
Article:
Keeble, C, Thwaites, PA, Barber, S et al. (2017) Adaptation of Chain Event Graphs for use with Case-Control Studies in Epidemiology. International Journal of Biostatistics, 13 (2). 20160073. ISSN: 1557-4679

https://doi.org/10.1515/ijb-2016-0073

---

1 Introduction

Case-control studies are an epidemiological study design which typically compare two groups of people retrospectively; those with a disease of interest (cases) and those without (controls), to try to explore the cause of the disease (Schlesselman, 1982). Participation rates in epidemiology studies have declined over recent years, with efforts to improve participation proving unsuccessful (Hartge, 2006). Case-control studies can incur bias (i) due to problems with non-participation, (ii) as a result of their retrospective nature, or (iii) due to the different methods used to collect data from cases and controls.

Chain event graphs (CEGs) form part of a family of probabilistic graphical models (PGMs) whereby a graph expresses the conditional dependence structure between variables. This family includes Bayesian networks as shown in Cooper (1990), Heckerman (1995), acyclic probabilistic finite automata (APFAs) as demonstrated by Ron, Singer, and Tishby (1998) and chain graphs for example in Buntine (2013). Briefly, CEGs are a form of directed graph which can be used to order and group combinations of categories of variables with respect to their probability of an outcome of interest (Smith and Anderson, 2008).

CEGs will be adapted here for use specifically with case-control data and presented in three sections. In §3.1, we discuss adaptations which could be used for a full case-control analysis. Firstly, to incorporate information regarding data missing as a result of non-participation (§3.1.1: Non-participation in case-control studies), and secondly to explore how exposure associations vary with the severity of a disease (§3.1.2: Associations by disease severity).

We then discuss, in §3.2, adaptation which would fully analyse a data set, but which could form a partial analysis and assist with an investigation of potential participation bias. We investigate how recruitment varies with the data collection approach used (§3.2.1: Recruitment by data collection method) report the characteristics of those who participate (§3.2.2: Participation as the outcome of interest), investigate how these characteristics differ between cases and controls (§3.2.3: Participation by disease group), and consider how data from similar (but not identical) studies can be combined regardless of data missing from non-participation or differing recorded variables (§3.2.4: Amalgamated case-control participation data).

Finally, in §3.3, we propose adaptations which are aimed at improving the analysis. Incorporating the reliability of different data sources is addressed in §3.3.1 (Data reliability), and how subsets of the data can be analysed separately depending on the outcome of interest is considered in §3.3.2 (Subset-chain event graphs).

2 Overview of chain event graphs

Terminology from Smith and Anderson (2008) which originates from graph theory is required to define a tree which forms a chain event graph (CEG). Variables are used to form *vertices*, with *edges* joining them. *Directed* edges have arrows starting at a *parent* and leading to a *child*. Vertices are either the *root vertex* (no parents), *leaves* (no children), or *situations* (not a leaf). A *tree* consists of a vertex and edge set, and is a connected directed graph which has no cycles, one root vertex and all other vertices have exactly one parent. Edges between vertices show the conditional probabilities along each path. A *subtree* is a tree with the child of a vertex as its root, and a *floret* is a subtree consisting of a single vertex and its children, plus the associated edges.

Initially an event tree represents the data, with the variables often given in chronological order. The study data in the event tree is then used in conjunction with the Bayesian agglomerative hierarchical clustering (AHC) algorithm detailed in Freeman and Smith (2011) to determine which vertices are in the same *stage*. Two situations are said to be in the same stage if the topology of their florets is the same under a bijection, as are the probability distributions associated with these florets (Smith and Anderson, 2008). Corresponding edges of two (or more) situations in the same stage are assigned the same colour forming a staged tree. Two situations are said to be in the same *position* if the topology of their subtrees is the same and the probability distributions associated with these subtrees are the same under a bijection (Smith and Anderson, 2008). A CEG is formed by collapsing a tree over its positions. A dashed line between CEG vertices signifies positions from the same stage and their corresponding edges adopt the same colours as in the staged tree. The CEG is interpreted from left to right and combinations of variable categories resulting in the same probability of outcome lead to the same position. An ordinal CEG is a CEG in which the vertices representing a given variable are ordered vertically with respect to the outcome (Barclay, 2014). For reference, a detailed explanation of CEGs is available (Smith and Anderson, 2008), as well as an example of their use with case-control data in Keeble, Thwaites, Baxter, Barber, Parslow, and Law (2017). The *R* code used throughout for the AHC algorithm is given in Barclay (2014). The first example which follows will be described in detail to demonstrate these steps.

3 Adaptations and examples

3.1 Adaptations for full analyses of case-control studies

3.1.1 Non-participation in case-control studies

Since non-participation is a recognised problem in case-control studies as summarised by Hartge (2006), analysis methods which can incorporate information relating to non-participation should be adopted. Non-participation in a case-control study may be just one variable from a missed question or an unavailable piece of data, it may be several variables if an individual is too ill to participate in an interview but their medical records are available, it could be the majority of variables when only an individual's disease status and basic demographics are available, or it may extend to all information except their disease status. Usually when individuals decline to participate in case-control studies they are excluded from the main analysis, and when variables are missing from an individual, those individuals are omitted from any analyses requiring such variables, through complete case analysis, which can lead to biased results.

Conclusions regarding missing data have been successfully reported using CEGs as developed in Barclay, Hutton, and Smith (2014) and this same approach can be used to tackle non-participation in case-control studies, since non-participation is one explanation for why data may be missing. However much data are missing, and by whichever means the data are missing, CEGs can be used by including an additional edge in the event tree for each variable which has missing data. Although, the more variables which have missing data, the more edges there will be in the event tree, thus extending the tree and possibly, although not always, complicating the CEG.

In a CEG, conclusions can be drawn using both the available and missing variables, including statements about the missingness mechanisms (Barclay et al., 2014). Missingness is often described using the three standard definitions given by Rubin; missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) (Little and Rubin, 2002). MAR means that any systematic difference between the missing and observed data can be explained by the observed data (Sterne, White, Carlin, Spratt, Royston, Kenward, Wood, and Carpenter, 2009). Knowledge about the missingness mechanism may be required to know the suitability of methods designed to adjust for non-participation. For example, often imputation requires the data to be MAR as demonstrated by Sterne et al. (2009) while other population-based methods can be used when data are MNAR such as that developed by Keeble, Barber, Baxter, Parslow, and Law (2014). If information regarding the missingness mechanism can be obtained using CEGs, this

can assist adjustment choice.

Hypothetical example

Let there be 100 participants in a case-control study. Some are full-participants (62%) providing their disease status, gender and smoking habits. Some are partial-participants (8%) providing their disease status and gender, but declining to provide their more personal smoking habits. Finally there are non-participants (30%) whose disease status is known from general practitioner records or a disease registry. Figure 1 shows the staged event tree for these data and Table 1 shows the data used. Uniform priors are assumed, meaning that it is not known whether a path containing male, female or unknown is more likely to be taken, and this also applies to smoker/non-smoker/unknown and case/control. For all examples here we use the Bayesian agglomerative hierarchical clustering (AHC) algorithm from Freeman (2010), to group vertices from the tree into stages as defined for CEGs (Smith and Anderson, 2008). Very simply, the Bayesian AHC algorithm is a *clustering* algorithm which starts with all vertices in the tree from a given variable separate and merges (or *agglomerates*) them, and is *hierarchical* since clusters have sub-clusters, which in turn have sub clusters and so on. The *Bayesian* element allows prior knowledge to be incorporated into the algorithm. More formally, the AHC algorithm is a local greedy search algorithm for finding the maximum a posteriori CEG. The algorithm starts with the finest partition of the vertices in the tree and seeks to combine vertices at each iteration which will result in the highest scoring CEG (Freeman, 2010). The initial CEG formed is identical to the event tree except the leaves are collapsed into one terminal vertex, and scored using the posterior probability of the CEG given the data. The algorithm continues by testing each pair of situations in the initial CEG from the same variable which have the same number of edges, to find which pair of situations in the same stage maximises the ratio of the scores from this CEG and the initial CEG. This is repeated until the coarsest partition has been achieved. The CEG with the highest score is then selected and the algorithm returns a list of vertices in the same stage.

Vertices in Figure 1 in the same stage are shown by colouring their corresponding edges in the same colour. For example situations s_1 and s_2 are in the same stage and hence their edges are assigned the same colours. This means it is plausible that the distribution of smoking may have come from the same population for males and for females, even though the numbers assigned to the smoking edges for males and females differs. Vertices are then said to be in the same position if each corresponding vertex along the remainder of the two paths is in the same stage, and vertices in the same position are collapsed over to form the CEG. Situations s_1

and s_2 are also in the same position as their subtrees are assigned the same colours, hence s_1 and s_2 are in the same position and therefore collapsed over to form the CEG shown in Figure 2. Full details for stages and positions are available in Smith and Anderson (2008). The CEGs in this paper are ordinal CEGs meaning that vertices from the same variable are ordered vertically by percentage of those possessing the outcome of interest. Each vertex has been labelled with the percentage of individuals who have the outcome of interest, from all those who have taken the same path. For example, the root vertex w_0 shows the entire sample and from Table 1 there are 15 cases, hence 15% of the individuals at w_0 are cases. At w_2, there are 30 individuals who have unknown gender, 4 of which are cases ($4/30 \approx 13\%$). The calculation for w_1 is a little more complicated since there are two paths, male and female, which lead to this vertex; 40 females and 30 males, of which 5 and 6 respectively are cases. Therefore the percentage of cases at w_1 is $11/70 \approx 16\%$. All percentages are calculated in the same way, which highest percentages placed above lower percentages in the CEG. Note that w_0 and w_∞ will always be the total number of individuals possessing the outcome of interest, and hence always be the same value.

Case	Non-smoker	Smoker	Unknown
Female	3	1	1
Male	3	2	1
Unknown	0	0	4
Control	Non-smoker	Smoker	Unknown
Female	30	2	3
Male	20	1	3
Unknown	0	0	26

Table 1: Non-participation data.

The CEG in Figure 2 suggests that there are similarities between males and females with respect to the disease of interest, since both genders lead to position w_1. The missing gender category resulting from non-participation may be due to data MNAR since it leads to position w_2 where individuals have a lower probability of disease (13%) than males and females (16%). Smokers generally have an increased probability of the disease ($w_3 = 36\%$) than non-smokers ($w_4 = 12\%$), and the unknown smoking status resulting from partial-participants or non-participants leads to both position w_3 and position w_4. Given known gender, the unknown smoking values may be mainly smokers, since both the smoker and unknown categories lead to w_3, hence suggesting the values may be MNAR.

Note that all 100 participants have been used to draw these conclusions about the associations between gender, smoking and the disease of interest, as well

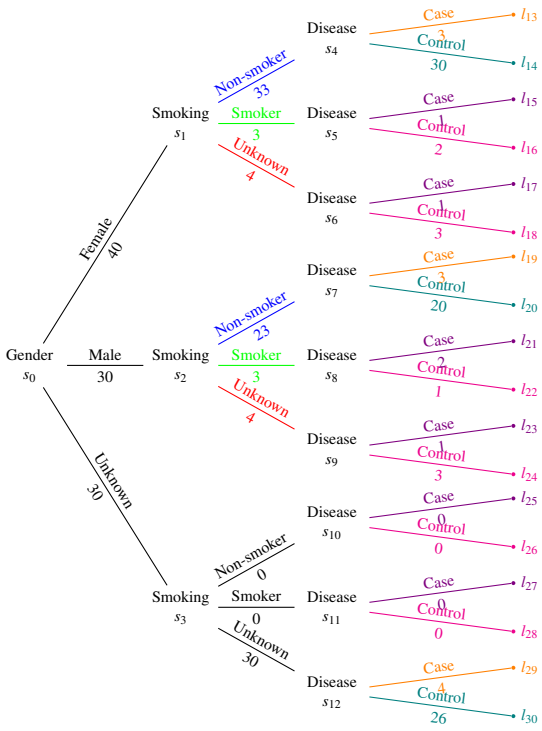

Figure 1: Non-participation staged tree. s denotes a situation and l denotes a leaf.

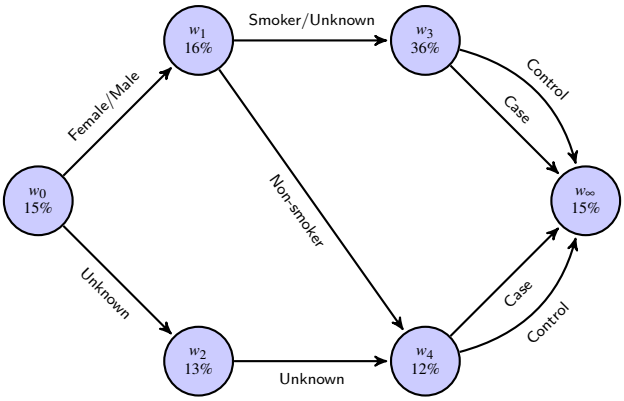

Figure 2: Chain event graph for non-participation. Percentage of cases shown at each position. Colouring is not required since all stages and positions (W) are equal. There is only one edge from w_2 as unknown gender results from non-participation, which also results in an unknown smoking category.

as conclusions about the missingness mechanisms. No individuals have been lost through complete case analysis or similar approaches. In this instance, since the data are likely to be MNAR, a population-based method could be appropriate to investigate non-participation such as that by Keeble et al. (2014), hence not introducing bias by using standard multiple imputation and assuming the data are MAR (Sterne et al., 2009).

3.1.2 Associations by disease severity

Case-control studies usually record a binary outcome of case or control. However, diseases often have different severities such as terminal or not, and this level of detail is likely to be clinically useful and hence recorded in the medical notes. The tree and corresponding CEGs can therefore have additional edges denoting the possible severities of the disease such as control, mild case or severe case.

The CEG can be formed in the usual manner as developed by Smith and Anderson (2008), and as before conclusions can be drawn regarding the combinations of variables associated with the range of severity outcomes. It is possible that only individuals with a particular characteristic or combination of characteristics are able to possess the most severe category of the disease and this information may otherwise be hidden in a standard case-control analysis. The case categories can of course be collapsed and the data analysed with a binary outcome for comparison with any previous analyses.

It may also be that the mechanism differs by disease severity. For instance, there may be more missingness amongst cases with the more severe version of the disease and the missingness may be in variables requiring input from the patient. These differences may be useful to highlight where missingness is occurring and in subsequent studies, different data collection strategies may be adopted, such as collection for all participants only through medical records. An example is given in Appendix A.

3.2 Adaptations for partial analyses: Investigating participation

Thus far, CEGs have been used with the disease status or similar as the outcome of interest. CEGs have not before been used to investigate participation as the outcome of interest. Therefore, this section proposes adaptations to the graphs where the final (outcome) variable represents participation, as well as approaches to investigate the success of data collection techniques by disease group. These CEGs are intended to form part of an interim step in the analysis of a case-control study when

investigating the possibility of bias resulting from non-participation, rather than the final analysis of a case-control study, where the final variable would represent the disease of interest.

Participation bias here is considered to be a subset of selection bias, using the definition from Hernan, Hernandez-Diaz, and Robins (2004), where the bias is caused by conditioning on a collider between the exposure and outcome of interest, meaning that participation is caused by both the exposure and outcome, with only study participants included in the analysis. Since case-control studies condition on disease status by design, by including only participants in the analysis and selecting participants by disease status, selection bias is more likely in case-control studies than other study designs.

3.2.1 Recruitment by data collection method

The effectiveness of different recruitment techniques or data collection strategies (such as web surveys, postal surveys, and electronic reminders) may be of interest. Rather than use a CEG which shows personal characteristics along each path, a CEG can be developed which contains solely information about data collection approaches. With the binary disease status forming the final vertices in the event tree, the CEG can be used to determine which approaches are more associated with cases and which are more associated with controls. Although this approach could highlight successful methods for recruiting case and controls respectively, adopting these approaches would assume a causal effect rather than an association, and could result in bias by recruiting the disease groups in different ways. Therefore, this approach should be used to identify bias which can then be adjusted for, rather than to suggest recruitment techniques. An example is given in Appendix B.

3.2.2 Participation as the outcome of interest

Typically in CEGs the final vertices of the event tree represent the outcome of interest. For case-control studies this is the disease status of the individuals and thus presents two options; case or control. If the outcome of interest is instead non-participation, the event tree can be restructured such that the final vertices represent participation (yes/no). Data collection techniques or individual characteristics can then form the paths in the tree. The CEG highlights which combinations of techniques or characteristics result in comparable probabilities of participants, and an ordinal CEG as introduced in Barclay (2014) can be used to order the combinations of categories from those associated with the lowest probability of participation to the highest. This approach provides a summary of the participation rates as well as

providing information on which factors, or their proxies, are associated with participation, and which should be included as good practice in the reporting of a study requiring participants. These findings can be used to form part of an investigation into possible selection bias in the study. Selection bias occurs in case-control studies when selection is conditioned on and affected by both the exposure and disease of interest (Hernan et al., 2004). Therefore if participation is affected by both the disease status of the individual and the exposure of interest in the study, then selection bias may be present.

This use of CEGs could be extended to more than two participation categories. For example the final vertices could have edges representing "no participation", "partial participation" and "full participation", where partial participation relates to those willing to give demographic data but not sensitive data, or for those willing to participate in a questionnaire but not in a subsequent interview. A similar approach could be used for recruitment phases, where each variable represents a study phase such as first contact, reminder, second reminder and so on, with the outcome of interest being whether the individual participated, refused or ignored, along with their reason for non-participation if given.

Hypothetical example

Let there be 50 copies of a survey distributed by mail and 100 distributed using a cheaper web option. Reminders are sent to 40% of the mail recipients and 50% of the web recipients, since electronic means are cheaper than postage costs. Some individuals return a completed survey, while others do not. Figure 3 shows these variables in a tree along with the number of individuals taking each edge; the corresponding CEG is shown in Figure 4.

The CEG shows that distribution by web and mail generally result in the same probability of receiving a completed survey (50%), but reminders are associated with increased participation rates. Those designing the survey may consequently choose to distribute web rather than mail surveys to save costs, but to include reminders to those who do not participate in the first phase. This is important, since often those who respond to reminders differ from those who responded to the initial survey and differ again to those who do not respond at all. It also assumes a causal association. Therefore, while this tactic appears to increase equality and reduce bias by increasing participation rates, it is possible that this may in fact lead to increased bias by recruiting different participant characteristics, possibly in each disease group. Therefore, this CEG may need to be used in conjunction with a CEG similar to that introduced in the next section (in §3.2.3), to compare the characteristics of the individuals. Knowing the recruitment potential of different techniques is

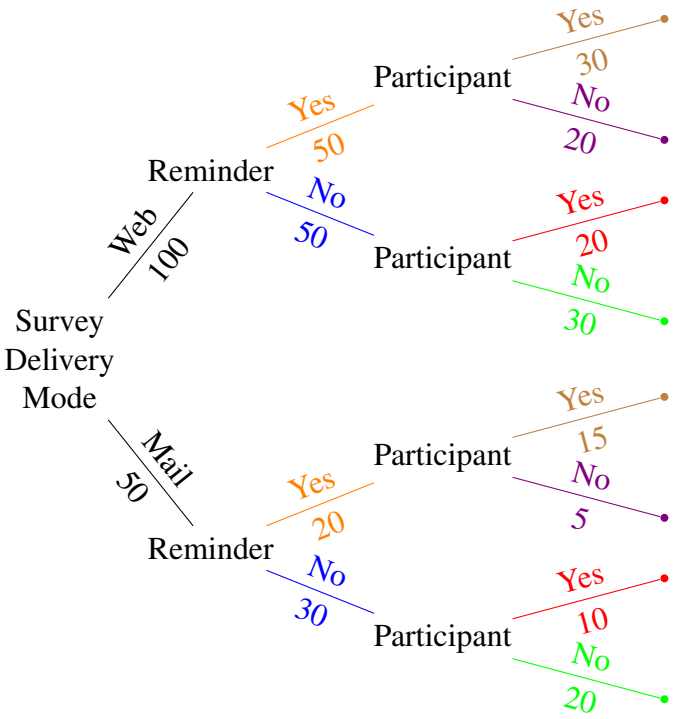

Figure 3: Participation staged tree.

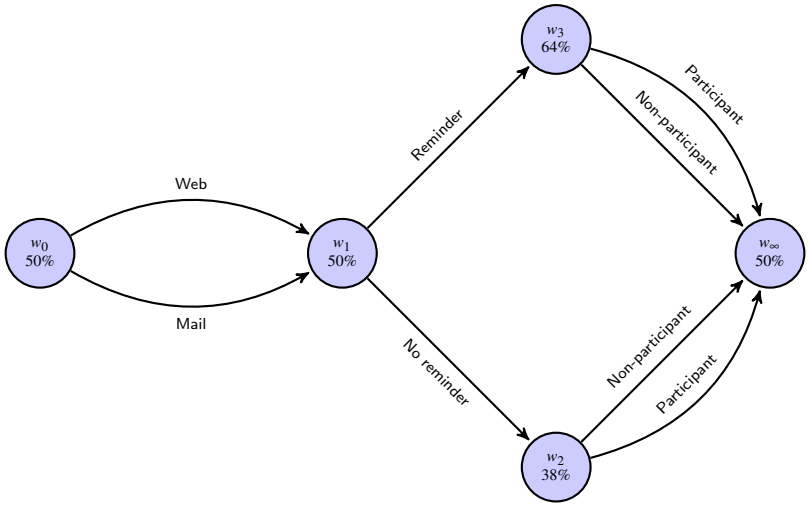

Figure 4: Chain event graph for participation. Percentage of participating individuals shown at each position (W).

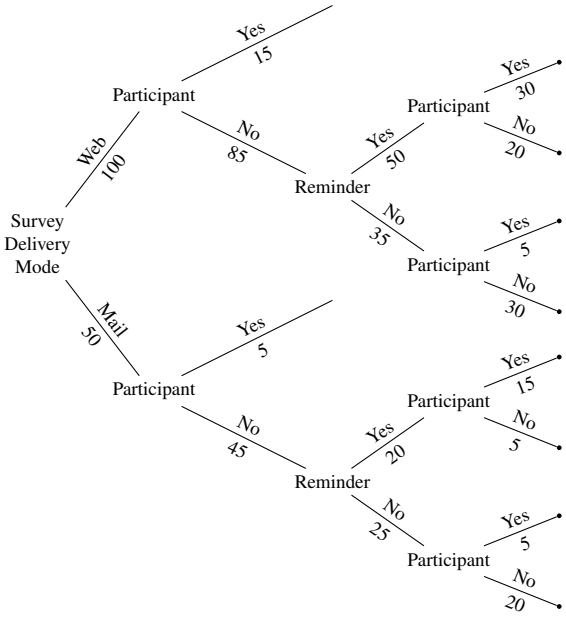

Figure 5: An example of an asymmetric tree.

informative and the associations found may be causal, hence these associations can be explored further. Additionally, potential biases can be unveiled by investigating the characteristics of those who are recruited using different techniques, particularly if these characteristics differ by disease group.

In addition, CEGs can be used for asymmetric problems, meaning that a particular decision at one variable can affect the choices available for later variables. This allows the tree here to be restructured such that the chronological ordering is web/mail, participant/non-participant, reminder/no reminder (only for those who were non-participants after the first phase) and finally participant/non-participant. Therefore the paths through the tree would be of varying lengths, as shown in Figure 5. It may be important to distinguish between participants from the first phase and participants recruited after the reminder phase, and this can be achieved by using different category names for the two groups of participants.

3.2.3 Participation by disease group

The factors associated with participation are required for the case and control group separately, since it is expected that the two disease groups will have different reasons for choosing to participate and since differences between these comparison groups can introduce bias. The CEGs used thus far have ordered the variables

chronologically. However, different orderings can be adopted by CEGs depending upon their use (Thwaites, Smith, and Riccomagno, 2010). If knowledge regarding the factors associated with participation for the cases and controls is required separately, disease status can be placed as the first variable in the tree, and participation as the final variable. This ensures the cases and controls are reported separately regarding participation. An example is given in Appendix C.

3.2.4 Amalgamated case-control participation data

If there are a series of case-control studies with similar variables recorded, the data from each study can be combined into one larger analysis using CEGs, where the outcome would be the disease status. Alternatively, the characteristics of cases or controls who are more likely to participate in a study of a particular topic may be of interest, or the most successful recruitment techniques may be sought. These findings can be achieved by having participation as the outcome variable, rather than disease status, as was demonstrated in Figure 4.

CEGs can be used in the same way as in §3.2.2 but the data are fed in from several studies. This may be particularly useful for studies of sensitive topics or those investigating very rare diseases, where the number of participants may be smaller. Conclusions can be drawn about the combination of factors associated with participation as demonstrated in §3.2.2. Patterns in the data can be used to form hypotheses regarding ways in which to increase participation in under-represented categories, or to inform future studies, although increased participation would not necessarily result in reduced bias. As CEGs can incorporate missing data as shown in Barclay et al. (2014), studies which record similar but not identical variables can be combined directly, with unrecorded variables included as an additional edge labelled 'unrecorded'. An example is shown in Appendix D.

This approach could also be used with the disease status as the outcome of interest, and with missing data as a result of non-participation or otherwise, included as extra edges. In this instance this adaptation could provide a full analysis for case-control studies. If desired, data missing from non-participation and data missing for other reasons, could be given separate edges in the event tree, such that differences between the types of missingness can be represented.

3.3 Adaptations to improve analyses

3.3.1 Data reliability

CEGs have been used previously with prospective data but rarely with retrospective data, which can have limitations such as recall bias. This feature of the case-control study data can be incorporated into the CEG framework to enhance the authenticity of the analysis and to include additional information about the reliability of the data obtained.

Data in retrospective studies may be recorded using a variety of means such as medical records, through interview, or using national databases. Some sources may be cross-checked and verified with other authorities, while other sources may depend upon a single handwritten report, such as in older medical records or in areas without electronic databases. In some instances the only source will be the memory of those present and will require the individual to recall specific details. Recall bias is known to differ between participants of different disease groups in case-control studies as stated in Health Knowledge (2016) and hence data reliability may differ by both source and disease status.

One way in which to allow for potentially less reliable study data is to form a CEG which has a greater dependency on prior knowledge, provided these data are collected from a more reliable source. Since CEG learning is Bayesian and combines prior knowledge with data, non-uniform priors can be specified during the agglomerative hierarchical clustering (AHC) algorithm phase to achieve this. The examples preceding this have used uniform priors since no additional information has been available from experts or previous studies about which paths are more likely. The equivalent sample size explained in Freeman and Smith (2011) is also specified and if a large equivalent sample size is used this suggests stronger prior beliefs and hence allows the priors to play a more dominant role, rather than depending strongly on the data.

The resulting tree will be structurally identical, but labelled with priors (or left unlabelled). The CEG may differ according to whether uniform or non-uniform priors are used, and this will depend upon the priors assigned and the data collected. This approach could be particularly useful in studies which suffer from non-participation, since the true population distributions of variables may not be apparent from the study data and this could potentially affect the conclusions generated.

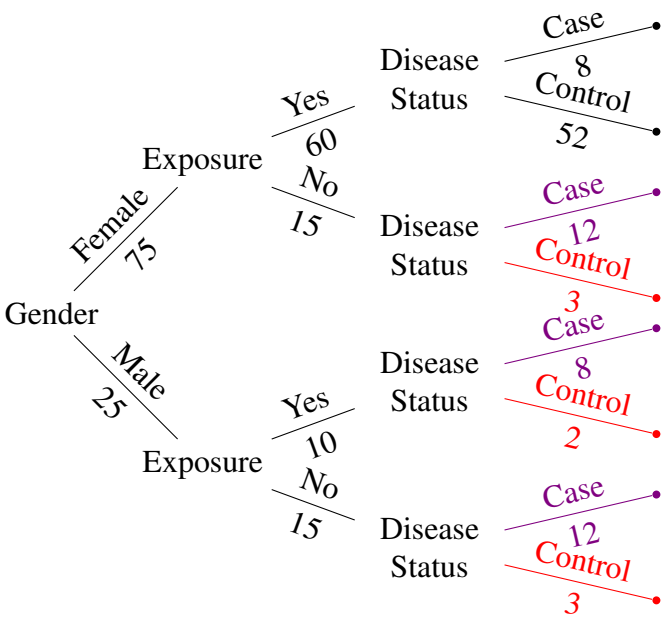

Figure 6: Data reliability: Staged tree with uniform priors.

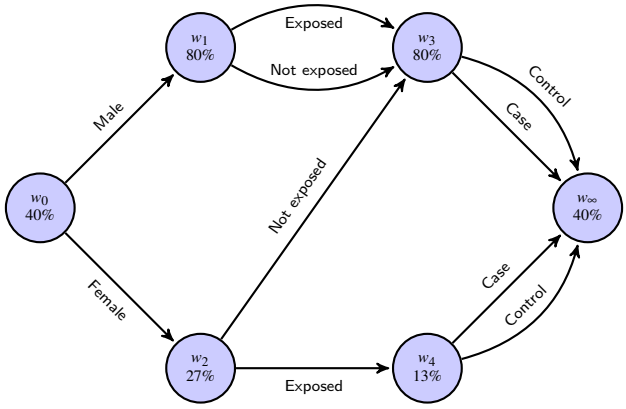

Figure 7: Data reliability: Chain event graph formed from uniform priors. W denotes positions.

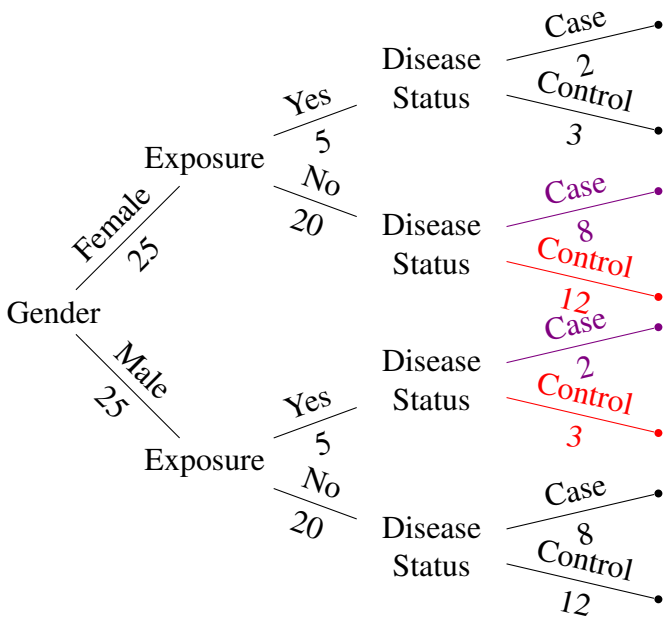

Figure 8: Data reliability: Staged tree with non-uniform priors. The numbers indicate priors rather than individuals; the number of individuals are shown in Figure 6.

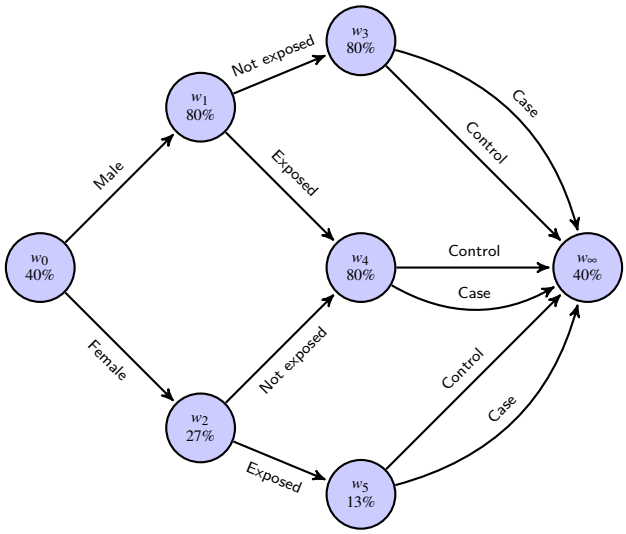

Figure 9: Data reliability: Chain event graph formed from non-uniform priors. W denotes positions.

Hypothetical example

Let there be a hypothetical study with two binary exposures, one of which is gender, plus a binary disease. First the analysis will be conducted using uniform priors and then with non-uniform priors constructed using hypothetical expert knowledge. Figure 6 shows the staged tree formed with uniform priors and Figure 7 shows the corresponding CEG. Figure 7 shows that males are more associated with case status than females (80% compared with 27%) and that for males, being exposed or not leads to the same position in the CEG (w_3) with 80% of the individuals being cases. However females differ by exposure, with 13% of exposed females (w_4) being cases and 80% of non-exposed females being cases (w_3).

Figure 8 shows the staged tree which uses the same data, but where non-uniform priors are allocated. The priors are the smallest possible such that each value is integer and these priors are shown along the edges of the tree in Figure 8. The priors have been assigned using the hypothetical prior knowledge of half males and half females in the population, with the exposure being as common amongst males as it is females, and with around 20% of the population being exposed. The ratio of cases to controls in the study has been maintained. The same data as in Figure 6 but with the priors in Figure 8, results in the CEG in Figure 9.

Figure 9 differs from Figure 7 in that position w_3 in Figure 7 is split into two positions (w_3 and w_4) in Figure 9, hence priors can affect the CEG produced. This split separates unexposed males, from exposed males and unexposed females, but returns two positions (w_3 and w_4) with the same proportion of cases (80%). Otherwise the CEGs are comparable and similar conclusions can be drawn.

In this example, the vertical ordering in the CEG is unchanged since the original position (Figure 7, w_3) and the two new positions (Figure 9, w_3 and w_4) each have 80% of individuals who are cases. However it is possible that in some instances the splitting of positions could result in the reordering of the variables associated with these positions, especially if the percentage of (in this example) cases is similar amongst the positions. The equivalent sample size corresponds to the strength of the prior beliefs and could also affect the CEG, hence it is advised to check the robustness of the CEG with respect to changes in the equivalent sample size (Barclay, 2014).

3.3.2 Subset-chain event graphs

CEGs are usually simple for a small number of variables, but quickly become complicated and more difficult to read when there are a large number of variables and/or each variable has a large number of categories. This includes where there are a large number of variables with missing data, where each variable includes an additional

edge denoting missingness. Here, for these instances which can occur in case-control studies, subset-chain event graphs (subset-CEGs) are proposed as a new variant of CEGs.

A subset-CEG is simply a subset of variables displayed in a CEG which relate to a particular aspect of the data, which can later be interpreted alongside other subset-CEGs. One such CEG could be constructed for individual characteristics and another for environmental factors, with the number of subset-CEGs dictated by the number of variables and categories. If desired, one final CEG can be constructed at the end of the analysis which contains all variables found to be important in the subset-CEGs.

Hypothetical example

Let there be a study where a total of 150 male and female individuals, who can be classed as either old or young by a given cut-off age, are asked to participate in a study by web or mail, with some receiving a reminder to participate and others not. The characteristics of the individuals are given in the staged tree in Figure 10 and the corresponding CEG is shown in Figure 11. The recruitment details for the study are as were shown in Figures 3 and 4.

Figure 11 shows that participation does not differ between males and females, but is more likely from those who are old than those who are young. Figure 4 showed reminders to be associated with participation, but not the survey delivery mode and, if desired, these variables can be used to construct a final CEG for the dataset. Figure 12 shows the staged tree for the variables of age and reminders on participation, and Figure 13 shows the corresponding CEG. The CEG suggests that old individuals are more likely to participate than young, and that reminders are associated with increased participation in old individuals, but are not as effective for young individuals.

Since the natural ordering of age and reminders is unclear, the analysis was also rerun with the reminder variable before age. This resulted in the CEG shown in Figure 14 which shows reminders to be associated with increased participation and old individuals to be more likely to participate than young individuals. Again reminders are not as effective for young individuals as they are old individuals. Therefore the same conclusions are drawn, regardless of the ordering of the age and reminder variables. This should be expected, since age is not affected by reminders and the allocation of reminders is not determined by age, hence the ordering of these two variables is less important than in other scenarios.

Here, subset-CEGs have been used to simplify the analysis into smaller steps and use variables thought to be associated with the outcome to form a fi-

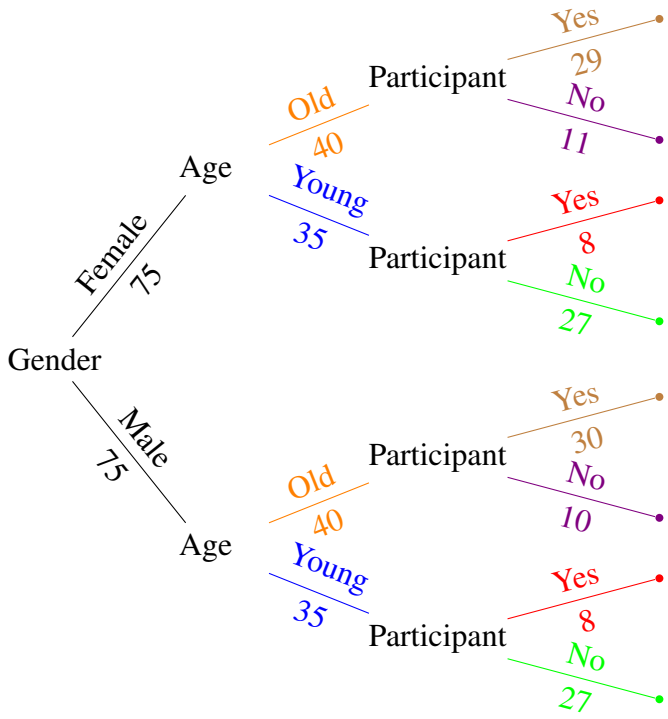

Figure 10: A staged tree used to form a subset-chain event graph.

nal CEG. Each subset-CEG shows a different aspect of the study, which might have been missed in a full CEG. This also improves the readability of the CEG. Another approach to improve readability, may be to present the CEGs against a grid representing the percentage of individuals at each vertex who have the outcome of interest (participation or disease status here). Rather than display the percentages within the vertices, here it is proposed that the vertices could be placed vertically against the grid to show their relative positioning. Figure 14 has been redrawn using a grid and is shown in Figure 15 as an example. This improves readability for spatial readers, but may cause the edges to be less clear, and could result in fewer planar graphs and hence a graph which is more difficult to interpret.

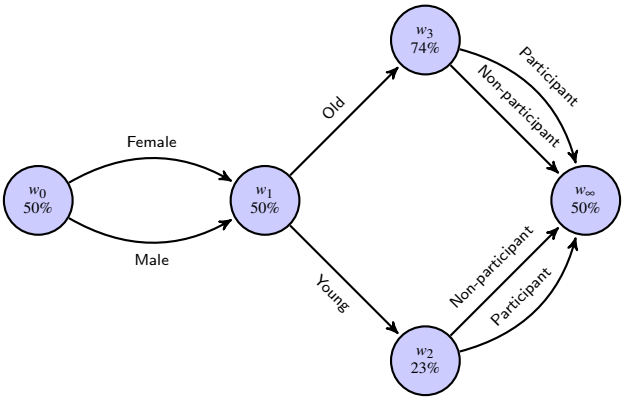

Figure 11: An example of a subset-chain event graph. Percentage of participating individuals shown at each position (W). Colouring is not required since stages and positions are equal.

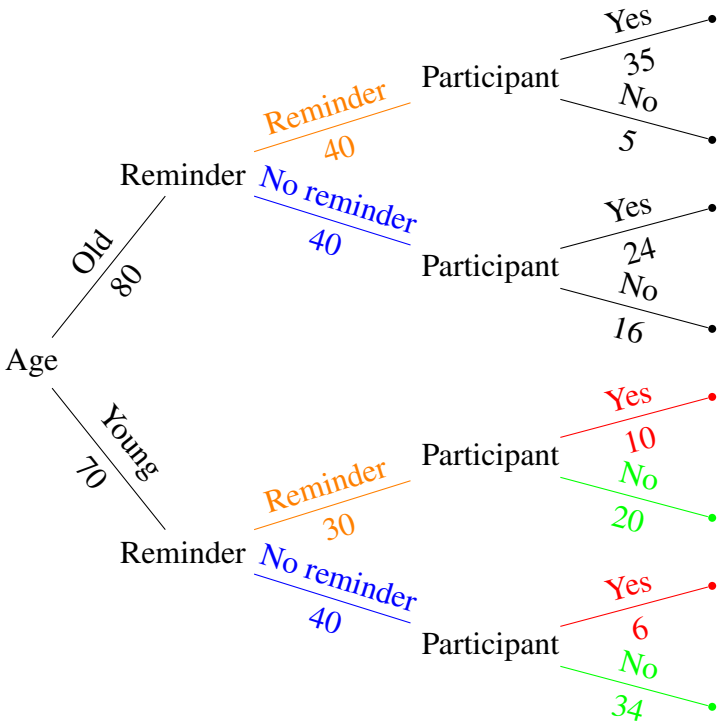

Figure 12: A staged tree with variables selected using subset-chain event graphs.

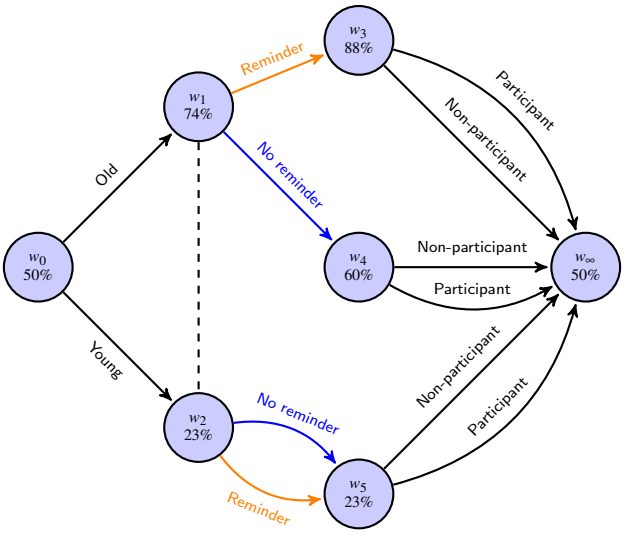

Figure 13: A final chain event graph with variables selected using subset-chain event graphs. Percentage of participating individuals shown at each position (W).

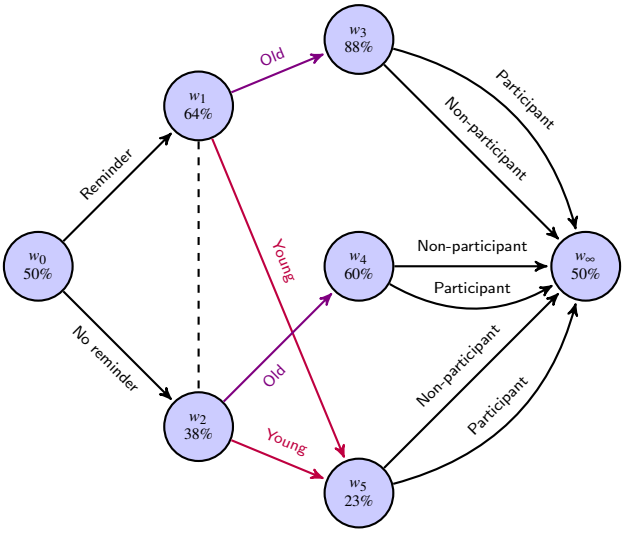

Figure 14: A final chain event graph with variables selected using subset-chain event graphs, age and reminder variables swapped. Percentage of participating individuals shown at each position (W). Example colouring has been used to highlight which positions were in the same stage, as the staged tree is not shown.

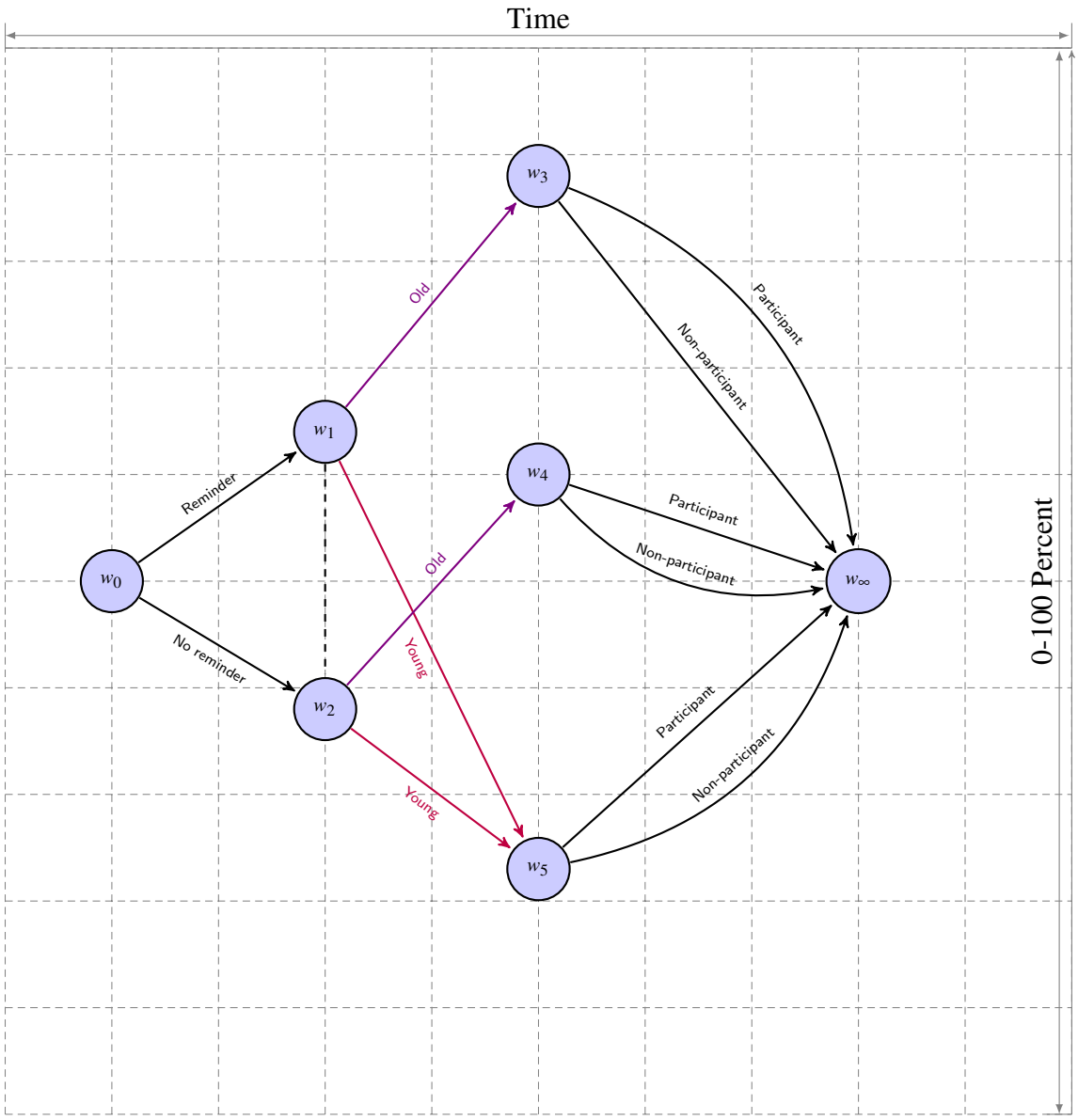

Figure 15: Example of a grid to position vertices vertically with respect to their percentage in an ordinal chain event graph. Each vertical line in the grid represents 10%.

4 Discussion

CEGs have not been used before to investigate biases such as those relating from non-participation, or to summarise data collection techniques. Eight adaptations have been proposed where the structure of the CEGs, the ordering of variables, or the outcome of interest have been changed, but the underlying methodology is the same as in previous publications such as Smith and Anderson (2008), Barclay et al. (2014). CEGs have therefore been applied to a new epidemiological area and used for new purposes which are applicable to case-control studies.

4.1 Conclusions for full analyses

Complete case analysis is commonly used when data are missing, leading to the exclusion of valuable study information. CEGs have been proposed here as an approach which can allow for data missing through non-participation, as often found in case-control study data, while retaining all data that are collected. Another area where data may be lost is through the (binary) categorisation of disease status. It is possible that the severity of the disease may be associated with different exposures and that the presence of multiple exposures may be associated with increased disease severity. CEGs have therefore been proposed here to investigate the association of exposures with disease severity. This approach does not prevent the categories from being collapsed so that traditional (binary outcome) analyses can be conducted. More traditional alternatives include participation flowcharts for missing data, but these approaches do not report conditional dependencies between the exploratory variables and do not provide information on the missingness mechanism.

4.2 Conclusions for partial analyses

Since non-participation is a recognised problem in case-control studies, and since cases and controls are often recruited using different methods, CEGs have been suggested here to explore the different recruitment techniques adopted within a study. CEGs have not previously been used for this purpose, but these findings may highlight differences between the disease groups being compared, particularly if cases and controls engage with the study in different ways, which may introduce bias which needs to be accounted for. Knowledge of these differences allows limitations for the case-control study findings to be highlighted, and for methods to reduce any biases to be implemented.

CEGs have been used here with participation as the outcome of interest which has not before been suggested. Information regarding the characteristics of those who do and do not participate, and comparisons between participating cases and controls, can be informative for determining whether participation bias is likely to have occurred in a study. CEGs can be used in conjunction with other graphical models such as directed acyclic graphs to investigate biases (Hernan et al., 2004). These CEGs can also act as an interim step in the analysis before other CEGs are used for full analyses, where the final edges represent disease status.

Alternative approaches to account for non-participation may include sensitivity analyses, imputation or weighting. However, it may not be known whether the data are missing at random and which types of individuals are missing, hence the assumptions of such approaches may not be fulfilled. CEGs offer an exploratory tool by which to investigate missingness through non-participation, and allow the analyst to further understand the data structure.

The basic characteristics of those who have declined to participate have been included in some of the hypothetical examples given here. This approach could be viewed as unethical, since they have not consented for their data to be analysed. However, the data included in the CEGs could be publicly available data, and is similar to the level of detail which could be used to conduct a sensitivity analysis, a participation flowchart, or a comparison between the characteristics of participants and non-participants.

The ability of CEGs to incorporate 'unrecorded' data is also used here to propose CEGs for combining data from multiple studies. Since recruitment rates in case-control studies have declined in recent years, and since case-control studies may focus on sensitive diseases such as sexually transmitted diseases or exposures such as illegal substance abuse, the ability to combine study data will help to increase the study sample size. This approach is also beneficial where the disease of interest is very rare.

4.3 Conclusions for adaptations to improve analyses

The retrospective nature of case-control studies can lead to problems such as recall bias or dependence upon unreliable historical data sources. Since CEGs are a Bayesian approach, prior information can be specified in the AHC algorithm which is used in the construction of the CEG. This allows expert opinion or potentially more reliable population-level data to be included in the analysis, to off-set any biases relating to poor or unreliable data sources.

Reduced ordinal CEGs have been introduced previously in Barclay et al. (2014) but here subset-CEGs have been suggested to simplify the graphs and im-

prove readability. Subset-CEGs can be used for any number of outcome categories and hence have an advantage over reduced ordinal CEGs which require the outcome to be binary. A grid-based background has also been suggested here for use with CEGs, such that the percentages do not need to be included in the vertices and instead the spatial layout of the CEG indicates these percentages.

4.4 Extensions and limitations

Each hypothetical example given here is relatively simple to demonstrate the new idea, but can of course be extended to include more variables and more variable categories as would likely be found in real data. For example, the data in §3.2.2 could also include variables relating to the characteristics of the individuals, such as their age category or gender. The CEG can include as many variables as required by including additional edges in the tree and increasing the length of the path. The ordering in the tree of survey and individual characteristics may not be obvious, so the ordering should be tested in a sensitivity analysis, by constructing two (or more) CEGs with different plausible orderings and testing whether the conclusions are sensitive to these orderings. It is also possible to include prior knowledge in all analyses, rather than assuming each path is equally likely, and to vary the strength of these prior beliefs. Real data would be analysed in the same way as the hypothetical examples here, although the trees and CEGs may become more complicated if there are more variables or a greater number of variable categories. In these situations, the subset-CEGs may be useful.

A limitation of CEGs is that they require the data to be categorical. While many continuous variables have clinically-relevant cut-off values, this may be problematic for some variables which do not have a sensible cut-off values, so the sensitivity to different cut-off values may need to be trialled. However, this should be less problematic in the scenarios listed here since many of the variables considered will be naturally categorical, such as the data collection method being face-to-face interviews or postal questionnaires.

CEGs can be used to explore the missingness mechanism as shown in Barclay et al. (2014) which can be used for data missing through non-participation. Further understanding the associations of recruitment techniques and disease status with non-participation can lead to the application of a suitable method to account for the missing data, such as multiple imputation in Sterne et al. (2009) or stratification in Schlesselman (1982), which in turn should lead to a more thorough analysis which returns more accurate results regarding the exposure-disease association.

4.5 Overview

CEGs offer a graphical (rather than numerical) approach to these analyses, which some researchers may prefer, and which may be easier to communicate to specialists who may not necessarily be statistically trained and who may be deterred by complicated models and notation. CEGs also have the advantage of being able to represent interactions between recruitment techniques or participant characteristics which would require more complicated interaction terms in traditionally modelling.

Another advantage of CEGs is that they can incorporate expert opinion or prior knowledge, which is not often possible in traditional analyses. Although methods such as Bayesian logistic regression are available, they are not in common practice in calculations following a case-control study. Regarding non-participation, it is possible to include 'missing' as an additional category for a variable, but again, this approach is not well practiced in logistic regression following a case-control study.

The examples given here include participation as an (interim) outcome of interest. Of course, case-control studies primarily aim to identify exposure-disease associations, therefore the CEGs where non-participation is the outcome of interest should be used as a tool during the analysis of a case-control study, or viewed as a subset CEG, rather than the full analysis of a case-control study. In addition to the presence of a disease or participation in a study, other outcomes of interest can be explored and in these instances the percentage shown at each vertex would correspond to the presence of the outcome of interest given the path thus far. The eight examples are a basis which can be adapted further to suit the nature of the study.

In summary, chain event graphs can be adapted to increase their suitability for used with case-control data. The unique features of a case-control study can be incorporated into the analysis to provide further insight, which can help to identify potential biases. In addition, adaptations can be used to improve the readability of the graphs and ease of analysis.

Figure legends

1. Non-participation staged tree. s denotes a situation and l denotes a leaf.
2. Chain event graph for non-participation. Percentage of cases shown at each position. Colouring is not required since all stages and positions (W) are equal. There is only one edge from w_2 as unknown gender results from non-participation, which also results in an unknown smoking category.
3. Participation staged tree.
4. Chain event graph for participation. Percentage of participating individuals shown at each position (W).
5. An example of an asymmetric tree.
6. Data reliability: Staged tree with uniform priors.
7. Data reliability: Chain event graph formed from uniform priors. W denotes positions.
8. Data reliability: Staged tree with non-uniform priors. The numbers indicate priors rather than individuals; the number of individuals are shown in Figure 6.
9. Data reliability: Chain event graph formed from non-uniform priors. W denotes positions.
10. A staged tree used to form a subset-chain event graph.
11. An example of a subset-chain event graph. Percentage of participating individuals shown at each position (W). Colouring is not required since stages and positions are equal.
12. A staged tree with variables selected using subset-chain event graphs.
13. A final chain event graph with variables selected using subset-chain event graphs. Percentage of participating individuals shown at each position (W).
14. A final chain event graph with variables selected using subset-chain event graphs, age and reminder variables swapped. Percentage of participating individuals shown at each position (W). Example colouring has been used to highlight which positions were in the same stage, as the staged tree is not shown.
15. Example of a grid to position vertices vertically with respect to their percentage in an ordinal chain event graph. Each vertical line in the grid represents 10%.

A1. Severity staged tree.
A2. Chain event graph for severity. Percentage of severe case (SC) and mild case (MC) individuals shown at each position (W).
B1. Data collection staged tree.

B2. Chain event graph for data collection. Percentage of cases shown at each position (W).

C1. Staged tree for participation by disease group. s denotes a situation and l denotes a leaf.

C2. Chain event graph for participation by disease group. Percentage of participating individuals with given characteristics are shown at each position (W).

D1. Staged tree formed from amalgamated data.

D2. Chain event graph for the amalgamated data. Percentage of participants shown at each position (W).

A Associations by disease severity

A.1 Hypothetical example

Let there be a case-control severity study which consists of a control group plus two categories of cases; mild case (MC) and severe case (SC). Let there be two independent exposures of interest, each of which is binary. The staged tree for the data is shown in Figure A.1 and the corresponding CEG is given in Figure A.2. The CEG shows that when only one exposure is present (w_4), the individuals have generally the same probability of the three disease categories as when no exposures are present (SC: 8%, MC: 11%). However when both exposures are present (w_3), there is an increased probability of being a severe case (SC: 80%). Exposure 1 alone (w_1) shows an increased probability of being a case, for both severities (SC: 33%, MC: 20% compared with SC: 7% , MC: 7%). A similar CEG could be constructed with missing values, to investigate missingness with respect to the disease severities. For example, missing edges may only lead to a severe disease status, while recorded edges may lead to any of the three disease categories. These severity CEGs allow the research team to understand where missingness is most common and hence where bias may be occurring, for example through less input from the most severe cases.

B Recruitment by data collection method

B.1 Hypothetical example

Let there be a hypothetical survey conducted. Figure B.1 shows that 25 of the participants were recruited by mail, and 50 were recruited by a web survey. The quantities of reminders required and the disease status recorded from the survey are also shown. This tree can be used to summarise which data collection techniques are more associated with case recruitment and which techniques are more associated with control recruitment.

The CEG formed using this information is given in Figure B.2. Mailed surveys, or web surveys without reminders, recruited a group consisting of around 62% cases and 38% controls. Web surveys with reminders recruited a greater proportion of controls (around 83% controls and 17% cases). Mailing alone was more successful at recruiting cases (around 72%), while web surveys were more successful at recruiting controls (around 70%). These percentages can be compared directly, since the study consisted of approximately half cases and half controls. This information can then be used to identify where bias may have occurred, for example, previous

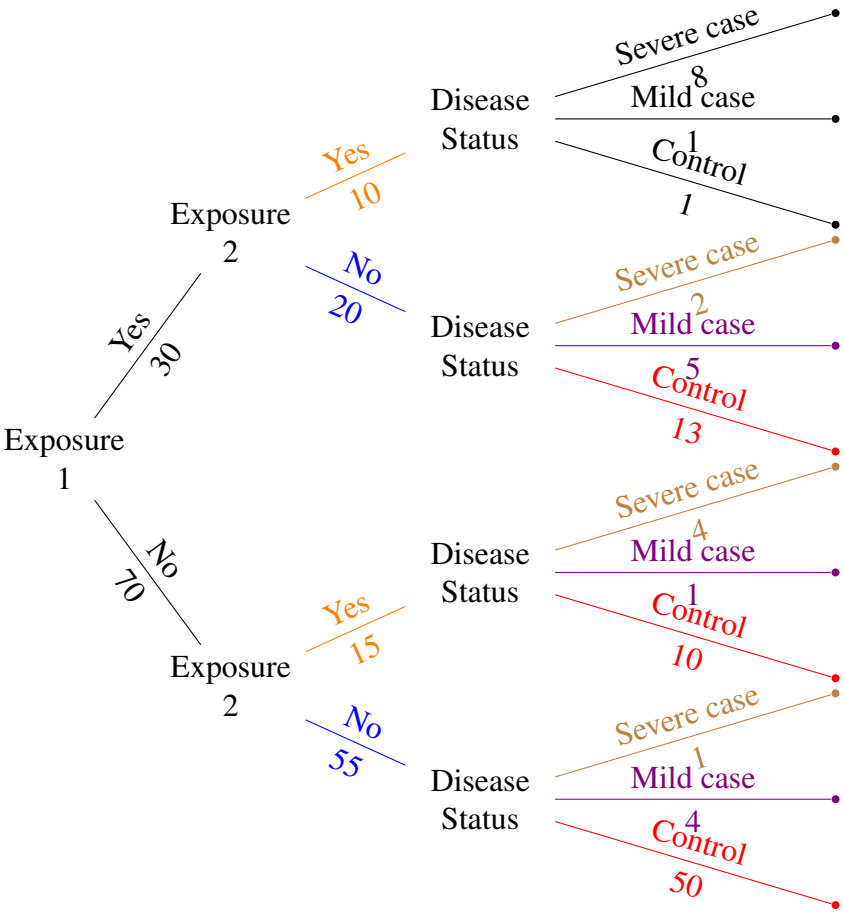

Figure A.1: Severity staged tree.

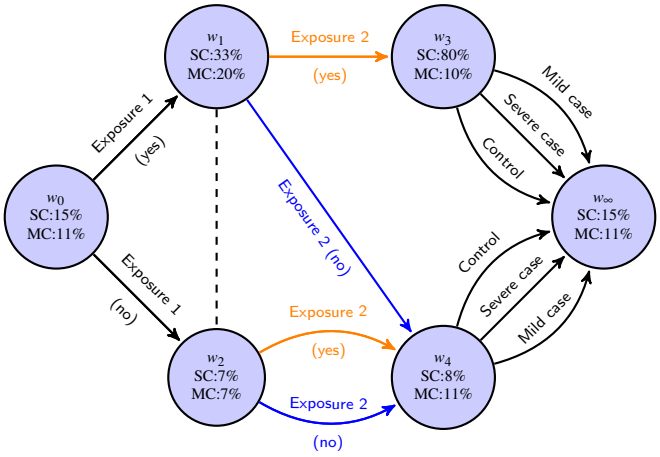

Figure A.2: Chain event graph for severity. Percentage of severe case (SC) and mild case (MC) individuals shown at each position (W).

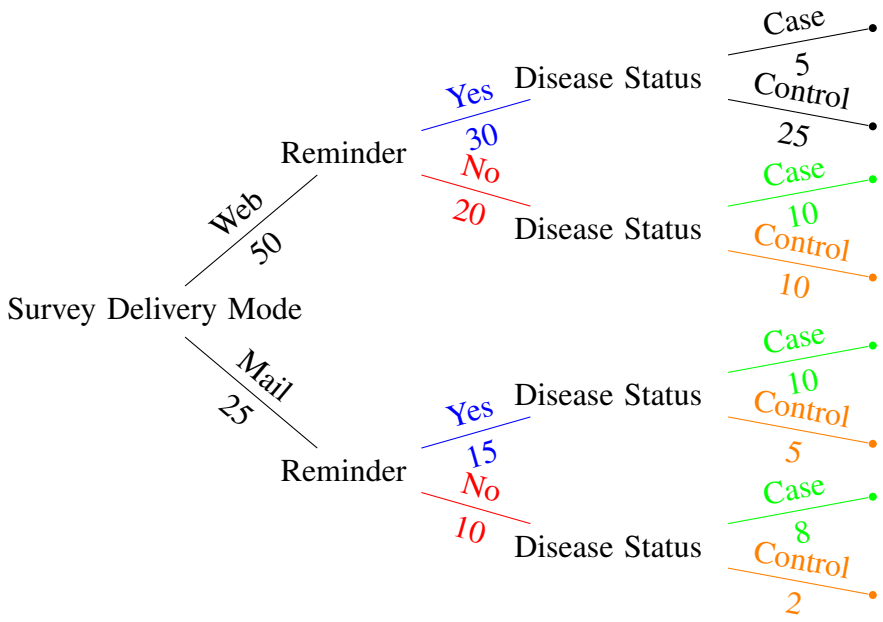

Figure B.1: Data collection staged tree.

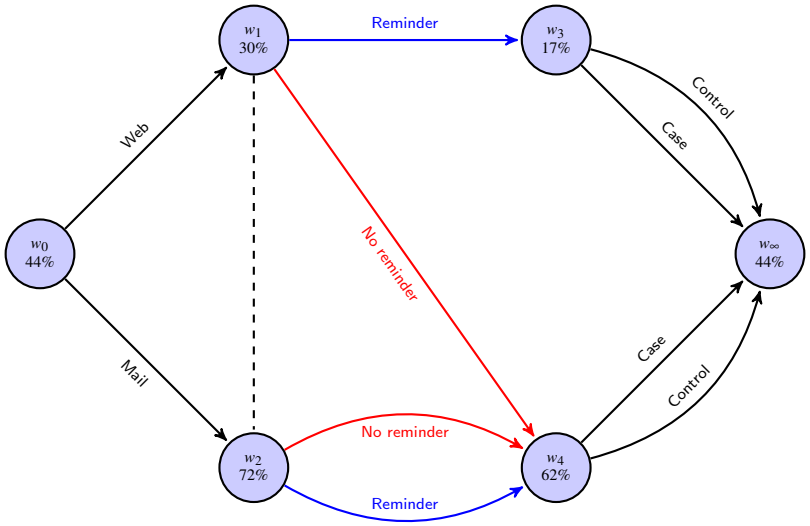

Figure B.2: Chain event graph for data collection. Percentage of cases shown at each position (W).

studies suggest that participants who respond to mailed surveys differ from those who respond to web surveys, and those who respond to reminders differ to those who respond to initial requests (Dillman, Phelps, Tortora, Swift, Kohrell, Berck, and Messer, 2009, Parsons and Manierre, 2014).

C Participation as the outcome of interest

C.1 Hypothetical example

Let there be 100 cases and 200 controls who are asked to participate in a hypothetical study, where the variables of interest are gender (male or female) and age (under 50 years, or 50 years and over). The staged tree is shown in Figure C.1 and the corresponding CEG is given in Figure C.2.

Figure C.2 shows cases are more likely to participate (80%) than controls (37%) regardless of gender or age. There are gender differences in the control group, with females (44%) participating more than males (30%), and age group differences with older males (40%) participating more than younger males (20%). Older male controls have a similar probability of participating as female controls of any age (43%). These findings can be used to explore differences between the disease groups for the consideration of methods to reduce participation bias, some of which are given in Keeble, Law, Barber, and Baxter (2015), hence producing more accurate case-control study results.

D Amalgamated case-control participation data

D.1 Hypothetical example

Let there be three hypothetical studies; one which recorded age and gender, another which recorded age and ethnicity, and a third which recorded ethnicity and gender. These data could be used to investigate the general characteristics of those more likely to participate in a case-control study by having the final vertices showing the participation status of the individuals.

Assume these variables are non-sensitive and hence were available to the researchers within a given study whether the individuals chose to participate or not. Sensitive data would only have been recorded for participants but could be investigated by including 'missing' edges for non-participants. Depending upon the purpose of investigating participation, the tree could be constructed for the entire study

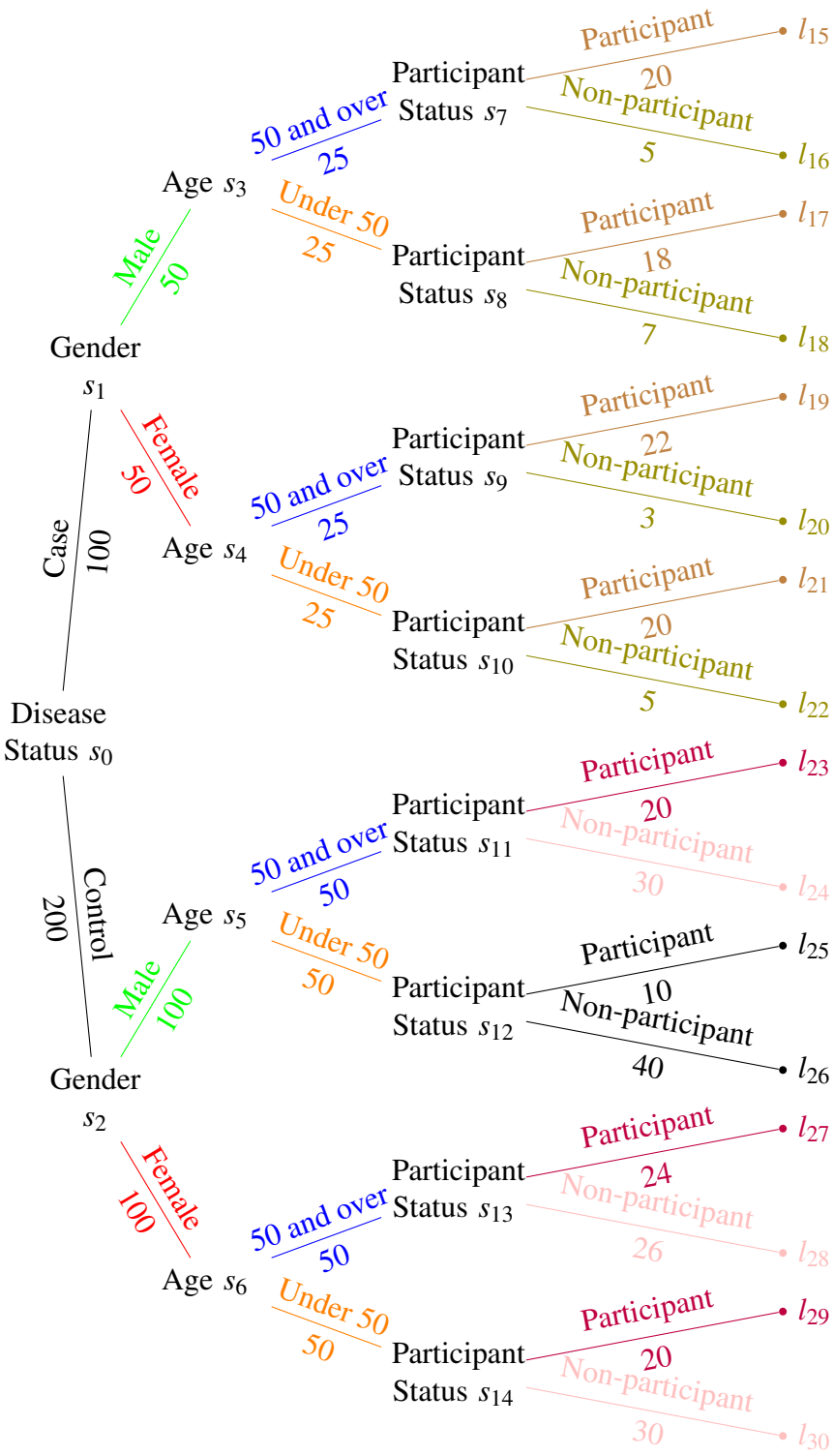

Figure C.1: Staged tree for participation by disease group. s denotes a situation and l denotes a leaf.

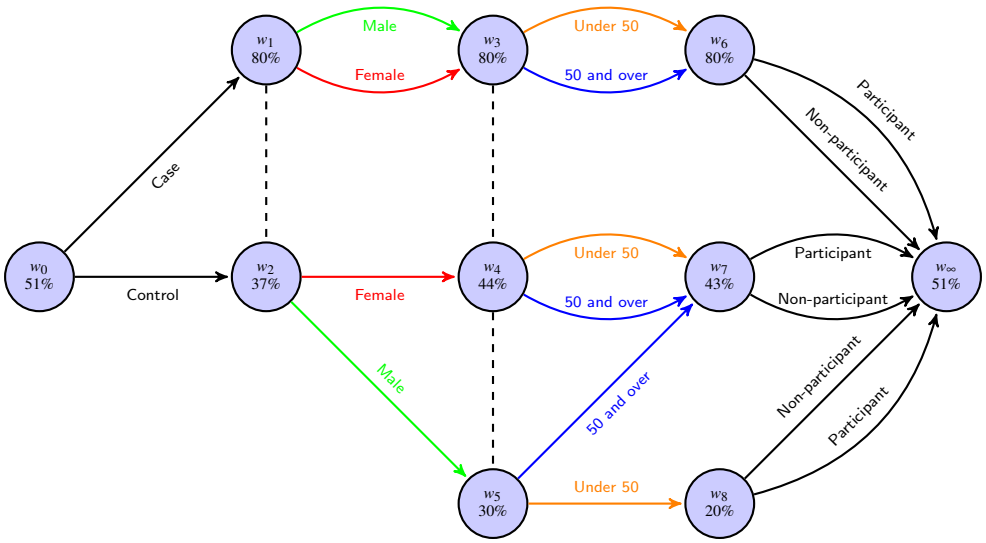

Figure C.2: Chain event graph for participation by disease group. Percentage of participating individuals with given characteristics are shown at each position (W).

group, just controls, or separate trees could be constructed for cases and controls for comparison.

Let the event tree be as in Figure D.1, with the corresponding CEG as in Figure D.2. For each of the studies, there are two variables recorded and one variable considered to be missing. The CEG shows that males are less likely to participate than females, and the unknown gender category may be missing at random (MAR) as defined by Little and Rubin (2002) as it is positioned between males and females (see Barclay et al. (2014) for further details on missingness in CEGs). If the data are MAR with respect to this larger sample, this could suggest that methods such as multiple imputation as shown in Sterne et al. (2009) could be adopted if required.

The distribution of ethnicity is shown to be indistinguishable given known gender, since the same green colours are assigned to edges emanating from situations s_1 and s_2 in Figure D.1, hence ethnicity is distributed similarly amongst males as it is females, as would be expected in the population. When gender is known, white participants are less likely to participate than non-white, since the edges representing white participants lead to w_4 and w_8 which are positioned higher in the ordinal CEG in Figure D.2 than positions w_7 and w_{10}, which the edges representing non-white participants lead to. The unknown ethnicity edges lie between the two known ethnic groups and hence it is possible that the unknown category consists of both white and non-white individuals, suggesting that the unknown ethnicity values may be MAR.

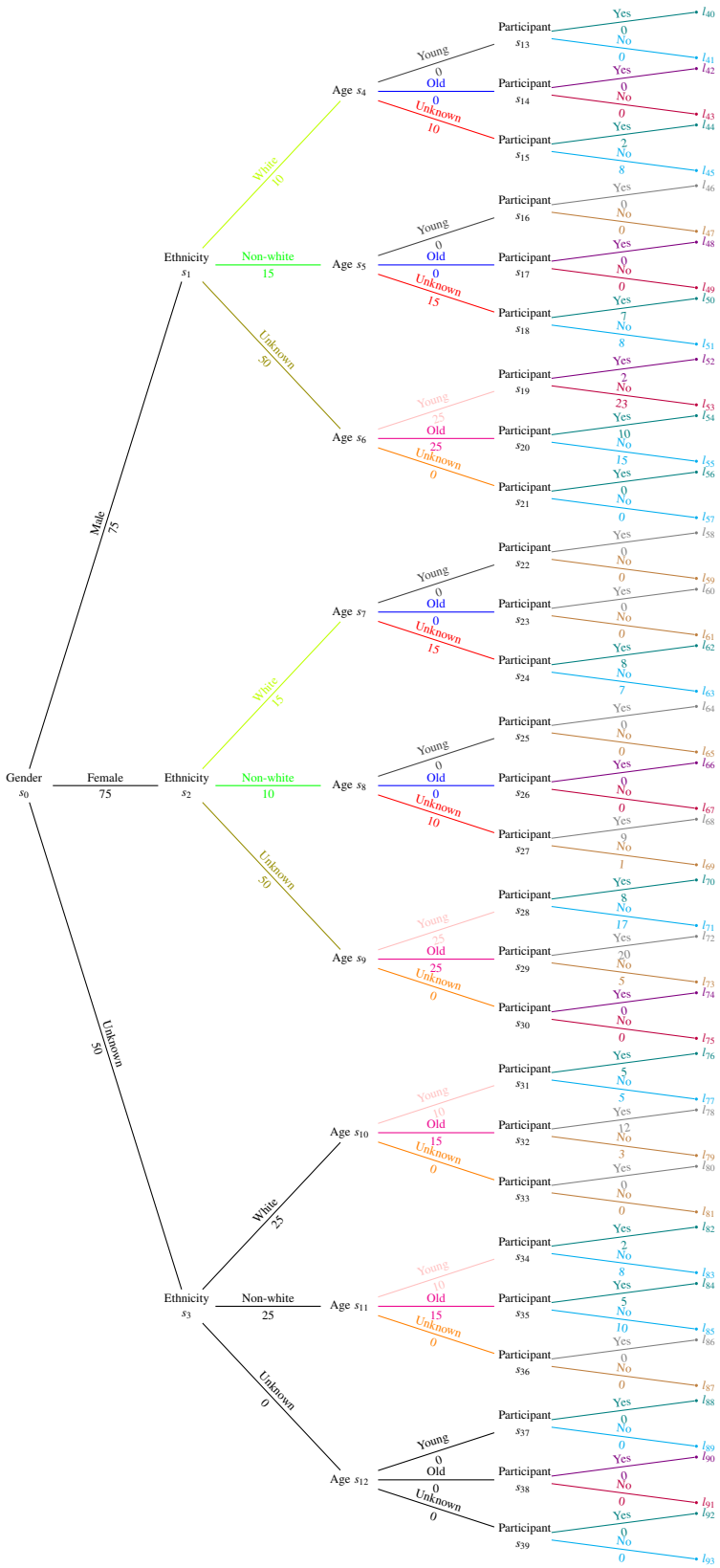

Figure D.1: Staged tree formed from amalgamated data.

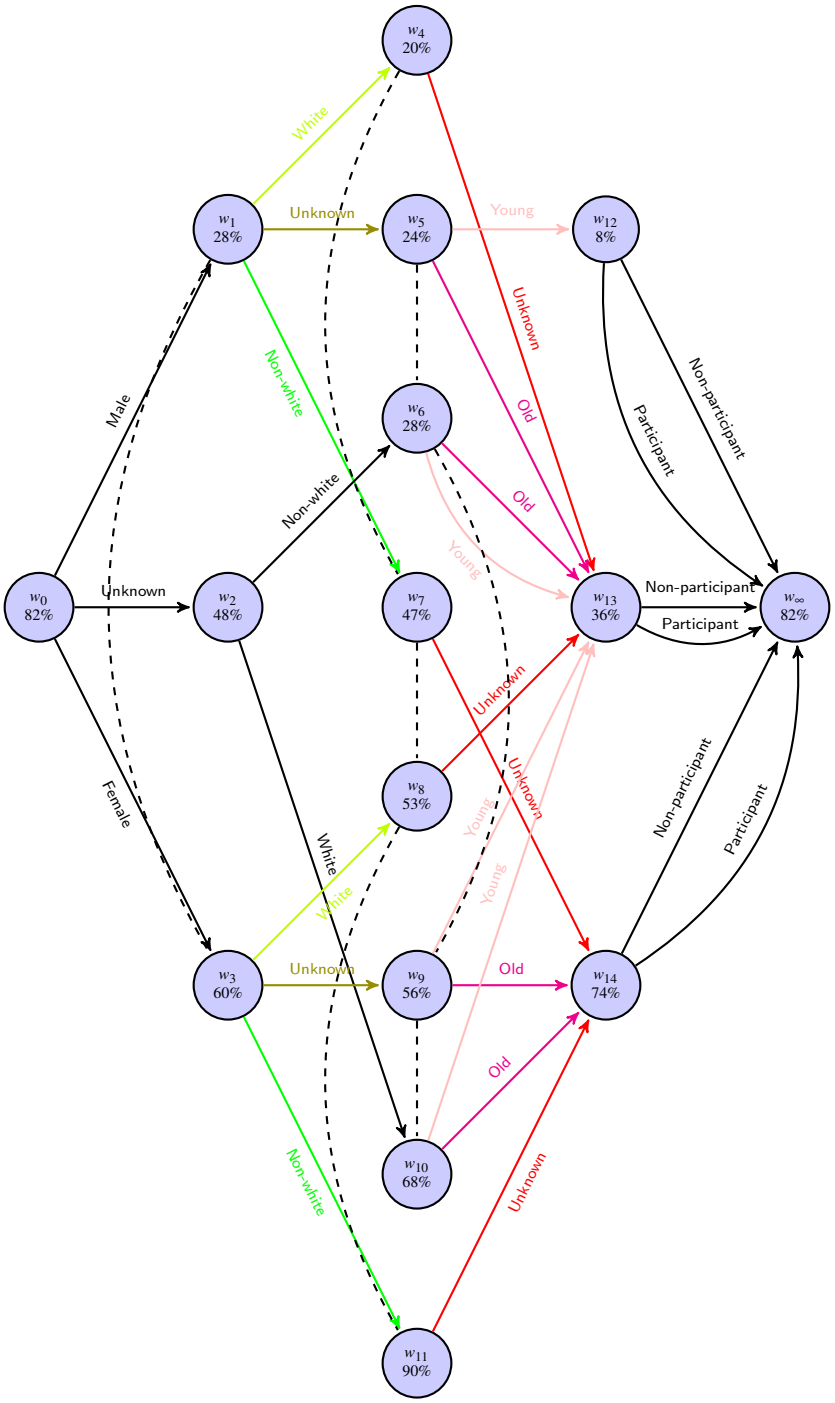

Figure D.2: Chain event graph for the amalgamated data. Percentage of participants shown at each position (W).

The distribution of age is indistinguishable given known gender and ethnicity, and it is also indistinguishable given unknown gender or ethnicity, as indicated by the corresponding colours and dashed lines between positions $\{w_4, w_7, w_8, w_{11}\}$ and $\{w_5, w_6, w_9, w_{10}\}$ in Figure D.2 respectively. If gender is unknown, non-white individuals are less likely to participate than white, since the non-white edge leads to position w_6, which is positioned higher than the white edge which leads to w_{10}.

The smallest probability of participation in the ordinal CEG ($w_{12} = 8\%$) can be reached only by one path; young males with unknown ethnicity. The greatest probability of participation in the CEG ($w_{11} = 90\%$) is reached only by non-white females. Those with older age are generally positioned lower in the ordinal CEG than those with younger age, suggesting older individuals are more likely to participate. Overall, age and gender appear to be associated with participation, with females and older individuals more likely to participate. There is no such association for ethnicity.

Initially there were three studies, one of which showed older females were more likely to participate, the second showed older white individuals were more likely to participate and the third showed non-white females were most likely to participate. Combining these studies into one overarching study allows for a larger sample size, since there are more participants, and a more generalisable conclusion, since these studies may have been located in different areas and with different research questions, and been affected by non-participation in different ways. Although of course if one of these studies already covers the research question and location of interest, it would be preferable to focus on that particular study. This approach of combining data may be useful in case-control studies which collect information regarding rare diseases and where participation rates have declined in recent years, to increase the overall sample size.

References

Barclay, L. (2014): *Modelling and reasoning with chain event graphs in health studies*, Ph.D. thesis, University of Warwick.

Barclay, L., J. Hutton, and J. Smith (2014): "Chain event graphs for informed missingness," *Bayesian Analysis*, 9, 53–76.

Buntine, W. (2013): "Chain graphs for learning," in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI1995)*, 46–54.

Cooper, G. (1990): "The computational complexity of probabilistic inference using Bayesian belief networks," *Artificial Intelligence*, 42, 393–405.

Dillman, D., G. Phelps, R. Tortora, K. Swift, J. Kohrell, J. Berck, and B. Messer (2009): "Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet," *Social Science Research*, 38, 3–20.

Freeman, G. (2010): *Learning and Predicting with Chain Event Graphs*, Ph.D. thesis, University of Warwick.

Freeman, G. and J. Smith (2011): "Dynamic staged trees for discrete multivariate time series : Forecasting, model selection and causal analysis," *Bayesian Analysis*, 6, 279–305.

Hartge, P. (2006): "Participation in population studies," *Epidemiology*, 17, 252–254.

Health Knowledge (2016): "Bias in epidemiological studies," http://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/biases, accessed online: 06/01/2016.

Heckerman, D. (1995): "A tutorial on learning Bayesian networks," Technical report, Microsoft Research.

Hernan, M., S. Hernandez-Diaz, and J. Robins (2004): "A structural approach to selection bias," *Epidemiology*, 15, 615–625.

Keeble, C., S. Barber, P. Baxter, R. Parslow, and G. Law (2014): "Reducing participation bias in case-control studies: Type 1 diabetes in children and stroke in adults," *Open Journal of Epidemiology*, 4, 129–134.

Keeble, C., G. Law, S. Barber, and P. Baxter (2015): "Choosing a method to reduce selection bias: A tool for researchers," *Open Journal of Epidemiology*, 5, 155–162.

Keeble, C., P. Thwaites, P. Baxter, S. Barber, R. Parslow, and G. Law (2017): "Learning through chain event graphs: The role of the mother in a childhood type I diabetes case control study," *American Journal of Epidemiology*, doi: 10.1093/aje/kwx171.

Little, R. and D. Rubin (2002): *Statistical Analysis with Missing Data*, Hoboken, New Jersey: Wiley.

Parsons, N. and M. Manierre (2014): "Investigating the relationship among prepaid token incentives, response rates, and nonresponse bias in a web survey," *Field Methods*, 26, 191–204.

Ron, D., Y. Singer, and N. Tishby (1998): "On the learnability and usage of acyclic probabilistic finite automata," *Journal of Computer and System Sciences*, 56, 133–152.

Schlesselman, J. (1982): *Case-Control Studies: Design, Conduct, Analysis*, New York: Oxford University Press.

Smith, J. and P. Anderson (2008): "Conditional independence and chain event graphs," *Artificial Intelligence*, 172, 42–68.

Sterne, J., I. White, J. Carlin, M. Spratt, P. Royston, M. Kenward, A. Wood, and J. Carpenter (2009): "Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls," *BMJ*, 338, 2393–2397.

Thwaites, P., J. Smith, and E. Riccomagno (2010): "Causal analysis with chain event graphs," *Artificial Intelligence*, 174, 889–909.