
Article:
Keeble, C orcid.org/0000-0003-1633-8842, Thwaites, PA orcid.org/0000-0001-9700-2245, Baxter, PD orcid.org/0000-0003-2699-3103 et al. (3 more authors) (2017) Learning Through Chain Event Graphs: The Role of Maternal Factors in Childhood Type I Diabetes. American Journal of Epidemiology, 186 (10). pp. 1204-1208. ISSN 0002-9262

https://doi.org/10.1093/aje/kwx171

Title Page

TITLE

Learning Through Chain Event Graphs: The Role of Maternal Factors in Childhood Type I Diabetes

AUTHOR LIST

Claire Keeble (BSc, MSc, PhD, GradStat), Peter Adam Thwaites (MSc, PhD), Paul David Baxter (BSc, PhD, PGCLTHE, CStat), Stuart Barber (PhD), Roger Charles Parslow (BSc, MSc, PhD), Graham Richard Law (BSc, PhD)

CORRESPONDING AUTHOR (Including approvals and answering questions)

Name: Dr Claire Keeble

Degrees: BSc, MSc, PhD, GradStat

Telephone: +44 (0)113 343 4315

Fax: +44 (0)113 343 4877

Email: c.m.keeble@leeds.ac.uk

Affiliation and address: Division of Epidemiology & Biostatistics, Leeds Institute of Cardiovascular and Metabolic Medicine (LICAMM), School of Medicine, 8.49 Worsley Building, University of Leeds, Leeds, West Yorkshire, LS2 9JT, ENGLAND

ABSTRACT

Chain event graphs are a graphical representation of a statistical model derived from event trees, previously applied to cohort studies but not to case-control studies. We apply the chain event graph framework to a Yorkshire case-control study of childhood type I diabetes, to examine four exposure variables associated with the mother, three of which are fully observed (her school-leaving-age, amniocenteses during pregnancy and delivery type) and one with missing values (her rhesus factor), while incorporating previous type I diabetes knowledge. We conclude that the unknown rhesus factor values are likely to be missing not at random, and are mainly rhesus positive. The mother's school-leaving-age and rhesus factor are not associated with the diabetes status of the child, whereas having at least one amniocentesis and, to a lesser extent, delivering by cesarean are, with the combination of both procedures further increasing the probability of diabetes. This application of chain event graphs to case-control data allows for the inclusion of missing data and prior knowledge, while investigating associations in the data. Communication of the analysis with the clinical expert is more straightforward than with traditional modelling, and this approach can be applied retrospectively or when assumptions for traditional analyses are not held.

Keywords: Case-Control Studies, Chain Event Graphs, Type 1 Diabetes Mellitus.

Abbreviations: CEG: chain event graph. OR: odds ratio.

Running head: The role of maternal factors in type I diabetes.

Word Count (abstract): 200 words.

Word Count (main text): 2829 words.

Chain event graphs (CEGs) are a graphical representation of a statistical model developed in statistics and artificial intelligence, which allow for different correlation structures in groups of data. Introduced in 2008, they are a form of directed graph which can be used to order and equate combinations of variable categories with respect to their probability of an outcome of interest(1-6). While the results of the analyses presented here are accessible, full understanding of the methods used will require understanding of the terminology for CEGs which can be found in Web Appendix 1; examples of these terms will later be given in Web Figure 1.

Case-control studies examine the possible association of variables with the disease of interest and the results usually report a subset of variables which are associated with the disease. However, it may be that certain categories of variables are associated (for example high and medium values), while others are not (such as low values). Alternatively it may be that combinations of categories from several variables are associated with the disease, such as a high value from two different variables, but traditional analyses often do not report this level of detail.

CEGs have been used for cohort studies(7) and causal analysis(8), but not to our knowledge with case-control studies. Here we apply the CEG framework to a type I diabetes data set to determine the association between variables linked to the mother and the development of type I diabetes in her child. These data have been analysed previously but these analyses have not been able to simultaneously identify variables and categories associated with diabetes, and draw conclusions with and about the missing data, while incorporating external information about the variables. We believe this approach is required for a thorough analysis of these data to address the research question.

METHODS

The process used to form the CEG and interpret the results is outlined below, with further details available in Web Appendix 1.

The diabetes data

The data were those relating to maternal factors present in a case-control study(9) recording cases of children under 16 years old diagnosed with type I diabetes while resident in the area of the former Yorkshire Regional Health Authority. The data consisted of 196 cases and 325 controls (129 matched triplets and 67 matched pairs)(9-13). The use of these data for this analysis was granted by the University of Leeds Research Ethics Committee HSLTLM/12/008.

The chronological ordering of the four categorical exposure variables and outcome is (i) rhesus factor of the mother, determined by the presence or absence of a protein in the blood (positive/negative/unknown), (ii) school-leaving-age of the mother, assuming the pregnancy begins after the mother has left school (16 or under/over 16) (iii) amniocentesis, usually during weeks 15–20 of the pregnancy(14) (yes - at least one with the study child/no - none), (iv) cesarean delivery at the end of the pregnancy (yes/no for the study child), and (v) diabetes status of the child, with type I diabetes diagnosis during childhood (case/control). Web Figure 1 displays the raw data and the ordering of the variables in the event tree, showing that all mothers with unknown rhesus factor also have delivery not by cesarean and do not have an amniocentesis, possibly suggesting an association between these

variables. Annotations on Web Figure 1 show examples of the terminology used. Chronological ordering allows for conditional or causal associations, although changes in the ordering of the variables should be tested where applicable, as will be included in the sensitivity analyses described in the Discussion.

A strength of CEGs being a Bayesian approach is that prior information from previous studies can be incorporated(13, 15-19). Population data which can be used as prior information for this study are given in Table 1. Table 1 shows that around 86% of the UK are rhesus positive(15, 16), but a limitation of population data are that they cannot be used to specify the expected percentage of unknown rhesus factor categories in a study. If the proportion of unknown rhesus factor from the data is used (3-4%) in conjunction with the data from Table 1, a split of negative:positive:unknown as 2:17:1 can be used as an approximation for the ratio of each category.

Table 1. Ratios of the Variable Categories in the Yorkshire Diabetes Data, Provided for the Time at Which the Study was Conducted. The True Case-Control Ratio is Simplified to 1:2 to Reduce the Equivalent Sample Size. For the Same Reason, the Rhesus Factor Ratio is Rounded.

First Author, Year (Reference No.)	Variable	Categories	Ratios	Assumptions
Home Health UK, 2015 (15). NHS, 2015 (16)	Rhesus factor	Negative:Positive: Unknown	-	Around 86% of the UK are rhesus positive.

Bolton, 2012 (17)	School-leaving-age	16 and under:Over 16	7:3	A parliamentary paper assuming the majority of mothers left school around 1970-1985.
Cambridge Fetal Care, 2013 (18)	Amniocentesis	Yes:No	1:49	Around 15,000 amniocentesis in Britain each year (about 2% of pregnancies).
Birth Choice UK, 2015 (19)	Caesarean	Yes:No	1:9	Around the time the children were born, around 10% of births were by cesarean.
-	Diabetes	Case:Control	1:2	Participants are in matched pairs (×67) or triplets (×129). For simplicity, let us assume that controls are twice as common as cases in these data.

The priors are calculated using the equivalent sample size of 3,000 (as described in Web Appendix 1) and the ratios shown in Table 1. The sum of the equivalent sample size at each edge associated with a given variable equals the overall equivalent sample size specified, and the value along each edge is guided by the proportions from Table 1. The priors are parameters of the Dirichlet distribution (described further in Web Appendix 1 and as shown in Web Figure 2). The sensitivity of the results to the equivalent sample size (using values of 30, 5 and 300,000), and the priors, is investigated in the Discussion.

Statistical analysis: The staged tree

The prior knowledge and data are combined using the Bayesian analysis. The agglomerative hierarchical clustering algorithm(2) implemented in R (R Core Team, Vienna, Austria)(7, 20), is used to convert the event tree into a staged tree. The resulting (staged) tree is given in Web Figure 3, with colouring showing which situations are in the same stage and which edges correspond, and labels showing the number of individuals taking each edge.

RESULTS

The chain event graph

Web Figure 4 shows the pruned ordinal CEG resulting from collapsing Web Figure 3 over its positions (W), with the percentage of cases given at each vertex. Pruning here refers to the removal of edges solely to produce a clearer diagram, in this case those edges which are not taken by any individual. An ordinal graph is the ordering of the vertices within a variable vertically with respect to their association with the binary outcome. The process of converting an event tree to a CEG and for interpreting a CEG were provided in Web Appendix 1.

Chain event graph conclusions

Web Figure 4 shows the design results in 38% of the individuals being cases. There is little difference between the rhesus factor categories, since around 40% of the individuals at each vertex are cases. In addition, the categories for the school-leaving-age do not display any clear pattern in the ordinal graph, with the over 16 years category leading to both the highest (w_9 = 50%) and lowest (w_4 = 20%) proportion of cases. This finding suggests rhesus

factor and school-leaving-age of the mother are not associated with the disease status of the child.

Mothers with at least one amniocentesis are situated towards the bottom of the ordinal graph, suggesting a higher probability of their child being a case, whereas those with no amniocenteses are situated towards the top of the graph, suggesting a higher probability of their child being a control. Therefore amniocentesis is clearly associated with the diabetes status of the child.

For the delivery of the child, there is a less clear pattern. However, generally the children delivered by cesarean have a higher probability of being a case than those not. The edges from $w_{10\text{-}18}$ to $w_{19\text{-}23}$ are those which depict cesarean delivery, and all the 'yes' edges lead to lower positions in the ordinal graph than the 'no' edges, with only w_{12} and w_{14} as exceptions. For these exceptions, w_{12} has both edges leading to the same vertex and for w_{14} the edges are only one vertex apart in the next variable, hence the difference in probability of disease is small. For vertices w_{19}-w_{23}, the paths containing no amniocentesis and delivery not by cesarean are positioned at least as high as those with at least one amniocentesis and cesarean delivery, showing the combination of these two variables to be associated with diabetes. Where there is only one of amniocentesis during pregnancy or caesarean delivery, those with cesarean delivery are generally positioned higher on the ordinal CEG than those with at least one amniocentesis, suggesting at least one amniocentesis is more of a risk factor than cesarean delivery.

The vertex with the highest probability of being a case (w_{23} = 100%) can be reached via three paths; each of which require at least one amniocentesis and cesarean delivery. This finding suggests rhesus factor and the school-leaving-age are not strongly associated with

the disease, while the other two variables may act as risk factors. However, it must be noted that there are only five cases present at this vertex. The vertex with the lowest probability of being a case (w_{19} = 25%) can be reached by two paths, both containing no amniocenteses and a school-leaving-age of over 16 years, suggesting possible associations with the disease. Unpopulated paths also provide information, for example there are no paths with amniocenteses or caesarean and unknown rhesus, which may suggest the rhesus category is recorded for an amniocentesis or cesarean.

Rhesus factor conclusions

The position of the unknown rhesus factor category in the ordinal CEG can be used to draw conclusions about the missingness mechanism(6). Since the unknown category (w_3) is positioned at the bottom of the ordinal CEG, underneath w_1 and w_2 which represent known rhesus factor, it is assumed that the rhesus data are missing not at random, since those with missing values are associated with a (marginally) higher probability of being a case than either the rhesus positive or rhesus negative categories. Although, the small percentage differences in positions w_1-w_3 should be noted (37%, 40%, 41%). If these data had been missing at random, the missing values would be expected to be a combination of the recorded values in a proportion similar to those in the data, and hence the missing category would be positioned between the recorded categories on the ordinal CEG. Since the missing category leads to a vertex with a more extreme probability than rhesus positive, it is likely that the majority of the missing values are rhesus positive.

DISCUSSION

Diabetes dataset summary

Both delivery by cesarean and having at least one amniocentesis were found to be associated with type I diabetes in the child, with amniocentesis more strongly associated than cesarean delivery, and with the combination of both procedures further increasing the association. The unknown rhesus factor values in the diabetes data were likely to be missing not at random, and mainly rhesus positive. Other variables which may be associated with diabetes in the child include the age of the mother and whether she has diabetes, or variables unrelated to the mother such as the health of the child in early life. The purpose of this analysis was to investigate associations between maternal factors and diabetes in the child, using those maternal variables available in the data, whilst considering any interactions between such variables and the nature of any missingness.

These data have been analysed previously(9-12) using approaches such as logistic regression, which found that amniocenteses and delivery type were significant in univariate analysis (Odd Ratio (OR)=3.85, 95% confidence interval: 1.34, 11.04 and OR= 1.84, 95% confidence interval: 1.09, 3.10 compared with non-assisted birth, respectively) while the rhesus factor and school-leaving-age were not (OR=0.90, 95% confidence interval: 0.56, 1.47 and OR=0.67, 95% confidence interval: 0.43, 1.04 respectively)(9). However univariate analysis offers no information about the combination of two or more variables. Another article used multivariable analysis for variables found to be significant in univariate analysis(10), which included delivery type (OR=1.45, 95% confidence interval: 0.82, 2.55) but not amniocentesis, rhesus factor or school-leaving-age. One previous analysis included only delivery type and school-leaving-age (OR=1.59, 95% confidence interval: 0.98, 2.59 and

OR=1.50, 95% confidence interval: 1.02, 2.19 respectively) and excluded any missing data(12), while another analysed only other variables from the data set(11). These analyses were therefore either overviews of the entire study, or addressed research questions other than those focusing on factors related to the mother. We are not aware of previous analyses of these data to determine the association of solely variables relating to the mother with type I diabetes in children, and we are confident these data have not been analysed using CEGs, nor are we aware of any other case-control studies analysed using CEGs.

All conclusions drawn here are based upon the 521 individuals in the case-control study conducted in Yorkshire(9), and it is acknowledged that a different dataset may lead to a different CEG with different conclusions. Analyses are encouraged, prospectively or retrospectively, of other diabetes datasets as agreement between studies would strengthen findings.

Associations are reported here and of course there may be unrecorded variables which are more closely associated with type I diabetes for which these recorded variables are acting as a proxy. These results nevertheless offer additional insight into the factors associated with type I diabetes. The conclusions for both the variables and the missingness were found to be insensitive to changes in the priors, the strength of the priors, and the ordering of the first two variables where the chronological ordering was less clear (see Web Appendix 2). Further conclusions regarding the missing rhesus factor values and the clinical implications of these findings are discussed in Web Appendix 3.

Comparison with traditional methods

CEGs have advantages over traditional methods. For example, they allow prior information to be incorporated in the analyses, which approaches such as logistic regression do not as standard. While methods such as Bayesian logistic regression are available, they are not common practice in calculations following case-control studies.

The non-parametric nature of CEGs can be advantageous. For example, CEGs could be used when assumptions for traditional analysis methods are not met, such as the rare-disease assumption for odds ratios or regression assumptions in modelling. Sparsely populated categories can also be troublesome during numerical analyses, but there are procedures in place for CEGs such as pruning the tree, combining edges or representing sparse edges using dotted lines(7).

Case-control studies are retrospective and this is often considered to be a negative feature of the study design, but one which may be advantageous for CEGs. Firstly, the number of variables and time period covered is known before analysis, and hence avoids the need for more complex graphs such as dynamic CEGs(22). Expert knowledge gained over time can also be incorporated into the analysis and inform paths which are more likely, or eliminate any paths which are not clinically plausible, in the same way as the data in Table 1 were utilized. One disadvantage of the retrospective study design may be the unclear variable ordering, upon which case-control study CEGs depend. The ordering of some variables will be obvious, while others may have occurred at seemingly the same time. For example, two variables such as amniocentesis and the occurrence of x-rays during pregnancy may be difficult to order chronologically. One way to circumvent this issue could be to create a new variable combining the two; there could be categories of ‘x-ray and amniocentesis’, ‘x-ray but no amniocentesis’, ‘no x-ray but amniocentesis’ etc, covering all combinations of the

two variables. Another approach is to test the effect of changing the ordering of the variables, as was shown in Web Appendix 2 for the first two variables in the diabetes data.

More generally, CEGs are a graphical approach, which may be preferable to numerical approaches for some researchers. The event tree used in the formation of the CEG can act as the basis for the analyst and the clinical expert to consider the plausibility of variable combinations and possible orderings of the required variables, resulting in a realistic variable subset entering the agglomerative hierarchical clustering algorithm. After analysis, the results can be presented to the expert as a combination of the variables which are most/least likely to result in the disease status, which can be easily interpreted without the need for advanced statistical training. This avoids the need for discussion of, sometimes complex, modelling such as interaction terms which may have hindered the necessary input from the expert if not a statistician. Interaction terms are discussed further in Web Appendix 3.

Extensions to the agglomerative hierarchical clustering algorithm used in the formation of a CEG could be developed to incorporate further contextual information, such as prior knowledge that two vertices should not, clinically or otherwise, be in the same stage. Extensions such as this would further improve the analysis and conclusions drawn from CEGs.

Summary

This application of CEGs to case-control data has allowed for a concise analytic approach which incorporates missing data and prior knowledge.

We conclude that amniocenteses and cesarean delivery are associated with increased probability of a child with type I diabetes, and encourage investigation into the possible causal links. The occurrence of both procedures further increased the probability of diabetes, and cesarean was found to be less strongly associated with diabetes than amniocentesis. We found no such association for the school-leaving-age or rhesus factor of the mother. We also believe the missing rhesus categories are missing not at random and that those with unknown rhesus factor are likely to be rhesus positive. Case-control studies do not typically present this level of detail in their findings, hence demonstrating an advantage of chain event graphs.

ACKNOWLEDGMENTS

Author affiliations: Division of Epidemiology and Biostatistics, University of Leeds, United Kingdom (Claire Keeble, Paul David Baxter, Roger Charles Parslow, Graham Richard Law); Department of Statistics, University of Leeds, United Kingdom (Peter Adam Thwaites, Stuart Barber).

This work was supported by the Higher Education Funding Council for England and a Medical Research Council Capacity Building Studentship.

Conflict of interest: none declared.

References

1. Smith JQ, Anderson PE. Conditional independence and chain event graphs. *Artificial Intelligence.* 2008;172(1):42-68.

2. Freeman G, Smith JQ. Bayesian MAP model selection of chain event graphs. *Journal of Multivariate Analysis.* 2011;102(7):1152-1165.

3. Thwaites P. Causal identifiability via Chain Event Graphs. *Artificial Inteligence.* 2013;195:291-315.

4. Silander T, Leong T-Y. A Dynamic Programming Algorithm for Learning Chain Event Graphs. In: *Discovery Science, volume 8140 of Lecture Notes in Computer Science*. Berlin: Springer; 2013: 201-216.

5. Barclay LM, Hutton JL, Smith JQ. Refining a Bayesian Network using a Chain Event Graph. *International Journal of Approximate Reasoning.* 2013;54(9):1300-1309.

6. Barclay LM, Hutton JL, Smith JQ. Chain Event Graphs for Informed Missingness. *Bayesian Analysis.* 2014;9(1):53-76.

7. Barclay LM. Modelling and reasoning with chain event graphs in health studies [dissertation]. Warwick, UK: University of Warwick; 2014.

8. Thwaites P, Smith JQ, Riccomagno E. Causal analysis with Chain Event Graphs. *Artificial Intelligence.* 2010;174(12-13):889-909.

9. McKinney PA, Parslow R, Gurney K, et al. Antenatal Risk Factors for Childhood Diabetes Mellitus; A Case-Control Study of the Medical Record Data in Yorkshire, UK. *Diabetologia.* 1997;40(8):933-939.

10. McKinney PA, Parslow R, Gurney KA, et al. Perinatal and neonatal determinants of childhood type 1 diabetes. A case-control study in Yorkshire, UK. *Diabetes Care.* 1999;22(6):928-932.

11. Fear NT, McKinney PA, Patterson CC, et al. Childhood type 1 diabetes mellitus and parental occupations involving social mixing and infectious contacts: Two population-based case-control studies. *Diabet Med.* 1999;16(12):1025-1029.

12. McKinney PA, Okasha M, Parslow RC, et al. Early social mixing and childhood Type I diabetes mellitus: A case-control study in Yorkshire, UK. *Diabet Med*. 2000;17(3):236-242.

13. Keeble C, Barber S, Baxter PD, et al. Reducing Participation Bias in Case-Control Studies: Type 1 Diabetes in Children and Stroke in Adults. *Open Journal of Epidemiology.* 2014;4(3):129-134.

14. NHS choices. Amniocentesis. http://www.nhs.uk/conditions/Amniocentesis. Updated April 21, 2016. Accessed October 23, 2015.

15. Home Health UK. The Blood. http://www.homehealth-uk.com/medical/blood.htm. Accessed December 15, 2015.

16. NHS. Rhesus disease - Causes. http://www.nhs.uk/Conditions/Rhesus-disease/Pages/Causes.aspx. Updated July 31, 2015. Accessed December 15, 2015.

17. Bolton P. Education: Historical statistics - Parliament. http://www.parliament.uk/briefing-papers/SN04252.pdf. Updated November 27, 2012. Accessed December 2, 2015.

18. Cambridge Fetal Care. Amniocentesis Test. http://www.fetalcare.co.uk. Accessed August 13, 2013.

19. Birth Choice UK. Graphs Of Historical Cesarean Section Rates. http://www.birthchoiceuk.com. Updated March 10, 2015. Accessed August 13, 2013.

20. R Development Core Team. R: A Language and Environment for Statistical Computing. http://www.r-project.org/. Updated July 16, 2016. Accessed 2015.

21. Cardwell CR, Stene LC, Joner G, et al. Cesarean section is associated with an increased risk of childhood-onset type 1 diabetes mellitus: A meta-analysis of observational studies. *Diabetologia.* 2008;51(5):726-735.

22. Barclay LM, Collazo RA, Smith JQ, et al. The dynamic chain event graph. *Electronic Journal of Statistics*. 2015;9(2):2130-2169.