Building and Environment 250 (2024) 111145

Contents lists available at ScienceDirect

Building and Environment

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

Development and use of a directed acyclic graph (DAG) for conceptual framework and study protocol development exploring relationships between dwelling characteristics and household transmission of COVID-19 – England, 2020

Hannah Taylor [a,b,c,d,*], Helen Crabbe [b], Clare Humphreys [c], Gavin Dabrera [e], Anna Mavrogianni [f], Neville Q. Verlander [g], Giovanni S. Leonardi [b,h]

[a] UK FETP, UKHSA, UK
[b] Environmental Epidemiology, RCE, UKHSA, UK
[c] SE Health Protection, UKHSA, UK
[d] HQ AMS, Camberley, UK
[e] COVID-19 National Epidemiology Cell, UK
[f] Institute for Environmental Design and Engineering (IEDE), The Bartlett, University College London (UCL), UK
[g] Statistics Unit, UKHSA, UK
[h] Department of Public Health, Environments and Society, London School of Hygiene and Tropical Medicine, UK

ARTICLE INFO

Keywords:
Epidemiology
Directed Acyclic Graphs
DAG
Public Health
Environmental exposures
Housing
Dwelling characteristics
COVID-19
Study design

ABSTRACT

Background: Household settings are high risk for COVID-19 transmission. Understanding transmission factors associated with environmental dwelling characteristics is important in informing public health and building design recommendations. We aimed to develop a directed acyclic graph (DAG) to inform a novel analytical study examining the effect of dwelling environmental characteristics on household transmission of COVID-19.
Methods: Key demographic, behavioural and environmental dwelling characteristics were identified by a multidisciplinary team. Using the DAG to visually display risk factors, and using expert knowledge of available datasets we reached a consensus on the factors included and directionality of relationships to build the final conceptual framework. Factors were displayed as nodes and relationships as pathways.
Results: Of 34 potential factors, 16 were included in the DAG, with 13 causal and three biasing pathways. Three variables were not measurable using retrospective datasets. The DAG enabled us to select data sources for the pilot study period and to inform the analysis plan. Key exposure nodes were energy efficiency or dwelling age; dwelling type or number of storeys; and dwelling size. We determined direct and proxy confounders which we could adjust for, potential interactions terms we could test in model building, and co-linear variables to omit in the same model.
Conclusions: The DAG helped identify key variables and datasets. It prioritised key nodes and pathways to formalise complex relationships between variables. It was pivotal in identifying unobserved variables, confounders, co-linearity and potential interactions. It has supported data selection and design of a retrospective pilot study analysis plan.

* Corresponding author. UK FETP, UKHSA, UK.
E-mail addresses: hannah.taylor@ukhsa.gov.uk (H. Taylor), Helen.Crabbe@ukhsa.gov.uk (H. Crabbe), Clare.Humphreys@ukhsa.gov.uk (C. Humphreys), Gavin.Dabrera@ukhsa.gov.uk (G. Dabrera), a.mavrogianni@ucl.ac.uk (A. Mavrogianni), Neville.Verlander@ukhsa.gov.uk (N.Q. Verlander), Giovanni.Leonardi@ukhsa.gov.uk (G.S. Leonardi).

https://doi.org/10.1016/j.buildenv.2023.111145
Received 12 September 2023; Received in revised form 29 November 2023; Accepted 13 December 2023
Available online 27 December 2023
0360-1323/

1. Introduction

1.1. COVID-19 transmission

The Coronavirus-19 (COVID-19) pandemic has had a significant impact on global morbidity and mortality, as have the mitigation measures used to control and prevent the spread of COVID-19. Over 20 million cases of COVID-19 have been reported in England since it was first identified in Wuhan, China, in December 2019 [1]. Understanding the factors which contribute to the risk of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) transmission continues being a priority [2]. Numerous studies have shown that transmission within a household has been a key driver of the pandemic [3–5]. Being a household contact of a confirmed case is a risk factor for becoming a secondary case [6–8]. Secondary transmission rates (also known as a secondary attack rate) to household contacts has previously been reported to be between 6% and 37% depending upon the case definition used for a case (for example defining a case based on symptoms reported, results from a lateral flow device test, or a laboratory based test) and the availability of testing [9–11].

1.2. Household transmission risks

As household transmission has been identified as a key driver of the pandemic [3–5], understanding transmission risks within a household setting is of significant public health importance. Understanding the factors which contribute to and drive transmission within households will ensure policy and guidance on mitigating household exposure to COVID-19 targets the appropriate risk factors. Public health policy makers can use such knowledge to make informed recommendations which balance the risks of transmission within a household and then onwards into the community against the potential impacts of any control measures recommended on the household.

Previous studies have identified key principles of COVID-19 transmission risk in three main areas: viral dynamics, household composition and behaviours, and environmental factors [2]. The majority of existing studies which look at household transmission of COVID-19 have focused on household composition, including household size and age of household contacts [9,12]. Often studies have focussed on viral dynamics, the impact of vaccination and different COVID-19 variants on household transmission [13,14]. Whilst some studies have considered the impacts of human behaviours on COVID-19 transmission in the community, specific impacts of occupant behaviours within a dwelling, on household transmission alone, have not been studied. In terms of the built environment, there is limited research on the impact environmental characteristics of dwellings have on household transmission of COVID-19. One study did include dwelling type in their analysis. This study found that there was reduced risk of transmission in flats compared to semi-detached houses and no difference for terraced or detached houses compared to semi-detached houses [9]. Therefore, a novel study examining the associations of environmental characteristics of dwellings and household transmission of COVID-19 is required to support public health and environmental design decision makers in providing evidence based advice on the topic.

1.3. Environmental characteristics of the dwelling

Health outcomes are influenced by a number of factors that are not dependent upon just an individual's biology or behaviours. These wider determinants of health include the physical environment, community and dwelling conditions in which they live [15]. Dwelling conditions are a specific factor considered in the adapted Labonte model for wider determinants of health [16]. There are a number of dwelling characteristics which may impact on household transmission of COVID-19. These include the geometric form, number of floors, volume and ventilation design of the dwelling. All of these factors can impact the way in which an airborne pathogen such as COVID-19 is dispersed around the dwelling, and areas of potential viral concentration and the way in which household members utilise the space. Geometric form refers to the type of dwelling, for example an apartment, multi-storey house or single-storey bungalow, and whether the dwelling is detached (separate to any other dwelling), semi-detached (attached on one side) or terraced (dwellings attached in a row). Volume can refer to physical volume, but also the number of rooms as both affect ventilation flows, the volume of air available [17], and the way in which the dwelling space is used and building systems (in particular, ventilation) operated by household members. These factors can influence the likelihood of transmission of the virus within the dwelling environment, and also the way in which household members in the dwelling behave and therefore impact the risk of transmission.

Quality of dwelling stock is also an area of consideration. Maintenance and upkeep of the property, including the ability to keep it ventilated, dry and heated may impact on transmission, but also disease severity [18,19]. In England, damp and mould conditions are historically worse in rental properties [20]. Age of construction materials, energy efficiency of materials and windows, all influence the conditions which make presence and transmission of respiratory viruses, including COVID-19, more or less likely. Housing quality has been considered in the context of incidence and severity of respiratory illness, with mould and damp found to have an association with increased risk of hospital admission for acute respiratory illness [21]. There are also links between dwelling energy efficiency and thermal regulation, with cold homes and fuel poverty associated with increased morbidity and mortality, including from respiratory illness [22]. COVID-19 cases may be more susceptible to the cold and to fungal infections, and therefore dwelling conditions that are cold and encourage mould may not just affect transmission, but also severity of illness and outcome of a case [23]. It is therefore of public health importance that we better understand how environmental dwelling characteristics, alone and in combination with other transmission risk factors affect COVID-19 transmission within the dwelling and, therefore, impact secondary attack rate within a household.

1.4. The role of a directed acyclic graph

Directed Acyclic Graphs (DAGs) originated in computer science. DAGs can allow specialists from a range of fields to use a multidisciplinary approach to visually communicate causal structures, identify bias and build better analytical epidemiological models [24]. They can be used as a way of depicting the assumptions within a system [25]. DAGs are composed of variables, known as nodes and arrows known as pathways. These pathways depict known and suspected associations between the variables. The DAG must include exposures of interest, in this case the environmental dwelling characteristics, the outcome of interest (becoming a secondary household case of COVID-19), and the variables which might influence those variables. Nodes may be dependent upon other variables, or independent and the directionality and number of pathways between nodes helps epidemiologists and statisticians determine the relationships between variables. A DAG can also highlight discrepancies between those variables we would like to capture and the data actually available. The conceptual model created in the visual form of a DAG informs study design and the analysis plan.

Previous studies have shown the success of using DAGs to help identify, gather and utilise the most relevant factors in a conceptual framework to inform analytical study design [26]. As highlighted in the section above, the factors which influence transmission within a dwelling are complex and their effects overlap with each other. Identification, understanding and prioritisation of these factors and complex interactions need to come from expertise across a number of disciplines, which in this example, include infectious disease epidemiology, environmental epidemiology, environmental design and engineering design. To ensure a comprehensive consideration of all potential risk factors, the approach requires a multidisciplinary team (MDT) approach to firstly identify key risk factors for household transmission, and then determine the plausibility, likelihood and directionality of their relationship to each other, other co-variables, and the outcome of household contacts becoming a secondary case.

1.5. Use of a DAG in developing a pilot study

Understanding the risk factors pertaining to environmental dwelling characteristics for COVID-19 transmission is critical to informing the public health advice we give. However, the relationship is complex and cannot be looked at alone. Other factors relating to the virus itself, the geographical area the dwelling is in, the household composition of the dwelling and the behaviours of household members will all impact the risk of transmission. There is data relating to some of these factors, but not all, and knowledge on the risks associated with factors in each of these areas is held by different subject matter experts, not one individual or group of specialists. We therefore chose to build a DAG to inform our conceptual model for a pilot study. A small pilot study would allow us to test our DAG, the feasibility of data collection, data linkage and the delivery of the study design and analysis plan.

The COVID-19 pandemic was prolonged, and the UK like other countries, underwent different stages of intervention and lockdowns to prevent virus spread. The impact of such interventions have been reviewed [27]. However, modelling such changes on secondary attack rates is complex. A population level study would need to account for these changes and changes in behaviour. As a pilot we wanted to focus on environmental dwelling characteristics, in order to minimise the challenges of building a model to account for different non-pharmaceutical interventions, such as lockdowns, changing guidance on physical distancing and the use of face coverings, the changing viral dynamics of different COVID-19 variants, different levels of community transmission and the impact of vaccination. To focus on the effect of dwelling characteristics on transmission, the decision was made to design a DAG to suit a retrospective pilot using existing data from the start of the pandemic (February to March 2020) [28–30]. Using a DAG allows expert knowledge, consideration of the variables required, exploration of the datasets available, and relevant time period to model the relationships [25].

This paper describes the development of the DAG and its utility in mapping a conceptual framework, identifying variable required, datasets to use and feasibility developing a pilot study.

1.6. Aim and objectives

Our aim was to develop a DAG that could be used as the conceptual framework for designing a retrospective pilot study protocol and analysis plan to explore associations between environmental dwelling characteristics and household transmission of COVID-19.

Our objectives were to pursue a MDT approach to design a DAG. We could then use the DAG to explore and visualise the complexity of these relationships, and to inform data source selection and linkage. This in turn informs the design of a pilot, population level, analytical study exploring risk factors for becoming a secondary case of COVID-19 following household contact with a primary case. These pilot findings will inform the study team of the utility of the DAG and conceptual framework, the data selection and the feasibility of data linkages.

1.7. Scope of the paper

In this paper, we will describe how we worked as an MDT to select and build consensus on the variables for the DAG and the design or nodes and pathways contained within it. We present how we determined the DAG nodes, labels and pathways, and how we could use this to inform a pilot study design and analysis plan. We discuss the ways in which we check the utility of our DAG in the pilot and the feasibility of this approach for a larger study. The purpose of this paper is to present the methods utilised in developing the DAG and the outcome of the final DAG as the output of this piece of work. We do not present any data or the findings of analysis of data in this paper as conducting the pilot study and it's results are the focus of separate future papers.

2. Methods

2.1. DAG development - the outcome of interest

The development of a DAG must start with the identification of the outcome of interest and the exposure, or exposures of interest. We start by defining the outcome of interest. This was defined as a household member of a primary case becoming a secondary case of COVID-19 through household contact. Once this was defined, epidemiology specialists were able to inform inclusion and exclusion criteria which would be required. Contacts had to live in a dwelling with the primary case during the primary case's infectious period. They could not be a co-primary case (another household member infected from the same source). Reflecting the case definition used at the time and as described in the HOSTED study [9], contacts were defined as a secondary case if they had symptoms consistent with COVID-19 infection (cough, fever or loss of sense of taste or smell) or a positive test within 2–14 days of the primary household case. Once this had been determined, factors of interest which influenced this outcome then needed to be identified.

2.2. Identifying key factors in household transmission of COVID-19

A MDT consisting of experts in environmental design and environmental (dwelling) engineering, environmental and infectious disease epidemiology, health protection and health inequalities, was established to identify, agree and prioritise key factors for COVID-19 transmission in households. We wanted to include subject matter experts from the breadth of specialties relevant to this complex subject to ensure that no key variables or relationships were missed. This can make the process more time consuming and detailed in order to establish consensus, but it is a key aspect of collaborative practice in research and improves the reliability of any conceptual framework produced. Using expert knowledge and a rapid literature review, each group of specialists identified key viral transmission dynamic factors, household demographic and behavioural factors, and environmental dwelling factors for consideration.

2.3. DAG development – potential data sources and the prioritisation of nodes in the DAG

Using DAGitty 3.0 [31], a freely available browser-based environment for creating, editing and analysing DAGs, the variables identified in the process described above were displayed visually as nodes. DAGitty was accessed for this study by using Microsoft Windows version 10 on a laptop with internet access. Initially all key factors were included. In order to prioritise the most important factors, a series of consecutive online meetings were held with the MDT. In these meetings each factor and their relationships with other factors were discussed. Factors identified by different specialist experts were gathered, sometimes with different names, but ultimately the key variables were retained and duplicates or dependent risk factors removed. The temporal positioning of each factor from left to right was then considered in relation to the outcome of interest on the far right (becoming a secondary household case of COVID). Through discussion, knowledge from the rapid literature review, the use of the DAG as a visual aid for consensus building and through knowledge of potentially available data sources, the team were able to combine and rename factors which were known composite measures.

Potential data sources for each identified node were identified and the advantages and disadvantages of each discussed. Where no existing data sources could be identified through MDT discussion, factors were considered to determine whether it could have a probabilistic causal effect on the outcome based on the other variables included and findings of the rapid review. Where no new probabilistic pathway were identified, the factor was removed. Where a potential association was found it was considered critical to the DAG and the factor remained. This process allowed the team to reach consensus on the key factors for inclusion in a final conceptual framework.

2.4. DAG node labelling

The nodes which directly pertained to the environmental characteristics of the dwelling, as opposed to the geographical location of the dwelling or household members or their behaviours, were defined as exposures. All of the variables were then linked by paths to each other and the outcome node (defined above as becoming a secondary case of COVID-19 through household contact). The directionality of the paths and number of paths added was determined by the probabilistic causal effect of each node on another from the MDTs specialist knowledge and the rapid literature review.

Once the paths were completed, unobserved nodes which had been identified as critical and therefore kept in the DAG were labelled as unobserved. This then allowed the remaining causal and biasing pathways of the DAG to be determined. Paths leading directly from an exposure to the outcome (even if passing through other nodes) are causal. Paths that block the link between an exposure and the outcome are non-causal pathways. Where a node lies between an exposure and the outcome but has pathways with directionality towards both, the node is identified and labelled as a confounder. Where one of these non-causal pathways goes through another node it is a proxy confounder.

Other node types identified include ancestors of the exposure, which are nodes that occur temporally before the exposure. Where at least two arrow heads meet at a node, these nodes may be labelled as a collider. Variables that lead to the outcome but are otherwise unconnected to the other nodes in the DAG are independent variables and were labelled as such. As well as the node labels determined by the DAG pathways, nodes can also be labelled as potential interactions. These interacting nodes are determined by specialist knowledge and the results of the literature review. Nodes may have more than one label.

2.5. Understanding how node labels inform the analysis plan

Observable nodes were labelled as described in the section above as either exposure variables, independent covariates, direct or proxy confounders, ancestors of exposure, colliders or interaction nodes. This determined how they would be treated in the analysis plan. Confounders and proxy confounders can be adjusted for by including them in a final multivariable analysis model. Interacting terms can be tested in the model building process, and co-linearity of variables can be tested and one can ensure the model does not include both terms in the same model. Ancestors of exposure may need to be treated as a random effects and this can be tested for in the model building process.

From the final conceptual model represented in the DAG, minimum sufficient adjustment sets for analysis (the minimum number of variables needed to test a hypothesis) were able to be identified. This can consider all key exposure nodes, or each exposure one by one. All identified relationships gleaned from the DAG can then be used to inform the analysis plan and study protocol design.

2.6. Incorporating the findings into a pilot study protocol

The results of this DAG building methodology were used to present the final DAG in the study protocol, identify the data sources that will be used and the processes anticipated for data linkage between the different datasets. The analysis plan influenced by the DAG is documented in the study protocol and will be referred back to during the conduct of the pilot study which is the focus of a future paper.

2.7. Testing the utility of the DAG in the pilot study

The utility of the DAG can be tested and the feasibility of using the data sources and data linkage methods identified by undertaking a sensitivity analysis of our final model. The next steps in model building and testing (not covered in this paper), are to undertake multivariable analysis using the variables identified in the DAG. Inclusion variables will be tested with likelihood ratio testing. As DAGs are a newer tool applied in epidemiology, sensitivity analysis is an important part of considering utility and validity of using DAGs in both pilot and subsequent studies.

3. Results

3.1. Factor identification and prioritisation

Rapid literature review and expert opinion of the MDT team initially identified 34 factors for consideration. These included combinations of the domestic environment such as number of individuals in the household, composition of the household, the way in which they may use the space based on access to hand washing facilities and number of rooms per person and whether the dwelling is privately owned or rented. It also included combinations of environmental characteristics including when the dwelling was built, the design of the house, materials used to build it and geographical location. Finally, these considered combinations of behavioural factors including the ways in which the dwelling was ventilated, hand-washing practices and self-isolation behaviours. The full list of these is shown in Supplement 1. Through discussion and screening of these factors and knowledge of available datasets a number of these were combined to create composite nodes.

Overcrowding and the three potential measures of person density were combined into one node entitled ‘household density’. Floorspace and height were used as composites to create dwelling volume and therefore these were also combined into one node titled ‘volume’. The three factors relating to the composition of the household were compiled into one composite node called ‘household composition’. In the UK, houses which have been sold or built since 2007 or which are rented must have an Energy Performance Certificate (EPC) rating. This is a composite measure of nine factors which relate to the state of the property, building materials used, insulator and ventilatory properties and overall energy efficiency of the property. The decision to include a composite data point therefore reduced nine nodes to one. The node ‘dwelling type’ captured both the geometric form and type of house attachment in a composite measure of the two for example semi-detached house or (detached) bungalow. This use of four composite nodes reduced the total number of nodes from 34 to 18.

Five nodes were then identified as not being measurable from the datasets collected in February to March 2020. These included the ability of the household member to self-isolate from others, for example, if they were a young child or an adult requiring care and support with activities of daily living. Data was also unavailable on ‘access to appropriate handwashing facilities’ which includes if these are separate to those used by others, if there is soap and running water available and measures of the thermal comfort of the water. Ventilation of the property was required but unavailable. This included measures of natural or mechanical ventilation or air conditioning and then the behaviours that facilitate and direct the ventilation, such as opening windows and doors. These were all felt to be critical to the potential for transmission and were retained in the DAG as unobserved nodes.

Occupation of the case or contacts was not collected in the original dataset on cases. Healthcare workers (HCWs) were found to be at increased risk before community transmission was widespread [32]. This did result in increased risk of transmission to households of HCWs later in the pandemic – particularly with shortages of PPE. However, it is not understood if this increased risk was also seen in Feb 2020- early Mar 2020 when there was not widespread community transmission, PPE was available and COVID-19 was categorised as a High Consequence Infectious Disease (HCID) [33]. Because of the specific time frame we were looking at the evidence base for probabilistic causality associated with occupation (specifically being a HCW) and COVID-19 household transmission during this very early phase of the pandemic was not deemed to be sufficient.

Likewise, no data on dwelling price was held by the team. This was wanted to understand the potential size of the dwelling and socio economic status of the occupants. On discussion it was noted that unless house price value at the study time period was available all properties and that a comparison of what value meant in different parts of the country could be made, the geographical location of the property, size of the dwelling and number of habitable rooms were more relevant in terms of environmental characteristics. It was also deemed that the English indices of deprivation 2019 score (IMD-19) would indicate the socio-economic status of the household occupants. This is a score which measures the relative deprivation of different small geographical areas of England of around 400–1200 people [34]. It considers income, employment and barriers to housing – including property price in a composite score [35]. Dwelling price and occupation were therefore removed from the DAG leaving 16 key nodes.

The remaining16 key factors in the three key areas of transmission risk and non-directional links were visualised, prioritised and agreed upon. Table 1 highlights the final 16 factors chosen and how they fit into one or more key areas of transmission risk.

3.2. DAG development

Fig. 1 shows the base DAG (without node type allocation) developed in DAGitty 3.0 This DAG shows how each of the 16 factors listed in Table 1 is represented as a node and how the reach, connectedness and directionality of relationship assumptions between nodes was agreed visually by adding pathways. Table 2 highlights the rational for the directionality of the relationships between nodes as agreed by the MDT. Where no rational is shown between nodes no probabilistic causal pathway was identified in the literature or from expert knowledge. Fig. 1 then highlights how the rational in Table 2 is used to create the directional pathways for the DAG.

The DAGitty R code to reproduce this is available in Supplement 2.

3.3. Data source selection

The base DAG visually demonstrated the directionality of relationship assumptions and the pathways between the 16 key factors by using pathways and nodes. We were then able to use the expert knowledge of the MDT, and in reference to the study time period (2020), identify the optimal data source from those potentially available for each factor. Table 3 shows the data sources identified for each variable. Table 3 also shows the node name given and if no data source was identified the variable remains in the DAG, labelled as unobserved.

We noted that data for three of the nodes could only be collected with a targeted and specific prospective or retrospective survey (Table 3). These factors were case and household contact use of appropriate handwashing facilities, compliance with self-isolation and ventilation practices. Collection of this data prospectively was not undertaken as part of the routine First Few 100 (FF100) or Contact Tracing and Advisory Service (CTAS) datasets and was not feasible for our study within our time frame, or budget, given that contact tracing for COVID-19 ceased in England in February 2022. These datasets were the complete list of all COVID-19 cases and contacts notified to Public Health England (PHE) and contained details about the case, their contacts, and the onset and duration of illness. They did not capture information about behaviours. The FF100 ran during February and March 2020 and later became CTAS after data on the first few cases had been collated. These nodes were therefore identified as unobserved nodes in the DAG.

Table 1
Final 16 factors for inclusion as nodes, with outcome node and exposures in the three sub-domains of transmission risk.

Viral transmission dynamics	Household and behavioural transmission factors	Environmental dwelling characteristic transmission factors
• Exposure to primary case during infectious period • Access to appropriate handwashing facilities	• Socioeconomic status • Dwelling tenure • Self-isolation behaviours	• Energy efficiency of dwelling • Age of construction • Dwelling type • Dwelling volume • Number of storeys • Number of habitable rooms
Household composition: • Intergenerational composition and household demographics (case & contact age, gender, ethnicity) • Household size		• Household density (number of people by dwelling space) • Ventilation (natural strategies, mechanical strategies and occupant behaviour related to ventilation)
Outcome: Becoming a secondary case through household contact of a primary case		

3.3.1. Prospective self-reported case data

The rationale for using data from February to March 2020 is described in the introduction, whilst data from this time period was transferred to the national CTAS dataset [65], the best and most comprehensive dataset on cases and their contacts from the start of the pandemic in England (February–March 2020) is the first few hundred (FF100) cases dataset, which is described in more detail in the HOSTED paper [9]. This dataset captured data of the demographics of the primary case and data on hospitalisation. This enabled us to determine if household contacts could have been exposed to the primary case. Exposure to the primary case during their infectious period is required for them to be defined as a contact.

The FF100 dataset also captured information on household contacts. Household size was self-reported by cases contemporaneously and accuracy of reporting was believed to be good at the time [9]. This dataset was therefore agreed by the MDT to be more reliable than data estimated from population datasets which may not account for households coming together/students returning home. Other details about contacts in the household was captured, in more detail than on CTAS. The dataset will enable us to determine if there are other co-primary cases (See section 2.1 for definitions) in the household and determine the age composition of the household.

3.3.2. Limitations of a prospective dataset focussing on cases

A limitation of this dataset is that the demographics of household contacts, other than age, were better captured for contacts who went on subsequently to become cases and less reliably for those who did not become a secondary case. This is a limitation of this dataset, but no other accessible retrospective datasets on the demographics of the household were identified.

3.3.3. Retrospective datasets used to inform data not self-reported prospectively by cases

Without a bespoke survey for cases, the nearest estimate of socioeconomic status that could be used was an ecological measure called IMD-19. This is described in more detail in section 3.1. This provides a rated composite measure in deciles of the level of deprivation for around 400–1200 houses in a defined small geographical area known as a Lower Super Output Area (LSOA) [34]. Other environmental dwelling characteristics needed to be captured for the specific dwelling cases and household contacts lived in. This therefore needed to use dwelling data held at the unique property level, which includes those datasets held by Ordnance Survey and the DLUHC including building age, building type, height and EPC as detailed in Table 3 above [39,60–63].

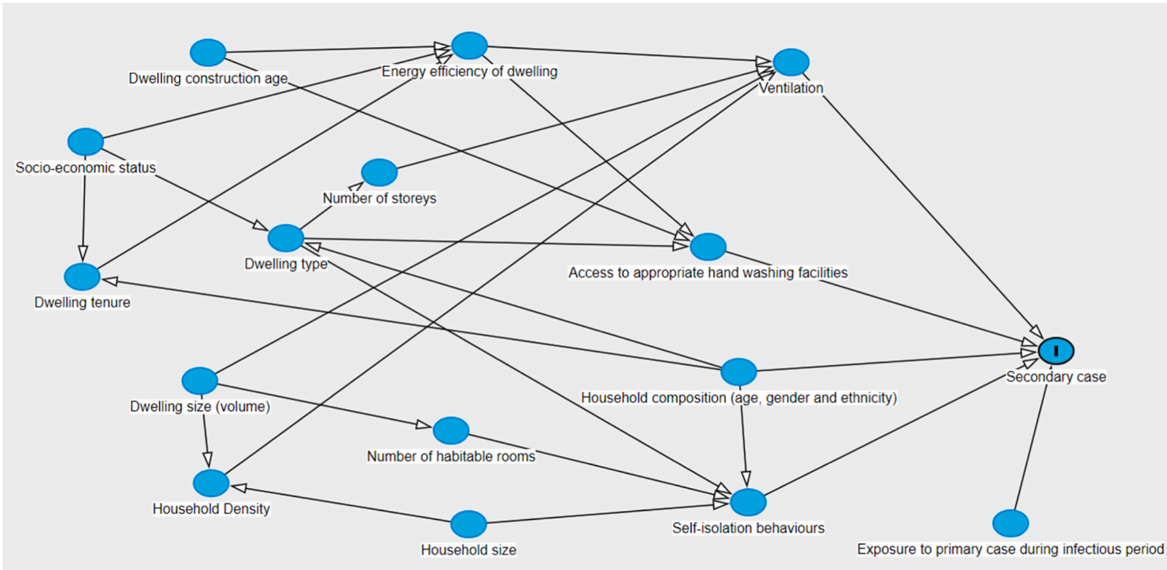

Fig. 1. Base Directed Acyclic Graph before data sources were identified and node types allocated.

3.4. Determining data linkage processes and feasibility

Part of the development of our pilot study protocol was then to determine if data linkage of these sources would be possible and what form of identifier could be used to complete this. The matching of public health surveillance and outbreak datasets is usually made to clinical hospital data using a hospital services number, but this is not possible for ecological and dwelling based datasets. Something relating to both the individual and the dwelling is therefore needed. An advantage of using the FF100 dataset is that the address of the case and contacts in the household was available. Each property in England is assigned a unique property reference number (UPRN), both at the building level, but also for each individual property in a larger building such as flats in an apartment block [66]. This can be used for data linkage to link the FF100 dataset to ecological and specific dwelling datasets. As Table 3 highlights access to the UPRN in each dataset is critical for data linkage between household data obtained from the FF100 and environmental dwelling characteristic datasets.

We were therefore able to note in the DAG development against the feasibility aims of this project, that if in the data collection stage of a pilot study, an overview of the feasibility of data linkage can be quickly established by the completeness of UPRN in the datasets. This was incorporated into the study protocol.

3.5. Allocating DAG node types

We were able to use the information we knew on each node to move from a base DAG (Fig. 1), to one where the nodes are labelled (Fig. 2) using the node allocation process described in the methods. We highlighted the unobserved nodes in the DAG, which are shown in Fig. 2 in white, and we were then able to visualise and identify the six key environmental dwelling exposure nodes (construction age band, energy efficiency rating, number of storeys, number of habitable rooms, dwelling type and dwelling volume). The six exposure nodes were determined as those most representative of environmental characteristics of dwellings and these are highlighted with an additional bold circle around the node. Other nodes were determined to be more linked to viral transmission dynamics, or the characteristics of the dwelling occupants (the household). This determined the structural inference of thirteen causal and three biasing pathways and allowed us to allocate node types to the remaining observed nodes.

In Fig. 2, the final DAG, household composition was identified as a confounder in pink with biasing pathways of dwelling type and the outcome of being a secondary case. Proxy confounders, in a different shade of pink, were identified as housing tenure and household size. Housing tenure has a biasing path from household composition through itself to energy efficiency and the outcome, whilst household size has biasing pathways through household density and the unobserved ability to self-isolate node to the outcome. Possible interactions were identified in the literature and from expert knowledge, and shown in turquoise, between type of dwelling and number of storeys (for example bungalows are always single level dwellings), and between dwelling volume and the number of habitable rooms. Household density was identified as a collider, shown in purple, with two arrowheads pointed towards it from household size and dwelling volume. IMD-19, shown in grey, was identified as an ancestor of two exposure nodes, both energy efficiency rating and dwelling type.

3.6. Using the DAG to inform the analysis plan

Identifying node types enabled the MDT to determine how to account for each specific node in the analysis plan. Dwelling volume, construction age, geometric form (dwelling type), number of storeys (as a marker of ventilation potential across floors/stack ventilation), number of habitable rooms, and energy efficiency were identified as the key environmental exposure nodes of interest. Dwelling age was identified as an ancestor of exposure of energy efficiency (a composite marker of the materials used to build the dwelling and heat retention of the dwelling including fuel and insulation). Therefore, as exposures on the same causal pathway either or both may be needed in the final model. Likewise dwelling type was identified as an ancestor of the exposure number of storeys (the number of levels of the dwelling) so as well as exploring the potential for any interaction terms for this pair in the final model design, the need for one or both should also be explored. This is also true for potential interaction between dwelling volume and number of habitable rooms.

Household composition was identified as a direct confounder, and housing tenure and household size as proxy confounders. The analysis plan developed for the pilot must therefore ensure these factors are adjusted for in the final model. We must also consider that household density is a collider of volume and household size. Therefore, it may be that one of volume or a measure or volume and household size are more appropriate to include than density alone with neither of these other factors. Deprivation was also identified as an ancestor of exposure and

Table 2
Description of rational for pathways shown in the base DAG.

Node from	Directionality of pathway	Node to	Rationale
Age of Construction	Towards outcome	Energy Efficiency	The age of construction of a property impacts on the construction materials used [36]. This is the single biggest factor in predicting the energy efficiency of dwelling and therefore, energy efficiency assessment made [37]. All those built after 2007 will take account of EPC requirements.
Age of Construction	Towards outcome	Access to appropriate handwashing facilities.	External ablutions and plumbing was common place until 1891. By 1949, 50% of dwellings had internal plumbing and by 1967 and 1991 respectively 25% and 1% of dwellings lacked an internal bath, shower, toilet or sink [38].
Energy Efficiency	Towards outcome	Access to appropriate handwashing facilities.	The need to have an EPC undertaken requires houses to have internal plumbing in order to be assessed for efficiency of hot water system [39]. Therefore, homes with a higher EPC assessment grade will more likely have accessible hot water.
Energy Efficiency	Towards outcome	Ventilation	The energy efficiency of a dwelling demonstrates the building fabric air tightness levels of a dwelling [39]. This impacts on how likely an occupant is to utilise mechanical ventilation or adopt behaviours to ventilate the dwelling based on how draughty the building is, where airflow is directed if windows are opened and if trickle vents are present [40]. Understanding of how EPC impacts occupants window opening behaviours is limited but some models have shown a probabilistic causal impact and is therefore included [41,42].
Ventilation	Towards outcome	Becoming a secondary case	The level and directionality of ventilation from physical, mechanical and behavioural factors will determine the number of air changes in a set time period and where areas of air stagnation occur. This will impact on how likely it is that a contact will be exposed to the SARS-CoV-2 virus inside the dwelling [43–46].
Access to appropriate handwashing facilities.	Towards outcome	Becoming a secondary case	Hand hygiene is a critical part of breaking the transmission cycle of any virus, including COVID-19 [47]. Access to appropriate hand washing facilities, including soap and hot water and separate facilities for the case and contacts is a key public health control measure for preventing onwards transmission.
Socio-economic status	Towards outcome	Energy Efficiency	Those with increased spending power may have newer and more energy efficient homes and where dwellings are older may have afforded to undertake more maintenance and repair or retrofit work to improve the energy efficiency of the property. This is noted to create inequalities in the energy efficiency and consumption in high income countries, such as the UK [48].
Socio-economic status	Towards outcome	Dwelling type	The socio-economic status of the household will determine the type of housing. Those with increased spending power may opt for larger detached dwellings which are more expensive compared to small, attached dwellings in a similar area. Those with more spending power also typically opt for whole units (houses or bungalows) compared to apartments and flats.
Socio-economic status	Towards outcome	Dwelling tenure	The socio-economic status of the household will determine the type of housing. Those with increased spending power are more likely to own their own home than rent [49].
Dwelling tenure	Towards outcome	Energy Efficiency	All dwellings which are rented or have been sold since 2007

(*continued on next page*)

Table 2 (*continued*)

Node from	Directionality of pathway	Node to	Rationale
			must have an EPC and more recent legislation since 2018, requires minimum standards for rented properties so energy efficiency should start increasing in rented properties [50].
Dwelling type	Towards outcome	Number of storeys	The type of dwelling will impact the number of storeys. A bungalow will almost always only have one recorded storey. Houses have two or three storeys, but may have basements. Flats are typically only one storey themselves but could be part of an apartment block and storey level is likely to be higher.
Number of storeys	Towards outcome	Ventilation	The number of storeys of a dwelling affects the flow of air up stairways or elevator shafts. Stack ventilation affects the way in which a property is ventilated and the routes of airflow through a dwelling [51].
Dwelling type	Towards outcome	Access to appropriate handwashing facilities.	The type and number of bathrooms available can be impacted by the type of dwelling, with flats and bungalows typically having less bathrooms than houses. Access to separate handwashing facilities for cases and contacts can therefore be less likely in such dwellings.
Dwelling type	Towards outcome	Self-isolation behaviours	The geometric layout of the house can impact on the likelihood of individuals to be able to self-isolate from other members of a household dwelling effectively.
Household composition	Away from outcome	Dwelling Type	The composition of a household influences dwelling type. A household composition of a couple over 65 is more likely to be resident in a bungalow than other household compositions [52]. Whilst a family with children is more likely to be resident in a house or larger dwelling.
Household composition	Away from outcome	Dwelling Tenure	Household composition with younger adult pairings is more likely to be renting than other household compositions. Ethnicity of the household can also impact on the likelihood of renting or owning a property.
Household composition	Towards outcome	Self-isolation behaviours	Households with children or significantly older occupants are likely to find it more difficult to adequately self-isolate as children and older adults may require support with activities of daily living or not understand the need to self-isolate.
Household composition	Towards outcome	Becoming a secondary case	Different household compositions have been shown to impact on the likelihood of becoming a secondary case [HOSTED]. The greatest risk of transmission was seen in older (over 65 years) pairs compared to adult (18–64 year old) pairs. The lowest risk was in houses with more than three adults, whilst in those with children, those houses with children and only one adult compared to two or more adults had higher secondary attack rates as the adult has physical closeness in providing care to other members of the household.
Self-isolation behaviours	Towards outcome	Becoming a secondary case	Well adhered to self-isolation will result in reduced likelihood of contacts within a household becoming a secondary case [53].
Exposure to primary case during infectious period	Towards outcome	Becoming a secondary case	The longer the duration of exposure the more chance of transmission and becoming a secondary case.
Dwelling size	Towards outcome	Ventilation	Taller properties have increased opportunity for circulating air to be above head height. Properties with a larger room volume also have a greater volume of air within which SARS-CoV-2 particles can be dispersed. This will impact the way in which air flows around the dwelling and the rate at which air

(*continued on next page*)

Table 2 (*continued*)

Node from	Directionality of pathway	Node to	Rationale
			changes per hour occur.
Dwelling size	Towards outcome	Number of habitable rooms	The number of habitable rooms will be impacted by the size of the property. There is a limit to how small a room can be and the larger the dwelling, the more potential for more rooms there is.
Number of habitable rooms	Towards outcome	Self-isolation behaviours	The number of habitable rooms determines whether it is possible for an infectious individual to be in a different living space to other contacts within the household and the likelihood of interaction.
Dwelling size	Towards outcome	Household density	Household density is determined by dwelling and household size. The greater the volume (size) for the same number of people the lower the density.
Household size	Away from outcome	Household density	Household density is determined by dwelling and household size. The greater the volume (size) for the same number of people the lower the density.
Household density	Towards outcome	Ventilation	The density of household members to volume of the property determines how much air is available to each individual [54]. Air quality may be negatively affected in overcrowded and highly dense environments [55].
Household size	Towards outcome	Self-isolation behaviours	The number of people in a household will determine how likely it is that the case comes into contact with other members of the household, be that for one or all members of the household.

may need to be considered as a random effect in a multilevel model. However, the use of households as the unit of measure, means that the household will already be used as a random effect and therefore may negate the need for other less specific and more ecological adjustments to be made. This should be tested by considering the impact deprivation has in a multilevel model on variance. Alongside using the DAG to understand how to treat each node in the analysis plan, the minimum sufficient adjustment sets can be identified in DAGitty 3.0 and used to inform the analysis plan. This was considered for inclusion of all six exposure nodes in a model, and by focussing on the impact of an individual exposure node. These minimum sufficient adjustment sets are shown below in Table 4.

3.7. Sensitivity analysis to determine the utility of the DAG in the pilot study

In using the DAG to determine if it had utility in informing our study design, protocol development and analysis plan for a pilot study we wanted to see how this method of model building compared to forwards and backwards stepwise regression. We therefore agreed to include models using these methods in the analysis plan to determine the utility, reliability and feasibility of using DAGs in such MDT pieces of work in the future.

4. Discussion

4.1. The DAG as a visual tool for consensus building and conceptualising a framework

We showed that both the process of developing a DAG and then the DAG itself were useful for study design and data collection. The process of developing the DAG enabled the complex relationships between transmission risk factors to be described and displayed visually. This included the directionality of assumptions being made. As a dispersed team, reliant on digital methods of communication, the ability to rapidly adapt the DAG in real time enabled transparent and open discussions and consensus building on both the nodes to be included and the directionality of pathways. It allowed specialists from all areas of the multi-disciplinary team to be involved. By agreeing on the design of the final DAG, we are able to show to other epidemiologists, who may want to use or build on this model, the final thought process we used in a transparent way. This is important as there may have been selection biases incorporated into our DAG in node selection by the knowledge and expertise brought from the MDTs own fields of specialist knowledge. This is a fallibility in DAGs and other study designs that has previously been noted [25]. Our use of sensitivity analyses in the analysis plan will help us to identify any such biases. However, the use of a DAG allows adaptations based on new knowledge to transparently be incorporated. This is particularly important when dealing with a relatively new hazard where not everything is known about it and there is a degree of uncertainty in causal inferences.

The process itself enabled frank discussion and agreement on the prioritisation of risk factors to be made. Whilst this is not the purpose for which DAGs were developed in the computer industry, their use in informing consensus has been described as one of the benefits of utilising DAGs in applied health research [67]. This example shows the utility of DAGs in complex multidisciplinary fields for supporting consensus building. Identifying tools which support our analytical decision making and improve transparency and consensus building is important as we move into a new normal of hybrid working.

Using the DAG to build the conceptual framework also enabled the study team to describe and successfully prioritise key factors and relationships. This was a key part of the process in reducing the key variables from 34 to 16, ensuring those factors which were less significant, duplicated or which were identified as unobserved with no likely causal effect already captured in the model were removed. This is important to ensure that we are not including redundant variables and data and reduce the possibility of finding an effect by chance. However, there is a debate and argument that some of these should have been left in as unobserved variables in the model. There is a risk that by doing so, or in the MDT's own selection biases, we could have missed an important variable elsewhere on the causal or biasing pathway. We tried to overcome this by ensuring a wide range of expertise was involved in the DAG development process and all had the opportunity to challenge and make additions or removals.

4.2. The DAG as a tool to identify data sources

The DAG was used to consider what sources of data were available

Table 3
Data source selection and combinations from data considered.

Transmission Outcome	Data sources considered	Combinations considered	Data source selected	Data linkage	Node name
Secondary case following household contacts	• FF100[a] (Jan–Mar 2020) & CTAS[b] (Mar 2020–Feb 2022) – Details of household contacts identified by primary cases • Epidemiology cell line list – Details of COVID-19 cases through notifications and laboratory records	• Co-primary • Secondary case • Not a secondary case	Linelist & FF100	FF100 contact details & Linelist case details.	Secondary case
Transmission Risk Factor	**Data sources considered**	**Combinations considered**	**Data source selected**	**Data linkage**	**Node name**
Socio-economic status	• Bespoke prospective or retrospective survey • Index of Multiple Deprivation score 2019 (IMD-19) [35] - ecological level data at lower super output area (LSOA)[c]	A composite of income, employment, education, health, crime, barriers to housing and services and the living environment averaged for a population of around 400–1200 people	IMD-19	Unique Property Reference Number (UPRN) [56] from FF100 address	IMD-19 by LSOA
Dwelling tenure	• Energy Performance Certificate (EPC)[d] data for all houses built, sold or rented since 2007 (therefore not all housing stock is included)	• Rented (privately or social renting) • Privately owned	EPC	UPRN from FF100 address	Dwelling tenure
Household size	• Self-reported on FF100[a]/CTAS[b] • National Population dataset (an estimate of household size, however self-reported figure should be prioritised) [57] • Pop24/7 population dataset– a spatiotemporal estimate of household size [58,59]	Number (children and adults counted equally)	FF100		Household size
Household composition	• FF100[a]/CTAS[b]/Epidemiology cell line list primary & secondary case demographics (age, gender, ethnicity). • Bespoke survey would be required for contact demographics.	• Adult pair, older pair, >3 adults, single parent with children, 2 parents with children, multigenerational • Age, gender and ethnicity of primary case and contacts	Linelist & FF100		Household composition
Exposure during primary case infectious period	• Hospitalisation of primary case variably reported on FF100. • Hospital Episodes Statistics (HES)-hospital admissions. Requires additional approvals & NHS Number for linkage [60].	Dates of hospitalisation compared to dates of infectious period	FF100		Primary case hospitalisation
Dwelling size (volume)	• EPC • Ordnance Survey (OS) floor area attribute (not available in study period) could be combined with height to estimate volume	Measure in m^3	EPC	UPRN from FF100 address	Dwelling volume
Dwelling construction age	• EPC • Ordnance Survey building age attribute [61]	Age bands as defined by OS	1st EPC; 2nd OS building age	UPRN from FF100 address	Construction age band
Dwelling type	• Ordnance Survey Basic Land and Property unit (BLPU) reference database [62]	• Houses, bungalows, apartments and flats • Detached, semi-detached and terraced properties	BLPU	UPRN from FF100 address	Dwelling type
Energy efficiency of dwelling	• EPC	A composite of building materials, window type and number, energy efficiency of walls, roof and windows, type of heating, hot water and lighting	EPC	UPRN from FF100 address	Energy efficiency rating
Number of habitable rooms	• EPC	Number (not including kitchens and bathrooms)	EPC	UPRN from FF100 address	Number of habitable rooms
Number of storeys	• EPC – gives exact number of storeys at time of assessment • Ordnance Survey building height attribute – storeys calculated from building height using a geometric model [63,64].	Number of flights of stairs	1st EPC; 2nd OS building height	UPRN from FF100 address	Number of storeys
Household density	• A calculation from household size and volume or number of habitable rooms.	Calculated as number of people/volume in m^3	FF100 & EPC		Household density
Ventilatory behaviours	Bespoke prospective or retrospective survey No retrospective data sources available	• Window opening frequency, duration and combinations • Use of mechanical ventilation (fans/air conditioning) and types and positions of mechanical ventilation			UNOBSERVED node
Access to appropriate hand washing facilities	Bespoke prospective or retrospective survey No retrospective data sources available	• Access to separate facilities • Access to hot water and soap • Able to use facilities without support			UNOBSERVED node

(continued on next page)

Table 3 (*continued*)

Transmission Outcome	Data sources considered	Combinations considered	Data source selected	Data linkage	Node name
Self-isolation behaviours	Bespoke prospective or retrospective survey No retrospective data sources available	• Frequency and duration of compliance with self-isolation guidance • Self-isolation for all or only some activities of daily living (sleeping, self-care, meals)			UNOBSERVED node

[a] First Few 100 cases dataset [9].
[b] Contact Tracing and Advisory Service for England [65].
[c] IMD-19 comprising of income, employment status, education, health, crime, barriers or housing and services and living environment [35].
[d] EPC rating is a composite measure of how efficient by volume a dwelling is at using energy produced. If considers the materials used to build the property, the types of windows used, the source of heating and hot water, and the size of the dwelling. It is held by the UK Department for Levelling Up, Housing and Communities (DLUHC) [39].

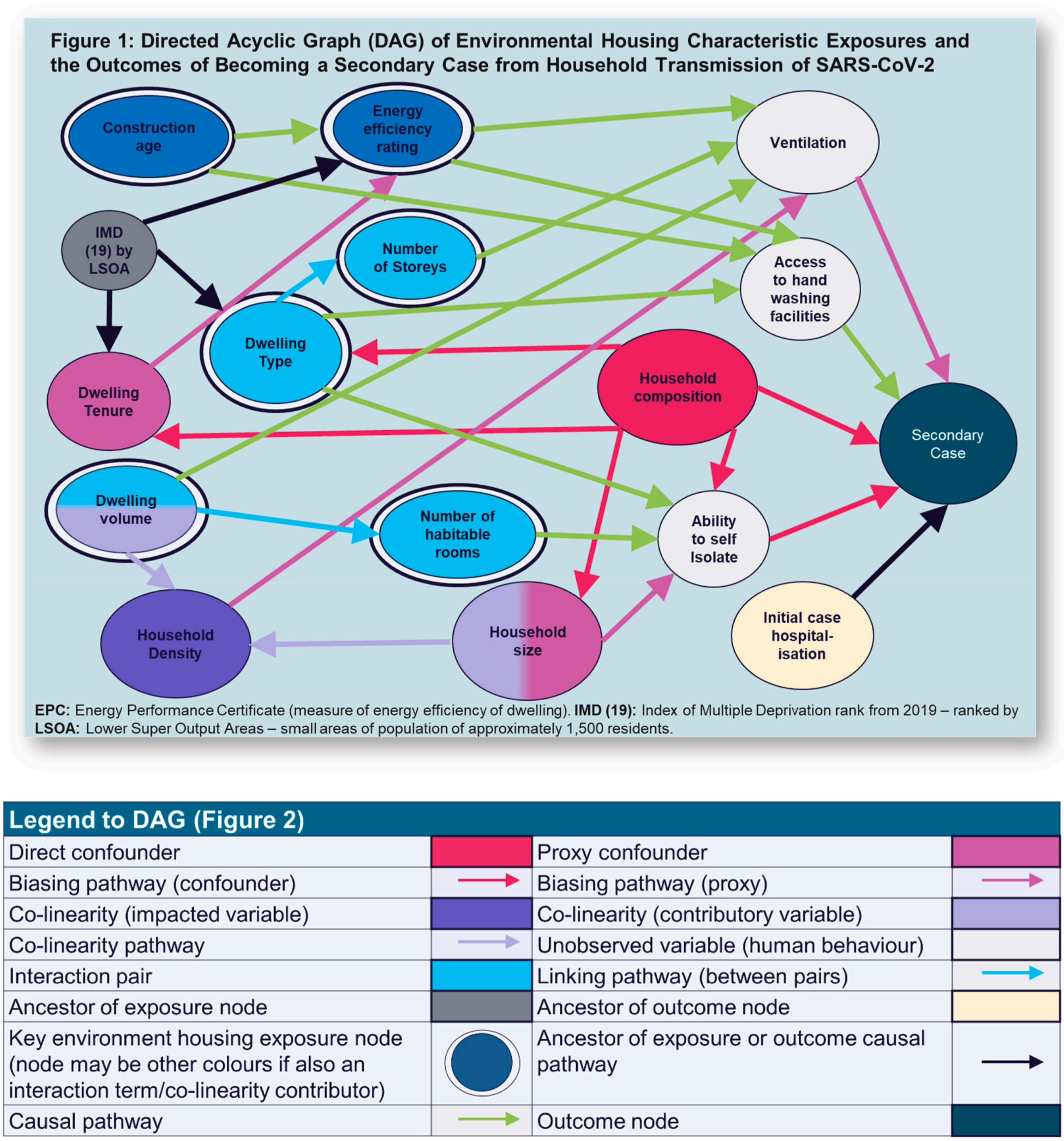

Fig. 2. DAG with legend identifying how nodes were classified.

for each of the nodes and to select the optimum variables and data sources for the time period of the pilot study. To collate sufficient data on households during COVID-19 focussing on environmental dwelling characteristics for the purposes of a pilot, we wanted minimal impact from changing transmission dynamics of community circulation and changing variants. We also wanted to use existing disease and dwelling

Table 4
Minimum sufficient adjustment sets identified from the DAG.

Key exposure node(s) considered	Minimal sufficient adjustment set	Comments
• Dwelling age • Energy efficiency • Number of storeys • Dwelling type • Dwelling volume • Number of habitable rooms	• Household size • Household composition	Requires EPC dataset (which has missing data) for energy efficiency, number of storeys and number of habitable rooms. If insufficient data on household composition, will need data on dwelling tenure.
• Dwelling age	DIRECT effect predicted. No adjustment necessary.	
• Dwelling volume	DIRECT effect predicted. No adjustment necessary.	
• Energy efficiency	• Dwelling age • Dwelling type • Household composition. May in addition need socio-economic status and geographical location.	Dwelling age is an ancestor of energy efficiency exposure and could be used in its place as a simpler and more direct model. Requires EPC dataset. If insufficient data on household composition, will need dwelling tenure and household size.
• Dwelling type (geometric form)	• Dwelling age • Energy efficiency • Household composition. May in addition need socio-economic status and geographical location.	Requires EPC dataset. If insufficient data on household composition, will need dwelling tenure and household size.
• Number of storeys	• Dwelling type OR • Dwelling age • Household composition	Requires EPC dataset for number of storeys.
• Number of habitable rooms	• Dwelling volume	Requires EPC dataset for number of habitable rooms and to calculate volume. Volume can be used in lieu of number of habitable rooms as an ancestor of number of habitable rooms.

datasets from a time period with only one variant, with limited community transmission (due to national public health lockdowns) allowing the datasets to be linked. We, therefore, chose a retrospective study design, using the DAG to identify the variables we did not have data for (unobserved variables) and identify the necessity to use nodes higher up the causal chain (ancestors of exposure) in the analysis. These unobserved behavioural factors may have had a more significant impact on the outcome than the ancestor variable chosen and therefore there is a risk that we may over-estimate the effect size of an ancestor variable.

As we developed this DAG for a specific time period, if it were to be repeated for a larger study across a longer time period, some of those variables we dismissed may need to be included and other variables relating to changing policies and guidance, levels of community transmission, variants and vaccination status may need to be added to the DAG and conceptual model. The use of composite measures in place of specific ones, based on the data available, may also mask the specific effect of one variable which contributed to the composite measure. This may result in further studies being required to identify if a specific element of the composite result is driving any effect seen. This DAG does however, provide a framework to build from which focuses on the environmental dwelling characteristics.

4.3. Using the tool to inform study design and the analysis plan

From the remaining 13 observed nodes, one of which was the outcome node, we were able to use the DAG to rapidly identify the type of node of each risk factor and the directionality of relationships. This enabled us to identify confounders, collinear and interaction terms and thereby inform the analysis plan based on evidence and expert consensus *a priori*. In considering all of the exposure nodes, household composition must be adjusted for *a priori*, and if insufficient data is available, one or both of the two proxy confounders identified (housing tenure and household size) will need to be adjusted for. Household density is causally influenced by dwelling volume and household size, so to prevent collinearity cannot be included in the same model. Interaction terms for number of habitable rooms and dwelling volume and for dwelling type and number of storeys should be tested in model development.

Dwelling age and dwelling volume were identified as having direct effects on outcome. These were identified as ancestors of exposure of energy efficiency and number of habitable rooms, respectively. It may therefore be more appropriate to use dwelling age over energy efficiency and dwelling volume over number of habitable rooms in model building. One risk of this, is that if a dwelling has undergone an energy or building material retrofit, the measure of energy efficiency may be more reliable from an EPC than from dwelling age. Likewise, dwelling type and number of storeys are on the same causal pathway. The choice of whether to use dwelling type or number of storeys may need to be made during model development. Use of the DAG is therefore not only useful in developing the analysis plan, but also in selecting the most appropriate variables during the model building and testing phase. However whilst DAGs can help support the analysis plan they are not an analysis approach and do not replace the need for testing analytical models and making decisions about which variables to choose, particularly where a choice or the potential for interaction terms is presented in a DAG.

4.4. Summary of DAG utility

Developing a DAG and gaining consensus on it can be time-consuming. As Rodrigues also found, this process therefore does not always result in one DAG, but a number of DAGs [68]. However, as well as its uses in consensus building, it can also serve many other purposes and be of benefit to the rigour of a study. This study has confirmed that DAGs can be used effectively to prioritise key risk factors for transmission across a multidisciplinary field, to develop consensus on a conceptual framework and then to inform data source selection, study design and analysis, whilst also ensuring transparency throughout the process. This DAG had practical utility and was used to inform the protocol for a pilot study in which we will test this as a conceptual framework. Based on the results of the pilot study, including the sensitivity analysis we will adjust the conceptual framework as needed for future studies.

Any changes to the DAG can be described to ensure a robust and transparent process for a larger, population level study which would be based on the same conceptual framework. The need to sometimes create and test multiple alternative DAGs or models, particularly when dealing with uncertainty, is widely recognised [25]. The role of using DAGs as an opportunity for researchers to be transparent, particularly when dealing with high levels of uncertainty and potential error, is highlighted by Ellison [69].

We have shown the benefit of using a DAG to develop a study protocol for this particular analytical pilot study with complex causal pathways and multiple interdependent exposures. This reflects the work of others who have shown the benefit of DAGs in informing the conduct of other observational studies [70]. However, DAGs are not always necessary for study design and analysis plan development. Using a DAG to test a simple hypothesis may not be necessary and can be a time-consuming step which may not be appropriate, for example, in acute outbreak response. There is also a time element in learning to develop and then use a DAG. Not all epidemiologists are trained to use DAGs. This can therefore be a limitation in the willingness of teams to design or then use them. Lack of knowledge of DAGs has previously been reported as a barrier to their use in research [71]. Learning about DAGs has now been incorporated into the Operational Research module of the UK Field Epidemiology Training Programme [72]. Other researchers have highlighted similar limitations of DAG usage in applied health research [67].

4.5. Recommendations for use of this DAG in the pilot

- The pilot study protocol developed from this work should be delivered upon and the utility of the DAG tested in informing the analysis plan and final model reviewed against the sensitivity analysis models.
- The utility of the DAG in identifying the data sources needed and the need for UPRN for data linkage should be reviewed throughout the conduct of the pilot.
- The utility of this DAG for a wider study should be highlighted at the conclusion of the pilot study with recommendations for adaptations to the DAG based on a longer time period made clear.

4.6. Recommendations for DAGs in other epidemiological studies

- DAGs are a useful tool in building agreement on a conceptual framework, in this example, for disease transmission risk. DAGs should be used in other multidisciplinary epidemiological studies, particularly when working remotely, to visually display ideas and gain consensus.
- Study teams exploring complex patterns of disease transmission risk, particularly where expertise from different fields relating to the pathogen, human behaviours, dwelling and environmental factors is needed, should consider using a DAG to describe and prioritise the relationships to explore.
- We recommend that DAGs are used by epidemiological study teams to contribute to the development of analytical studies with complex causal pathways and multiple interdependent exposures, both at the pilot and final study phases. DAGs produced for pilot studies can be updated based on the findings of the pilot.
- DAGs should be used in study designs using multidisciplinary multi-faceted analytical studies.
- The development and use of DAGs should continue to be taught, or added as a learning outcome, to relevant field epidemiology programmes. This will expand workforce knowledge, skills and expertise.
- DAGitty is a useful and accessible tool that is freely available as a browser-based environment and does not require any specialist hardware or software in its use.

5. Conclusions

5.1. Outcomes against aim and objectives

We were able to meet our aim and develop a DAG that can be used as a conceptual framework for a pilot study. The DAG has supported us in developing the protocol and analysis plan for a retrospective pilot study exploring the associations between environmental dwelling characteristics and household transmission of COVID-19.

We were able to pursue a MDT approach to design the DAG and use it to explore and visualise complex relationships, support optimal data source selection for the time period of the study and highlight the need for UPRN for data linkage.

5.2. Further goals

We now plan to utilise the study protocol developed from this DAG in undertaking the pilot study and to report the findings of that pilot both in terms of and potential associations identified, but also the utility of the DAG development process and using a DAG to inform this type of study. We also plan to support teaching programmes within UK Field Epidemiology Training Programmes on the use of DAGs.

5.3. Wider applicability of this work

Whilst we have focused on how the development of the DAG supported the development of our pilot study protocol, our learning can be used for future development of studies on this topic, and is applicable to other multidisciplinary epidemiological studies. The process of creating a DAG enabled the study team, which cut across multiple specialities, to work together to identify and prioritise key variables and relationships for interactions. DAG development was a useful tool for our MDT to build a conceptual framework and transparently form consensus on variable selection. The DAG itself was a way of visualising and formalising the complex relationships between the multiple variables included. By visualising relationships, the study team were able to identify variables which are key to the conceptual framework and *a priori* variables which are confounders, collinear terms and potential interactions. These *a priori* classifications informed the pilot study analysis plan, including the confounders and collinear variables to be compared in separate models and interaction terms considered in analyses. We would therefore recommend that other study teams consider using a DAG in their study design development process, particularly when the topic is complex, requires input form multidisciplinary subject matter experts and where there is a degree of uncertainty about causal inference and pathways.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The organisation worked for by Gavin Debrera, Public Health England, received an unrestricted grant from GSK to undertake a study on the outcome of patients who received parenteral zanamavir. The funder received data and interim reports from PHE but did not influence analysis and reporting of the study. Gavin Dabrera had no involvement in the GSK-funded study on parenteral zanamavir. Furthermore, the currently submitted work was part of the public health response activities to COVID-19 and had no relationship to GSK nor the study on parenteral zanamavir.

Data statement

The R code used for the design of the DAG is available in supplement 1. No data was used in the development of the DAG.

CRediT authorship contribution statement

Hannah Taylor: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Conceptualization. **Helen Crabbe:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Conceptualization. **Clare Humphreys:** Writing – review & editing, Supervision, Investigation, Conceptualization. **Gavin Dabrera:** Writing – review & editing, Supervision, Conceptualization. **Anna Mavrogianni:** Writing – review & editing, Methodology, Conceptualization. **Neville Q. Verlander:** Writing – review & editing, Methodology, Conceptualization. **Giovanni S. Leonardi:** Writing – review & editing, Methodology, Conceptualization.

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.

Data availability

No data was used for the research described in the article.

Acknowledgements

We would like to acknowledge the support of Alicia Barassa, UK FETP Front Line Co-ordinator for her support with this piece of work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.buildenv.2023.111145.

References

[1] UKHSA, Coronavirus (COVID-19) in the UK, June 2023. Available from: https://coronavirus.data.gov.uk/.
[2] PHE Transmission Group, Factors Contributing to the Risk of SARS-CoV-2 Transmission Associated with Various Settings, Dec 2020. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/945978/S0921_Factors_contributing_to_risk_of_SARS_18122020.pdf.
[3] M. Shen, Z. Peng, Y. Guo, L. Rong, Y. Li, Y. Xiao, et al., Assessing the effects of metroplitan-wide quarantine on the spread of COVID-19 in public spaces and households, Int. J. Infect. Dis. S1201–9712 (2020) 30326–X.
[4] Z. Wang, W. Ma, X. Zheng, G. Wu, R. Zhang, Household transmission of SARS-CoV-2, J. Infect. 81 (1) (2020) 179–182, https://doi.org/10.1016/j.jinf.2020.03.040.
[5] S. Denford, K. Morton, J. Horwood, et al., Preventing within household transmission of Covid-19: is the provision of accommodation to support self-isolation feasible and acceptable? BMC Publ. Health 21 (2021) 1641, https://doi.org/10.1186/s12889-021-11666-z.
[6] EMG and SPI-B, EMG and SPI-B: Mitigating Risks of SARS-CoV-2 Transmission Associated with Household Social Interactions, 29 Nov 2020. Available online: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/939073/S0922_EMG_and_SPI-B_-_Mitigating_risks_of_SARS-CoV-2_transmission_associated_with_household_social_interactions.pdf.
[7] SAGE Ethnicity Subgroup, Consensus Statement: housing, household transmission and ethnicity, Available online: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/943178/S0923_housing_household_transmission_and_ethnicity.pdf, November 2020.
[8] PHE, SARS-CoV2 Susceptibility and Transmission Risk in Children: an Overview of Current Evidence from Public Health England's Surveillance Work, August 2020. Available online: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/916891/phe-susceptibility-transmission-children-s0717-sage-53-200819.pdf.
[9] J.A. Hall, et al., HOSTED-England's Household Transmission Evaluation Dataset: preliminary findings from a novel passive surveillance system of COVID-19, Int. J. Epidemiol. (2021). Available online: https://pubmed.ncbi.nlm.nih.gov/33837417/.
[10] N.L. Boddington, et al., Epidemiological and clinical characteristics of early COVID-19 cases, United Kingdom of Great Britain and Northern Ireland, Bull. World Health Organ. 99 (3) (2021 Mar 1) 178–189, https://doi.org/10.2471/BLT.20.265603. Epub 2020 Nov 30. PMID: 33716340; PMCID: PMC7941108. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7941108/.
[11] E. Miller, et al., Transmission of SARS-CoV-2 in the household setting: a prospective cohort study in children and adults in England, J. Infect. (2021). Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8325949/.
[12] J. Lopez Bernal, et al., Transmission Dynamics of COVID-19 in Household and Community Settings in the United Kingdom, Medrix, 2020. Available online: https://www.medrxiv.org/content/10.1101/2020.08.19.20177188v1.
[13] H. Allen, et al., Household transmission of COVID-19 cases associated with SARS-CoV-2 delta variant (B.1.617.2): national case-control study, Lancet Regional Health (2021). Available online: https://www.sciencedirect.com/science/article/pii/S2666776221002386?via%3Dihub.
[14] A. Singanayagam, et al., Community Transmission and Viral Load Kinetics of the SARS-CoV-2 Delta (B.1.617.2) Variant in Vaccinated and Unvaccinated Individuals in the UK: a Prospective, Longitudinal, Cohort Study, Lancet Infectious Diseases, 2021. Available online: https://www.sciencedirect.com/science/article/pii/S1473309921006484?via%3Dihub.
[15] G. Dahlgren, M. Whitehead, Policies and Strategies to Promote Social Equity in Health, Institute for Futures Studies, Stockholm, Sweden, 1991.
[16] R. Labonte, Health promotion and empowerment: practice frameworks, in: Health Promotion and Empowerment: Practice Frameworks: Centre for Health Promotion, University of Toronto, 1993. Available online.
[17] Jones, B. Why space volume matters in OVID-19 transmission. Chartered Institution of Building Services Engineers. May 21. Available online: Why Space Volume Matters in Covid-19 Transmission – CIBSE Journal.
[18] L. Moses, K. Morrissey, R.A. Sharpe, T. Taylor, Exposure to indoor mouldy odour increases the risk of asthma in older adults living in social housing, Int. J. Environ. Res. Publ. Health 16 (14) (2019 Jul 22) 2600, https://doi.org/10.3390/ijerph16142600. PMID: 31336583; PMCID: PMC6679100.
[19] D. Caillaud, et al., Indoor mould exposure, asthma and rhinitis: findings for systematic reviews and recent longitudinal studies, Eur. Respir. Rev. 27 (148) (Jun 2018) 170137, https://doi.org/10.1183/16000617.0137-2017.
[20] Regulator of Social Housing, Damp and Mould in Social Housing: Initial Findings, UK Government, 2 Feb 2023. Available online: https://www.gov.uk/government/publications/damp-and-mould-in-social-housing-initial-findings/damp-and-mould-in-social-housing-initial-findings-accessible-version.
[21] Ingham, et al., Damp Mouldy Housing and Early Childhood Hospital Admissions for Acute Respiratory Infection: a Case Control Study, Thorax, 2019. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6824607/.
[22] Marmot Review Team, The Health Impacts of Cold Homes and Fuel Poverty. Friend of the Earth, May 2011. Available online: https://www.instituteofhealthequity.org/resources-reports/the-health-impacts-of-cold-homes-and-fuel-poverty/the-health-impacts-of-cold-homes-and-fuel-poverty.pdf.
[23] CDC, Fungal Diseases and COVID-19, Fungal Diseases and COVID-19 | CDC, March 2022. Available online:.
[24] R. Foraita, J. Spallek, H. Zeeb, Directed acyclic graphs, in: W. Ahrens, I. Pigeot (Eds.), Handbook of Epidemiology, Springer, New York, NY, 2014, https://doi.org/10.1007/978-0-387-09834-0_65.
[25] J.C. Digitale, J.N. Martin, M.M. Glymour, Tutorial on directed acyclic graphs, J. Clin. Epidemiol. 142 (2022 Feb) 264–267, https://doi.org/10.1016/j.jclinepi.2021.08.001. Epub 2021 Aug 8. PMID: 34371103; PMCID: PMC8821727. Available online: Tutorial on Directed Acyclic Graphs - PMC (nih.gov).
[26] T.H. Ciesielski, et al., Transdisciplinary approaches enhance the production of Translational knowledge, Transl. Res. (2016). Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5362296/.
[27] UKHSA, Effectiveness of Non-pharmaceutical Interventions to Reduce Transmission of COVID-19 in the UK A Rapid Mapping Review, UKHSA, Sep 2023. Available online: https://assets.publishing.service.gov.uk/media/65144556b1bad400144fd910/NPI_and_COVID_mapping_review.pdf.
[28] Institute for Government Analysis, Timeline of UK Government Coronavirus Lockdowns and Measures, March 2020 to December 2021, Dec 2021. Available online: timeline-coronavirus-lockdown-december-2021 (instituteforgovernment.org.uk).
[29] UKHSA. COVID-19 variants identified in the UK. Last updated 22 Sep 2023. Available online: https://www.gov.uk/government/news/covid-19-variants-identified-in-the-uk-latest-update.
[30] UKHSA. COVID-19 vaccination programme. Last updated 19 Sep 2023. Available online: https://www.gov.uk/government/collections/covid-19-vaccination-programme.
[31] Johannes Textor, Benito van der Zander, Mark K. Gilthorpe, Maciej Liskiewicz, T. George, H. Ellison, Robust causal inference using directed acyclic graphs: the R package 'dagitty', Int. J. Epidemiol. 45 (6) (2016) 1887–1894. Available online: https://pubmed.ncbi.nlm.nih.gov/28089956/.
[32] P. Kohler, T. Dörr, A. Friedl, et al., SARS-CoV-2 risk in household contacts of healthcare workers: a prospective cohort study, Antimicrob. Resist. Infect. Control 12 (2023) 98, https://doi.org/10.1186/s13756-023-01300-5.
[33] D. Martin, S. Platt, P. Hampshire, C. Meadows, COVID-19: an update from England's high consequence infectious diseases intensive care unit leads, J. Intensive Care Soc. 21 (2) (2020 May) 99–101, https://doi.org/10.1177/1751143720918517. Epub 2020 Apr 5. PMID: 32489403; PMCID: PMC7238469.
[34] ONS, Census, geographies (2021). Available online: https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeographies/census2021geographies.
[35] MHCLG, National Statistics English Indices of Deprivation 2019, 2019. Available online: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019.
[36] R. Vagtholm, A. Matteo, B. Vand, L. Tupenaite, Evolution and current state of building materials, construction methods, and building regulations in the U.K.: implications for sustainable building practices, Buildings 13 (2023) 1480, https://doi.org/10.3390/buildings13061480.
[37] ONS. Age of the property is the biggest single factor in energy efficiency of homes. ONS. Jan 22. Available online: https://www.ons.gov.uk/peoplepopulationandcommunity/housing/articles/ageofthepropertyisthebiggestsinglefactorinenergyefficiencyofhomes/2021-11-01.
[38] Nesta, The Rise of Indoor Plumbing in the UK: Lessons for the Green Transition, Sep 2023. Available online: https://www.nesta.org.uk/feature/stories-change/rise-indoor-plumbing-uk-lessons-green-transitio/.
[39] DLUHC, Find an Energy Certificate, May 2023. Available online: https://www.gov.uk/find-energy-certificate.
[40] Y. Gu, et al., Study on influencing factors for occupant window-opening behavior: case study of an office building in Xi'an during the transition season, Build. Environ. (2021), https://doi.org/10.1016/j.buildenv.2021.107977.
[41] Lie, Y et al. Rethinking the limitations of research on occupants' window-opening behavior: A review. Energy Build.. Dec 22; vol 277. https://doi.org/10.1016/j.enbuild.2022.112552.
[42] K. Cheng, J. Yao, R. Zheng, Energy performance of occupant behaviors on windows: a green building based study, Energies 16 (2023) 2209, https://doi.org/10.3390/en16052209.

[43] EMG, Role of fVentilation in COntorlling SARS-COV-2 Transmission, SAGE, 2021. Available online: https://assets.publishing.service.gov.uk/media/5f917153d3bf7f35e3900820/S0789_EMG_Role_of_Ventilation_in_Controlling_SARS-CoV-2_Transmission.pdf.
[44] Y. Lin, et al., A Systematic Review on COVID-19 Related Research in HVAC System and Indoor Environment, Energy and Built Environment, Aug 2023, https://doi.org/10.1016/j.enbenv.2023.07.009.
[45] HPSC, Guidance on non HCbuilding ventilation during COVID, HSE (Jan 2022). Available online: https://www.hpsc.ie/a-z/respiratory/coronavirus/novelcoronavirus/guidance/educationguidance/Guidance%20on%20non%20HCbuilding%20ventilation%20during%20COVID.pdf.
[46] EMG, Ventilation Actions Summary, SAGE, 2021. Available online: https://assets.publishing.service.gov.uk/media/5fdb7f62e90e074527d9fece/S0973_Ventilation_Actions_Summary_16122020_V2.pdf.
[47] WHO, Handwashing an Effective Tool to Prevent COVID-19, Other Diseases, WHO, Oct 2020. Available online: https://www.who.int/southeastasia/news/detail/15-10-2020-handwashing-an-effective-tool-to-prevent-covid-19-other-diseases#:~:text=%27Handwashing%20has%20always%20been%20one,cornerstones%20of%20COVID%2D19%20prevention.
[48] R. Galvin, M. Sunikka-Blank, Economic inequalities and household energy consumption in high-income countries, A Challenge for Social Science Based Energy Research. Ecological Economics 153 (2018) 78–88.
[49] J.J.B. Seng, Y.H. Kwan, H. Goh, et al., Public rental housing and its association with mortality – a retrospective, cohort study, BMC Publ. Health 18 (2018) 665, https://doi.org/10.1186/s12889-018-5583-6.
[50] Department for Energy Security and Net Zero, Domestic private rented property: minimum energy efficiency standard, Available online: https://www.gov.uk/guidance/domestic-private-rented-property-minimum-energy-efficiency-standard-landlord-guidance, April 2023.
[51] International Energy Agency. Ventilation and Infiltration Characteristics of Lift Shafts and Stair Wells. AVIC. Available online: Ventilation and infiltration characteristics of lift shafts and stair wells (aivc.org).
[52] MHCLG, Better Homes and Bungalows for Britain's Older People, 2015. Available online: https://www.gov.uk/government/news/better-homes-and-bungalows-for-britains-older-people.
[53] EMG, Reducing within- and Between-Household Transmission in Light of New Variant SARS-CoV-2, SAGE, 2021. Available online: https://assets.publishing.service.gov.uk/media/60019403d3bf7f33bc62466a/s1020-Reducing-within-between-household-transmission.pdf.
[54] J.G. Allen, A.M. Ibrahim, Indoor air changes and potential implications for SARS-CoV-2 transmission, JAMA 325 (20) (2021) 2112–2113, https://doi.org/10.1001/jama.2021.5053.
[55] Karl A. Holden, Alice R. Lee, Daniel B. Hawcutt, Ian P. Sinha, The Impact of Poor Housing and Indoor Air Quality on Respiratory Health in Children, vol. 19, European Respiratory Society, 2023 230058, https://doi.org/10.1183/20734735.0058-2023.
[56] Central Digital and Data Office, Identifying Property Ad Street Information, 2022. Available online: https://www.gov.uk/government/publications/open-standards-for-government/identifying-property-and-street-information.
[57] HSL, National Population Database, 2022. Available online: https://www.hsl.gov.uk/what-we-do/data-analytics/national-population-database.
[58] University of Southampton, Population 24/7, 2022. Available online: http://pop247.geodata.soton.ac.uk/.
[59] D. Martin, S. Cockings, S. Leung, Developing a Flexible Framework for Spatiotemporal Population Modelling Annals of the Association of American Geographers, 2015, https://doi.org/10.1080/00045608.2015.1022089.
[60] NHS Digital, Hospital episode statistics (HES), Available online: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics, March 2023.
[61] J.F. Rosser, D.S. Boyd, G. Long, et al., Predicting residential building age from map data (3 more authors), Comput. Environ. Urban Syst. (2018), https://doi.org/10.1016/j.compenvurbsys.2018.08.004. ISSN 0198-9715.
[62] OS, OS Open Linked Identifiers Relationships, 2023. Available online: https://www.ordnancesurvey.co.uk/business-government/tools-support/open-linked-identifiers-relationships.
[63] OS, OS MasterMap building height attribute, Available online: https://www.ordnancesurvey.co.uk/products/os-mastermap-building-height-attribute, 2023.
[64] Capman, P. A geometric model of dwellings for use in simple energy calculations. Energy Build.. 21 91994) 83-91.
[65] NHS Digital, Delivering Digital Service for the Test, Trace, Contain Ad nEnable (TTCE) Programme, June 2020. Available online: https://digital.nhs.uk/coronavirus/nhs-digital-coronavirus-programme-updates/programme-updates-19-june-2020/delivering-digital-services-for-ttce.
[66] OS, OS Open UPRN, 2023. Available online: https://www.ordnancesurvey.co.uk/products/os-open-uprn.
[67] Tennant, P et al. Use of Directed Acyclic Graphs (DAGs) in Applied Health Research: Review and Recommendations. May 21. Available online: Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations - PMC (nih.gov).
[68] D. Rodrigues, et al., Reflection on modern methods: constructing directed acyclic graphs (DAGs) with domain experts for health services research, Int. J. Epidemiol. 51 (4) (2022 Aug 10) 1339–1348, https://doi.org/10.1093/ije/dyac135. PMID: 35713577; PMCID: PMC9365627.
[69] G.T.H. Ellison, Using directed acyclic graphs (DAGs) to represent the data generating mechanisms of disease and healthcare pathways: a guide for educators, students, practitioners and researchers, in: D.J.J. Farnell, R. Medeiros Mirra (Eds.), Teaching Biostatistics in Medicine and Allied Health Sciences, Springer, Cham, 2023, https://doi.org/10.1007/978-3-031-26010-0_6.
[70] M.L. Watson, et al., Directed acyclic graphs in perioperative observational research–A systematic review and critique against best practice recommendations, PLoS One (Feb 2023), https://doi.org/10.1371/journal.pone.0281259. Available online:.
[71] R. Barnard-Mayers, E. Childs, L. Corlin, E.C. Caniglia, M.P. Fox, J.P. Donnelly, E. J. Murray, Assessing knowledge, attitudes, and practices towards causal directed acyclic graphs: a qualitative research project, Eur. J. Epidemiol. 36 (7) (2021 Jul) 659–667, https://doi.org/10.1007/s10654-021-00771-3. Epub 2021 Jun 10. PMID: 34114186; PMCID: PMC8609501.
[72] UKHSA, UK Field Epidemiology Training Programme (UK FETP), Jan 2023. Available online: https://www.gov.uk/guidance/field-epidemiology-training-programme-fetp.