons included in this systematic review.

4.6.1. Comparison: Worked 41–48 h/week compared with worked 35–40 h/week

We had serious concerns regarding risk of bias in the body of evidence for this comparison. First, we were concerned about risk of selection bias due to incomplete outcome data (domain 6 in our risk of bias assessment), as the majority of studies assessed depression at baseline and then again some years later at follow-up, which means that these studies likely missed cases of depression with onset between baseline and follow-up that were in remission at the time of the follow-up assessment. Second, most studies did not assess whether study participants had ever experienced an episode of depression before assessment at baseline. For these studies, it is unclear whether incident depression during the study period was first time incidence of depression or recurrent incidence of depression. We are uncertain if the incomplete outcome data during follow-up or the failure to assess life-time depression before baseline have biased the estimate of the association between baseline working hours and risk of depression in the studies. But since we cannot rule out risk of bias with confidence, we had serious concerns for risk of bias and consequently downgraded the quality of evidence by one level (−1).

We had no serious concerns for inconsistency, as the I^2 of 46% indicated only moderate heterogeneity (+/−0 levels). We also did not have any serious concerns for indirectness, because the body of evidence reasonably well matched the populations, exposures, comparators and outcomes of our interest; we therefore did not downgrade the quality of evidence for this consideration (+/− 0 levels). We had serious concerns for imprecision, because if the lower confidence limit for our best OR estimate (from the meta-analysis) represented the truth (0.86) there would have been an appreciable beneficial effect of exposure to long working hours. If the upper confidence limit represented the truth (OR 1.29), there would have been an appreciable harmful effect of exposure to long working hours. Consequently, we downgraded for imprecision by one level (−1). We did not have any serious concerns for publication bias (+/− 0 levels; funnel plot not calculated because there were fewer than ten studies). There was no large effect estimate, no evidence for a dose–response and no evidence suggesting that residual confounding, bias or effect modification had led to an underestimation of the effect, and consequently we did not upgrade the body of evidence (+/−0 levels). In conclusion, we started at "moderate" quality of evidence for observational studies and downgraded by a total of two levels (−2) and did not upgrade (+/−0) to arrive at a final rating of the quality of evidence as "low"; further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.

Table 7
Summary of results from pre-defined sensitivity analyses on long working hours and (≥55 h) and acquiring depression.

Sensitivity analysis	Pooled odds ratio (95% confidence interval)
Depression measurement (n = 17 studies)	
Clinical diagnostic interview	1.28 (0.71 to 2.29)
Other assessment methods	1.07 (0.94 to 1.22)
p for subgroup differences:	0.56
Risk of bias (n = 17 studies)	
Only "low"/"probably low"	1.17 (0.93 to 1.48)
Any "high"/"probably high"	1.07 (0.91 to 1.25)
p for subgroup differences	0.52

4.7. Comparison: Worked 49–54 h/week compared with worked 35–40 h/week

As for weekly working hours of 41–48 (described above), we also had serious concerns regarding risk of bias in the body of evidence on the comparison of weekly working hours 49–54 versus 35–40 for incident depression, because of possible incomplete outcome reporting (domain 6 of the risk of bias assessment) and failure to measure episodes of depression prior to baseline. Consequently, we downgraded the quality of evidence by one level (−1) for risk of bias.

We had no serious concerns for inconsistency, as the I^2 of 40% indicated only moderate heterogeneity (+/−0 levels). We did not have any serious concerns for indirectness because the body of evidence reasonably well matched the populations, exposures, comparators and outcomes of our interest, and we therefore did not downgrade the quality of evidence for this consideration (+/− 0 levels). We however had serious concerns for imprecision, because if the upper confidence limit for our best estimate (from the meta-analysis) represented the truth (OR 1.21), there would have been an appreciable harmful effect of exposure to long working hours, and we therefore downgraded by one level (−1). We did not have serious concerns for publication bias (+/− 0 levels, funnel plot not calculated because there were fewer than ten studies). There was neither a large effect estimate, nor evidence for a dose–response, nor evidence suggesting that residual confounding, bias or effect modification had led to an underestimation of the effect, and consequently we did not upgrade the body of evidence (+/−0 levels). In conclusion, we started at "moderate" quality of evidence for observational studies; downgraded by a total of two levels (−2); did not upgrade (+/−0); and therefore arrived at a final rating of the quality of evidence of: "low".

4.7.1. Comparison: Worked ≥55 h/week compared with worked 35–40 h/week

As for the previous two comparisons (described above), we also had serious concerns regarding risk of bias in the body of evidence on the comparison of weekly working hours of ≥55 compared with 35–40 for incident depression, because of possible incomplete outcome reporting (domain 6 of the risk of bias assessment) and failure to measure episodes of depression prior to baseline. Consequently, we downgraded the quality of evidence by one level (−1) for risk of bias.

We had no serious concerns for inconsistency, as the I^2 of 46% indicated only moderate heterogeneity (+/−0 levels). We did not have any serious concerns for indirectness, because the body of evidence reasonably well matched the populations, exposures, comparators and outcomes of our interest, and we therefore did not downgrade the

quality of evidence for this consideration (+/−0 levels). We had serious concerns for imprecision, because if the upper confidence limit of our best OR estimate (from the meta-analysis) represented the truth (1.24), there would have been an appreciable harmful effect of exposure to long working hours, and we therefore downgraded by one level (−1). We did not have any serious concerns for publication bias, as our funnel plot (Fig. 7) looked reasonably symmetrical and therefore provided no evidence for the presence of publication bias (+/−0 levels). There was no large effect estimate, no evidence for a dose–response and no evidence that residual confounding, biases or effect modification had led to an underestimation of the association; we did not upgrade the quality of evidence (+/−0 levels). In conclusion, we started at "moderate" quality of evidence for observational studies and downgraded by a total of two levels (−2) and did not upgrade (+/−0) to arrive at a final rating of "low" quality of evidence for this comparison.

4.8. Assessment of strength of evidence

According to our protocol, we rated the strength of evidence based on a combination of the four criteria outlined in the Navigation Guide: (1) Quality of the entire body of evidence; (2) Direction of the effect estimate; (3) Confidence in the effect estimate; and (4) Other compelling attributes.

4.8.1. Quality of the entire body of evidence

This systematic review found no studies with evidence on two outcomes: "Has depression" (depression prevalence) and "Died from depression" (depression mortality). For the only outcome with studies included in the systematic review, "Acquired depression" (depression incidence), we judged the quality of the bodies of evidence for all three comparisons to be "low" (see Section 4.6 *Assessment of quality of evidence*). This is the lowest quality of evidence rating within the Navigation Guide framework. Consequently, we consider the quality of evidence for all three included outcomes as insufficient for assessing the strength of evidence for all three outcomes.

4.8.2. Direction of the effect estimate

The bodies of evidence for all three outcomes in all three included comparisons are insufficient to assess the direction of the effects on the outcome of exposure to long working hours. For the only outcome with any included evidence, "Acquired depression", for all three included comparisons (weekly working hours of 41–48, 49–54 and ≥55, compared with those of 35–40), the CIs of the pooled effect estimates included unity, with the lower limit and the higher limit suggesting moderate to minor decreases and increases, respectively in the odds of the outcome (41–48 h/week: 0.86 to 1.29; 49–54 h/week: 0.93–1.21; ≥55 h/week: 0.94 to 1.24). Thus, it was not possible to conclude whether exposure to long working hours decreased, increased or had no effect on risk of acquiring depression.

4.8.3. Confidence in the effect estimate

Considering that there is no evidence on two included outcomes, "Has depression" and "Died from depression", there are no effect estimates for which we could judge our confidence in them. For the outcome with any included evidence, "Acquired depression", the quality of the body of evidence for all three included comparisons was "low", and the directions of the pooled effect estimates were also unclear for all three included comparisons, as delineated above. Consequently, we have low confidence in the effect estimates for all included comparisons for this outcome.

In addition to this, our confidence in these effect estimates is also low for the following reasons:

- First, no additional data are available on causal pathways explaining the mechanisms that may link exposure to long working hours to depression.
- Second, the assumption of a dose–response relationship between the three exposure categories and the outcome was not supported by our findings.
- Third, even if there was an effect of long working hours on depression, the strength of such a harmful effect would be modest, as even the upper limits of the CI of the pooled effect estimates did not exceed an OR of the size of 1.29. Although even a modest increase in risk can be relevant for policy under conditions of high prevalence of the exposure (which is the case with long working hours, as per (Pega et al., 2021b)), this low level does not increase confidence in the effect estimate.
- Fourth, no intervention studies are available that demonstrate a reduction of the effect estimate because of a reduction of the exposure to a minimal level.

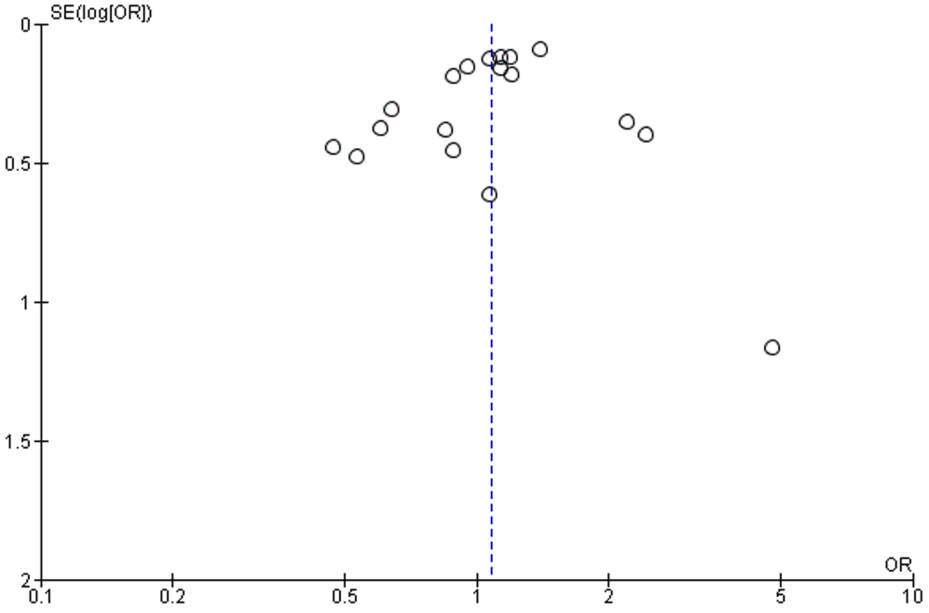

Fig. 7. Funnel plot for acquired depression, worked ≥55 h/week compared with worked 35–40 h/week.

4.8.4. Other compelling attributes

We were not able to access data that could offer evidence for a discussion of other compelling attributes in assessing the strength of evidence. Additional assessment of strength of evidence based on the Bradford Hill considerations (Bradford Hill 1965) is presented in Appendix 9 in the Supplementary data (though note that this is already covered via our approach to evaluating the quality of evidence as described above).

4.8.5. Rating by outcome and comparison

Based on the considerations presented above, we judged the existing bodies of evidence as:

- "Inadequate evidence of harmfulness" for the exposure categories 41–48, 49–54 and ≥ 55h/week for "Has depression" (depression prevalence); the available evidence is insufficient to assess effects of the exposure.
- "Inadequate evidence of harmfulness" for the exposure categories 41–48, 49–54 and ≥55 h/week for "Acquired depression" (depression incidence).
- "Inadequate evidence of harmfulness" for the exposure categories 41–48, 49–54 and ≥55 h/week for the outcome "Died from depression" (depression mortality).

5. Discussion

5.1. Summary of evidence

As shown in the table of summary of findings (Table 8), our systematic review found no eligible study on the outcomes of depression prevalence and depression mortality, resulting in ratings of "low quality of evidence" and "inadequate evidence of harmfulness" for all three exposure categories for these outcomes. Our systematic review found 22 eligible studies for depression incidence, of which 17 studies were included in one or more meta-analyses. We found low quality of evidence and inadequate evidence of harmfulness for the effect of working 41–48, 49–54 and ≥ 55 h/weeks on the outcome of depression incidence, when compared with working 35–40 h/week.

5.2. Comparison to previous systematic review evidence

To our knowledge, three systematic reviews have in recent years examined the association between exposure to long working hours and risk of depression (Theorell et al., 2015; Virtanen et al., 2018; Watanabe et al., 2016). The Theorell et al. (2015) systematic review included six cohort studies and did not conduct a meta-analysis, and the authors concluded that the evidence for any increase of depression from exposure to long working hours was "limited" for women and "very limited" for men. The Watanabe et al. (2016) systematic review included seven cohort studies and reported a modestly elevated pooled RR that included unity (RR 1.24, 95% CI 0.88 to 1.75) for the association between working ≥50 h/week and depression based on a sensitivity analyses of four cohort studies. The Virtanen et al. (2018) systematic review and meta-analysis included 28 studies (ten published and 18 unpublished ones) and reported a RR of 1.14 (95% CI 1.03 to 1.25) for the outcome "depressive symptoms" (combining clinical depression, depressive symptoms and psychological distress) and a RR of 1.09 (95% CI 0.94 to 1.26) for the outcome "depression" (combining clinical depression and depressive symptoms but excluding psychological distress). Our systematic review and meta-analysis covered the three outcomes "Has depression", "Acquired depression" and "Died from depression", identifying 22 studies for the outcome "Acquired depression", while identifying no studies for the outcomes "Has depression" and "Died from depression". Our review included three comparisons (41–48 vs 35–40 h/week, 49–54 vs 35–40 h/week and ≥55 h/week, compared with 35–40 h/week), and it found that the body of evidence for all three outcomes and all three comparisons was inadequate to assess harmfulness.

Compared to the previous systematic reviews, our systematic review and meta-analysis had some added value. First, we published a protocol detailing all methods of the systematic review and meta-analysis before commencing literature search (Rugulies et al., 2019). Second, by searching seven academic databases and two grey literature databases, our search was more comprehensive than the searches in the previous reviews. Third, we searched the literature up to July 2018 with a supplementary top-up search of Medline shortly before finalizing the article (November 2019) resulting in the most up-to-date review. We considered all studies included in the previous reviews and identified two additional studies (Ahn 2018; Zadow et al., 2019) not included in the previous reviews. Fourth, we were able to re-analyse six studies (Ahn 2018; Kim 2013; NLSY OA Cohort, 2019; Wang et al., 2012a; Wang et al., 2012b; Zadow et al., 2019), allowing us to harmonize exposure data and statistical modelling. Fifth, we applied state-of-the-art methods for assessing quality of evidence and strengths of evidence, using both GRADE (Morgan et al., 2016; Schünemann et al., 2011a) and Navigation Guide methodology (Woodruff and Sutton, 2014).

5.3. Limitations and strengths of this review

5.3.1. Limitations

Our systematic review has several limitations. First, we may have missed eligible studies, for example due to them being published in languages other than English. However, as we searched seven academic and two grey literature databases using a comprehensive search strategy, conducted an extensive hand search and consulted additional experts who also did not identify any additional eligible studies, we believe it is unlikely that we missed any important study.

Second, we did not receive a substantial amount of the missing data we requested for the studies included in this systematic review. We requested missing data from principal study authors at least three times, but the principal study authors generally did not share these requested missing data with us or only shared selected data. As a result, subgroup analyses by sex, age group and SES could only be conducted for a limited number of studies.

Third, all identified eligible studies were observational studies and therefore vulnerable to bias due to unmeasured confounding. Experimental studies, quasi-experimental studies or natural experiments could have provided stronger evidence, but we did not find any of these study types that fulfilled our inclusion criteria. We found one remarkable uncontrolled before-and-after study with first-year medical residents in Japan that measured depressive symptoms and clinical depression before and after entering residency (Ogawa et al., 2018). Entering residency led to exposure to very long working hours, with 45% of residents reporting working ≥ 80 h/week. The study showed a marked increase of depressive symptoms in the study participants after 3 months in residency, particularly amongst residents working 80–99 h/week and ≥100 h/week, suggesting that very long working hours, exceeding the predefined working hours categories in our review, might substantially increase risk of depression. However, as the study was conducted without a control group, it did not fulfil our predefined eligibility criteria, and we had to exclude it from our systematic review.

Fourth, the validity of exposure assessment was somehow restricted, not only due to concerns of self-reported exposure measurements, but also since exposure to long working hours was assessed at baseline only. This has likely resulted in exposure misclassification, as average hours worked per week may change over time. Further, the lack of repeated measures of exposure made it impossible to analyse the potential effect of changes in exposure.

Fifth, we provided subgroup analyses stratified by WHO region, but national health, labour and other social policies, as well as welfare state regimes, may vary considerably within these regions. It is conceivable that these variations modify association between working conditions, including exposure to long working hours, and risk of depression

Table 8
Summary of findings.

Effect of exposure to long working hours on depression among workers

Population: all workers of working age (≥15 years)
Settings: all countries and work settings
Exposure: worked 41–48, 49–54 or ≥55 h/week
Comparator: worked 35–40 h/ week

Outcomes	Exposure category	Illustrative comparative risks (95% CI)		Relative effect (95% CI)	No. of participants (studies)	Navigation Guide (Woodruff and Sutton 2014) quality of evidence rating	Navigation Guide strength of evidence rating for human evidence	Comments
		Assumed risk Unexposed workers (worked 35–40 h/week)	**Corresponding risk** Workers in the exposure category					
Has depression	–	–	–	–	–	–	Inadequate evidence of harmfulness	No evidence was found on this outcome.
Acquired depression (measured with clinical diagnostic interview, validated self-administered rating scales or self-reported doctor-diagnosed or treated depression) Follow-up: 1–10 years	**Worked 41–48 h/ week**	**349 cases per 10,000 person-years**	**366 cases per 10,000 person-years (300 to 450)**	**OR 1.05** (0.86 to 1.29)	49,392 (8 studies)	⊕⊖⊖ Low [b]	Inadequate evidence of harmfulness	Better indicated by lower values. Additional evidence from four studies not included in the meta-analysis also provided no evidence for an effect of this comparison on the outcome. We are very uncertain about the effect of this exposure category on this outcome.
	Worked 49–54 h/ week		**370 cases per 10,000 person-years (325 to 422)**	**OR 1.06** (0.93 to 1.21)	49,392 (8 studies)	⊕⊖⊖ Low [b]	Inadequate evidence of harmfulness	Better indicated by lower values. Additional evidence from four studies not included in the meta-analysis also provided no evidence for an effect of this comparison on the outcome. We are very uncertain about the effect of this exposure category on this outcome.
	Worked ≥55 h/ week		**377 cases per 10,000 person-years (328 to 433)**	**OR 1.08** (0.94 to 1.24)	91,142 (17 studies)	⊕⊖⊖ Low [b]	Inadequate evidence of harmfulness	Better indicated by lower values. Additional evidence from five studies not included in the meta-analysis also provided no evidence for an effect of this comparison on the outcome. We are very uncertain about the effect of this exposure category on this outcome.
Died from depression	–	–	–	–	–	–	Inadequate evidence of harmfulness	No evidence was found on this outcome.

CI: confidence interval; **OR**: odds ratio.

Navigation Guide quality of evidence ratings:

High quality: Further research is very unlikely to change our confidence in the estimate of effect.
Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Low quality: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.

Navigation Guide strength of evidence ratings

Sufficient evidence of harmfulness: The available evidence usually includes consistent results from well-designed, well-conducted studies, and the conclusion is unlikely to be strongly affected by the results of future studies. For human evidence a positive relationship is observed between exposure and outcome where chance, bias, and confounding, can be ruled out with reasonable confidence.
Limited evidence of harmfulness: The available evidence is sufficient to determine the effects of the exposure, but confidence in the estimate is constrained by such factors as: the number, size, or quality of individual studies, the confidence in the effect, or inconsistency of findings across individual studies. As more information becomes available, the observed effect could change, and this change may be large enough to alter the conclusion. For human evidence a positive relationship is observed between exposure and outcome where chance, bias, and confounding cannot be ruled out with reasonable confidence.
Inadequate evidence of harmfulness: Studies permit no conclusion about a toxic effect. The available evidence is insufficient to assess effects of the exposure. Evidence is insufficient because of: the limited number or size of studies, low quality of individual studies, or inconsistency of findings across individual studies. More information may allow an estimation of effects.
Evidence of lack of harmfulness: The available evidence includes consistent results from well-designed, well-conducted studies, and the conclusion is unlikely to be strongly affected by the results of future studies. For human evidence more than one study showed no effect on the outcome of interest at the full range of exposure levels that humans are known to encounter, where bias and confounding can be ruled out with reasonable confidence. The conclusion is limited to the age at exposure and/or other conditions and levels of exposure studied.

Footnotes:

[a] As assumed risk we extracted the risk in the study by Wang et al. (2012a) because this study was rated with "low"/"probably low" risk of bias in all domains and further used the gold standard measure, a clinical diagnostic interview, to measure incident depression

[b] Downgraded by two levels (−2) in total, comprising downgrading by one level each for serious concerns for risk of bias (−1) and imprecision (−1).

(Dragano et al., 2011; Lunau et al., 2013), yet our data did not allow us to examine such a possible effect modification.

Sixth, the subgroup analyses by sex, age and SES could not been performed with all studies, but only with a subsample of those studies that provided estimates stratified by these variables. We cannot rule out that subgroup analyses would have provided different results, had we been able to conduct them with all included studies.

Seventh, as in all systematic reviews, we included studies that were conducted in the past. Work environment conditions, including working time arrangements, from the time period of the studies that we included in the review might be different from current or future work environment conditions, and changes might sometimes occur rapidly and to a dramatic extent, for example in the global financial crisis that emerged in 2007 (Torá et al., 2015) or in the current COVID-19 pandemic (Burdorf et al., 2020; Sim, 2020). Whether this would affect the applicability of the results of this systematic review needs to be investigated in future studies.

5.3.2. Strengths

Our systematic review and meta-analysis have several strengths, including:

- Strictly speaking, previous systematic reviews have not undergone all steps of a systematic review (see Fig. 1 in (Woodruff and Sutton, 2014)), but our systematic review and meta-analysis have done so, including having pre-published a protocol (Rugulies et al., 2019) and having assessed the strength of evidence; this presents a substantial improvement in systematic review methods on the topic.
- Previous systematic reviews have not commonly and not comprehensively provided detailed analyses across all analytic steps of the systematic review and meta-analysis for comparisons of standard categories of exposure to long working hours compared with standard working hours, but we have provided such analyses for three such comparisons commonly used in the epidemiological literature across all steps of the systematic review, and again this provides an improvement in accuracy of systematic review evidence on this topic.
- Whereas previous systematic review evidence has not commonly and comprehensively assessed risk of bias and quality of evidence using established systematic review frameworks with dedicated tools and approaches, we have applied the Navigation Guide framework in this systematic review, which should have ensured rigor and transparency in this systematic review.
- In previous systematic reviews, strength of the evidence was not commonly assessed, but in our systematic review we have applied pre-specified criteria to rate the strength of evidence for each included comparison for each included outcome; again, this is a novel contribution to the systematic review and meta-analytic body of evidence on the topic.
- Finally, to our knowledge, this is amongst the first systematic reviews and meta-analyses conducted specifically for a global occupational burden of disease study, and as such it provides a model for future systematic reviews that will help ensure that these global health estimates adhere fully with the *GATHER Guidelines for Accurate and Transparent Health Estimates Reporting* (Stevens et al., 2016).

6. Use of evidence for burden of disease estimation

This systematic review and meta-analysis was conducted by WHO and ILO, supported by a large number of individual experts, for the development of the WHO/ILO Joint Estimates ((Pega et al., 2021a) Ryder, 2017). More specifically, it provides the evidence base for the organizations to consider producing estimates of the burden of deaths and DALYs from depression attributable to exposure to long working hours. The systematic review found a considerable number of studies, but the body of evidence was rated as "inadequate evidence of harmfulness" of exposure to long working hours for having, acquiring and dying from depression; the available evidence is insufficient to assess effects of the exposure. Producing estimates of the burden of depression attributable to exposure to long working hours appears therefore not evidence-based at this point.

To improve the evidence base for WHO and ILO's future considerations of producing WHO/ILO Joint Estimates of the depression burden from exposure to long working hours, better studies are needed that examine whether and to what extent working long hours increases risk of depression. Such studies should address the limitations in the literature identified in this systematic review, including:

- Assessing depression not only at baseline and follow-up, but also continuously monitoring the onset of depression between baseline and follow-up, which would allow to conduct time-to-event analyses. This can be done for example by using register data for hospital-treated depression, as has been shown in research on the association between job strain and clinical depression (Madsen et al., 2017).
- Assessing lifetime prevalence of depression prior to baseline, to distinguish between first-time onset of depression and onset of recurrent depression during follow-up. This can be done either by diagnostic interviews (Wang et al., 2012a) or by use of health register data (Svane-Petersen et al., 2020).
- Measuring exposure to long working hours not only once but repeatedly, taking changes in exposure into consideration. This would not only give more precise exposure data but also allow analysing observational data as a non-randomized pseudo trial; this has been done for example in research on onset of impaired sleep and change in health-related behaviours (Clark et al., 2015).
- Conducting experiments (if found to be ethical and feasible) and quasi-experimental studies, including natural experiment studies, on the effect of exposure to long working hours on the risk of depression.

7. Conclusions

We did not identify studies providing evidence on the association between long working hours and depression prevalence and depression mortality. Regarding depression incidence, we identified 22 studies. From these bodies of evidence, we are very uncertain regarding the effect of exposure to long working hours on having, acquiring and dying from depression and judged the existing bodies of evidence from human data as "inadequate evidence for harmfulness" for the exposure categories 41–48, 48–54 and $\geq$55 h/week, compared with 35–40 h/week.

8. Differences between protocol and systematic review

- In the protocol, we planned to convert ORs into RRs, if possible. For such conversions, information on the "prevalence of outcome in reference group or baseline risk" is required. However, such information was not available from the included studies. As all included studies reported ORs, we meta-analysed these ORs.
- There were no deviations from the Medline search strategy that was published in the protocol other than corrections of typing errors (e.g., missing parentheses).

Financial support

All authors are salaried staff members of their respective institutions. The publication was prepared with financial support to the World Health Organization from its cooperative agreement with the Centres for Disease Control and Prevention National Institute for Occupational Safety and Health of the United States of America (Grant 1E11 OH0010676-02; Grant 6NE11OH010461-02-01; and Grant 5NE11OH010461-03-00); the German Federal Ministry of Health (BMG Germany) under the BMG-WHO Collaboration Programme 2020-2023 (WHO specified award ref. 70672); and the Spanish Agency for International Cooperation (AECID) (WHO specified award ref. 71208).

Sponsors

The sponsors of this systematic review are the World Health Organization (WHO) and the International Labour Organization (ILO).

Author contributions

Had the idea for this systematic review: FP, Ivan Ivanov (WHO), Nancy Leppink (ILO).

Selected the lead reviewers and gathered the review teams: FP, Ivan Ivanov, Nancy Leppink.

Coordinated the entire series of systematic reviews: FP, YU.

Was the lead reviewer of this systematic review: RR.

Led the design of the systematic review, including developed the standard methods: FP.

Contributed substantially to the design of the systematic review: RR, KS, CDT, MB, BMR, YU, ED, JLA-M, MC, AD, ND, JL, QD-M, HE, JG, LG, IEHM, DVP, GS, JS, KT, AZ, SI.

Conducted additional analyses of original data: KS, SA, LG, JK, JW, AZ.

Conducted the search: RR, KS, CDT, MB, BMR.

Selected studies: RR, KS, CDT, MB, BMR, ED, JLA-M, MC, ND, QD-M, HE, JG, IEHM, GS, KT AZ, SI.

Extracted data: KS, CDT, MB, BMR.

Requested missing data: RR.

Assessed risk of bias: RR, KS, CDT, MB, BMR, SI.

Conducted the meta-analyses: RR, KS, FP.

Assessed quality of evidence: RR, KS.

Assessed strength of evidence: RR, KS.

Developed the standards and wrote the template for all systematic reviews in the series: FP.

Wrote the first draft of the manuscript using the template: RR, KS.

Revised the manuscript critically for important intellectual content: All authors.

Ensured tailoring of the systematic review for WHO/ILO estimation purposes: FP.

Ensured harmonization across systematic reviews in the series: FP.

Approved the final version of the systematic review to be published: All authors.

Agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved: All authors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Dr Paul Whaley (Associate Editor for Systematic Reviews, *Environment International;* and Lancaster Environment Centre, Lancaster University) and Professor Tim Driscoll (University of Sydney) for the editorial guidance and support. We thank Dr Ivan Ivanov (WHO) and Nancy Leppink for their coordination and other support for this systematic review; research librarian Elizabeth Bengtsen for her help in developing the search strategy and searching the electronic databases; and Dr Natalie Momen (WHO) for technically editing the manuscript. Professor Anne H. Garde was a member of the Working Group for Systematic Review 13 on the effect of exposure to long working hours on depression from 17 October 2017 to 25 March 2020. WHO gratefully acknowledges her participation at the meetings and her contribution to the work of the Working Group. The authors alone are responsible for the views expressed in this article, and they do not necessarily represent the views, decisions or policies of the institutions with which they are affiliated.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2021.106629.

References

10th International Labour Conference. Transition from the Informal to the Formal Economy (Recommendation No. 204). Geneva: International Labour Organization 2015.

Ahn, S., 2018. Working hours and depressive symptoms over 7 years: evidence from a Korean panel study. Int. Arch. Occup. Environ. Health 91, 273–283.

Anderson, L.M., Petticrew, M., Rehfuess, E., Armstrong, R., Ueffing, E., Baker, P., Francis, D., Tugwell, P., 2011. Using logic models to capture complexity in systematic reviews. Res. Synth. Methods 2, 33–42.

Arditi, C., Burnand, B., Peytremann-Bridevaux, I., 2016. Adding non-randomised studies to a Cochrane review brings complementary information for healthcare stakeholders: an augmented systematic review and meta-analysis. BMC Health. Serv. Res. 16, 598.

Baglioni, C., Battagliese, G., Feige, B., Spiegelhalder, K., Nissen, C., Voderholzer, U., Lombardo, C., Riemann, D., 2011. Insomnia as a predictor of depression: a meta-analytic evaluation of longitudinal epidemiological studies. J. Affect. Disord. 135, 10–19.

Bannai, A., Tamakoshi, A., 2014. The association between long working hours and health: a systematic review of epidemiological evidence. Scand. J. Work Environ. Health 40, 5–18.

Bannai, A., Yoshioka, E., Saijo, Y., Sasaki, S., Kishi, R., Tamakoshi, A., 2016. The risk of developing diabetes in association with long working dours differs by shift work schedules. J. Epidemiol 26, 481–487.

Barroga, E.F., Kojima, T., 2013. Research study designs: an appraisal for peer reviewers and science editors. Eur. Sci. Ed. 44–45.

Bech, P., Rasmussen, N.A., Olsen, L.R., Noerholm, V., Abildgaard, W., 2001. The sensitivity and specificity of the Major Depression Inventory, using the Present State Examination as the index of diagnostic validity. J. Affect. Disord. 66, 159–164.

Beller, E.M., Glasziou, P.P., Altman, D.G., Hopewell, S., Bastian, H., Chalmers, I., Gotzsche, P.C., Lasserson, T., Tovey, D., Group, P.f.A. PRISMA for Abstracts: reporting systematic reviews in journal and conference abstracts. PLoS Med 2013;10: e1001419.

Bergs, Y., Hoofs, H., Kant, I., Slangen, J., Jansen, N.W., 2018. Work-family conflict and depressive complaints among Dutch employees: examining reciprocal associations in a longitudinal study. Scand. J. Work Environ. Health 44, 69–79.

Berthelsen, M., Pallesen, S., Magerøy, N., Tyssen, R., Bjorvatn, B., Moen, B.E., Knardahl, S., 2015. Effects of psychological and social factors in shiftwork on symptoms of anxiety and depression in nurses: a 1-year follow-up. J. Occup. Environ. Med. 57, 1127–1137.

Boden, J.M., Fergusson, D.M., 2011. Alcohol and depression. Addiction 106, 906–914.

Bradford Hill, A., 1965. The environment and disease: association or causation? Proc. R. Soc. Med. 58, 295–300.

Burcusa, S.L., Iacono, W.G., 2007. Risk for recurrence in depression. Clin. Psychology Rev. 27, 959–985.

Burdorf, A., Porru, F., Rugulies, R., 2020. The COVID-19 (Coronavirus) pandemic: consequences for occupational health. Scand. J. Work Environ. Health 46, 229–230.

Clark, A.J., Salo, P., Lange, T., Jennum, P., Virtanen, M., Pentti, J., Kivimäki, M., Vahtera, J., Rod, N.H., 2015. Onset of impaired sleep as a predictor of change in health-related behaviours; analysing observational data as a series of non-randomized pseudo-trials. Int. J. Epidemiol. 44, 1027–1037.

Commission of Social Determinants on Health, 2008. Closing the gap in a generation: health equity through action on the social determinants of health. Final report of the Commission on Social Determinants of Health. World Health Organization, Geneva.

Dahlgreen, G., Whitehead, M., 2006. European strategies for tackling social inequalities in health. Levelling up: Part 2. WHO Regional Office for Europe, Copenhagen.

Deeks, J., Higgins, J., Altman, D., 2011. Chapter 9: Analysing data and undertaking meta-analyses in: Higgins J., Green S., eds. Cochrane Handbook for Systematic Reviews of Interventions Version 510 [updated March 2011]: The Cochrane Collaboration. Available from www.handbook.cochrane.org.

Dembe, A.E., Yao, X., 2016. Chronic disease risks from exposure to long-hour work schedules over a 32-year period. J. Occup. Environ. Med. 58, 861–867.

Descatha, A., Sembajwe, G., Baer, M., Boccuni, F., Di Tecco, C., Duret, C., Evanoff, B.A., Gagliardi, D., Ivanov, I.D., Leppink, N., Magnusson Hanson, L.L., Marinaccio, A., Ozguler, A., Pega, F., Pico, F., Prüss-Üstün, A.M., Ronchetti, M., Roquelaure, Y., Sabbath, E., Stevens, G.A., Tsutsumi, A., Ujita, Y., Iavicoli, S., 2018. WHO/ILO work-related burden of disease and injury: protocol for systematic reviews of exposure to long working hours and of the effect of exposure to long working hours on stroke. Environ. Int. 19, 366–378.

Descatha, A., Sembajwe, G., Pega, F., Ujita, Y., Baer, M., Boccuni, F., Di Tecco, C., Duret, C., Evanoff, B.A., Gagliardi, D., Godderis, L., Kang, S.K., Kim, B.J., Li, J., Magnusson Hanson, L.L., Marinaccio, A., Ozguler, A., Pachito, D., Pell, J., Pico, F., Ronchetti, M., Roquelaure, Y., Rugulies, R., Schouteden, M., Siegrist, J., Tsutsumi, A., Iavicoli, S., 2020. The effect of exposure to long working hours on stroke: a systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environ. Int. 142, 105746.

Dragano, N., Siegrist, J., Wahrendorf, M., 2011. Welfare regimes, labour policies and unhealthy psychosocial working conditions: a comparative study with 9917 older employees from 12 European countries. J. Epidemiol. Community Health 65, 793–799.

Drazen, J.M., de Leeuw, P.W., Laine, C., Mulrow, C., DeAngelis, C.D., Frizelle, F.A., Godlee, F., Haug, C., Hebert, P.C., James, A., Kotzin, S., Marusic, A., Reyes, H., Rosenberg, J., Sahni, P., Van der Weyden, M.B., Zhaori, G., 2010a. Toward more uniform conflict disclosures: the updated ICMJE conflict of interest reporting form. JAMA 304, 212–213.

Drazen, J.M., Van der Weyden, M.B., Sahni, P., Rosenberg, J., Marusic, A., Laine, C., Kotzin, S., Horton, R., Hebert, P.C., Haug, C., Godlee, F., Frizelle, F.A., de Leeuw, P. W., DeAngelis, C.D., 2010b. Uniform format for disclosure of competing interests in ICMJE journals. JAMA 303, 75–76.

Drill, R., Nakash, O., DeFife, J.A., Westen, D., 2015. Assessment of clinical information: Comparison of the validity of a Structured Clinical Interview (the SCID) and the Clinical Diagnostic Interview. J. Nerv. Ment. Dis. 203, 459–462.

Ebrahim, S., Davey Smith, G., 2013. Commentary: Should we always deliberately be non-representative? Int. J. Epidemiol. 42, 1022–1026.

Ezzati, M., Lopez, A.D., Rodgers, A., Murray, C.J.L., 2004. Comparative Quantification of Health Risks: Global and Regional Burdedn of Disease Attributable to Selected Major Risk Factors. World Health Organization, Geneva, Switzerland.

Figueroa, J.L., 2014. Distributional effects of Oportunidades on early child development. Soc. Sci. Med. 113, 42–49.

Forsyth, S.R., Odierna, D.H., Krauth, D., Bero, L.A., 2014. Conflicts of interest and critiques of the use of systematic reviews in policymaking: an analysis of opinion articles. Syst. Rev. 3, 122.

Fujimura, Y., Sekine, M., Tatsuse, T., 2014. Sex differences in factors contributing to family-to-work and work-to-family conflict in Japanese civil servants. J. Occup. Health 56, 485–497.

Godderis, L., Bakusic, J., Boonen, E., Delvaux, E., Ivanov, I.D., Lambrechts, M.-C., Latorraca, C.O., Leppink, N., Martimbianco, A.L., Pega, F., Prüss-Üstün, A.M., Riera, R., Ujita, Y., Pachito, D.V., 2018. WHO/ILO work-related burden of disease and injury: protocol for systematic reviews of exposure to long working hours and of the effect of exposure to long working hours on alcohol use and alcohol use disorder. Environ. Int. 120, 22–33.

Gold, P.W., 2015. The organization of the stress system and its dysregulation in depressive illness. Mol. Psychiatry 20, 32–47.

Goodman, J.E., Lynch, H.N., Beck, N.B., 2017. More clarity needed in the Navigation Guide systematic review framework. Environ. Int. 102, 74–75.

Greenland, S., Daniel, R., Pearce, N., 2016. Outcome modelling strategies in epidemiology: traditional methods and basic alternatives. Int. J. Epidemiol. 45, 565–575.

Greenland, S., Pearce, N., 2015. Statistical foundations for model-based adjustments. Annu. Rev. Public Health 36, 89–108.

Gunasekara, F.I., Richardson, K., Carter, K., Blakely, T., 2014. Fixed effects analysis of repeated measures data. Int. J. Epidemiol. 43, 264–269.

Harmer, C.J., O'Sullivan, U., Favaron, E., Massey-Chase, R., Ayres, R., Reinecke, A., Goodwin, G.M., Cowen, P.J., 2009. Effect of acute antidepressant administration on negative affective bias in depressed patients. Am. J. Psychiatry 166, 1178–1184.

Higgins, J., Altman, D., Sterne, J., 2011. Chapter 8: Assessing risk of bias in included studies. in: Higgins J., Green S., eds. Cochrane Handbook for Systematic Reviews of Interventions Version 510 [updated March 2011]: The Cochrane Collaboration. Available from http://handbook.cochrane.org.

Hulshof, C.T.J., Colosio, C., Daams, J.G., Ivanov, I.D., Prakash, K.C., Kuijer, P.P.F.M., Leppink, N., Mandic-Rajcevic, S., Masci, F., van der Molen, H.F., Neupane, S., Nygård, C.H., Oakman, J., Pega, F., Proper, K., Prüss-Üstün, A.M., Ujita, Y., Frings-Dresen, M.H.W., 2019. WHO/ILO work-related burden of disease and injury: protocol for systematic reviews of exposure to occupational ergonomic risk factors and of the effect of exposure to occupational ergonomic risk factors on osteoarthritis of hip or knee and selected other musculoskeletal diseases. Environ. Int. 125, 554–566.

Hulshof, C.T.J., Pega, F., Neupane, S., Colosio, C., Daams, J.G., Kc, P., Kuijer, P.P.F.M., Mandic-Rajcevic, S., Masci, F., van der Molen, H.F., Nygard, C.H., Oakman, J., Proper, K.I., Frings-Dresen, M.H.W., 2021a. The effect of occupational exposure to ergonomic risk factors on osteoarthritis of hip or knee and selected other musculoskeletal diseases: A systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environ. Int. 106349.

Hulshof, C.T.J., Pega, F., Neupane, S., van der Molen, H.F., Colosio, C., Daams, J.G., Descatha, A., Kc, P., Kuijer, P.P.F.M., Mandic-Rajcevic, S., Masci, F., Morgan, R.L., Nygard, C.H., Oakman, J., Proper, K.I., Solovieva, S., Frings-Dresen, M.H.W., 2021b. The prevalence of occupational exposure to ergonomic risk factors: A systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environ. Int. 146, 106157.

Imai, T., Kuwahara, K., Miyamoto, T., Okazaki, H., Nishihara, A., Kabe, I., Mizoue, T., Dohi, S., 2016. Japan Epidemiology Collaboration on Occupational Health Study, G. Validity and reproducibility of self-reported working hours among Japanese male employees. J. Occup. Health 58, 340–346.

International Labour Organization, 1999. ILO estimates over 1 million work-related fatalities each year. International Labour Organization, Geneva.

International Labour Organization, 2014. Safety and health at work : a vision for sustainable prevention: XX World Congress on Safety and Health at Work 2014: Global Forum for Prevention, 24–27 August 2014, Frankfurt, Germany. International Labour Organization, Geneva.

Johnson, P.I., Koustas, E., Vesterinen, H.M., Sutton, P., Atchley, D.S., Kim, A.N., Campbell, M., Donald, J.M., Sen, S., Bero, L., Zeise, L., Woodruff, T.J., 2016. Application of the Navigation Guide systematic review methodology to the evidence for developmental and reproductive toxicity of triclosan. Environ. Int. 92–93, 716–728.

Johnson, P.I., Sutton, P., Atchley, D.S., Koustas, E., Lam, J., Sen, S., Robinson, K.A., Axelrad, D.A., Woodruff, T.J., 2014. The Navigation Guide - evidence-based medicine meets environmental health: systematic review of human evidence for PFOA effects on fetal growth. Environ. Health Perspect. 122, 1028–1039.

Karasek, R., 1979. Job demands, job decision latitude, and mental strain: implications for job redesign. Administration Science Quarterly 24, 285–307.

Kato, R., Haruyama, Y., Endo, M., Tsutsumi, A., Muto, T., 2014. Heavy overtime work and depressive disorder among male workers. Occup. Med. (Lond) 64, 622–628.

Kessler, R.C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K.R., Rush, A.J., Walters, E.E., Wang, P.S., 2003. The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). JAMA 289, 3095–3105.

Kim, J., 2013. Depression as a psychosocial consequence of occupational injury in the US working population: findings from the medical expenditure panel survey. BMC Public Health 13, 303.

Kim, W., Park, E.C., Lee, T.H., Kim, T.H., 2016. Effect of working hours and precarious employment on depressive symptoms in South Korean employees: a longitudinal study. Occup. Environ. Med. 73, 816–822.

Kivimäki, M., Jokela, M., Nyberg, S.T., Singh-Manoux, A., Fransson, E.I., Alfredsson, L., Bjorner, J.B., Borritz, M., Burr, H., Casini, A., Clays, E., De Bacquer, D., Dragano, N., Erbel, R., Geuskens, G.A., Hamer, M., Hooftman, W.E., Houtman, I.L., Jockel, K.H., Kittel, F., Knutsson, A., Koskenvuo, M., Lunau, T., Madsen, I.E.H., Nielsen, M.L., Nordin, M., Oksanen, T., Pejtersen, J.H., Pentti, J., Rugulies, R., Salo, P., Shipley, M. J., Siegrist, J., Steptoe, A., Suominen, S.B., Theorell, T., Vahtera, J., Westerholm, P. J., Westerlund, H., O'Reilly, D., Kumari, M., Batty, G.D., Ferrie, J.E., Virtanen, 2015. M.; for the 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 386, 1739–1746.

Koustas, E., Lam, J., Sutton, P., Johnson, P.I., Atchley, D.S., Sen, S., Robinson, K.A., Axelrad, D.A., Woodruff, T.J., 2014. The Navigation Guide - evidence-based medicine meets environmental health: systematic review of nonhuman evidence for PFOA effects on fetal growth. Environ. Health Perspect. 122, 1015–1027.

Kronfeld-Schor, N., Einat, H., 2012. Circadian rhythms and depression: human psychopathology and animal models. Neuropharmacology 62, 101–114.

Lam, J., Koustas, E., Sutton, P., Cabana, M., Whitaker, E., Padula, A., Vesterinen, H., Daniels, N., Woodruff, T.J., 2016a. Applying the Navigation Guide: Case Study #6. Association Between Formaldehyde Exposures and Asthma, In preparation.

Lam, J., Koustas, E., Sutton, P., Johnson, P.I., Atchley, D.S., Sen, S., Robinson, K.A., Axelrad, D.A., Woodruff, T.J., 2014. The Navigation Guide - evidence-based medicine meets environmental health: integration of animal and human evidence for PFOA effects on fetal growth. Environ. Health Perspect. 122, 1040–1051.

Lam, J., Lanphear, B., Bellinger, D., Axelrad, D., McPartland, J., Sutton, P., Davidson, L. I., Daniels, N., Sen, S., Woodruff, T.J., 2017. Developmental PBDE exposure and IQ/ADHD in childhood: A systematic review and meta-analysis. Environ. Health Perspect. 125.

Lam, J., Sutton, P., Halladay, A., Davidson, L.I., Lawler, C., Newschaffer, C.J., Kalkbrenner, A., Joseph J. Zilber School of Public Health; Windham, G.C., Daniels, N., Sen, S., Woodruff, T.J. Applying the Navigation Guide Systematic Review Methodology Case Study #4: Association between Developmental Exposures to Ambient Air Pollution and Autism. PLoS One 2016b;21.

Lam, J., Sutton, P., Padula, A.M., Cabana, M.D., Koustas, E., Vesterinen, H.M., Whitaker, E., Skalla, L., Daniels, N., Woodruff, T.J., 2016c. Applying the Navigation Guide Systematic Review Methodology Case Study #6: Association between Formaldehyde Exposure and Asthma: A Systematic Review of the Evidence: (Protocol registered in PROSPERO, CRD42016038766). University of California at San Francisco, San Francisco, CA.

Larsen, A.D., Hannerz, H., Moller, S.V., Dyreborg, J., Bonde, J.P., Hansen, J., Kolstad, H. A., Hansen, A.M., Garde, A.H., 2017. Night work, long work weeks, and risk of accidental injuries: a register-based study. Scand. J. Work Environ. Health 43, 578–586.

Lee, D.W., Hong, Y.C., Min, K.B., Kim, T.S., Kim, M.S., Kang, M.Y., 2016. The effect of long working hours on 10-year risk of coronary heart disease and stroke in the Korean population: the Korea National Health and Nutrition Examination Survey (KNHANES), 2007 to 2013. Ann. Occup. Environ. Med. 28, 64.

Li, J., Brisson, C., Clays, E., Ferrario, M.M., Ivanov, I.D., Landsbergis, P., Leppink, N., Pega, F., Pikhart, H., Prüss-Üstün, A.M., Rugulies, R., Schnall, P.L., Stevens, G.A., Tsutsumi, A., Ujita, Y., Siegrist, J., 2018. WHO/ILO work-related burden of disease and injury: protocol for systematic reviews of exposure to long working hours and of the effect of exposure to long working hours on ischaemic heart disease. Environ. Int. 119, 558–569.

Li, J., Pega, F., Ujita, Y., Brisson, C., Clays, E., Descatha, A., Ferrario, M.M., Godderis, L., Iavicoli, S., Landsbergis, P.A., Metzendorf, M.I., Morgan, R.L., Pachito, D.V., Pikhart, H., Richter, B., Roncaioli, M., Rugulies, R., Schnall, P.L., Sembajwe, G., Trudel, X., Tsutsumi, A., Woodruff, T.J., Siegrist, J., 2020. The effect of exposure to long working hours on ischaemic heart disease: a systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environ. Int. 142, 105739.

Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gotzsche, P.C., Ioannidis, J.P., Clarke, M., Devereaux, P.J., Kleijnen, J., Moher, D., 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med. 6, e1000100.

Lorant, V., Deliege, D., Eaton, W., Robert, A., Philippot, P., Ansseau, M., 2003. Socioeconomic inequalities in depression: a meta-analysis. Am. J. Epidemiol. 157, 98–112.

Lunau, T., Wahrendorf, M., Dragano, N., Siegrist, J., 2013. Work stress and depressive symptoms in older employees: impact of national labour and social policies. BMC Public Health 13, 1086.
Laaksonen, M., Lallukka, T., Lahelma, E., Partonen, T., 2012. Working conditions and psychotropic medication: a prospective cohort study. Soc. Psychiatry Psychiatr. Epidemiol. 47, 663–670.
Madsen, I.E.H., Nyberg, S.T., Magnusson Hanson, L.L., Ferrie, J.E., Ahola, K., Alfredsson, L., Batty, G.D., Bjorner, J.B., Borritz, M., Burr, H., Chastang, J.F., de Graaf, R., Dragano, N., Hamer, M., Jokela, M., Knutsson, A., Koskenvuo, M., Koskinen, A., Leineweber, C., Niedhammer, I., Nielsen, M.L., Nordin, M., Oksanen, T., Pejtersen, J.H., Pentti, J., Plaisier, I., Salo, P., Singh-Manoux, A., Suominen, S., Ten Have, M., Theorell, T., Toppinen-Tanner, S., Vahtera, J., Väänänen, A., Westerholm, P.J.M., Westerlund, H., Fransson, E.I., Heikkilä, K., Virtanen, M., Rugulies, R., Kivimäki, M., 2017. For the IPD-Work Consortium. Job strain as a risk factor for clinical depression: systematic review and meta-analysis with additional individual participant data. Psychol. Med. 47, 1342–1356.
Mandrioli, D., Schlünssen, V., Ádám, B., Cohen, R.A., Colosio, C., Chen, W., Fischer, A., Godderis, L., Göen, T., Ivanov, I.D., Leppink, N., Mandic-Rajcevic, S., Masci, F., Nemery, B., Pega, F., Prüss-Üstün, A., Sgargi, D., Ujita, Y., van der Mierden, S., Zungu, M., Scheepers, P.T.J., 2018. WHO/ILO work-related burden of disease and injury: protocol for systematic reviews of occupational exposure to dusts and/or fibres and of the effect of occupational exposure to dusts and/or fibres on pneumoconiosis. Environ. Int. 119, 174–185.
Martikainen, P., Bartley, M., Lahelma, E., 2002. Psychosocial determinants of health in social epidemiology. Int. J. Epidemiol. 31, 1091–1093.
McEwen, B.S., 2004. Protection and damage from acute and chronic stress: allostasis and allostatic overload and relevance to the pathophysiology of psychiatric disorders. Ann. N. Y. Acad. Sci. 1032, 1–7.
McEwen, B.S., 2012. Brain on stress: how the social environment gets under the skin. Proc. Natl. Acad. Sci. U. S. A. 109 (Suppl 2), 17180–17185.
Moffitt, T.E., Caspi, A., Taylor, A., Kokaua, J., Milne, B.J., Polanczyk, G., Poulton, R., 2010. How common are common mental disorders? Evidence that lifetime prevalence rates are doubled by prospective versus retrospective ascertainment. Psychol. Med. 40, 899–909.
Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L.A., Group, P.-P., 2015. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev 4:1.
Morgan, R.L., Thayer, K.A., Bero, L., Bruce, N., Falck-Ytter, Y., Ghersi, D., Guyatt, G., Hooijmans, C., Langendam, M., Mandrioli, D., Mustafa, R.A., Rehfuess, E.A., Rooney, A.A., Shea, B., Silbergeld, E.K., Sutton, P., Wolfe, M.S., Woodruff, T.J., Verbeek, J.H., Holloway, A.C., Santesso, N., Schunemann, H.J., 2016. GRADE: assessing the quality of evidence in environmental and occupational health. Environ. Int. 92–93, 611–616.
Morgan, R.L., Whaley, P., Thayer, K.A., Schunemann, H.J., 2018. Identifying the PECO: a framework for formulating good questions to explore the association of environmental and other exposures with health outcomes. Environ. Int. 121, 1027–1031.
Munafò, M.R., Tilling, K., Taylor, A.E., Evans, D.M., Davey Smith, G., 2018. Collider scope: when selection bias can substantially influence observed associations. Int. J. Epidemiol. 47, 226–235.
Murray, C.J.L., Ezzati, M., Lopez, A.D., Rodgers, A., Vander Hoorn, S., 2004. Comparative Quantification of Health Risks: Conceptual Framework and Methodological Issues. In: Ezzati, M., Lopez, A.D., Rodgers, A., Murray, C.J.L. (Eds.), Comparative Quantification of Health Risks: Global and Regional Burden of Disease Attributable to Selected Major Risk Factors. World Health Organization, Geneva.
National Academies of Sciences, Engineering, and Medicine. Application of Systematic Review Methods in an Overall Strategy for Evaluating Low-Dose Toxicity from Endocrine Active Chemicals. Washington (DC): National Academies Press (US); 2017. https://www.ncbi.nlm.nih.gov/pubmed/28896009.
NLSY OA Cohort. Analysis on long working hours and depression using the the National Longitudinal Survey of Youth open access database. Unpublished manuscript 2019.
O'Reilly, D., Rosato, M., 2013. Worked to death? A census-based longitudinal study of the relationship between the numbers of hours spent working and mortality risk. Int. J. Epidemiol. 42, 1820–1830.
Ogasawara, K., Nakamura, Y., Aleksic, B., Yoshida, K., Ando, K., Iwata, N., Kayukawa, Y., Ozaki, N., 2011. Depression associated with alcohol intake and younger age in Japanese office workers: a case-control and a cohort study. J. Affect. Disord. 128, 33–40.
Ogawa, R., Seo, E., Maeno, T., Ito, M., Sanuki, M., Maeno, T., 2018. The relationship between long working hours and depression among first-year residents in Japan. BMC Med. Educ. 18, 50.
Olsen, J., 2014. Who needs selection bias? Scand. J. Work Environ. Health 40, 103.
Organisation for Economic Co-operation and Development (OECD). OECD.Stat: Average usual weekly hours worked on the main job 2018. Available from: http://stats.oecd.org/Index.aspx?DatasetCode=AVE_HRS. (Accessed: 21 March 2018).
Pachito, D.V., Pega, F., Bakusic, J., Boonen, E., Clays, E., Descatha, A., Delvaux, E., De Bacquer, D., Koskenvuo, K., Kröger, H., Lambrechts, M.C., Latorraca, C.O.C., Li, J., Cabrera Martimbianco, A.L., Riera, R., Rugulies, R., Sembajwe, G., Siegrist, J., Sillanmäki, L., Sumanen, M., Suominen, S., Ujita, Y., Vandersmissen, G., Godderis, L., 2020. The effect of exposure to long working hours on alcohol consumption, risky drinking and alcohol use disorder: a systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related burden of disease and injury. Environ. Int.
Pariante, C.M., Lightman, S.L., 2008. The HPA axis in major depression: classical theories and new developments. Trends Neurosci. 31, 464–468.
Paulo, M.S., Adam, B., Akagwu, C., Akparibo, I., Al-Rifai, R.H., Bazrafshan, S., Gobba, F., Green, A.C., Ivanov, I., Kezic, S., Leppink, N., Loney, T., Modenese, A., Pega, F., Peters, C.E., Prüss-Üstün, A.M., Tenkate, T., Ujita, Y., Wittlich, M., John, S.M., 2019. WHO/ILO work-related burden of disease and injury: Protocol for systematic reviews of occupational exposure to solar ultraviolet radiation and of the effect of occupational exposure to solar ultraviolet radiation on melanoma and non-melanoma skin cancer. Environ. Int. 126, 804–815.
Pega, F., Blakely, T., Glymour, M.M., Carter, K.N., Kawachi, I., 2016. Using marginal structural modeling to estimate the cumulative impact of an unconditional tax credit on self-rated health. Am. J. Epidemiol. 183, 315–324.
Pega, F., Chartres, N., Guha, N., Modenese, A., Morgan, R.L., Martínez-Silveira, M.S., Loomis, D., 2020. The effect of occupational exposure to welding fumes on trachea, bronchus and lung cancer: a protocol for a systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environ. Int. 145, 106089.
Pega, F., Liu, S.Y., Walter, S., Lhachimi, S.K., 2015. Unconditional cash transfers for assistance in humanitarian disasters: effect on use of health services and health outcomes in low- and middle-income countries. Cochrane Database Syst. Rev 9, CD011247. https://doi.org/10.1002/14651858.CD011247.pub2.
Pega, F., Liu, S.Y., Walter, S., Pabayo, R., Saith, R., Lhachimi, S.K., 2017. Unconditional cash transfers for reducing poverty and vulnerabilities: effect on use of health services and health outcomes in low- and middle-income countries. Cochrane Database Syst. Rev. 11 (11), CD011135. https://doi.org/10.1002/14651858.CD011135.pub2.
Pega, F., Momen, N.C., Ujita, Y., Driscoll, T., Whaley, P., 2021a. Systematic reviews and meta-analyses for the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environment International 155. https://doi.org/10.1016/j.envint.2021.106605.
Pega, F., Náfrádi, B., Momen, N.C., Ujita, Y., Streicher, K.N., Prüss-Üstün, A.M., Descatha, A., Driscoll, T., Fischer, F.M., GodderiS, L., Kiiver, H.M., Li, J., Magnusson Hanson, L.L., Rugulies, R., Sørensen, K., Woodruff, T.J., 2021b. Global, regional, and national burdens of ischemic heart disease and stroke attributable to exposure to long working hours for 194 countries, 2000-2016: a systematic analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environment International. https://doi.org/10.1016/j.envint.2021.106595.
Pega, F., Norris, S.L., Backes, C., Bero, L.A., Descatha, A., Gagliardi, D., Godderis, L., Loney, T., Modenese, A., Morgan, R.L., Pachito, D., Paulo, M.B.S., Scheepers, P.T.J., Schlünssen, V., Sgargi, D., Silbergeld, E.K., Sørensen, K., Sutton, P., Tenkate, T., Corrêa, Torreão, da Silva, D., Ujita, Y., van Deventer, E., Woodruff, T.J., Mandrioli, D., 2020. RoB-SPEO: a tool for assessing risk of bias in studies estimating the prevalence of exposure to occupational risk factors from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environ. Int. 135, 105039. https://doi.org/10.1016/j.envint.2019.105039.
Pittenger, C., Duman, R.S., 2008. Stress, depression, and neuroplasticity: a convergence of mechanisms. Neuropsychopharmacology 33, 88–109.
Prüss-Üstün, A., Wolf, J., Corvalan, C., Bos, R., Neira, M., 2017. Preventing Disease Through Healthy Environments: A Global Assessment of the Burden of Disease From Enviornmental Risks. In: Department of Public Health, Environmental and Social Determinants of Health. World Health Organization, Geneva.
Radloff, L.S., 1977. The CES-D Scale: a self-report depression scale for research in the general population. Appl. Psychol. Meas. 1, 385–401.
Rehfuess, E.A., Booth, A., Brereton, L., Burns, J., Gerhardus, A., Mozygemba, K., Oortwijn, W., Pfadenhauer, L.M., Tummers, M., van der Wilt, G.J., Rohwer, A., 2018. Towards a taxonomy of logic models in systematic reviews and health technology assessments: a priori, staged, and iterative approaches. Res. Synth. Methods 9, 13–24.
Rooney, A.A., Cooper, G.S., Jahnke, G.D., Lam, J., Morgan, R.L., Boyles, A.L., Ratcliffe, J. M., Kraft, A.D., Schunemann, H.J., Schwingl, P., Walker, T.D., Thayer, K.A., Lunn, R. M., 2016. How credible are the study results? Evaluating and applying internal validity tools to literature-based assessments of environmental health hazards. Environ. Int. 92–93, 617–629.
Rothman, K.J., Gallacher, J.E., Hatch, E.E., 2013a. Rebuttal: When it comes to scientific inference, sometimes a cigar is just a cigar. Int. J. Epidemiol. 42, 1026–1028.
Rothman, K.J., Gallacher, J.E., Hatch, E.E., 2013b. Why representativeness should be avoided. Int. J. Epidemiol. 42, 1012–1014.
Rugulies, R., Ando, E., Ayuso-Mateos, J.L., Bonafede, M., Cabello, M., Di Tecco, C., Dragano, N., Durand-Moreau, Q., Eguchi, H., Gao, J., Garde, A.H., Iavicoli, S., Ivanov, I.D., Leppink, N., Madsen, I.E.H., Pega, F., Prüss-Üstün, A.M., Rondinone, B. M., Sørensen, K., Tsuno, K., Ujita, Y., Zadow, A., 2019. WHO/ILO work-related burden of disease and injury: protocol for systematic reviews of exposure to long working hours and of the effect of exposure to long working hours on depression. Environ. Int. 125, 515–528.
Rugulies, R., Aust, B., Syme, S.L., 2004. Epidemiology of health and illness. A socio-psycho-physiological perspective. In: Sutton, S., Baum, A., Johnston, M. (Eds.), The Sage handbook of health psychology. Sage, London.
Ryder, G., 2017. Welcome address from the Director General of the International Labour Organization. XXI World Congress on Safety and Health at Work. Sands Expo and Convention Centre, Singapore.
Schünemann, H., Hill, S., Guyatt, G., Akl, E.A., Ahmed, F., 2011. The GRADE approach and Bradford Hill's criteria for causation. J. Epidemiol. Community Health 65, 392–395.
Schünemann, H., Oxman, A., Vist, G., Higgins, J., Deeks, J., Glasziou, P., Guyatt, G., 2011b. Chapter 12: Interpreting results and drawing conclusions. In: Higgins J., Green S., eds. Cochrane Handbook for Systematic Reviews of Interventions Version 510 [updated March 2011]: The Cochrane Collaboration; 2011b. Available from www.handbook.cochrane.org.

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L.A., Group, P.-P., 2015. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ 350:g7647.

Shields, M. Long working hours and health. Health Rep 1999;11:33-48(Eng); 37-55(Fre).

Sim, M.R., 2020. The COVID-19 pandemic: major risks to healthcare and other workers on the front line. Occup. Environ. Med. 77, 281–282.

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

Stevens, G.A., Alkema, L., Black, R.E., Boerma, J.T., Collins, G.S., Ezzati, M., Grove, J.T., Hogan, D.R., Hogan, M.C., Horton, R., Lawn, J.E., Marusic, A., Mathers, C.D., Murray, C.J., Rudan, I., Salomon, J.A., Simpson, P.J., Vos, T., Welch, V., 2016. Guidelines for Accurate and Transparent Health Estimates Reporting: the GATHER statement. Lancet 388, e19–e23.

Svane-Petersen, A.C., Holm, A., Burr, H., Framke, E., Melchior, M., Rod, N.H., Sivertsen, B., Stansfeld, S., Sorensen, J.K., Virtanen, M., Rugulies, R., Madsen, I.E.H., 2020. Psychosocial working conditions and depressive disorder: disentangling effects of job control from socioeconomic status using a life-course approach. Soc. Psychiatry Psychiatr. Epidemiol. 55, 217–228.

Teixeira, L.; Pega, F.; de Abreu, W.; de Almeida, M.; de Andrade, C.; Azevedo, T.; Dzhambov, A.; Hu, W.; Macedo, M.; Martínez-Silveira, M.; Sun, X.; Zhang, M.; Zhang, S.; da Silva, D. The prevalence of occupational exposure to noise: A systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environment International, 2021a, vol. 154, page: 106380. Doi: 10.1016/j.envint.2021.106380.

Teixeira, L.R., Azevedo, T.M., Bortkiewicz, A., Corrêa da Silva, D.T., de Abreu, W., de Almeida, M.S., de Araujo, M.A.N., Gadzicka, E., Ivanov, I.D., Leppink, N., Macedo, M.R.V., de S Maciel, E.M.G., Pawlaczyk-Luszczynska, M., Pega, F., Prüss-Üstün, A. M., Siedlecka, J., Stevens, G.A., Ujita, Y., Braga, J.U., 2019. WHO/ILO work-related burden of disease and injury: Protocol for systematic reviews of exposure to occupational noise and of the effect of exposure to occupational noise on cardiovascular disease. Environ. Int. 125:567-578.

Teixeira, L.R., Pega, F., Dzhambov, A.M., Bortkiewicz, A., da Silva, D.T.C., de Andrade, C.A.F., Gadzicka, E., Hadkhale, K., Iavicoli, S., Martínez-Silveira, M.S., Pawlaczyk-Luszczynska, M., Rondinone, B.M., Siedlecka, J., Valenti, A., Gagliardi, D., 2021. The effect of occupational exposure to noise on ischaemic heart disease, stroke and hypertension: a systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-Related Burden of Disease and Injury. Environ. Int. 154, 106387. https://doi.org/10.1016/j.envint.2021.106387.

Tenkate, T., Adam, B., Al-Rifai, R.H., Chou, B.R., Gobba, F., Ivanov, I.D., Leppink, N., Loney, T., Pega, F., Peters, C.E., Prüss-Üstün, A.M., Silva Paulo, M., Ujita, Y., Wittlich, M., Modenese, A., 2019. WHO/ILO work-related burden of disease and injury: protocol for systematic reviews of occupational exposure to solar ultraviolet radiation and of the effect of occupational exposure to solar ultraviolet radiation on cataract. Environ. Int. 125, 542–553.

Theorell, T., Hammarström, A., Aronsson, G., Träskman Bendz, L., Grape, T., Hogstedt, C., Marteinsdottir, I., Skoog, I., Hall, C., 2015. A systematic review including meta-analysis of work environment and depressive symptoms. BMC Public Health 15, 738.

Theorell, T., Karasek, R., 1996. Current issues relating to psychological job strain and cardiovascular disease research. J. Occup. Health Psychol. 1, 9–26.

Tokuyama, M., Nakao, K., Seto, M., Watanabe, A., Takeda, M., 2003. Predictors of first-onset major depressive episodes among white-collar workers. Psychiatry Clin. Neurosci. 57, 523–531.

Torá, I., Martínez, J.M., Benavides, F.G., Leveque, K., Ronda, E., 2015. Effect of economic recession on psychosocial working conditions by workers' nationality. Int. J. Occup. Environ. Health 21, 328–332.

Veritas Health Innovation. Covidence systematic review software, Veritas Health Innovation, Melbourne, Australia. Available at www.covidence.org.

Vesterinen, H., Johnson, P., Atchley, D., Sutton, P., Lam, J., Zlatnik, M., Sen, S., Woodruff, T., 2014. The relationship between fetal growth and maternal glomerular filtration rate: a systematic review. J. Maternal Fetal Neonatal Med. 1–6.

Virtanen, M., Ferrie, J.E., Gimeno, D., Vahtera, J., Elovainio, M., Singh-Manoux, A., Marmot, M.G., Kivimäki, M., 2009. Long working hours and sleep disturbances: the Whitehall II prospective cohort study. Sleep 32, 737–745.

Virtanen, M., Jokela, M., Madsen, I.E.H., Magnusson Hanson, L.L., Lallukka, T., Nyberg, S.T., Alfredsson, L., Batty, G.D., Bjorner, J.B., Borritz, M., Burr, H., Dragano, N., Erbel, R., Ferrie, J.E., Heikkilä, K., Knutsson, A., Koskenvuo, M., Lahelma, E., Nielsen, M.L., Oksanen, T., Pejtersen, J.H., Pentti, J., Rahkonen, O., Rugulies, R., Salo, P., Schupp, J., Shipley, M.J., Siegrist, J., Singh-Manoux, A., Suominen, S.B., Theorell, T., Vahtera, J., Wagner, G.G., Wang, J.L., Yiengprugsawan, V., Westerlund, H., Kivimäki, M., 2018. Long working hours and depressive symptoms: systematic review and meta-analysis of published studies and unpublished individual participant data. Scand. J. Work Environ. Health 44, 239–250.

Virtanen, M., Jokela, M., Nyberg, S.T., Madsen, I.E.H., Lallukka, T., Ahola, K., Alfredsson, L., Batty, G.D., Bjorner, J.B., Borritz, M., Burr, H., Casini, A., Clays, E., De Bacquer, D., Dragano, N., Erbel, R., Ferrie, J.E., Fransson, E.I., Hamer, M., Heikkilä, K., Jockel, K.H., Kittel, F., Knutsson, A., Koskenvuo, M., Ladwig, K.H., Lunau, T., Nielsen, M.L., Nordin, M., Oksanen, T., Pejtersen, J.H., Pentti, J., Rugulies, R., Salo, P., Schupp, J., Siegrist, J., Singh-Manoux, A., Steptoe, A., Suominen, S.B., Theorell, T., Vahtera, J., Wagner, G.G., Westerholm, P.J., Westerlund, H., Kivimäki, M., 2015. Long working hours and alcohol use: systematic review and meta-analysis of published studies and unpublished individual participant data. BMJ 350, g7772.

Virtanen, M., Stansfeld, S.A., Fuhrer, R., Ferrie, J.E., Kivimäki, M., 2012. Overtime work as a predictor of major depressive episode: a 5-year follow-up of the Whitehall II study. PLoS ONE 7, e30719.

Viswanathan, M., Ansari, M.T., Berkman, N.D., Chang, S., Hartling, L., McPheeters, M., Santaguida, P.L., Shamliyan, T., Singh, K., Tsertsvadze, A., Treadwell, J.R., 2008. Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions. Methods Guide for Effectiveness and Comparative Effectiveness Reviews, Rockville (MD) http://www.ncbi.nlm.nih.gov/pubmed/22479713.

Wang, J.L., Patten, S.B., Currie, S., Sareen, J., Schmitz, N., 2012a. A population-based longitudinal study on work environmental factors and the risk of major depressive disorder. Am. J. Epidemiol. 176, 52–59.

Wang, J.L., Patten, S.B., Currie, S., Sareen, J., Schmitz, N., 2012b. Predictors of 1-year outcomes of major depressive disorder among individuals with a lifetime diagnosis: a population-based study. Psychol. Med. 42, 327–334.

Watanabe, K., Imamura, K., Kawakami, N., 2016. Working hours and the onset of depressive disorder: a systematic review and meta-analysis. Occup. Environ. Med. 73, 877–884.

Wirtz, A., Lombardi, D.A., Willetts, J.L., Folkard, S., Christiani, D.C., 2012. Gender differences in the effect of weekly working hours on occupational injury risk in the United States working population. Scand. J. Work Environ. Health 38, 349–357.

Wittchen, H.U., Jacobi, F., 2005. Size and burden of mental disorders in Europe-a critical review and appraisal of 27 studies. Eur. Neuropsychopharmacol. 15, 357–376.

Woodruff, T.J., Sutton, P., 2014. The Navigation Guide systematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes. Environ. Health Perspect. 122, 1007–1014.

World Health Organization, 2015. ICD-10: International Statistical Classification of Diseases and Related Health Problems: 10th Revision. World Health Organization, Geneva.

World Health Organization. in: Department of Information, Evidence and Research, ed. WHO Methods and Data Sources for Global Burden of Disease Estimates 2000–2015 Global Health Estimates Technical Paper WHO/HIS/IER/GHE/20171. Geneva: World Health Organization; 2017.

World Health Organization. Disease burden and mortality estimates: disease burden, 2000–2016 Geneva, Switzerland: World Health Organization 2019. Available from: https://www.who.int/healthinfo/global_burden_disease/estimates/en/index1.html. (Accessed: 20 May 2019).

Zadow, A., Dollard, M.F., Dormann, C., Landsbergis, P., 2021. Predicting new major depression symptoms from long working hours, psychosocial safety climate and work engagement: a population based cohort study. BMJ Open. In press.

Zadow, A., Dollard, M., Dormann, C., Landsbergis, P., 2019. Predicting new clinical depression from long working hours and psychosocial safety climate. Unpublished manuscript.