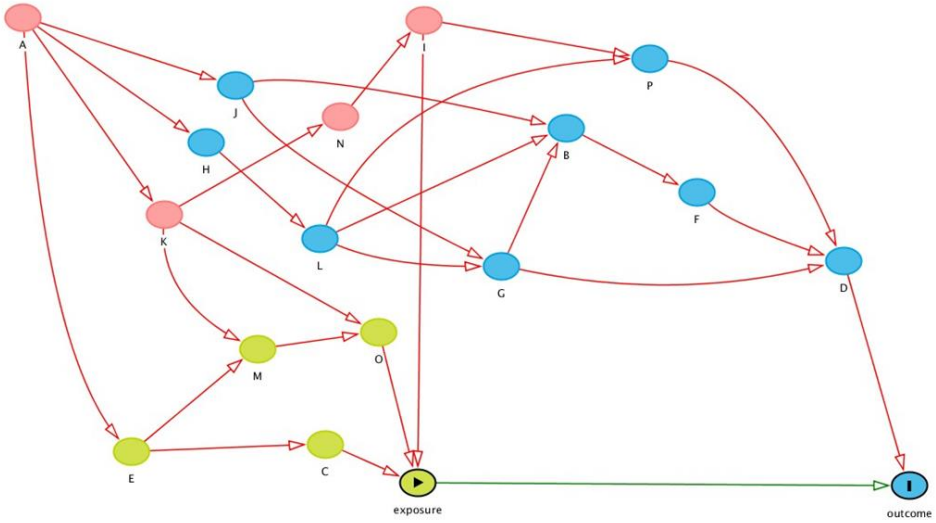
NiPN-Ethiopia
National Information Platform for Nutrition

Ethiopian Public Health Institute

Using DAGs to Inform Analyses: Identifying Drivers of Non-Communicable Diseases in Ethiopia



Addis Ababa

June 2021

Contents

Background

The National Information Platform for Nutrition (NIPN) is a global initiative launched by the European Commission to support Scaling Up Nutrition countries that have a high malnutrition burden. It supports the generation of evidence that is used by nutrition stakeholders to develop policy, design programs and to allocate investments. Global support to the NIPN is managed by Capacity for Nutrition (C4N) which is part of the German Federal Ministry for Economic Cooperation and Development's Knowledge for Nutrition Program and is implemented by the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ).

The NIPN in Ethiopia was launched in 2018. It is hosted at the Ethiopian Public Health Institute (EPHI) within the Food Science and Nutrition Research Directorate (FSNRD). Technical assistance to the NIPN is provided by the International Food Policy Research Institute (IFPRI) under the Ethiopia NIPN Technical Assistance Project (ENTAP). Both institutions are funded by the European Union (EU) Delegation to Ethiopia, with support from the Foreign Commonwealth and Development Office and the Bill and Melinda Gates Foundation.

The NIPN in Ethiopia aims to strengthen national capacity to monitor progress towards the reduction in under-nutrition and to implement more cost-effective and evidence-based nutrition policies and programs. It promotes evidence-based decision making for nutrition and supports the implementation of the National Nutrition Program (NNP) and the Food and Nutrition Policy (FNP).

From the analysis of available and shared data, the NIPN generates evidence that is used by national stakeholders to influence policies and programs, through an operational cycle consisting of three elements that constantly revolve and feed into each other: question formulation based on government priorities; analysis of data to inform the questions; communication of the findings back to policy and decision makers. The NIPN in Ethiopia works under the national multisectoral nutrition governance structures. It constantly engages multisectoral stakeholders to build monitoring, evaluation, and research capacities to support evidence-based decision making.

Introduction

During the 2019 NIPN policy question formulation process, two research questions were prioritized; one was on the "Progress in Water, Sanitation and Hygiene (WASH) coverage and its contributions to the reduction in stunting and diarrhea", and the other was on the "Drivers of non-communicable diseases, i.e., overweight, obesity, diabetes, and hypertension, in Ethiopia".

To respond to the second research question and as part of its technical assistance to the NIPN, IFPRI procured the services of a collaborator who brought specific expertise in the analyses of secondary data, particularly from cohort studies. Of relevance to the priorities of the NIPN, the collaborator has a portfolio of work related to obesity, cardio-metabolic risk factors and

obesity-related non-communicable diseases such as type 2 diabetes, hypertriglyceridemia and hypertension. The collaborator has experience of working with large complex datasets and in the application of a variety of statistical methods including causal pathway diagrams (aka directed acyclic graphs, DAGs) to comprehend the causal framework relating exposures, covariates and outcomes so that suitable analysis multivariable models can be compiled.

The IFPRI collaborator proposed incorporating DAGs into the analysis plan when he began working with the NIPN team in April 2020. He shared several articles and resources to provide some background and rationale for using the approach. On the 25th of May 2020, he delivered a presentation to the group on the basics of causal path diagrams and in using DAGitty.

Based on these learnings, the NIPN team decided to adopt this new methodological technique (causal path diagrams) into the analysis plan. In adopting this approach, the NIPN team would not only acquire a novel research skill, but it would also expose the team to a different theoretical framework, both of which can be utilized in related work in the future.

The IFPRI collaborator, following the consultation with the experts (see below) and in preparation of the team constructing the formal DAGs, delivered a tutorial to the team on the use of DAGitty. The material covered in this tutorial included: adding and deleting variables, connecting variables with arcs, adjusting for a variable, changing the exposure, adding unmeasured variables, identifying the minimal adjustment set, saving the DAG as a picture/pdf for use in publications, and saving the model code.

What are Causal Path Diagrams?

Causal path diagrams are a visual summary of causal links amongst variables based on a priori contextual knowledge and understanding. This visual summary of variable interrelationships is used in causal analysis and such diagrams are becoming increasingly employed in epidemiological analyses (Hoggart et al. 2003). Despite their increasing use however, they remain considerably underutilized.

There are various ways in which causal diagrams may be used, e.g., to think clearly about how the exposure, outcome, and potential confounding variables are causally related; to communicate these causal inter-relationships to the audience; or to indicate which variables are important to measure (prior to data collection). In the case of the NIPN team, the purpose was to inform the statistical modelling process, particularly the identification of confounders, to ultimately obtain unbiased (or less biased) estimates of the association between specified exposure and outcome variables.

What are DAGs and DAGitty?

Causal path diagrams are the basis of a formal theoretical framework in which causal relationships can be identified and evaluated. The simplest kind of causal path diagram is a directed acyclic graph (DAG). DAGs are based on graphical model theory and identify covariates or ‘adjustment sets’ that upon adjusting for, remove all confounding (from the

specified DAG). Specifically, the graphical rule used to identify these adjustment sets is the 'back-door criterion' (Pearl 2009)[1].

A DAG consists of 'nodes' that represent variables (e.g., X, M, Y) and arrows that depict direct causal effects (e.g., X→M). To describe relationships between variables in a diagram, we often read them like an ancestry tree. For example, in the diagram X→M→Y, M is a direct descendent of X and X is a direct ancestor of M; M and Y are descendants of X, and X and M are ancestors of Y. A causal diagram is called a directed acyclic graph if no variable is an ancestor of itself (i.e., no loops exist). Arrows in a DAG reflect *a priori assumptions about cause and effect in a particular context*, based either on firm knowledge and understanding of actual or likely relationships between variables, or based on speculative hypotheses. Critically, these assumptions are specified a priori (i.e., before looking at any data) and *cannot be inferred empirically from the data on which the analyses are to be conducted*.

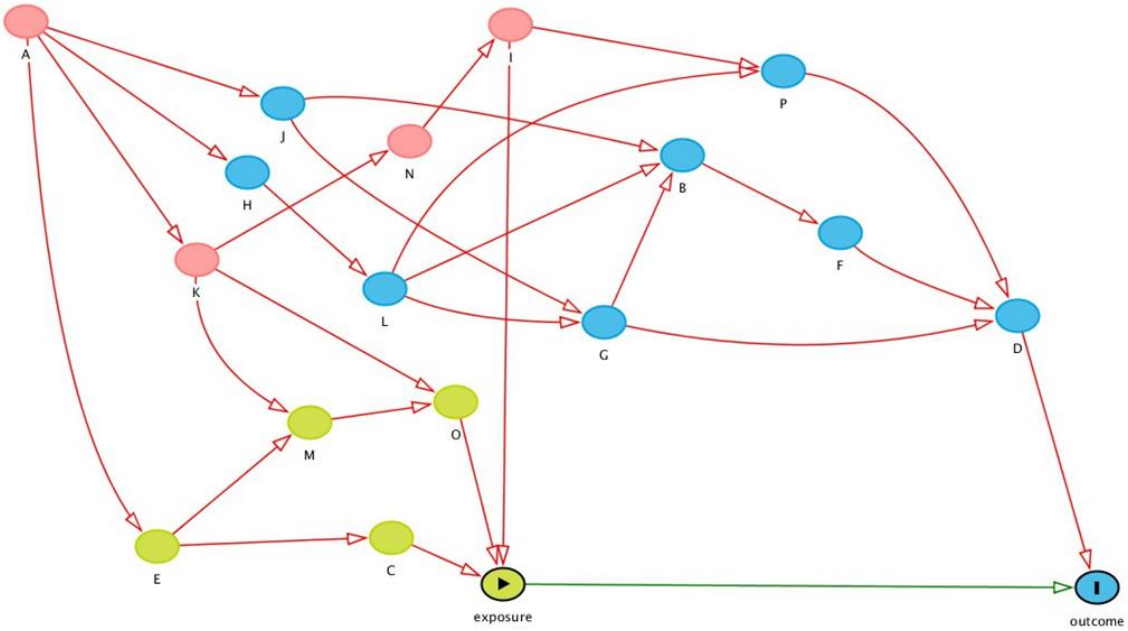
Despite their obvious use in building and interpreting correct statistical models, DAGs remain an oversimplification of the causal relationships between variables. For example, DAGs are non-parametric in that there are no assumptions made about how included variables are measured or of their distributions. Furthermore, DAGs do not indicate the direction (positive or negative) or magnitude of effects between variables or whether interactions exist (Hernan et al. 2004). Nonetheless DAGs enable researchers to think clearly and logically about their research question, and to make explicit any assumptions being made about the relationships amongst variables. This DAG and the underlying assumptions can then be communicated to other researchers, which increases research transparency and potentially increase the replicability of findings.

Drawing DAGs is not straightforward, especially for the uninitiated. However, there exists an online tool, DAGitty (http://www.dagitty.net/) or the R package –dagitty- (Textor et al. 2016), which can be used to construct DAGs in a systematic and robust way.

An example DAG produced using DAGitty is presented below. In the model the exposure is called 'exposure' (yellowish green node (variable) with a triangle in), the outcome is called 'outcome' and is (blue node with a line in it). Direct ancestors of the outcome are in blue, and direct ancestors of the exposure are in green. The nodes in pink are the ones which are common causes. Red arcs are the biasing paths, and these are the ones which have been identified as backdoor paths from the exposure to the outcome and which will have to be removed in order to remove bias from common causes. In this example, DAGitty has identified that the minimal sufficient adjustment sets for estimating the total effect of 'exposure' on 'outcome' are variable A and I. The DAG can then either be saved in figure format or translated into code which can be saved in a text file and then re-entered into the software later, thus simplifying the process of editing the DAG.

---

[1] '*Backdoor criterion: Given an ordered pair of variables (X, Y) in a directed acyclic graph G, a set of variables Z satisfies the backdoor criterion relative to (X, Y) if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X* (Pearl, Glymour & Jewell 2009).



Nonetheless, as DAGitty is reliant on the information included by the user, i.e., the DAG is only as good as the things you put in, before attempting to construct the DAG in the program, it is important to not only conduct a comprehensive literature review of the research topic (e.g., drivers of non-communicable diseases) but crucially, obtain specialist and contextual knowledge from experts in the fields related to the research topic (here nutrition, chronic disease epidemiology and public health).

The NIPN's Application and Use of the DAGs

The Systematic Review and Expert Consultation

The first step the NIPN team took, as part of the process of constructing the DAGs which would inform the analyses, was to conduct a systematic literature review to identify nutritional, behavioral, and socio-economic factors associated with overweight/obesity, diabetes, and hypertension in Sub-Saharan Africa. Using the variables identified from the systematic review, a set of provisional causal path diagrams, specific to each of the outcomes, were constructed to visualize the relationships between drivers, outcomes, and covariates. These preliminary causal path diagrams were then taken forward to facilitate discussion during a planned consultation with several invited experts from the fields of nutrition, non-communicable disease epidemiology and public health.

On 5th June 2020, the NIPN team hosted an online seminar with 14 of these invited experts. While this seminar primarily served to provide an opportunity for the NIPN team to draw upon the expert knowledge of these collaborators which could then be incorporated into a

final set of DAGs, it also represented an opportunity for the IFPRI collaborator to provide some background of the methodology to a new audience. He made a presentation on the DAGs, which also provided another opportunity for the team to understand the approach.

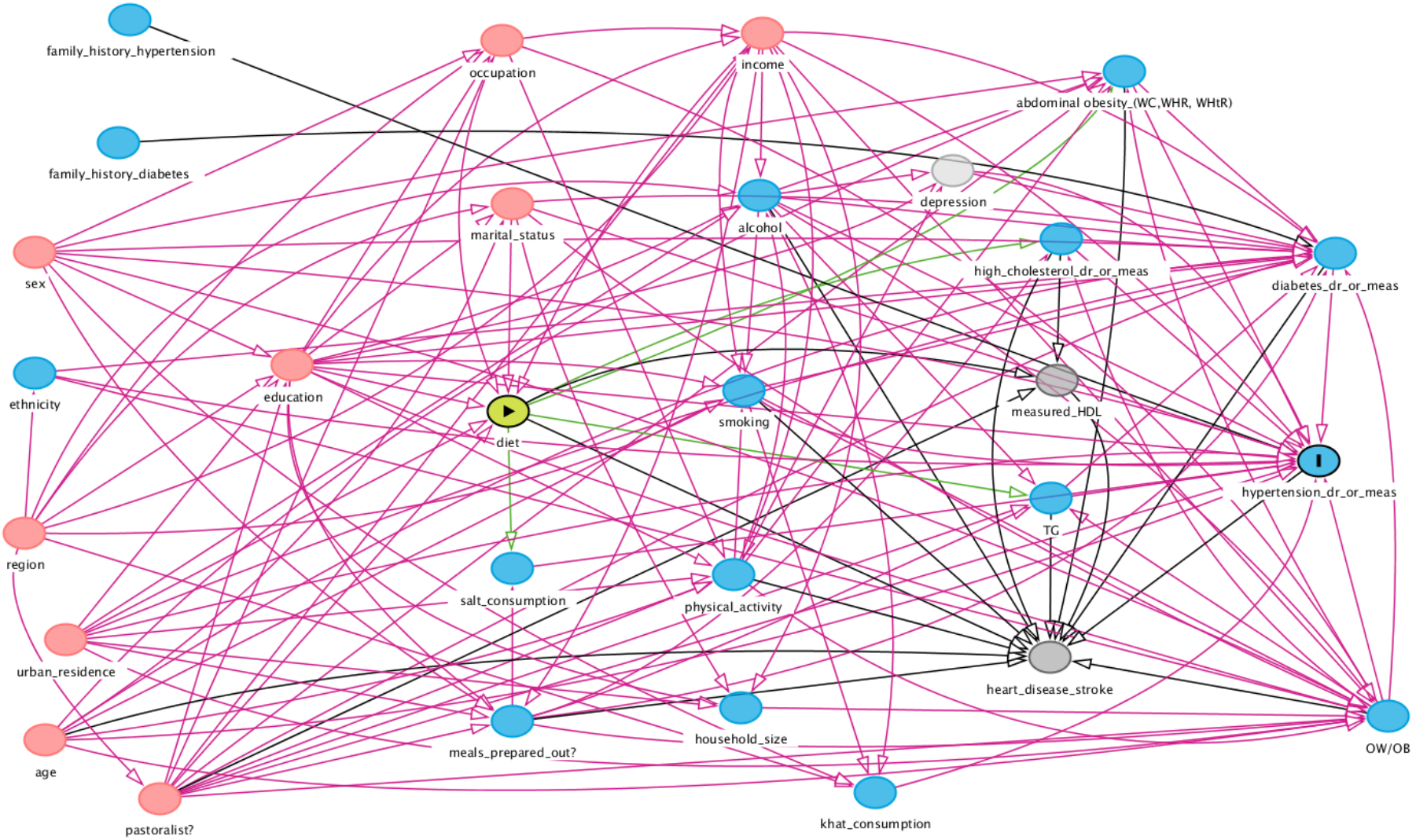
Using the provisional outcome-specific DAGs constructed using the variables identified in the systematic review, the invited experts were invited to thoroughly inspect the DAGs. The experts were firstly asked to identify any variables they believed to be missing from the DAGs and after identifying these, adding all of the associations between this new variable and the other variables in the DAG. They were then asked to identify any missing associations between variables already in the DAG and finally, asked whether they disagreed with any of the associations already drawn in the DAG. It was explained that the removal of a connection between two variables is an explicit assumption that there is no direct causal relationship between the variables (in either direction) and not simply that it is unknown whether a relationship exists. As such, the omission of a connection between two variables is a strong assumption which should be supported by a strong evidence base. In the absence of such support, it was explained that connections should be drawn.

After the consultation, the NIPN team, with support from the IFPRI collaborator, redrew the DAGs incorporating the amendments suggested by the experts and the model code was saved. Figures of the DAGs were then shared with the experts for a final time to ensure that they captured all of the modifications.

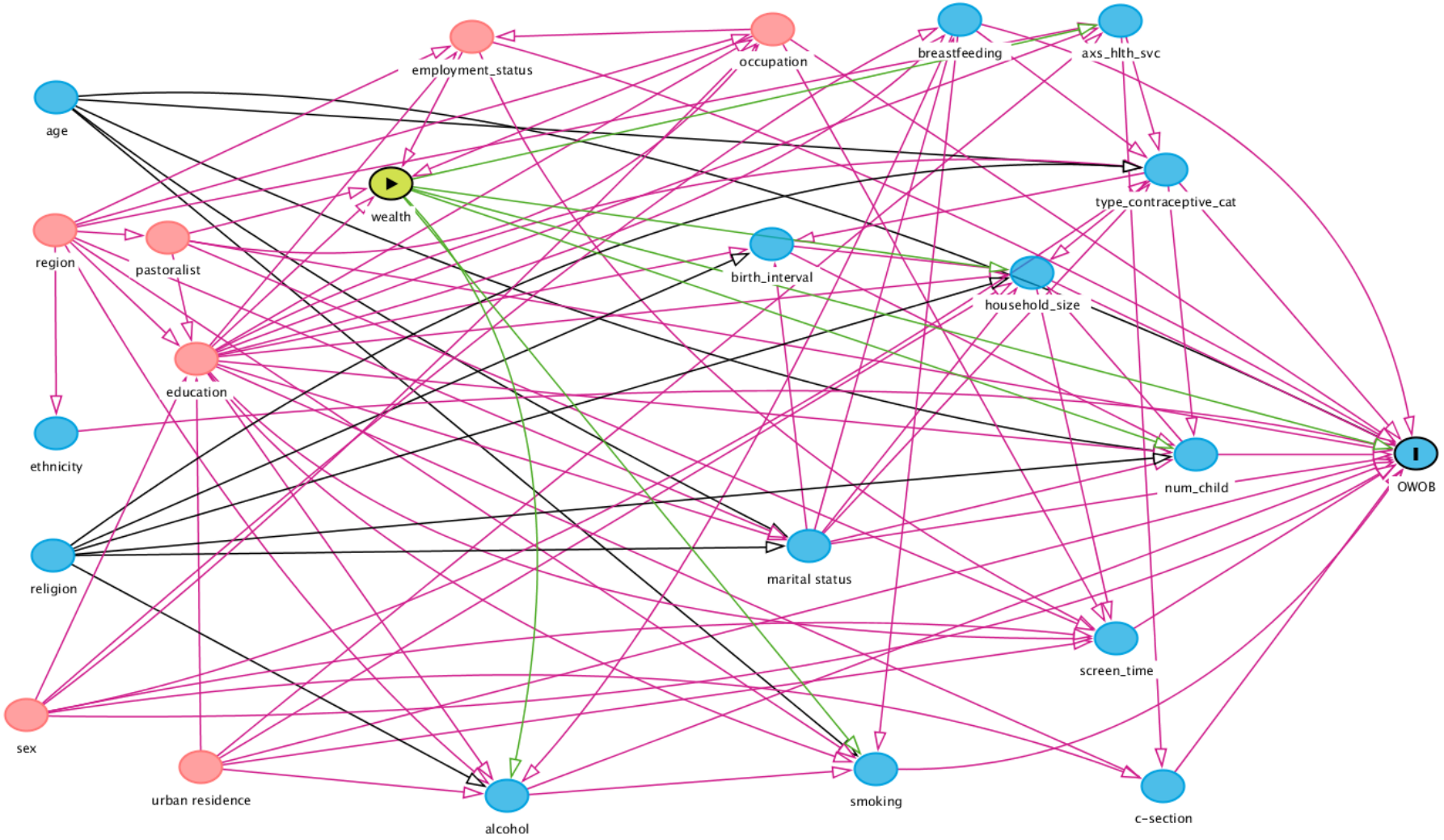
How Was the DAG Used in the NCD Analysis?

Once the outcome specific DAGs (i.e., overweight/obesity and hypertension/diabetes) had been created, the team set about identifying the relevant variables in the two datasets identified, to address the study aims, namely multiple sweeps of the Ethiopia Demographic Health Survey (EDHS) and the Ethiopia non-communicable diseases (NCD) Steps survey. DAGs should be constructed without knowledge of whether a particular variable is available or not. Nonetheless, it is highly likely that some of the variables included in the DAG are unavailable. In this instance, instead of removing the unavailable variables from the DAG, these variables should be retained but modified to denote that they were unavailable. This transparency enables reviewers of the DAG to see which variables were unavailable and thus assess how realistic the DAG is. As such, after reviewing the data available in the respective datasets, the NIPN team went about amending the DAGs to highlight any missing (or ‘unmeasured’) variables. The final DAGs can be seen below, firstly for the EDHS data and then for NCD Steps.

NCD Steps DAG



EDHS DAG


employment_status
occupation
breastfeeding
axs_hlth_svc
age
wealth
type_contraceptive_cat
region
pastoralist
birth_interval
household_size
education
ethnicity
num_child
OWOB
religion
marital status
screen_time
sex
urban residence
alcohol
smoking
c-section

Once complete, the team was then able to begin using the DAGs to inform the analysis. Specifically, the final model code was uploaded into DAGitty and then for each of the overweight/obesity, hypertension, and diabetes outcomes and for each of the identified predictor variables specified *a priori* (above), the team set about identifying the minimal adjustment sets necessary to remove confounding bias from the constructed DAGs. This was a straightforward process and simply entailed re-specifying the exposure variable within the DAG. DAGitty then automatically updated the adjustment sets and reported these to the user. This process was repeated for each of the exposures, resulting in a set of outcome-exposure specific adjustment sets to be taken forward into the multivariable regression models. As mentioned previously however, DAGs are non-parametric and do not specify the nature of the relationship between variables (e.g., linear vs non-linear) or whether transformations of variables are required. As such, the team conducted relevant data preparation before the final models were run. These DAGs and data preparations and analyses were saved on a shared Dropbox folder, easily accessible to the team members working on it.

The collaborator provided technical assistance during the analysis, via email correspondence and weekly virtual meetings. In accordance with availability of data in the NCD Steps and EDHS datasets and the DAGs produced, the team finalized which variables would be used in the analyses. At this stage, the collaborator provided assistance with regard to the cleaning of the anthropometric and cardio-metabolic data and the derivation of clinically relevant variables such as overweight, obesity, hypertension, diabetes which were based on published criteria. He advised on the use of sex-stratified analyses, discussed the possibility of conducting multiple imputations to handle missing data, and proposed a set of sensitivity analyses to test the robustness of findings to different assumptions (e.g., combining different body mass index (BMI) categories, changing the definition of diabetes and hypertension). The collaborator provided guidance and shared literature regarding the use of non-linear decomposition methods which were not previously employed by the NIPN team (linear probability models were previously favoured). After decomposition models and multivariable logistic models were finalized, he contributed to the interpretation of findings, provided recommendations for the presentation of the results, and assisted in the production of these.

Strengths and Limitations of the DAG Approach

The key strength of a DAG is that it forces researchers to think clearly and logically about their research question and in doing so, make explicit assumptions about the relationships between exposure(s), outcome(s) and covariates(s). By providing the DAG as part of the published output, external parties are able to understand and evaluate the assumptions which have been made and then test these assumptions in a separate sample. As such, the publication of the adopted DAG represents an opportunity to increase the reproducibility of research findings and thus provide more robust findings to guide the development of policy recommendations.

Nonetheless, DAGs remain an oversimplification of the relationships between variables. As a non-parametric tool, DAGs make no assumptions made about how included variables are measured or of their distributions. Furthermore, DAGs do not indicate the direction (positive or negative) or magnitude of effects between variables or whether interactions exist. It is the job of the researchers to carefully think about the parameterization of these variables, using expert subject knowledge and a comprehensive review of the literature to guide the process. Sensitivity analyses should also be performed to test the robustness of findings to any modifications to the DAG or to the parameterization of any variables within the DAG.

References

Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004;15:615-25.

Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG et al. Control of confounding of genetic associations in stratified populations. Am J Hum Genet. 2003;72:1492-504.

Pearl J. Causality: Models, Reasoning, and Inference. Cambridge: Cambridge University Press, 2009.

Textor J, van der Zander B, Gilthorpe MS, Liskiewicz M, Ellison GT. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. Int J Epidemiol. 2016;45:1887-94.