

Directed acyclic graphs for the study of effects of occupation on risk of COVID-19-related outcomes.

Jack Wilkinson^{1*}, Sarah Beale², Mark Cherrie³, Rhiannon Edge⁴, David Fishwick⁵, Matt Gittins¹, Srinivasa Vittal Katikireddi⁶, Damien McElvenny³, Vahé Nafilyan^{7,8}, Neil Pearce⁹, Sarah Rhodes¹, Martie van Tongeren¹⁰

* jack.wilkinson@manchester.ac.uk

¹ Centre for Biostatistics, Manchester Academic Health Science Centre, University of Manchester.

² Institute of Health Informatics, University College London.

³ Institute of Occupational Medicine, Edinburgh.

⁴ Lancaster Medical School, Lancaster University.

⁵ Centre for Workplace Health.

⁶ MRC/CSO Social and Public Health Sciences Unit, Institute of Health and Wellbeing.

⁷ Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine.

⁸ Office for National Statistics.

⁹ Faculty of Epidemiology and Population health, London School of Hygiene and Tropical Medicine.

¹⁰ Centre for Occupational and Environmental Health, School of Health Sciences, University of Manchester.

Version 1: 21/10/21

Overview of document: Directed acyclic graphs (DAGs) may be used to represent our knowledge (or assumptions) about a data-generating process (1). A DAG may then be used to identify which variables should or should not be adjusted for in a statistical analysis in order to answer particular questions relating to the effect of an exposure on an outcome. The purpose of this document is to collate directed acyclic graphs (DAGs) that have been used in the study of the effects of occupation on COVID-19-related health outcomes (e.g. infection, severity, mortality). By collating DAGs, we hope to identify points of consensus and of contention. We also hope that this collection might form the basis for critical discussion, which may in turn lead to new proposals.

This document is intended to be updated as new DAGs concerning occupation and COVID-19-related outcomes emerge. However, we have not undertaken any sort of systematic search

strategy in order to identify relevant DAGs, and so this collection is unlikely to be comprehensive. The document represents the beginning of a live and ongoing data collection exercise, rather than one which is complete.











We present DAGs in alphabetical order (first author surname) together with limited details relating to the objectives of the study, the design of each study, the methodology for constructing the DAG, and details of how the DAG was used.

We then present a table of minimal sufficient adjustment sets for the total effect of occupation on outcome implied by each DAG.

The document concludes with some brief comments and observations about the DAGs, and information about how to leave feedback is provided. We have redrawn some of the DAGs using Daggity (daggity.net, (2)) for consistency of presentation.

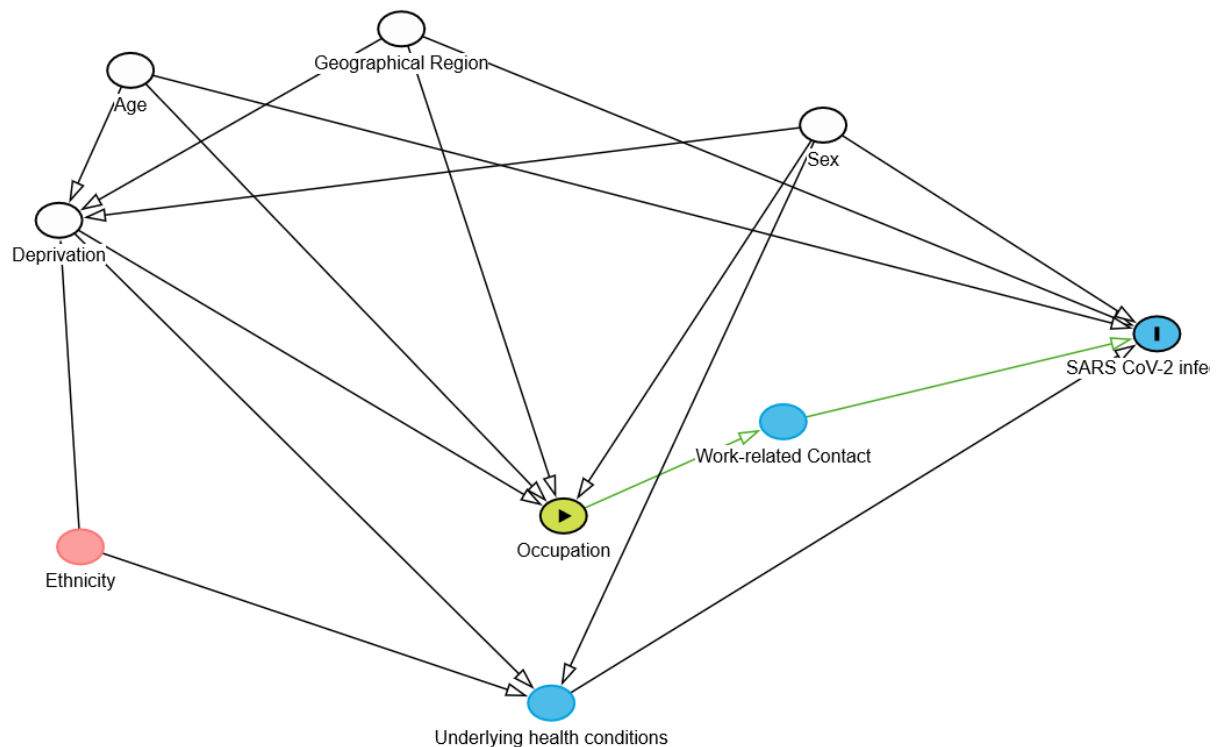
DAGs identified in the literature

Throughout, we refer to DAGs by the surname of the first author and publication year (e.g. Smith, 2021). Daggity uses the following visual code (taken from daggity.net):

-  exposure
-  outcome
-  ancestor of exposure
-  ancestor of outcome
-  ancestor of exposure *and* outcome
-  adjusted variable
-  unobserved (latent)
-  other variable
-  causal path
-  biasing path

1. Beale, et al., 2021. Occupation, Work-Related Contact, and SARS-CoV-2 Anti Nucleocapsid Serological Status: Findings from the Virus Watch prospective cohort study.

DAG, as appearing in preprint:



Redrawn from Supplementary Figure 1 from Beale, et al., 2021, using Daggity (daggity.net).

Reference: (3)

Study objectives/ estimands: “...to estimate the **total effect** of occupation on SARS-CoV-2 serological status, **whether this is mediated by** frequency of close contact within the workplace, and how exposure to poorly ventilated workplaces varied across occupations”[abstract, our emphasis]. In the text, the first objective is described differently: “1) How do odds of SARS-CoV-2 anti-nucleocapsid seropositivity vary across occupations?(primary objective)” [pg. 4]. The second description corresponds to a descriptive analysis, rather than a causal one, but the analysis indicates that the causal objective is the main interest of the authors.

Exposure: Occupation group (see page 5 and Supplementary Table 1, Beale, et al., 2021).

Outcomes: Primary outcome was serological status (based on a cut-off index of ≥ 0.1) for SARS-CoV-2 anti-nucleocapsid antibodies acquired through natural infection. “Participants who provided samples across multiple months were coded as seronegative if all samples were

below the cut-off value or as seropositive if any sample was above the cut-off value.” Secondary outcome was frequency of workplace exposure to poor ventilated environments, based on a survey questionnaire with an ordinal response (never, intermediate, every day).

Study design details: UK-based Cohort sub-study including adults who conducted monthly at-home self-administered blood antibody testing. To be eligible for the sub-study, participants had to have reported their occupation at study registration, to have had a valid antibody test result conducted between 1st Feb 2021 and 28th April 2021, and to have responded to the February 2021 monthly survey regarding features of work during the pandemic.

Method for DAG construction: Not reported.

How was DAG used (and what analyses were performed)?

The DAG was used to identify potential confounders of the relationship between occupation, work-related contact, and SARS-CoV-2 infection risk, in conjunction with VanderWeele principles of confounder selection (4). On this basis, the adjustment set identified as leading to a minimally-adjusted unbiased estimate of the total and direct effects of occupation were age, sex at birth, geographic region, and deprivation based on household income. On the basis of the DAG, it was judged that it was not necessary to additionally adjust for other socio-demographic confounders, such as ethnicity and underlying health conditions.

The authors investigated whether frequency of work-related close contact with other individuals was a mediator of the effect of occupation on serological status, and this was incorporated in the DAG. This was an ordinal variable (never, intermediate, every day) which was asked of participants who had reported being employed or self-employed at time of the survey.

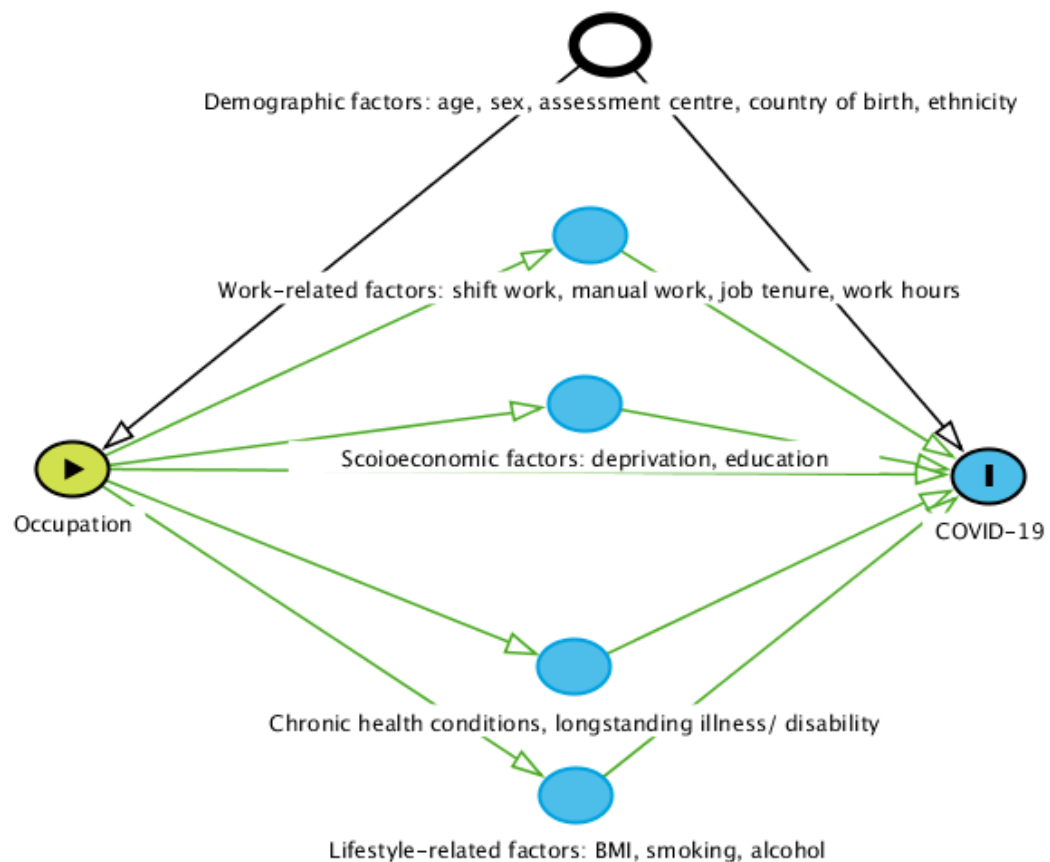
For the first two objectives, analysis was performed using Buis logistic decomposition (*ldecomp* in Stata V.16) to obtain total, direct, and indirect effects (5). Authors also used ordered logistic regression to investigate the relationship between occupation and frequency of exposure to poorly ventilated workplaces. Authors note that poor ventilation is a plausible moderator of the mediated effect (through frequent close contact), but was not analysed as such (and not included in DAG) as it was unclear from the data whether reported contacts occurred in poorly ventilated spaces. The latter model was not adjusted for sociodemographic factors, because any association between occupation and poorly ventilated workplace was assumed to be causal.

Other relevant comments: Authors note in discussion “despite using a directed acyclic graph, to inform sociodemographic confounder adjustment, the complex interrelationships between these factors make excluding these effects challenging. Notably, the relationship between

occupation, workplace contact, and serological status may be confounded by occupation-related non-workplace contacts, e.g., using public transport to reach work and contacts outside the workplace that may be increased through attending work. We also did not control for vaccination, although the timing of the antibody tests was such that, other than for healthcare workers, most of those in working age groups will not have been vaccinated. Our mediation model was constrained by statistical power and lack of available data on other relevant workplace factors, such as crowding, ventilation during periods of contact, and PPE.”

2. Mutambudzi et al., 2021. Occupation and risk of severe COVID-19: prospective cohort study of 120 075 UK Biobank participants.

DAG as supplied by authors [adjustment for 'Model 1' shown].



Reference: (6)

Study objectives/ estimands: “to examine the risk of infection by (1) broad essential occupational groups, (2) detailed essential occupational groups, and (3) Standard Occupational Classification (SOC) 2000 major groups, while accounting for baseline sociodemographic, socioeconomic, work-related, lifestyle and health factors.”

Exposure: Occupation group (see Online Supplementary Table S1)

Outcomes: Severe COVID-19 “severe COVID-19, defined by a positive test result for SARS-CoV-2 in a hospital setting (i.e., participants whose tests were taken while an inpatient or attending an emergency department) or death with a primary or contributory cause reported as COVID-19”.

Study design details: Analysis of UK Biobank data, linked to SARS-CoV-2 test results from Public Health England microbiology database, Second Generation Surveillance System and mortality records from NHS Information Centre. To be eligible, participants had to be (1) working at baseline, (2) below retirement age (<65 years) in 2020, and (3) had their baseline assessment in England. Baseline data were collected between 2006–2010, Public Health England data from 16 March to 26 July 2020 were used, linked to mortality data up to 28 June 2020.

Method for DAG construction: Not reported.

How was DAG used (and what analyses were performed)? “To assess the potential to which different covariates might be confounding or mediating differences in occupational exposure we estimated six nested models, sequentially adjusting for all covariates. Model 1 included sociodemographic factors, that is, age, sex, assessment centre, country of birth, and ethnicity. Model 2 included all covariates in model 1, plus socioeconomic factors, that is, area-level socioeconomic deprivation quartile, and education level. Model 3 included all covariates in model 2, plus work-related factors, that is, shift work, manual work, job tenure, and work hours. Model 4 included all covariates in model 2, plus number of chronic conditions, and longstanding illness/disability. Model 5 included the covariates from model 2 as well as lifestyle-related factors, that is, BMI, smoking, and alcohol. Model 6 was fully adjusted for all the above covariates. In post-hoc analyses to examine potential effect modification by race, we grouped people into white/non-essential worker, non-white/non-essential worker, white/essential worker, and non-white/essential worker, and repeated the models above.”

Authors then observed how much the estimates for occupation group changed when purported confounders or mediators were added to the models.

Other relevant comments: Authors note in comments “our sample is mostly people aged 50–64 years and so is affected by survival bias. Low-skilled workers are disproportionately affected by socioeconomic disadvantage, which is associated with poorer health outcomes and higher mortality rates overall”.

3. Nafilyan, et al., 2021. Occupation and COVID-19 mortality in England: a national linked data study of 14.3 million adults.

DAG from preprint, redrawn in Daggity [adjustment for 'fourth model' shown]:

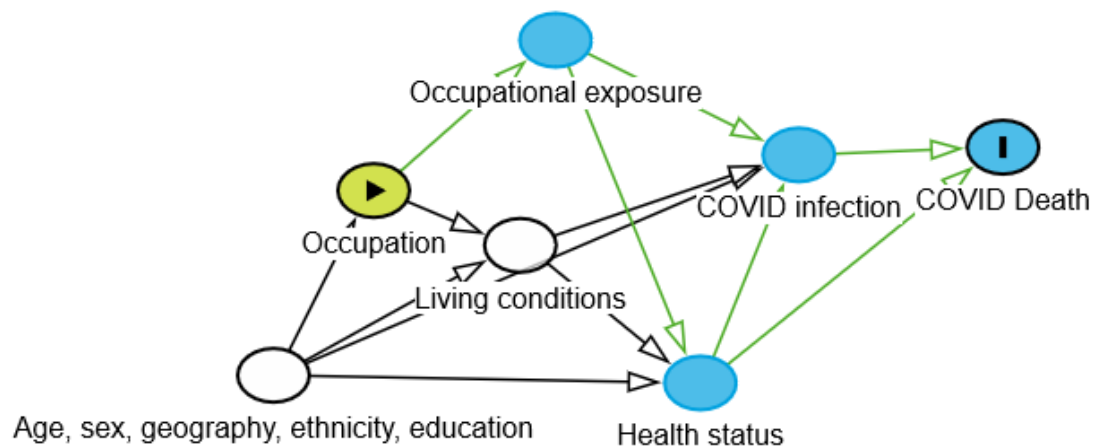


Figure 1 from Nafilyan, et al., 2021.

Reference: (7)

Study objectives/ estimands: “To estimate occupational differences in COVID-19 mortality, and test whether these are confounded by factors, such as regional differences, ethnicity and education or due to non-workplace factors, such as deprivation or pre-pandemic health.”[abstract]

“In this study, we estimated occupational differences in COVID-19 mortality in England and Wales during 2020. We have examined how much these differences changed after adjustment for non-workplace factors, using Cox proportional hazard models.”

Exposure: Occupation group at the time of the 2011 Census (see Supplementary Table S1 in preprint).

Outcomes: COVID-19 death, “defined as confirmed or suspected COVID-19 death as identified by one of two ICD10 (International Classification of Diseases, 10th revision) codes (U07.1 or U07.2) derived from the medical certificate of cause of death.”

Study design details: Analysis of data from the Public Health Data Asset. “This dataset is based on the 2011 Census in England, linked with the NHS number to death records, Hospital Episode Statistics and the General Practice Extraction Service (GPES) data for pandemic planning and research... excluded individuals (12.4%) who did not have a valid NHS number or were not linked to GPES primary care records. We used data on 14,295,900 individuals who were aged 31-55 years at the time of the 2011 Census and were therefore likely to be in stable employment both in 2011 and 2020 (by which time they were aged 40-64 years). We examined the differences between occupation groups in the risk of death involving COVID-19 during the 11 months from 24 January to 28 December 2020.”

Method for DAG construction: Not reported.

How was DAG used (and what analyses were performed)? Used Cox regression in order to estimate the effect of occupation due to work-related exposures on COVID-19 mortality. Five models were fitted sequentially, in order to assess how variables might confound or mediate the effect. “Our first model was only adjusted for age. The second model also adjusted for geographical factors (region, population density, rural urban classification) to account for the differential spread of the virus in different areas. The third model further adjusted for other confounding factors, ethnicity and education, which are related both to occupation and COVID-19 risk. The fourth model also controlled for non-workplace factors (living conditions), including socio- economic factors (Index of Multiple Deprivation, household deprivation, household tenancy and house type) and household composition (household size, children in the household, overcrowding). Finally, the last model adjusted for pre-pandemic health (BMI, chronic kidney disease, learning disability, cancer or immunosuppression, and other conditions; see Supplementary Table S2 for details on all the covariates). We used corporate managers and directors as the reference category, because it is a large group with a low absolute risk”. Each of these models were fitted to men and women separately. The variables to adjust for may have been selected using the DAG confounders and mediators were chosen).

Authors then observed how much the hazard ratios for occupation group changed when confounders or the possible mediator were added to the models.

Summary of minimal sufficient adjustment sets for total effect of occupation on COVID-19 outcome

| DAG name | Outcome | Minimal sufficient adjustment set implied by DAG | Adjustment set 1 | Adjustment set 2 | Adjustment set 3 | Adjustment set 4 | Adjustment set 5 | Adjustment set 6 |
|-----------------|-------------------------------|---|--|---|--|---|--|---------------------------------|
| Beale 2021 | SARS-CoV-2 serological status | Age, deprivation, geographical region, sex | Age, sex, geographic region, deprivation | | | | | |
| Mutambudzi 2021 | Severe COVID-19 | Sociodemographic factors (age, sex, assessment centre, country of birth, and ethnicity) | Age, sex, assessment centre, country of birth, and ethnicity | Set 1 plus area-level socioeconomic deprivation quartile, and education level | Set 2 plus shift work, manual work, job tenure, and work hours | Set 2 plus number of chronic conditions, and longstanding illness/disability | Set 2 plus BMI, smoking, and alcohol. | All variables in previous sets. |
| Nafilyan 2021 | COVID-19 mortality | Age, sex, geography, ethnicity, education | Age | Set 1 plus region, population density, rural urban classification | Set 2 plus ethnicity and education | Set 3 plus Index of Multiple Deprivation, household deprivation, household tenancy and house type) and household composition (household size, children in the household, overcrowding | Set 4 plus BMI, chronic kidney disease, learning disability, cancer or immunosuppression, and other health conditions. | |

Table 1: Minimal sufficient adjustment sets implied by DAG, and adjustment sets used.

Discussion points

- All DAGs contain super-nodes (representing multiple variables). Is there anything to be gained from separating these out, and representing causal relationships between them?
- Differences include whether (groups of) variables are considered to be confounders or mediators. For example, deprivation is a confounder in Beale, but is subsumed as a

mediator under ‘living conditions’ in Nafilyan, 2021 and under ‘socioeconomic factors’ in Mutambudzi, 2021. Health status is incorporated differently in the three DAGs also: it is not directly or indirectly affected by occupation in Beale, 2021; it is directly affected (a mediator) by occupation in Mutambudzi, 2021; and indirectly affected (via occupational exposure and living conditions) in Nafilyan, 2021. Also, ethnicity is considered to directly affect occupation in Nafilyan, 2021 and in Mutambudzi, 2021. In Beale, 2021, ethnicity affects occupation via deprivation. We might consider the possibility that e.g. structural racism or cultural expectations could influence representation of different groups in different occupations (8, 9); ethnicity would then serve as a proxy for these constructs.

- How to include non-work-related contact? Is this part of the effect of interest? For example, occupation affects income, which in turn affects e.g. ability to socialise? Another example would be contact due to transport to work. See comments below around refining the research question.
- Part of the disagreement between DAGs might result from the fact that occupation can be construed as a prolonged, continuous, exposure, which is both affected by and a cause of other factors over time. For example, extended tenure in a physically demanding (or sedentary) occupation may adversely affect health. Health problems may then influence the likelihood of staying in that occupation or switching to another. This would be an example of time-varying confounding. The examples collected here represent constructs relating to health and living conditions as either causally influencing occupation, or as being affected by occupation, but we haven’t identified any examples so far which considered the possibility that both these things might be true. Again, see discussion of the research question below (also see Robins (10), on the ‘Healthy Worker’ effect.)
- There may be a danger in over simplifying the term ‘occupation’. In essence, occupation/job title is just a label for a (sometimes very) complicated set of tasks that make up work. Having a single node for “occupation” might miss complexity relating to (i) physical characteristics of the environment, ii) numbers of contacts with other workers, (iii) contacts with known cases of COVID 19. Trying to capture this complexity with a single node may force many assumptions in subsequent models relying on the DAG. The actual behaviours and interactions experienced as part of one’s occupation might be highly variable even within a particular Standard Occupational Classification code (SOC, a classification system used in the UK, for example). As an example, the term ‘nurse’ covers Intensive Therapy Unit nurses with daily COVID patient contact, right through to district nurse doing research from home, and everything in between.

- Similarly, it might be important not to oversimplify health, by including it as a single node. Health conditions differ in their nature. Some are caused specifically by work, others not. In both categories, there are health conditions that will (i) NOT influence the transmission of COVID e.g arthritis of the knee, (ii) conditions that might increase the risk of contracting infection (e.g. diabetes, immunocompromised conditions) and also those (iii) that may dictate a poorer outcome should one become infected (e.g. cardiovascular conditions, elevated BMI and so on). One node may not sufficiently articulate how these should be modelled, particularly in relation to occupational exposures and occupational related ill health.”
- ‘The effect of occupation on COVID-related outcomes’ is compatible with a variety of research questions (and estimands). It is anticipated that different DAGs (or different adjustment sets) will be appropriate for different questions. It is also possible that some questions might not be easily represented by a DAG. It is important to clearly specify the effect of interest before constructing a DAG and selecting an approach to analysis. Some examples of possible variations are hinted at in the comments above. For example, we might be interested in the impact of being in an occupation for a prolonged period of time, allowing that this could have an impact on long-term health and living conditions. Or, we might be interested in answering a question like “if you were to start a new occupation (When? At the start of the pandemic? After COVID measures are in place?), how would this affect your COVID risk?” This would imply that effects mediated by long-term health are of lesser interest, since short exposure to an occupation would be anticipated to have minimal effects on long-term health. A related possibility would be to consider the effect of restarting to go to work, when you had been working at home. We must also specify whether we are interested purely in risks caused by workplace contact, or whether we are also interested in e.g. contacts outside the workplace, which might be affected by your employment (including travel to work, or ability to socialise, which is determined by salary, working hours, and social contact with colleagues).
- It is possible that different DAGs would be appropriate for different points in time. The DAGs presented here relate to studies concerning time periods before there were high levels of vaccination in the working-age population. Changes in vaccination levels, as well as changes in other restrictions to life and work (such as mandates to work from home) need to be considered, further complicating the study of occupation effects.

Feedback

Feedback is encouraged both with respect to the DAGs presented here and to the design and content of the current document. We have emulated the approach of Barnard-Mayers

and colleagues (11) in this regard. Interactive versions of the three DAGs presented here are available at dagitty.net/mGDcq_a (Beale, 2021), dagitty.net/mie77DC (Mutambudzi, 2021) and dagitty.net/mdV78B7 (Nafilyan, 2021). These links may be copied into your browser to access the DAGs. We have set up a Google form where readers may submit comments or links to their own DAGs (<https://forms.gle/B5FuSk5LNnE4FH2H6>).

Funding statement

This work was supported by funding from the PROTECT COVID-19 National Core Study on transmission and environment, managed by the Health and Safety Executive on behalf of HM Government.

References

1. Pearl J. Causal diagrams for empirical research. *Biometrika*. 1995;82(4):669-88.
2. Textor J, van der Zander B, Gilthorpe MS, Liskiewicz M, Ellison GT. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *Int J Epidemiol*. 2016;45(6):1887-94.
3. Beale S, Patel P, Rodger A, Braithwaite I, Byrne T, Fong WLE, et al. Occupation, Work-Related Contact, and SARS-CoV-2 Anti-Nucleocapsid Serological Status: Findings from the Virus Watch prospective cohort study. *medRxiv*. 2021.
4. VanderWeele TJ. Principles of confounder selection. *Eur J Epidemiol*. 2019;34(3):211-9.
5. Buis ML. Direct and indirect effects in a logit model. *Stata J*. 2010;10(1):11-29.
6. Mutambudzi M, Niedwiedz C, Macdonald EB, Leyland A, Mair F, Anderson J, et al. Occupation and risk of severe COVID-19: prospective cohort study of 120 075 UK Biobank participants. *Occup Environ Med*. 2020.
7. Nafilyan V, Pawelek P, Ayoubkhani D, Rhodes S, Pembrey L, Matz M, et al. Occupation and COVID-19 mortality in England: a national linked data study of 14.3 million adults. *MedRxiv*. 2021.
8. Hawkins D. Differential occupational risk for COVID-19 and other infection exposure according to race and ethnicity. *Am J Ind Med*. 2020;63(9):817-20.
9. Katikireddi SV, Lal S, Carrol ED, Niedzwiedz CL, Khunti K, Dundas R, et al. Unequal impact of the COVID-19 crisis on minority ethnic groups: a framework for understanding and addressing inequalities. *J Epidemiol Community Health*. 2021;75(10):970-4.
10. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical Modelling*. 1986;7:1393-512.
11. Barnard-Mayers R, Kouser H, Cohen JA, Tassiopoulos K, Caniglia EC, Moscicki A, et al. A case study and proposal for publishing directed acyclic graphs: The effectiveness of the quadrivalent HPV vaccine in perinatally HIV exposed girls OSF. 2021.