Centre of Excellence for Development Impact and Learning

March 2023

Research project paper 5

Process Outcome Integration with Theory (POInT): academic report

Matthew Juden, Tichaona Mapuwei, Till Tietz, Rachel Sarguta, Lily Medina, Audrey Prost, Macartan Humphreys, Alan Jacobs, Elizabeth Allen, Henry Mwambi, Calum Davey

LIDC London International Development Centre

UKaid from the British people

Oxford Policy Management

About CEDIL

The Centre of Excellence for Development Impact and Learning (CEDIL) is an academic consortium supported by the UK government through UK Aid. The mission of the Centre is to test innovative methodologies in evaluation and evidence synthesis and promote evidence-informed development. CEDIL-supported projects fall into three programmes of work: evaluating complex interventions, enhancing evidence transferability, and increasing evidence use.

CEDIL research project paper

CEDIL research project papers are the research outputs of the projects supported by CEDIL. They include impact evaluations, systematic reviews, evidence and gap maps and exploratory papers on research methods and approaches.

About this research project paper

This paper was submitted to CEDIL by the "Process Outcome Integration with Theory" L.292 team.

Please direct any comments or queries to the corresponding author, Calum Davey at c.davey@niot.org.uk.

Suggested citation: Juden, M., Mapuwei, T., Tietz, T., Sarguta, R., Medina, L., Prost, A., Humphreys, M., Jacobs, A., Allen, E., Mwambi, H. and Davey, C. (2023) 'Process Outcome Integration with Theory (POInT): academic report', CEDIL Research Project Paper 5. Centre of Excellence for Development Impact and Learning (CEDIL), London and Oxford. Available from: https://doi.org/10.51744/CRPP5

This project was funded by the Centre of Excellence for Development Impact and Learning (CEDIL), supported by UK aid from the UK Government. The views expressed in this research project paper do not necessarily reflect the UK Government's official policies or CEDIL.

All content is the sole responsibility of the authors. CEDIL has not commissioned a professional copy-edit of this paper. Any errors and omissions are the sole responsibility of the authors.

Abstract

We have developed and tested a novel approach to evaluating development interventions - the POInT approach. We used Bayesian causal modelling to integrate process and outcome data to generate insights about all aspects of the theory of change, including outcomes, mechanisms, mediators and moderators. We partnered with two teams who had evaluated or were evaluating complex development interventions: The UPAVAN team had evaluated a nutrition-sensitive agriculture intervention in Odisha, India, and the DIG team was in the process of evaluating a disability-inclusive poverty graduation intervention in Uganda.

We adapted the partner teams' theory of change into a formal causal model, depicted as a directed acyclic graph (DAG). We specified the DAG in the statistical software R, using the CausalQueries package, having extended the package to handle large models. Using a novel prior elicitation strategy to elicit beliefs over many more parameters than has previously been possible, we elicited the partner teams' beliefs about the nature and strength of causal links in the causal model (priors) and combined these into a single set of shared prior beliefs. We updated the model on data alone as well as on data plus priors to generate posterior models under different assumptions. Finally, we queried the prior and posterior models to learn about estimates of interest, and the relative role of prior beliefs and data in the combined analysis.

Our study finds multiple gains from representing theories of change as causal models, with these providing a way to link practitioner beliefs to estimation strategies; in addition, we identify specific lessons for this elicitation process. We also developed simplified methods for eliciting research beliefs---priors---over these models in a way that can be used for analysis. In both applications we found that the models that were elicited from practitioners were generally extremely complex, and considerably more complex than current methods can handle. We show that many queries can be fruitfully asked using parts of these models. In such cases we often found estimated effects were smaller than would be expected given practitioner priors. However, queries that depend on the *full* model are generally difficult to calculate even with our extensions to existing tools. We thus see a tension between the goal of faithfully representing practitioners' theories as models and the goal of accessing minimally complex models that can be effectively analysed, with implications both for the development of tools and for procedures for eliciting models. One approach is to identify queries earlier in the POInT process and elicit models that are minimally complex for the queries of interest.

1. Introduction

Description of the project

We have developed and tested a novel approach to evaluating development interventions. This approach allows evaluators to integrate process and outcomes data to generate insights about all aspects of their theory of change, including outcomes, mechanisms, mediators[1] and moderators[2]. The method uses a Bayesian causal modelling approach which we have adapted and extended from proof-of-concept work previously conducted in political science. We call this method POInT: Process-Outcome Integration with Theory.

We partnered with two teams who had evaluated or were evaluating complex development interventions. The UPAVAN team had evaluated a nutrition-sensitive agriculture intervention in Odisha, India, and the DIG team was in the process of evaluating a disability-inclusive poverty graduation intervention in Uganda. With these teams we were able to:

1. Adapt the partner teams' theory of change into a formal causal model, depicted as a directed acyclic graph (DAG).
2. Specify the DAG in the statistical software R using the CausalQueries package.
3. Elicit the partner team's beliefs about the nature and strength of causal links in the causal model (priors).
4. Combine expert priors into a single set of shared prior beliefs.
5. Update the model on data alone or data and priors to generate posterior models.
6. Query the prior and posterior models to learn about estimates of interest, and the relative role of prior beliefs and data in the combined analysis.

[1] Mediation occurs when the effect of one variable on another passes through a third variable, e.g., A→M→Y. M is a mediator of the relationship between A and Y. A may also have a direct effect on Y or other indirect effects through other mediators.

[2] Moderators act to either intensify or frustrate the activation of a mediating node on the causal chain.

To achieve 2., 5. and 6. we extended and improved CausalQueries to make it suitable for moving beyond small examples to handle the larger models implied by the kinds of theories of change being used by development practitioners.[3] It is important to note that CausalQueries was not developed for this project but was considerably expanded through the experiences of working with real life evaluation examples considered in this project. An additional package was developed by this project CausalQueriesTools which expands CausalQueries capabilities to handle larger models.

To achieve 3. we developed a novel prior elicitation strategy that allowed us to elicit an order of magnitude more parameters than was previously possible.

To achieve 4. and 5. we developed an approach to interpreting prior beliefs as the data which, if observed, would justify those beliefs, and then update the model on that data.

Objectives of the project

The aim of the project was to develop a rigorous method for integrating process and outcome data with programme theory that will help evaluators make better inferences about the theory of change and the transportability of the results to other contexts.

We sought to unify three strands of modern best-practice evaluation research: theories of change to describe complex interventions (Breuer et al., 2015; Davies, 2018; De Silva et al., 2014) , mixed-methods process and impact evaluations (Aarons et al., 2012; Cartwright, 2019; Creswell et al., 2011; Jimenez et al., 2018), and the use of middle-range theory as the basis for transporting findings to other contexts (Giddens, 1984; Leviton, 2017; Pearl & Bareinboim, 2014). In summary, the key objectives of the project were:

1. To adapt theories of change into formal causal models and to investigate and report the strengths and weaknesses of such an approach.

2. To demonstrate how process and outcome information can be combined to make optimally-informed inferences about the effects that occurred and the underlying processes that drove those effects.

3. To develop and validate an approach that may help to answer the key question facing policy makers and forward-looking intervention designers: how can we best learn from an evaluation about what to do next time.

The research therefore sought to answer the following questions:

[3] More work of this sort is still required to fully implement step 6. (queries) for the UPAVAN model, due to its size. We are confident that we have taken good steps in this direction that we should complete within the project funding period and be able to write up for academic publication. If we are not able to complete these steps, the progress made so far means that they should be taken up by ourselves or other researchers after this funding period ends.

1. How can theories of change (ToCs) be adopted into formal causal models that can be used as standard diagrammatic language in the causal inference literature?
2. How can qualitative and quantitative information be combined to make optimally-informed inferences about the effects that occurred and the underlying processes that drove those effects?
3. How can we best learn from an evaluation about what to do next using the proposed POInT approach?

We did this by adapting a method that has been proposed in political science. The method had been validated against proof-of-concept examples but had not been used in the evaluation of real development interventions. We explored this in partnership with teams evaluating development interventions who wished to learn more about their theory of change, the mechanisms active in their context, and the transportability of their results to other settings. We were aware that this process of applying a proof-of-concept approach in real-world evaluation would pose significant conceptual, practical, methodological and computational challenges. We believe this work provides a strong foundation on which other researchers, evaluators and programme designers can build.

Contribution to the literature

Evaluators usually estimate causal effects of programmes by comparing outcomes among programme participants with an estimate of what would have been observed had they not taken part in the programme (e.g. using a control group) (Gertler et al., 2010). As part of process evaluations, evaluators often also collect information on the theory of change (De Silva et al., 2014) that explains how the programme was able (or unable) to change outcomes, and about important contextual factors (Moore et al., 2015). Process data could include monitoring indicators (e.g. quantitative) and the subjective experiences of participants (e.g. qualitative). Process data may be collected from all participants, for example as part of an endline survey, or in only a subset, for example for in-depth qualitative interviews.

Current best practice in the evaluation of public health and development interventions requires the use of a process evaluation as well as an impact evaluation. However, formal analysis processes keep analysis of effects on outcomes (i.e. the impact) and investigation of processes separate from one another (C. H. Davey et al., 2017). We lack tools for bringing together these two sources of information in a formal, transparent, replicable way. Yet if we want to draw valid conclusions about an intervention's effectiveness, and to understand the implication of the results for policy elsewhere, we require methods that can integrate outcome- and process-related evidence(C. Davey et al., 2018). The need to combine these sources of information to make inferences not just about whether but about how the intervention worked is particularly pressing for more complex, or complicated, interventions.

In practice, evaluation teams take an informal approach to combining process and outcome data to produce inferences that apply across contexts; they draw conclusions that feel right based on an ad hoc application of the logic of causal inference as the evaluation team

understands it(Polit & Beck, 2010). These inferences become the basis of policy recommendations but are not based on a reproducible analysis or optimal use of the data. They are also susceptible to bias. We need to be able to integrate data in ways that are transparent, reproducible, and that maximise use of available data (Miguel et al., 2014).

The goal of the POInT project was to respond to this demand for a formal method for integrating different data sources using programme theory. POInT stands for: Process-Outcome Integration with Theory. We aimed to develop and implement a flexible method for combining data from the impact evaluation and data from the process evaluation to make better inferences about the theory of change and the transportability of the results as predictions for other contexts.

Theories of change are conceived of and represented in a bewildering variety of ways (Davies, 2018). This renders them unsuitable for grounding transparent, replicable integration of theory with data. Our approach begins by working with existing evaluation teams to translate their theory of change into a directed acyclic graph (or DAG). We can then use the mathematics of causal inference developed to describe and analyse relationships in the DAG, updating our understanding on the basis of expert knowledge and diverse data sources(Greenland et al., 1999). We do this by adapting an existing method that builds on formal causal inference methods and a Bayesian framework to accommodate various types of data (e.g. within-case and between-case) on various aspects of the underlying theory (Humphreys and Jacobs, 2015).

The approach combines thinking on “Bayesian process tracing” in the political science literature (Bennett & Checkel, 2014) and work in computer science that formalises causal models as Bayesian networks (Spirtes et al., 2000) with insights from work in statistics that seeks to estimate complex causal relations by partitioning populations into a finite set of groups that respond in similar ways to a finite variety of conditions (Frangakis & Rubin, 2002). In principle these approaches make it possible to take any causal model, represent it with a finite set of parameters, provide priors, update the model with any mixture of treatment, output, and process data, and then pose arbitrary causal questions to the updated model. While in theory the benefits are great, in practice there are real questions about whether the theories of change in the minds of practitioners can in fact be represented as causal models, whether the beliefs of practitioners can be accessed, and whether—given finite computation constraints—the resulting models can in practice be updated and queried. These are questions that POInT sought to answer.

Policy relevance

From its inception, the POInT project was about bridging the gap between promising methodological innovations and the practical realities of evaluating real-world programmes. We anticipated that we could help advance the role of middle-range theory in the evaluation and learning from interventions by adapting existing methods from other fields. To do so, we had to develop our own innovations, but the focus was always on the relevance to policy through better evaluation.

Although this project intended to develop a coherent process, we can highlight three areas of methodological development that will have particular resonance with the practice of evaluation:

1. **Theories of change as DAGs.** DAGs are powerful ways to present underlying theory, as well as becoming the standard diagrammatic language in the causal inference literature. This report provides intervention designers and evaluators with the tools to build their theory as a DAG. We found that it is possible to do, and reveals areas where data collection could be more parsimonious and where data should certainly be collected. We observed that the rigour of developing the theory using the DAG language encouraged clarity and explicit communication of assumptions among the partner teams.
2. **Revealing expectations in advance of the evaluation**. We developed and implemented a method to elicit priors – peoples' beliefs about the theory before seeing the data – that could be compared between experts and used to estimate the expected effect of the intervention. This revealed that the experts held very different beliefs about the theory, in particular about the presence (or absence) of important contextual factors. The presence of contextual factors could be explored in advance of starting the project and the data used to inform the expected impacts, reducing the risk of implementing an intervention that is not believed to have sufficient impact. Crowd-sourced and transparent estimates of the expected impact of the intervention could inform sample size calculations, reducing the risk of underpowering the study and showing the sensitivity to the perspectives held by different experts.
3. **Asking the right questions**. The method we adapted for evaluations uses the power of explicit causal modelling to allow for precise and interesting questions. We have shown how to ask "for those who had the outcome, in what proportion of cases was this due to the intervention?" Mediation and moderation questions are simple, within a single framework. More complex mediation questions would be possible, looking at how mediated effects are moderated by features of context – reminiscent of the 'context-mechanism-outcome' configuration in the Realist evaluation tradition. This is a significant advance for the evaluation of complex interventions for which the more difficult questions are often the most interesting and important for translating the evidence into policy in other settings.

Innovation and relevance to CEDIL

The POInT project is the first time that a formal structured causal modelling approach has been attempted for models that capture a real intervention's theory of change. Previous examples in the literature are all very small models with very few nodes that are best characterised as 'proof-of-concept'. We have worked with real trial teams to implement this approach for the full ToC for one arm of the UPAVAN trial, and for a model that captures a specific subquestion of interest to the DIG team.

These first steps have been challenging but have resulted in an enormous amount of methods and tool development. As discussed above, to encode, update and query our partner teams' models we had to extend and improve CausalQueries to make it suitable for moving beyond small examples to handle the larger models implied by real theories of change. This work is covered in more detail in Annex A, Section A2.

To elicit expert beliefs from the partner teams we had to develop a novel prior elicitation strategy that allowed us to elicit an order of magnitude more parameters than was previously possible. This is documented in detail in the Methodology section of this report. The prior elicitation worksheets that we used with our partner teams are available online for others to inspect and adapt. These links are in Part 3 of the Results section of this report.

To update a formal Bayesian causal model in CausalQueries on the expert beliefs elicited, we had to develop an approach to interpreting prior beliefs as the data which, if observed, would justify those beliefs, and then update the model on that data. This is documented in detail in the Methodology section of this report and all our code will be freely available from LSHTM's code and data repository.

We hope that other researchers can build on this method and tool development to further test the approach, further assess the costs and benefits, and find the applications to which it is best suited.

2. Methodology

The first step in the POInT method is to form partnerships with teams evaluating a social intervention. Once the partnership has been established, for each application to a partner trial there are six major steps in the POInT approach:

1. Adapt the partner teams' informal theory of change into a formal causal model in the form of a directed acyclic graph (DAG) (or multiple DAGs if consensus cannot be reached).
2. Specify the DAG in the statistical software R using a freely downloadable package called CausalQueries.
3. Elicit the partner team's prior beliefs about the nature and strength of causal links in the DAG.
4. Combine these expert priors into a single set of shared prior beliefs.
5. Update the model on those prior beliefs, update the model on data alone or data and priors to generate posterior models.
6. Query the prior and posterior models to learn about estimates of interest, and the relative role of prior beliefs and data in the combined analysis.

In this section we present the theory that underlies this approach and then the empirical methods used to operationalise it.

Theory

Part 1: DAGs

The POInT method begins by working with existing evaluation teams to translate their theory of change into a directed acyclic graph (or DAG). In this section we explain what a DAG is and then why we need to use DAGs for the POInT approach.

What is a DAG?

A DAG or Directed Acyclic Graph is more or less what the name suggests. It is a sort of graph that contains nodes (or vertices, boxes) connected by directed edges (arrows with one or two

heads) and which is not cyclic (it contains no loops). For example, consider Figure 1, below:

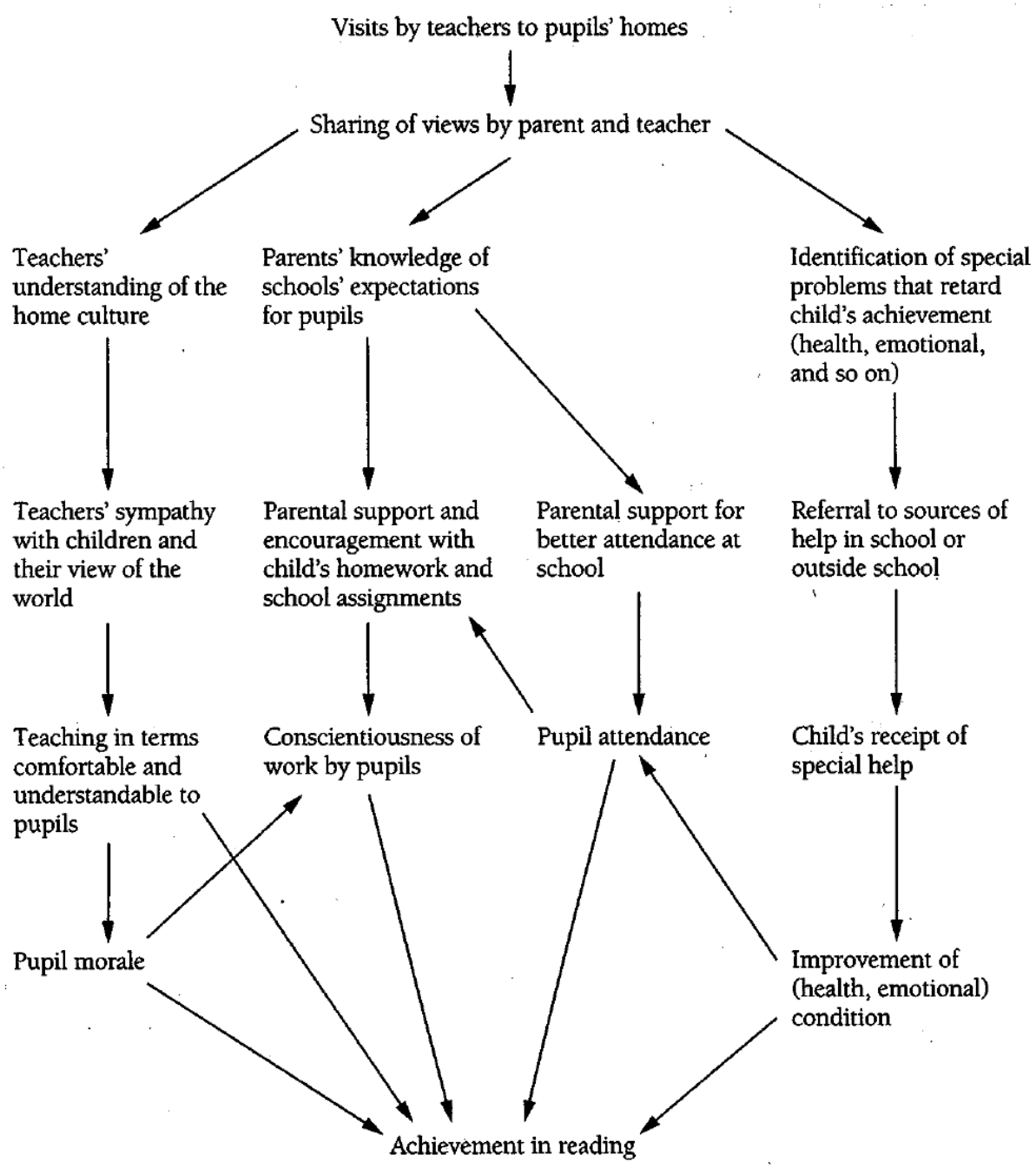


Figure 1: Theory of a program of teacher home visits in (Weiss, 1972, p. 50)

Figure 1 is reproduced from Weiss' first discussion of theory-based evaluation, in 1972. Despite being produced before the term 'directed acyclic graph' came into regular use in the late 1980s, this representation of intervention theory obeys the simple and intuitive rules of DAGs.[4] Nodes in the DAG are connected by directed edges, and there are no loops.

[4] Querying Google Ngrams' corpus of language from scanned books for 'directed acyclic graph' returns a handful of hits in 1880 and in 1979, but does not return consistent usage for several years in a row until a period starting in 1989.

If we interpret Figure 1 as a DAG, then it precisely and succinctly represents a large number of causal claims which add up to a complicated theory of intervention causation.[5] For example, the DAG has one edge between 'Visits by teachers…' to 'Sharing of views…' This encodes the claim that visits by teachers to pupils' homes are causally connected. The fact that the arrow has one head pointing to the second node encodes the claim that changes in the construct underlying the first node cause changes in the second, to some extent, and that changes in the second do not cause changes in the first. As in ToCs more generally, what is excluded from a DAG is just as important as what is included. For example, consider that in Figure 1, no other edges terminate at 'Sharing of views…,' only the edge from 'Visits by teachers….' This encodes the fact that the model excludes other causes of 'Sharing of views….'

Figure 1 has three edges from 'Sharing of views…' to three other nodes: 'Teachers' understanding of the home culture,' 'Parents' knowledge of schools' expectations for pupils,' and 'Identification of special problems that retard child's achievement.' This encodes the claim that sharing of views by parent and teacher is expected to cause achievement in reading through three main channels. The claim that all and only these three main channels are important is a claim that can be evaluated and which might explain relative success or failure of a particular implementation of the intervention. Looking at the first of these three channels, we can see that 'Teaching in terms comfortable and understandable to pupils' is a node through which all of the effect of that first channel is expected to run. There is no other path in the DAG from 'Teachers' understanding of the home culture' to 'Achievement in reading' except through that node. This is a strong claim that can be evaluated and which might explain a failure of a particular implementation of the intervention or for a particular subgroup (i.e. schools that forced teachers to teach using particular terms, perhaps through adherence to a strict curriculum and specified teaching materials should be expected not to benefit from the first channel of intervention causation.)

One convention of DAG notation is not present in Figure 1: the representation of unobserved confounding[6]. Curved, bi-headed edges (arrows) between nodes are conventionally used to represent confounding of those nodes by an unobserved third node, as in Figure 2, below:

[5] As (Dawid, 2010) reminds us, it is important to distinguish between probabilistic DAGs that encode (only) conditional independence relationships and might have many alternative compatible sets of edges for the same nodes, and causal DAGs, which make much stronger claims and generally have only one compatible set of edges for the same nodes. These are sometimes called 'structural causal models' to distinguish them from probabilistic DAGs. However, we follow the alternative, more widespread convention of calling them causal DAGs, and usually just DAGs.

[6] Confounding is the distortion of the association between the independent and dependent variables because a third variable is independently associated with both.

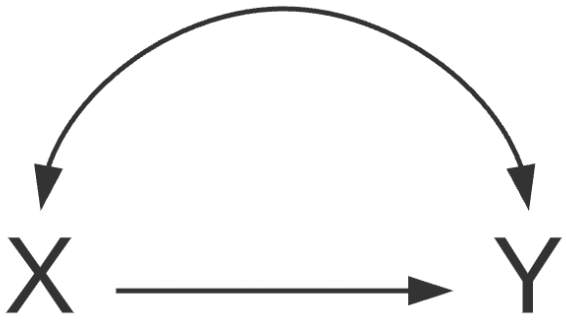


Figure 2: Confounding

Simple presence and particularly absence of nodes and edges can encode complicated and powerful causal claims. The simplicity of DAG notation is the root of its great strength, but also generates limitations and challenges that must be managed. Most importantly, leaving an important node or an important arrow off the DAG can bias estimates of effects.

Why do we want to use DAGs?

In the literature, many claims are made for the power of DAGs. For example, it is often said that working with DAGs can help to identify instruments, confounding and sources of bias, demonstrate inappropriate statistical adjustments for presumed biases, and understand threats to validity for particular study designs (Angrist & Krueger, 2001; Brito & Pearl, 2012; Rohrer, 2018; Tennant et al., 2019; Williams et al., 2018). It is even often said that DAGs can be used to determine cause-and-effect relationships from observational data without needing to set up randomised controlled experiments (for example (Arif & MacNeil, 2022) citing Pearl (2009)). It is outside the scope of this project to assess such claims in detail, though below we demonstrate that we have found some support for the claim that developing a DAG can be extremely useful to evaluation teams, even if they do not intend to use a full structural causal modelling approach such as the POInT approach. We will also say that the boldest claims about the power of DAGs do not reflect the extremely restrictive set of assumptions which must be met for them to be true and whose satisfaction introduces a large amount of model uncertainty to such causal estimates. There are two theoretical reasons that we use DAGs in our approach which we discuss below.

As Davies (2018) notes, graphical ToCs, even just those specific to development, are represented in a bewildering variety of ways. Figure 3 is reproduced from Davies (ibid, p.9) and shows a Google image search result for 'theories of change' which illustrates the many and varied forms that ToCs can take.

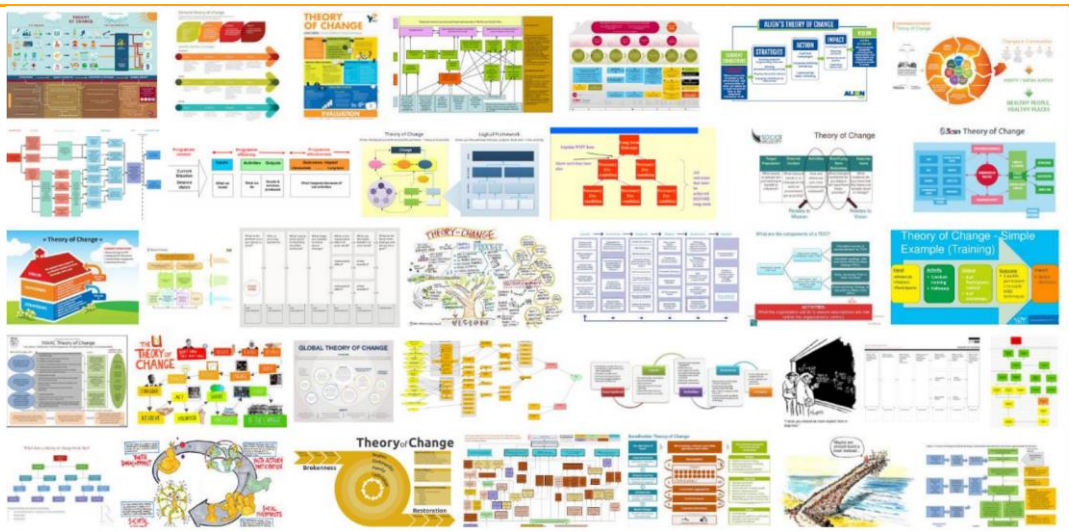


Figure 3: Varieties of ToCs in Davies (2018, p.3)

As Figure 3 demonstrates, understanding the visual language of one ToC often does not help much with understanding the next. By contrast, DAG notation is fairly standard. Because it is so simple, changes to the ways in which nodes and edges are represented do not much alter the fundamental presentation and do not undermine legibility as presented in Figure 4.

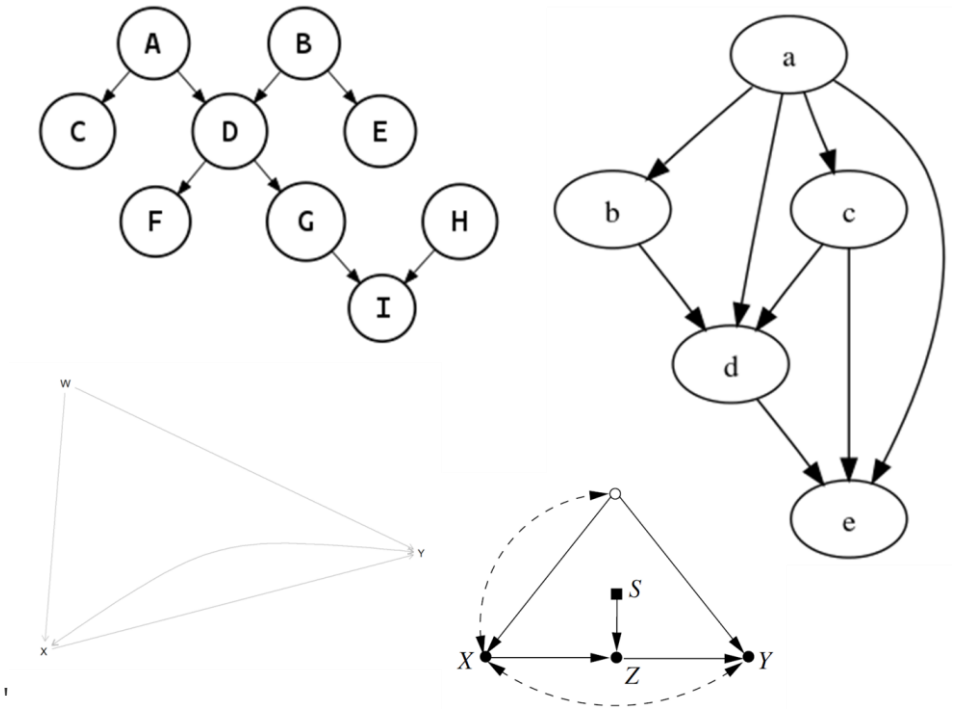


Figure 4: Varieties of DAGs

The simplicity of DAG notation does not only help legibility. One major advantage of using a DAG is that it is necessarily unambiguous. DAGs are unambiguous in the relationships that they communicate because the rules of DAG notation do not permit any ambiguous relationships. This lack of ambiguity means that all relationships in a DAG are evaluable given data on nodes in the graph, whereas an ambiguous presentation in a traditional ToC might not be evaluable regardless of the amount of data available. In order to use a partner team's

theory of change to guide the integration of process and outcomes data, that ToC needs to be represented unambiguously. Changing a ToC into a DAG forces this disambiguation and this is the first reason that we need to use DAGs for the POInT approach.

The further reason we need to use DAGs for our approach is that it allows us to encode the unambiguous model elicited from the partner team as a formal causal model in Bayesian modelling software such as CausalQueries. We can then update that model on expert beliefs (priors) and on process and impact data from the partner trial to create a combined model that incorporates the theory of change, the partner team's beliefs and data from different sources. This model can then be interrogated to generate estimates of causal effects that use all of this information. How this full process works is described in much more detail below, but encoding a partner team's theory of change as a DAG is the necessary first step.

Part 2: Priors

The process of converting the partner team's theory of change into a DAG goes some way to capturing their theory of intervention causation. However, this job is not yet complete. It is also important to the POInT approach to elicit more detailed beliefs from the partner team about the nature and strength of causal relationships in the DAG. In this section we explain why this is so and explain why doing it in practice was very challenging and required a novel approach.

Why elicit prior beliefs?

In the Bayesian approach to inference, data are used to update beliefs about the values of parameters in the model. For beliefs to be updated, there must be a 'prior' belief that is combined with observed data to produce a new belief about the parameter value and the associated uncertainty (called the 'posterior'). 'Flat' or 'uninformative' priors are assumed when there is little or no reason to think the parameters have any particular value in advance of seeing the data. In practice, uninformative priors are often used even when there is previous data on the likely effects and strong expert beliefs about what values parameters should take. This represents a loss of information if inferences drawn from the trial data are based only on a naive reading of the data and not on the existing theory which it may support or contradict. More usually, a model updated on flat priors will then be interpreted in the context of existing beliefs and theory in an informal way which does not make clear the relative role of theory and data in the combined inference. It is preferable for precision, transparency, and replicability to include these existing beliefs as prior beliefs formally encoded in the model.

The place of prior beliefs in evaluation research is contested. The concept of 'equipoise' is central to the ethics of conducting a randomised controlled trial; that is, there is uncertainty about the likely effect of the intervention to the point of equivalence between the treatment and the control. This amounts to uninformative priors or informative prior beliefs that there should be no effect. However, in practice and especially for development (as opposed to medical) interventions it is difficult to defend a trial and to win funding for it unless there is some reason to think that the intervention will work. In international development and public

health, programmes are designed, proposed, funded, and delivered with a degree of belief that they should have an effect on the outcomes of interest. In the language of priors, this implies informative prior beliefs about the average treatment effect.

For a complex intervention specified as a DAG, it is also possible – and perhaps less controversial – for experts to hold prior beliefs about the steps and dependencies in the DAG, rather than about the effect of the whole intervention. Specifying priors about each component of the theory of change could imply an overall prior about the average treatment effect. We believe that such prior beliefs are valuable information that should be used in causal inference and that the use made of them should be explicit, formal, transparent and replicable where this is possible at a reasonable cost. We therefore sought to elicit priors from the members of the partner teams to include in our analysis.

The most consequential priors that we collected from experts are reflected in the DAG, in the choice to include or exclude nodes and edges. Next-most consequential are the decisions to accept or reject assumptions of monotonicity of causes and non-interaction between causes. These priors were elicited from partner experts using a DAG-elicitation process described in detail in the Empirical Methods section below.

In addition, we wanted to elicit priors from experts about the strength and direction of causal effects in the DAG, including the extent of any interactions[7] between multiple parents causing the same child. These beliefs are also consequential. A belief that a particular channel of causation will only have very weak effects and another will have very strong effects could entirely change recommendations about the contexts in which the intervention can be expected to have strong effects or none.

How can we elicit the required prior beliefs?

Eliciting detailed prior beliefs about the nature and strength of causal relations in the DAG is challenging if there are several multi-parent nodes in the DAG because it requires collecting a very large number of estimates conditioned on the different combinations of values of parent nodes that are possible. To explain why this is so, we must first explain the concept of a nodal type. Consider the following DAG in Figure 5:

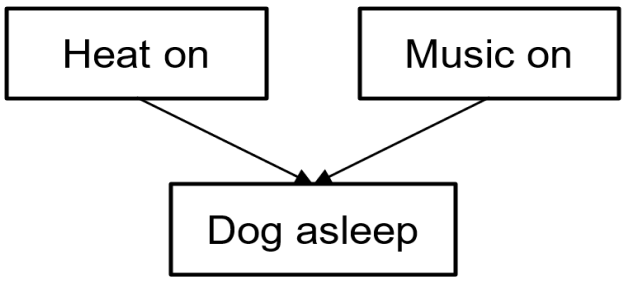


Figure 5: A simple DAG

[7] Interaction may arise when considering the relationship among 3 or more variables, and it describes a situation in which the effect of one causal variable on an outcome variable depends on the state of a second causal variable.

Assume all three nodes in this DAG are binary. The node 'Dog asleep' has a set of nodal types that describe all of the different ways that it could respond causally to the four possible combinations of values of its parents. These nodal types are expressed in CausalQueries as collections of four binary digits such as 0000 or 0101. Each of the four digits encodes how the node in question responds to a given combination of parent node values.

So, for 'Dog asleep', the first digit encodes the value that the node would take if 'Heat on' = 0 and 'Music on' = 0. A four digit nodal type that began with '0' would mean that when the heating is off and the music is off, the dog is not asleep ('Dog asleep' = 0.) The second digit of the nodal type encodes the value that the node would take if the first parent equals 1 but the second equals 0. So, a nodal type with a 1 in the second position would mean that when the heating is on and the music is off, the dog is asleep ('Dog asleep' = 1.)

The nodal type 0000 encodes the situation where the dog is not asleep regardless of the values of the parent nodes. The nodal type 1000 encodes the situation where the dog is asleep only when the heating is off and the music is off. Otherwise, it is not asleep. The nodal type 0101 encodes the situation where the dog is asleep when the heating is on and the music is off, or when the heating and music are both on, but otherwise is not asleep. Table 1, summarises the meaning of each digit in the nodal types for the 'Dog asleep' node.

Table 1: Digit interpretation for nodal types for 'Dog asleep'

Nodal type digit	Natural language definition
Y [*]***	Heating off and music off
Y *[*]**	Heating on and music off
Y **[*]*	Heating off and music on
Y ***[*]	Heating on and music on

Table 2, below, summarises the sixteen nodal types for 'Dog asleep' if it is assumed that 'Heat on' and 'Music on' are interacting causes. There are only six nodal types if these causes are assumed to be independent.

Table 2: Natural language translations of all nodal types of 'Dog asleep'

Nodal type	Natural language definition
0000	Never asleep, regardless of heating or music.[8]
1000	Asleep only when both off.
0100	Asleep only when heating on and music off.
1100	Asleep when neither on, or when heating on and music off.
0010	Asleep only when heating off and music on.
1010	Asleep when neither on, or when heating off and music on.
0110	Asleep when either on, but not both or neither.
1110	Asleep when either or neither on, but not both.
0001	Asleep only when both on.
1001	Asleep when both or neither on, but not if only one on and other off.
0101	Asleep when heating on and music off, or both on.
1101	Asleep except when heating off and music on.
0011	Asleep when heating off and music on, or both on.
1011	Asleep except when heating on and music off.
0111	Asleep except when both off.
1111	Always asleep, regardless of heating or music.

Even in this very simple DAG there are sixteen nodal types for 'Dog asleep.' There are many more nodal types for nodes with more parents, as the total number of nodal types for a

[8] This general type ('0000' in a two binary parent one binary child case but '00' in a one binary parent case) is sometimes referred to in the process tracing literature as 'chronic' or 'cursed'. The '11' or '1111' type is conversely sometimes referred to as 'destined' or 'blessed.' In the simplest one binary parent, one binary child case the other two possible outcomes ('10' and '01') are referred to as 'adverse' (those who would get better if and only if they do not receive the treat) and 'beneficial' (those who would get better if and only if they do receive the treatment) respectively.

binary node with k binary parents is 2^2^k. A binary node with four binary parents and no restrictions on interactions or monotonicity therefore has 65,536 nodal types.

Eliciting prior beliefs about the relative share of each of these nodal types for units to whom the model applies is extremely challenging because of this large number of parameters. We were able to overcome this challenge through constraining the size of the models with which we worked and through the development of a novel prior elicitation strategy as discussed in the Empirical Methods section below.

Part 3: Data

As part of the DAG-elicitation process, we gathered definitions of the nodes in the DAG from participants. We were then able to work with the partner teams to identify measures in the process and impact evaluation data collected for each intervention for each node in the DAG. These data could then be extracted and combined to create a dataset of node values for recipients of the intervention and members of the control group.

A limitation of the POInT approach is that CausalQueries was designed to work for fine-grained analysis of the causal relationships between binary nodes. We have conducted extensive work extending CausalQueries to handle ternary, quaternary and arbitrarily many-valued nodes. This work is now complete and is presented in Annex A, Section A1. Thanks to this work, this limitation is not strict for applications of the POInT method or other users of CausalQueries in the future. However, the number of nodal types in a model is an important conceptual and computational constraint on applying our method, as alluded to in the previous section and throughout this document. Applying the methods to handle non-binary nodes presented in Annex A implies large increases in the number of nodal types in a model where a non-binary node is the child of another node and even larger increases when a non-binary node is a parent of another node. This means that in the future, as it was for our application, it will be desirable to encourage partner teams to express their theory in terms of binary nodes where possible. This meant that we encouraged partner teams to think in terms of binary nodes where possible and tried to find solutions using only ternary or quaternary nodes where it was impossible to create a model containing only binary nodes. This limitation did not prove to be an impediment to expressing the partner teams' models for our application, as will be discussed in the Results section. The extent to which working with low-valued nodes is a limitation of the POInT approach is discussed further in the Limitations subsection of the Conclusions section of this report.

As well as avoiding nodes and measures that have too many possible values, it is also important to the POInT approach to avoid measures defined in ways that are unnecessarily context-specific. For example, it might be tempting to specify a node such as 'Woman in household has completed primary school.' This might undermine the transportability of the insights generated by the analysis process by masking the more fundamental and more widely-applicable construct that the partner expects to be causally significant. It might be the case, in this example, that what is important is that women are functionally literate. Defining the node as '…has completed primary school' would make the causal insight implicitly reliant

on the fact that completing primary school generally leads to functional literacy in this context (and not completing primary school implies illiteracy), though it may not in others. In order to best facilitate transferability, it is necessary to drill down the most fundamental concept available.

Once measures have been defined for all nodes in the DAG, and the available sources of data for these measures have been identified, the model can be updated on the data. The way in which this is done in the software CausalQueries is discussed in the Empirical Methods section. In using data to update a model, CausalQueries seeks to learn from the data about the (possibly joint) population-level distribution of nodal types (i.e., causal effects) across all nodes in the model. When confronted with data on some or all nodes in a model, for some set of cases, CausalQueries will put more weight on those nodal types that are more consistent with the data.

For instance, to return to the sleeping dog example, if we observe a case with the heat on, the music off, and dog asleep, this observation is consistent with nodal types 0100, 1100, 0110, 1110, 0101, 1101, 0111, and 1111, but inconsistent with the other 8 nodal types, and CausalQueries will now increase the weight placed on these first 8 types relative to the others. These first 8 include types in which the music and the heat do and do not have effects. However, additionally observing a case in which the heating off, the music on, and the dog awake would constitute evidence that the heat contributed to the dog's sleep, leading to more weight being placed on nodal types in which the heating has a positive effect (0100, 0110, 0101, and 0111) and less weight on the other types. In line with Bayesian logic, the posterior distribution that CausalQueries generates over the nodal types will reflect a compromise between our priors and the data. While priors will play a larger role and updating will be more modest with data in only a handful of cases, greater shifts in beliefs and greater reductions in posterior uncertainty are possible with larger sample sizes.

At the inception of this project we hoped to be able to make use of qualitative data collected through focus groups and interviews as well as quantitative data collected through surveys and management information systems. This proved too challenging a project to be incorporated into this research programme, but our initial steps in this direction, the reasons why it was challenging, and the future possibilities for incorporating qualitative information into the POInT approach are detailed in the Conclusions section of this report.

Part 4: Queries

Having created a DAG and updated it on the data to create a posterior model, we can query the model to produce estimates of effects before and after exposure to data along with credibility intervals for those estimates. This allows us to describe how the data increase or decrease confidence in aspects of a model and the model as a whole. The way in which this is done in the software CausalQueries is discussed in the Empirical Methods section.

To query a model means to formulate a descriptive or causal question about the model and extract an answer to that question from the beliefs about nodal types that are embedded in

the model. For instance, in the sleeping-dog example, we might want to know what the probability is that a randomly selected case (a randomly drawn living room with dog, heater, and stereo system) is one in which music has a positive effect on the dog's sleep when the heating is on. The first step in posing this query is to identify the one or more nodal types in this model in which this specific effect holds. There are four nodal types that satisfy this query: 0001, 0011, 1001, and 1011. Recall now that, after updating, we have a posterior distribution over the shares of nodal types in the population. So the second step is to extract from this posterior distribution over all nodal types a posterior over the proportion of cases in the population that are of one of the nodal types that satisfies the query. This proportion is the answer to our query since the probability that a randomly selected case satisfies the query is identical to the proportion of cases in the population that satisfy it.

In sum, for any query, we can define the query in terms of a collection of nodal types that satisfy it and then "ask" the model about the relative share of those types relative to all others. We can also condition queries on observations of nodes in a case: for instance, if we see that the music is on and the dog is asleep, we would begin by eliminating all nodal types inconsistent with this observation, and then our query becomes the share of the population with nodal types that satisfy the query and are consistent with the observed divided by the share with nodal types consistent with the case observations.

Queries can be case-level (as in this example) or population-level (e.g., about the average effect of one node on another), and can be about a range of descriptive (how often are dogs asleep?) and causal quantities, including questions about effects and questions about the pathways through which effects occur. For all of these we are simply asking the model about how common the nodal types, or collections of nodal types, are that satisfy the query. Note that we can also, in the very same way, query a model using our *prior* beliefs. By comparing the answer to our query using our priors and the answer using our posteriors (updated on the data), we can assess how much the data have changed our beliefs about the query and how much the data have reduced our uncertainty about the query.

Empirical methods

Part 1: Developing partnerships and selecting experts

Our development of the POInT method through this programme of research was designed as a series of partnerships with partner trials. These trials were identified opportunistically through our networks. The partner trials engaged are summarised in Table 3 and described in more detail below.

Table 3: A summary of partner trials engaged for the POInT method

Project	Intervention summary	Location	Team lead
UPAVAN (pilot)	Videos and group discussions to improve maternal and child nutrition	Rural Odisha, India	LSHTM
Disability-Inclusive Graduation (DIG)	Modification of an ultra-poor graduation intervention to be more inclusive of recipients with disabilities	Kiryandongo District, Gulu District and Oyam, Uganda	BRAC, LSHTM
Adapted Microplanning to Eliminate Transmission of HIV in Sex Transactions (AMETHIST)	Risk-differentiated intervention for FSW, centred around the implementation of microplanning and self-help groups	Zimbabwe (National)	LSHTM

In developing and testing the POInT approach we partnered with three partner evaluations. The first of these partnerships was with LSHTM colleagues working on the UPAVAN evaluation and was a pilot project used to develop the method summarised in this report. The UPAVAN analysis was an exclusively retrospective reanalysis of existing trial data. All of our proposed partnerships involve no data collection on our part other than data on our colleagues' recollections of their initial understandings of intervention causation (their 'priors' about causation in the intervention). We had initially hoped to augment this approach with data collection on the priors' of recipients of the intervention, but COVID-related delays and the impossibility of securing a no-cost extension to our project meant that we were limited to exclusively secondary re-analysis.

For all partner evaluations, the first stage in building the partnership is to contact the lead evaluators and build a relationship with them, establishing what the POInT approach might be able to contribute to their evaluation and what their involvement with us would involve. Next,

working with the lead evaluators we build a group of intervention designers and implementers who we can work with to build the DAG.

UPAVAN

The UPAVAN trial was an evaluation of three nutrition-sensitive agriculture (NSA) interventions seeking to positively affect maternal and child nutrition. The trial was implemented as a parallel, four-arm, observer-blind, cluster-randomised trial in Keonjhar district, Odisha, India (Kadiyala et al., 2018). The interventions combined information videos, household visits from outreach workers, and the establishment of participatory-learning-and-action groups in different configurations. For the purposes of the partial reanalysis reported on here, we focused on the simplest of the three active arms (referred to as AGRI in our DAGs), which provided women with the opportunity to watch videos about agricultural techniques that were tailored to be appropriate to the local context as well as follow-up household visits from an outreach worker for support with implementing the agricultural innovations covered in the videos.

Clusters were defined as one or more villages with a combined population of at least 800 residents, and 37 were allocated to each arm. Data were collected from a stratified random sample of on average 32 households per arm at baseline and again at endline; the random sampling was independent in each round and therefore only a small minority of households would have participated in both rounds. The survey included data on a number of domains, including diet, crop production, and about the physical (e.g. water supply) and social (e.g. family relations) environment. The trial had four main outcome measures, one of which was maternal dietary diversity. To test our method against the minimally complex case, we focused on one outcome. We chose that outcome in consultation with the UPAVAN team, deciding to focus on maternal dietary diversity because this was the only main outcome on which the trial had detected an effect in the AGRI arm that meets the conventional threshold for statistical significance in frequentist analysis and was therefore the most interesting outcome to the UPAVAN team for which to further investigate the distribution, mediation and moderation of effects of the AGRI intervention(Kadiyala et al., 2021).

A process evaluation collated data on the delivery of the programmes' components and collected qualitative data on the experiences of project participants and those responsible for delivery(Prost et al., 2022).

DIG

The DIG trial was an evaluation of a disability-inclusive ultra-poor graduation intervention delivered by the international NGO BRAC to three districts of Uganda. The intervention was delivered in Kiryandongo District (Western Region), Gulu District (Northern Region), and Oyam. These districts are characterised by higher than national average levels of poverty, social deprivation, and prevalence of people with disabilities. The programme aim was to reduce extreme poverty among extremely poor families, and in particular those with a female head of household and/or with an adult living with a disability.

The 'standard' ultra-poor graduation intervention as implemented by BRAC is well-studied, with randomised controlled trials (RCTs) undertaken in Bangladesh, Ethiopia, Ghana, Honduras, India, Pakistan and Peru(Banerjee et al., 2015). The standard programme combines support to address immediate needs with longer-term investments in life skills and technical skills training, asset transfers, enterprise development, savings and planning for the future to transition ultra-poor families into sustainable livelihoods. By addressing the social, economic and health needs of families simultaneously, these programmes aim to provide holistic support to participants that enables them to exit extreme poverty and equips them with the level of resilience necessary not to re-enter.

The DIG programme is an adaptation of the standard graduation intervention with added intervention components intended to overcome barriers to people with disabilities benefitting. Adaptations of the programme for people with disabilities were developed in collaboration with the NGO Humanity & Inclusion (HI) in Uganda and the National Union of Women with Disabilities of Uganda (NUWODU). These adaptations included the provision of physiotherapy, psychosocial support, access to assistive technology and eight other disability-inclusive intervention components.

Individuals within households were the unit of participation with a maximum of one project participant per household. Clusters were defined as groups of villages such that each cluster contains at least 10 participating individuals. 233 clusters were assigned to two arms: the DIG intervention or a 'business as usual' control arm. All participants were interviewed at the start of the trial. Shortly after the intervention delivery was complete, the participants who were identified at the baseline to have a disability were interviewed in a first of two endline surveys; the second will be conducted one year after the programme ended.

AMETHIST

The AMETHIST intervention (Adapted Microplanning to Eliminate Transmission of HIV in Sex Transactions) was a risk-differentiated intervention for female sex workers (FSW), centred around the implementation of microplanning and self-help groups. It was designed to support uptake of, and adherence to, HIV prevention, testing and treatment behaviours among FSW. Twenty-two towns in Zimbabwe were randomised to receive either the Sisters programme (usual care) or the Sisters programme plus AMETHIST.

The composite primary outcome was defined as the proportion of all FSW who are at risk of either HIV acquisition (HIV-negative and not fully protected by prevention interventions) or of HIV transmission (HIV-positive, not virally suppressed and not practising consistent condom use). The outcome was assessed after 2 years of intervention delivery in a respondent-driven sampling survey (total n = 4400; n = 200 FSW recruited at each site). Primary analysis used the 'RDS-II' method to estimate cluster summaries and will adapt Hayes and Moulton's '2-step' method to produce adjusted effect estimates. An in-depth process evaluation was also undertaken.

Selecting experts

Once these partnerships had progressed to the stage where senior members of the evaluation and implementation teams had agreed to partner with the POInT team, the next stage in the partnership was identifying the group of experts who would participate in the DAG development and prior elicitation processes. This was done by consulting with the PI and other senior members of the trial evaluation and implementation teams, to identify any experts in the theory of change of the intervention who were available to participate. The first DAG elicitation workshop served to cement links with this group of people and to provide an opportunity for the group of experts to be refined by participants choosing to drop out or to recruit other people they thought had relevant information to join the process at the next workshop.

Part 2: DAGs

Once a partnership had been established, the next step in applying the POInT approach to a given partner trial was to elicit from the partner team their model of intervention causation, expressed as a DAG.

Aggregate existing information to create a candidate DAG

Ideally, the DAG elicited from the partner team would reflect only the partner team's beliefs and would be minimally biassed by the involvement of the analysis team. However, the model also had to be elicited quickly so as to not exhaust the patience and enthusiasm of the partner team. Optimising over these two conflicting considerations, we decided to develop a candidate DAG for each partner trial based on our analysis of documents such as any trial protocol and papers, with particular attention to theory of change diagrams and supporting text. These candidate DAGs were able to serve as a teaching aid and as a starting point for the elicitation process that followed.

Workshop the candidate DAG with expert participants and iterate towards the approved DAG

For each partner trial, having developed a candidate DAG we organised a workshop with the partner team and employed the following process to elicit a consensus DAG or DAGs:

Box 1: DAG Elicitation Process

1) **Teach DAG concepts and notation**
2) **Introduce POInT-specific considerations**
3) **Introduce candidate DAG, reinforcing learning of 1. and 2.**
4) **Facilitated discussion with live editing of the working DAG**
5) Privately tidy up working DAG post-discussion, highlighting areas of uncertainty or disagreement
6) Email working DAG to participants for comments and corrections
7) Repeat 4.-6. until final DAG or DAGs are agreed

Steps in bold take place in a facilitated workshop with the partner team.

Some parts of this process require further specification, which is briefly given in the following paragraphs.

Box 1, point 2. mentions some POInT-specific considerations that had to be explained to participants alongside the rules of DAGs. The first of these was that POInT requires the DAG to be specified in terms of binary or ternary nodes, rather than allowing continuous variables as is standard practice in the evaluation of development and public health interventions. This is a limitation of the approach imposed by the CausalQueries software, discussed in more detail in the Limitations subsection of this report.

The second POInT-specific consideration that had to be communicated to the partner team was that nodes must be specified at the theoretical level rather than at the level of measurement. So, for example, 'women eat a sufficiently diverse diet' rather than 'women described eating at least one item from five or more of the food groups defined by the endline survey.' This is essential in order to integrate different data sources which may contain different empirical measurements of the same underlying theoretical construct.

The final POInT-specific requirement of the DAG explained to the partner team was that they would need to define some restrictions on the causal interaction between nodes where possible. Piloting suggested that this would be necessary to reduce the complexity of the DAG to a level within the computational limits that can easily be exceeded when analysing a large DAG with CausalQueries. Two restrictions were desirable where they were supported by the partner team's theory. Firstly, where possible, to restrict the causal effect of parent nodes on child nodes to be monotonic. That is, to assume that an increase in the parent causes an increase in the child. Secondly, where possible, to restrict the causal effects of different parent nodes on the same child node to be non-interacting.

As Box 1 describes, the first four stages in the elicitation process took place in a facilitated workshop with the entire partner team. A first workshop ran through stages 1.-3. before moving on to stage 4. Stages 5. and 6. were then executed by email. A second workshop was then required to begin a second iteration of stages 4.-6. Both workshops were facilitated by one member of the POInT team, with a second member 'holding the pen' for stage 4. This holding of the pen involved live editing a working DAG and adding annotations to reflect active areas of discussion. Facilitation focussed on two objectives: 1) progressing the discussion efficiently so as to cover as much ground as possible and 2) ensuring balance between contributions of different team members so that all perspectives within the partner team were explored as much as possible.

If necessary, reduce the approved DAG to create the analysis DAG

As discussed above, the models which underpin our analysis strategy explode in the number of parameters and exceed reasonable computational limits if any node has too many parents. Therefore, it may be necessary to simplify the approved DAG to make it tractable for analysis. In piloting we developed a three-stage approach to doing this:

1. **Remove nodes for which data cannot be generated of sufficient quantity and quality:** The approved DAG may contain nodes which the partner evaluation team had not planned to measure. Where possible, we work with the partner teams to generate data on all nodes in the DAG. However, this is not always possible. Nodes which cannot be measured can be removed from the analysis DAG, so long as the channels of causation that flow between included nodes via those excluded nodes are preserved. (See point 3.)
2. **Further prune least important nodes:** If the DAG remains too complex to handle in CausalQueries then we must further simplify the DAG in order to have a tractable analysis DAG. We can do this by pruning nodes participants believe to be least important. This reduces the complexity of our model, reducing its sensitivity to some contextual factors, mediating or moderating factors. However, in many cases this reduction may be a justifiable simplification in order to keep the model within computational limits.
3. **Propagate edges:** Despite having removed some nodes from the approved DAG, the channels of causation that flow via those nodes must be preserved in order for the analysis DAG to be consistent with the causal thinking in the approved DAG. So, all of the channels of causation that flowed from an earlier node to a later one must be preserved even if a node that they flowed via has been removed. Removing nodes may also create a situation in which a double-headed arrow must be added to the simpler DAG to represent confounding by a removed node.

Once complete, the analysis DAG can be created within the CausalQueries package for R, using the make_model command which takes as an argument a description of a DAG in the syntax of the R package dagitty.

Part 3: Priors

CausalQueries allows priors to be manually entered for models using the set_priors command. This command takes as arguments alpha values for the dirichlet distributions that describe the distribution of types, where each nodal type distribution has a corresponding alpha value. The relative sizes of the alpha values represent the expected causal effects by capturing the relative share of nodal types expected. The order of magnitude of all alpha values captures the certainty of the prior over those effects (Humphreys and Jacobs, 2020, chap.3). What this means for prior elicitation from experts is that in order to create a complete prior model using the set_priors command in CausalQueries, one must elicit from experts the relative shares of nodal types for every node in a DAG. This is a huge number of parameters for DAGs as complex as we expected would be necessary to capture theory about our partner trials, and asking so many questions is not possible.

Alternative prior elicitation strategies considered

Given that we could not ask participants about distributions of nodal types directly, we considered four alternative strategies for eliciting priors over causal effects:

1. A group-based, graphical approach using extended DAG notation
2. A survey asking about causal effects using summary questions such as these:

   > "Node LAND causes DIVY for at least some households, and so do nodes WATER, DESC and EDU. In addition, does the strength of the causal effect of LAND on DIVY depend on the value of other nodes? Pick any which are true:
   >
   > i. No, LAND independently effect DIVY.
   > ii. Yes, when WATER is true, the effect of LAND on DIVY is amplified.
   > iii. Yes, when WATER is true, the effect of LAND on DIVY is frustrated.
   > iv. Yes, when DESC is true, the effect of LAND on DIVY is amplified.
   > v. Yes, when DESC is true, the effect of LAND on DIVY is frustrated.
   > vi. Yes, when EDU is true, the effect of LAND on DIVY is amplified.
   > vii. Yes, when EDU is true, the effect of LAND on DIVY is frustrated."

3. Structured interviews asking about causal effects using question prompts such as these:

   > "Interactions or lack thereof: For [the node discussed], does the strength of the causal effect of [NODE] on DIVY depend on the value of other nodes? If so, how?"

4. Contingency tables of all possible parent node values with a column asking for point estimates of effects for the child node

Approach 1. did not work because we could not design a rich enough notation that we had any hope was intuitive for participants. However, approach 1. had a more serious problem that was shared with approaches 2. and 3. Responses to all of these exercises would have to be translated into beliefs about nodal types in order to be entered into causal queries. It was not at all clear to us how this could be accomplished. It seemed to us after some piloting that unless you ask about nodal types, one cannot reliably encode answers as nodal types.

The fourth approach was the only promising approach we identified because it would allow us to generate data upon which a model with flat priors could be updated, allowing CausalQueries to infer the nodal types that generated the data on which the model was being updated. This approach is explained in detail on the next page.

Using contingency tables to elicit priors

The intuition behind using contingency tables to elicit priors is that even when there are unmanageably many nodal types for a given child node, there is a much smaller and somewhat manageable number of possible combinations of binary parent node values. If we ask participants for an estimate of the probability of the child node taking the value 1 given each combination of parent nodes, those answers imply much richer information about the strength, direction and interactions of causal effects of the parent nodes. We can use these point estimates combined with information about the experts' confidence in their answers to infer the experts' beliefs about the distribution of true causal effects for each combination of parent node values. We can use those distributions to generate data on which we can update the Bayesian model in CausalQueries to create a prior model that can in turn be updated on data from the partner trial.

Although this approach reduces the number of parameters of interest massively compared to asking directly about nodal types, there are still very many parameters of interest. For example, in the UPAVAN DAG there are six exogenous nodes (nodes with no parents in the DAG) which can take the value 0 or 1 and for which we need to know the expected probability of taking 0 or 1. We need to know this to calibrate the prior model to the expected characteristics of the study context. In addition, there are nine endogenous nodes (nodes with parents in the DAG) which each have between two and five parents (mean 3.89). Seven of these nodes are binary, taking the value 0 or 1, whereas two are ternary, taking values 0, 1 or 2. There are between four and 32 possible combinations of parent node values for each child node. There are a total of 208 different combinations of parent node values for non-zero child node values (just 1 in the case of binary nodes, but 1 or 2 in the case of ternary nodes). To populate the UPAVAN prior model with expert beliefs, we needed to elicit distributions of the probability of a child node taking each possible non-zero value for each combination of parent node values, plus distributions for the six binary exogenous nodes. This was a total of 214 parameters for which we wanted to know distributions.

The high number of parameters that we needed to elicit meant that none of the methods for eliciting priors that have been described in the literature were suitable for our case. Standard approaches in the literature for the elicitation of priors focus on eliciting a limited number of parameters, generally a single parameter with a point estimate and associated distribution. Participants are often asked to engage in time-consuming tasks such as drawing distributions or distributing counters to approximate histograms(Errington et al., 1991; Freedman & Spiegelhalter, 1983; Hughes, 1991).

The prior elicitation process we developed was a series of contingency tables, one per child node in the DAG and another table for all exogenous nodes. For each row in these tables the participant was asked the proportion of units who would have an outcome under different conditions. In addition to these point estimates, we wished to capture information about how certain participants were about their answers and therefore some information about what they thought the distribution of true answers might be. We therefore additionally asked participants about their level of certainty in their answers for a contingency table. We used this level of certainty to select between possible frequency density functions centred around participants' point estimates to derive the distributions that we needed to elicit. Asking about certainty at the level of the contingency table and therefore of the child node rather than for each individual combination of possible parent values represented a small loss of accuracy for a large gain in speed, making the elicitation for a larger number of distributions feasible.

For each partner trial the prior elicitation exercises were undertaken by the same set of expert participants as took part in the DAG development workshops. The process is described in detail below using the training exercise that we developed for participants as an example.

Training exercise example

Using a training exercise to familiarise participants with the elicitation process is recommended best practice in the literature(Johnson et al., 2010). We used such a training exercise to familiarise participants with the form of the questions and with the way in which we would interpret each of the levels of certainty that we were asking participants to select between. All participants had already been through the DAG elicitation workshops which familiarised them with the concept of DAGs and with the DAG that they co-developed for the particular intervention which was the subject of the prior elicitation exercise. Figures 6 and 7 present the content of the first section of the training exercise.

Training example - PUNCTURE

In this example, imagine a group of people are taking part in a bicycle race. We're interested in whether individuals in the group are going to get punctures in their bicycle tires. Everyone will be riding about 100km using tyres inflated to an appropriate pressure. We think there are two main determinants of whether cyclists will get puntures: 1) whether they have cheap tyres; 2) whether they are are riding on gravel or on smooth roads. This thinking has been represented as a causal model, below. Boxes or 'nodes' are connected by arrows representing the direction of causal relationships. Nodes in the model are binary; they can take only the values true or false.

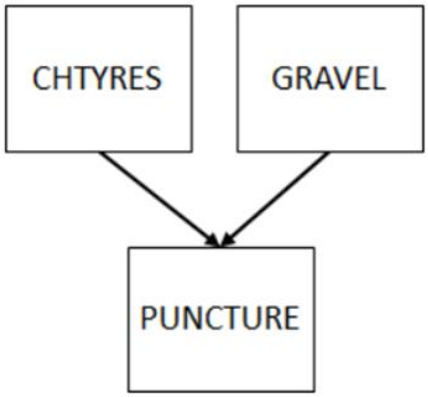


Figure 6: Simple DAG and explanation from training exercise

The simple example used for the training exercise was based on a three-node DAG made up of one binary child node with two binary parents. We used an everyday, non-technical and non-controversial example for which we were fairly sure that participants would not have divergent priors to which they were strongly attached about how causation worked in the example case.

The contingency table in the training exercise was presented alongside the simple DAG on a landscape format page layout designed to be easily read in a browser window maximised to take up all of a screen with a resolution of at least 1080 by 1920 pixels. For every contingency table in the later sections of the prior elicitation exercises this same formatting was used, with the child node or nodes mentioned in the contingency table being presented to the left of the contingency table on a version of the DAG with irrelevant nodes made translucent so as to highlight only the relevant nodes. Figure 7 shows the contingency table for the training exercise along with the table reminding participants of the meanings of each node and a paragraph of text introducing the question associated with each row in the contingency table.

Node meaning reminder

Node	Description of theoretical entity	Possible values
CHTYRES	Using cheap tyres, rather than quality tyres	TRUE, FALSE
GRAVEL	Gravel road surface rather than smooth tarmac road	TRUE, FALSE
PUNCTURE	Cyclist gets a puncture in at least one of their tyres at least once	TRUE, FALSE

Collecting your priors

For 100 cyclists with the following combinations of values for the CHTYRES and GRAVEL nodes, how many would you expect to get at least one puncture during the race (for how many would PUNCTURE = TRUE)? For each answer, respond with the value you think most likely to be correct. In addition, begin by telling us how certain you are about your answers for this node.

Certainty

CHTYRES	GRAVEL	Number PUNCTURE = TRUE
TRUE	TRUE	
TRUE	FALSE	
FALSE	TRUE	
FALSE	FALSE	

Figure 7: Contingency table, explanation and question used for eliciting priors over causal effects in the training exercise

The expected values of the child node were defined in the question as the expected proportions of 100 units who would have a positive outcome for that node under each configuration of the parents. We asked about proportions rather than probabilities in order to make questions more concrete and less reliant on mathematical fluency of the participant (Johnson et al., 2010).

In the second section of the training exercise, the participants were shown a series of beta distributions with different values of n, and corresponding natural language descriptions of the level of certainty that each distribution could be interpreted as representing. Figure 8 shows the beta distributions shared with expert participants as part of the prior elicitation exercise.

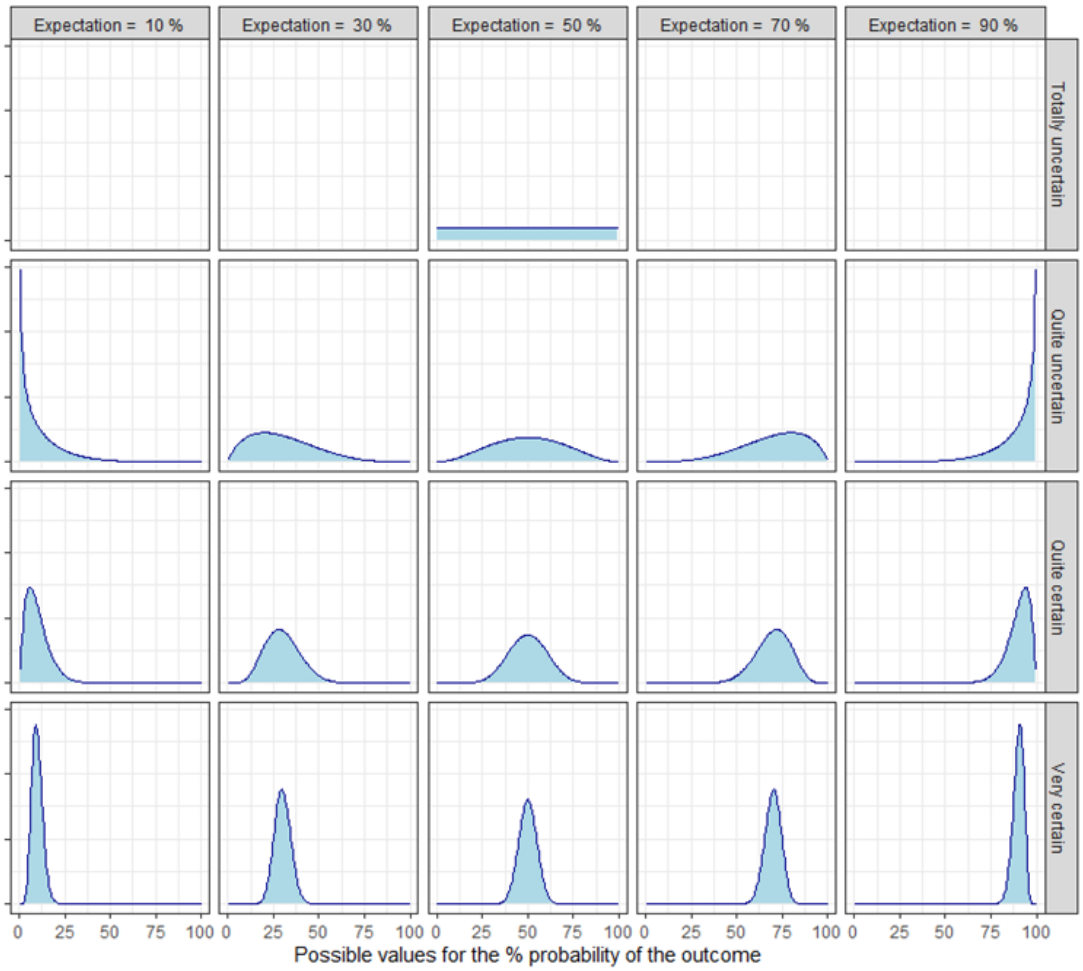


Figure 8: Beta distributions used to illustrate our interpretation of different levels of certainty

By using natural language descriptions of levels of certainty combined with the graphical representation of our interpretation of those natural language descriptions, we attempted to maximise the ability of participants to respond quickly to a large number of questions while communicating their beliefs somewhat accurately. We established that this method was acceptably user-friendly using piloting within our team and by soliciting the feedback of participants for each prior elicitation exercise. We also gave ourselves the option of giving participants feedback on their answers if we detected anomalies in those answers that suggested a misunderstanding of the question. Several individual participants for UPAVAN were given such feedback and one email was sent to all the DIG expert participants when it appeared that all were making the same error in understanding the first four tables of the exercise.

We conducted various statistical tests on the priors elicited to check them for the homogeneity of experts' beliefs and for the violation of restrictions on the model established during the DAG elicitation process. These are detailed below and their results reported in the Results section. The results of these tests and the responses to feedback make us reasonably confident that the prior elicitation tool has generated acceptably accurate results for the prior beliefs of experts in our samples. However, if this sort of prior elicitation tool were in regular

use, it would be wise to conduct research validating the answers generated by comparing them to other, more time-consuming tools that are more standard in the literature.

The prior elicitation exercise was conducted by creating a Google Sheets workbook for each partner trial containing three tabs. Copies of this workbook were created, one for each participant, and the participant's workbook was shared with them over email using a secure link unique to them. The first sheet (tab) in the workbook contained a brief introduction and instruction to cycle through sheets. The second sheet contained the training exercise and the third sheet contained the contingency tables to be filled out. Protected data ranges, data validation and conditional formatting were used to ensure that participants answered all and only the relevant questions and that their answers fell within acceptable ranges. Answers from all participants were aggregated and anonymised using a further google sheet that updated in real time based on the values of the participant sheets. These answers were then exported to be analysed in R. Read-only copies of the workbooks for UPAVAN and DIG are linked to in the relevant results section.

Expressing the elicited priors as data

It was not possible to enter information about the distributions of experts' priors directly into CausalQueries, creating a prior model through data entry. Rather, in order to create the prior model, we had to express experts' priors as a set of data which, if observed, would justify the experts' beliefs. Each expression of a prior belief in response to the prior elicitation exercise was transformed into a dataset of N observations where those observations would justify the prior belief expressed. N is set by the level of certainty in their prior belief that the expert has expressed, and the proportion of true values is set by the point estimate that they have expressed.

The beta distributions in Figure 8 are based on a set of observations of a binary variable whose mean corresponds to a point estimate consistent with the column heading (0.1, 0.3, 0.5, 0.7, 0.9) and whose N is set to a different number of data points for each level of certainty. The N for each row is 0, 4, 20 and 100 from top to bottom of the table. In order to create a prior model in CausalQueries we represented the answer to each row of each contingency table for each participant as a small dataset whose mean was as close as possible to the point estimate given by the participant given that the N was set by the level of certainty the participant had selected for the contingency table. All of the row datasets per contingency table could be combined to create a dataset for the contingency table. This resulted in several datasets of hypothetical observations: one per contingency table per expert.

Combining the priors

When combining expert priors from different individuals we face a modelling choice relating to the extent to which we model experts' insights as independent from each other or as overlapping. If we believe that experts' have formed their beliefs on the basis of different observations of different phenomena, then when combining experts' beliefs, we are combining independent data sources and the more beliefs we are able to aggregate, the more certain we should be about our aggregated findings. At the other extreme, if we believe that

experts have conducted many of the same observations as each other, observing many of the same phenomena, and/or have already undergone a process of aggregation by which they have adjusted their beliefs in response to information about the beliefs of the others, then we are not combining independent data sources. If in this case we treated experts' beliefs as independent, we would generate an aggregated model that overstated the level of certainty in the model that was warranted.

This concern is particularly pressing in the case of experts who are part of the same trial team, as is the case with the experts in our partner teams. These experts have gone through a process of aggregation of priors already in that they have discussed their beliefs about causation in the trial context at length in the process of designing the intervention implementation and evaluation. If we treat these experts' beliefs as independent, then we will overstate the level of confidence that is warranted in the aggregated prior model. However, treating these experts' priors as entirely overlapping, based on all of the same data is also incorrect, understating the level of confidence in the aggregated prior model that is warranted.

In order for our approach to be robust to disagreement with assumptions about the independence of partner experts' priors, we combined the priors in two different ways. These two methods each represent an extreme assumption about the level of independence of experts' priors in our samples.

1. **Exclusive priors:** The first assumed that the observations that the experts were referring to in their certainty about their priors were independent of one another. As such, the total N that was notionally observed for any child-parent component was the sum of the Ns reported by the experts that had completed the prior elicitation process for this component of the DAG.
2. **Inclusive priors:** The second assumed that the observations were all inclusive, such that people who based their beliefs on fewer observations than others had nonetheless observed the same cases but that another might have observed more in addition. Thus, the total N was the maximum N observed by any one expert. In both cases the overall prior value was the certainty-weighted average of the prior proportions.

Assessing the priors for assumption violations

Checking for homogeneity of experts priors

To assess the homogeneity of priors of different experts on the same child nodes, proportions of 1's for each person for each child node were plotted using a composite graph, with the level of certainty with which the expert answered reflected by the size of the dot on the graph.

Checking for consistency with the monotonicity assumption

We assessed the monotonicity assumption and interactions in the priors, i.e. the assumptions that were made in the DAGs. For this we used the first of the two methods for aggregating the

priors above. We used method one, 'exclusive priors' since this method preserves the maximum information in the elicited priors and therefore were most likely to show if either of the assumptions had been violated.

Monotonicity in the context of a causal relationship is that effects of the parent node can only be in one direction. That is, to assume that an increase in the parent causes an increase in the child. The investigation of monotonicity was only applied to child nodes who have parent nodes in the causal chain presented in the DAG. Below is a summary of steps that were conducted in checking for monotonicity among the child nodes.

Steps:

1. Fit a logistic regression model with interactions to estimate the effect of parent nodes on the priors for a certain child node.
2. Add the effect estimate of each main factor (parent node) to the effect estimate of each of the corresponding two-factor interactions.
3. Check whether the sign of the sum obtained in (2) above, is the same as the sign of the estimate for the main effect. If the signs are the same, then the monotonicity assumption is not violated.

Checking for interactions

In order to reduce the complexity of the model, where possible, restrictions of causal effects of different parent nodes on the same child node were assumed to be non-interacting. However, it was necessary to verify these assumptions before considering this to be incorporated in the final causal analysis.

We used the likelihood ratio test to assess the contribution of interaction effects to the logistic regression model on the priors. A large p-value (>0.05) implies that the two models (with/without interactions) are not significantly different in their fits for data. In other words, interactions are not significant. These results were compared to the partner team's priors on interactions as stated during the DAG elicitation process.

Update the naïve model to create the prior models

Once the priors had been elicited, aggregated, and checked for violations of assumptions, the naïve model created in CausalQueries could be updated on the priors data to create a prior model. There are different approaches that can be used to represent priors given the data gathered from this exercise.

One approach, which we implement here, seeks to assess what data a practitioner would have had to have been exposed to in order to justify the beliefs they hold. For instance, if a practitioner is very confident that a node takes the value 1 around 50% of the time, this belief is justified by past exposure to a large amount of data on the node with this distribution. Once beliefs are represented as data in this way, node by node, beliefs over other quantities that are justified by these data can be calculated and these data can be combined with project data to form posteriors. Note however that while this approach can be used node by node, we do not have access to a single (implied) data set that simultaneously justify beliefs at all

nodes. A consequence of this was that it was not possible to update the entire model at once using the approach. Instead, each sub-DAG was updated separately and stitched together. This utilised the update_stitch command; an approach to fitting large models that we developed to allow us to fit models too large to update all at once, but which also proved invaluable for fitting models to priors interpreted as data. As we had two sets of priors for each partner team, reflecting the two different prior aggregation methods, we updated each naive model twice to create two alternative prior models per partner trial.

An alternative approach is to specify priors over practitioner beliefs and use a Stan model to update over these beliefs given the answers provided by practitioners.[9] This approach estimates parameters that represent expert beliefs, which can then be used to characterise prior beliefs over queries or to define the causal model prior to updating with program data. We did not implement this approach for this project but may do so in future work in order to overcome certain limitations of the approach we did use. These are discussed in relation to UPAVAN in Part 5 of the Results section, below.

[9] If beliefs over the response types of a child can be represented by a set of Dirichlet parameters $\alpha = \alpha 1, \alpha 2, \ldots, \alpha k$, then certainty corresponds to the sum of α and the only remaining uncertainty over beliefs relates to the *relative* size of the elements of α. But these relative sizes are exactly what is estimated by the CausalQueries package given large datasets —that is, the share of units with a given response type. Thus CausalQueries can itself be used directly to estimate our posterior beliefs about practitioner prior beliefs, given practitioner reports.

Part 4: Data

Prepare trial data for analysis

We worked with data owners in the partner teams to gain access to partner datasets identified when defining measures for nodes. We then coarsened data values to binary or ternary following the definitions agreed in the Measures for Nodes documents developed with each partner team. After this process, data was integrated from different data sources and coarsened to binary with some missingness as illustrated in Table 4.

Table 4: Example of trial data

Individual ID	Cluster ID	Process measure 1 (FG)	Process measure 2 (CS)	Endline measure 1	Endline measure 2	...etc.
1	1	0		1	0	...etc.
2	1	1		1	1	...etc.
3	2	1	0	0	1	...etc.
4	3		1	0	1	...etc.
...etc.	...etc.	...etc.	...etc.	...etc.	...etc.	...etc.

Coarsen non-binary values

The Theory subsection of this Methods section discussed the fact that CausalQueries was designed to handle binary nodes. However, we worked hard during this project to extend CausalQueries functionality to handle non-binary nodes. This work is now complete and is presented in Section A1 of Annex A. However, this work was not completed in time to incorporate into the analysis of the partner trials reported in the main body of this paper. Therefore, we were forced to coarsen non-binary measures to binary in order to complete the analysis. This was done by analysing the non-binary data to identify which coarsening strategy would lead to the least loss of information. Results are reported in the Results section.

Update the naïve model on data alone or data and priors to generate posterior models

As described above, the naïve model was updated on priors data to create two prior models, one for each alternative prior aggregation assumption. We also updated the naive model on trial data which had been mapped to node values. This created a data only model. To generate a posterior model informed by both priors and data for each partner trial, we also updated each naïve model on both priors data, and data from the trial, with one set of data simply appended to the other. The end result of this process was six updated models for each partner trial resulting from six different model specifications as described in Table 5.

Table 5: The six model specifications

Model name	Updated on priors dataset	Updated on trial data
Naive model	None	No
Prior model 1	Priors 1 (Exclusive)	No
Prior model 2	Priors 2 (Inclusive)	No
Data only model	None	Yes
Posterior model 1	Priors 1	Yes
Posterior model 2	Priors 2	Yes

The process of updating on data was completed in CausalQueries using the update_stitch command. This was necessary for the UPAVAN model because this model has too many causal types to be updated in a single Stan fitting on a consumer computer. Indeed, our experiments with institutional high-performance computers suggested that these computers might also not be capable of updating the model using Stan. Therefore, it was necessary to break the model down into independent sub-DAGs, fit these models and then stitch them back together to create an updated full model using the update_stitch function.

We were also forced to use this approach for the DIG model specifications that included prior data. This was because the data generated to represent experts' priors was created at the sub-DAG level where each sub-DAG captures only the relationships between a child node and its immediate parents. This data includes an even distribution of parent node values that has been artificially set by the prior elicitation exercise. The useful information contained in the data for each sub-DAG concerns the value the child node takes given each combination of parent node values. However, updating a model with flat priors on this prior data using update_model updates the model not only on the useful information contained on the relationships between parent and child nodes, but also erroneously updates the model so as to reinforce the expectation that parent node values will be evenly distributed. This problem arises because update_model treats the priors data as a rectangular dataset with missing values when in fact this is inappropriate. In fact, this data should be thought of as a collection of datasets with one for each child node in the DAG; treating it as a single dataset with high missingness is inappropriate. The update_stitch solves this problem because it updates each sub-DAG on the relevant data and then discards information about nodal types for parent nodes which are not exogenous nodes in the full model. It therefore does not erroneously reinforce an expectation that parent nodes' values will be evenly distributed.

One limitation of this approach is that it does not allow for missing data. Therefore, data which was missing in the trial data translated to node values had to be imputed. We used

multiple imputation to create a complete set of data. We adopted the MICE (Multiple Imputation by Chained Equations) method(Van Buuren, 2007; Van Buuren et al., 2006; Van Buuren & Groothuis-Oudshoorn, 2011). This is a robust, informative method of dealing with missing data in datasets. The procedure 'fills in' (imputes) missing data in a dataset through an iterative series of predictive models based on available information in the dataset. This procedure was conducted for the trial data for both partner trials, results of which are presented and validated in the Results section.

Before querying the models produced through updating, we plotted the posterior distributions of nodal types for all nodal types in each model as a way of 'eyeballing' and sense-checking our results. These plots are presented in the Results section. As discussed in the Conclusions section, we found this step in our method to be of surprising value and would encourage it for future researchers applying this method.

Part 5: Queries

Once the naïve model had been updated according to the six different model specifications in Table 4 above, we had six models of causation in each partner intervention in CausalQueries that could be queried for causal estimates. This was done using the query_model command, which takes three compulsory arguments: model, a causal model; queries, one or more statements about potential outcomes to be evaluated such as "Y[X=1] - Y[X=0]"; and using, a statement that specifies whether to evaluate the queries against the posterior, the prior or the naively parameterised distribution corresponding to the model. The command returns a data.frame with columns 'Query,' 'Given' and 'Using' defined by corresponding input values. Further columns are generated as specified in the stats argument, defaulting to the mean and standard deviation for the probability of the query evaluating to true across the distribution of causal types. This is calculated by assessing the distribution of causal types in a model. Additionally, a given argument can be specified which allows the estimand to be conditioned on an observational or counterfactual distribution for example by specifying "X==1 & Y==1". For example, the query "Y[X=1]> Y[X=0]" is asking whether, according to the model, X has a positive effect on Y. Using a list of givens, this query could return the probability of the statement being true under each of a list of assumptions about the values of other nodes.

Table 6, below, outlines possible queries and givens for different simple DAGs and gives their interpretation:

Table 6: Possible queries for simple models

DAG	Query	Given	Interpretation
X -> Y	Y[X=1] > Y[X=0]		Probability that X has a positive effect on Y
X -> Y	Y[X=1] < Y[X=0]	X == 1	Probability that X has a negative effect on Y among those for whom X=1
X -> Y	Y[X=1] > Y[X=0]	X==1 & Y==1	Probability that Y=1 is due to X=1 (Attribution)
X -> Y <- W	Y[X=1] > Y[X=0]	W == 1	Probability that X has a positive effect on Y for a case in which W = 1 (where W is possibly defined post treatment)
X -> Y <- W	Y[X=1, W = 1] > Y[X=0, W = 1]	W==0	Probability that X has a positive effect on Y if W were set to 1 for cases for which in fact W=0
X -> Y <- W	Y[X=1] > Y[X=0]	Y[W=1] > Y[W=0]	Probability that X has a positive effect on Y for a case in which W has a positive effect on Y
X -> Y <- W	(Y[X=1, W = 1] > Y[X=0, W = 1]) > (Y[X=1, W = 0] > Y[X=0, W = 0])	W==1 & X==1	Probability of a positive interaction between W and X for Y; the probability that the effect of X on Y is stronger when W is larger
X -> M -> Y <- X	Y[X = 1, M = M[X=1]] > Y[X = 0, M = M[X=1]]	X==1 & M==1 & Y==1	The probability X would have a positive effect on Y if M were controlled to be at the level it would take if X were 1 for units for which in fact M==1
X -> M -> Y <- X	(Y[M = 1] > Y[M = 0]) & (M[X = 1] > M[X = 0])	Y[X=1] > Y[X=0] & M==1	The probability that X causes M and M causes Y among units for which M = 1 and X causes Y

To evaluate the query statement, the query_model function looks at all of the possible ways in which a unit can react causally to the nodes in the causal model: the total set of causal types. Each causal type is a way in which the nodal types can sum to create a unique flow of causation through the causal model. In the relatively simple DIG model there are 594 possible causal types. The query_model function evaluates the query statement against each of these causal types, then returns a weighted average across all causal types. The weight given to causal types when returning this average is determined by the expected prevalence of that causal type. This expectation is set by the draws from the posterior distribution of nodal types returned by Stan when the model in question was updated. Based on these draws Stan infers

a distribution for the probability of each nodal type obtaining for units to whom the model applies. So, for example, many of the 594 possible ways in which units might react causally to the total model break down into a set of nodal types that are all expected to be rare. These will have very little weight in determining the value returned by the query. Other causal types will be summations of nodal types that are expected to be common and these will have a higher weight in the average returned by the query.

We queried all six model specifications described in Table 5, above for each partner intervention to generate alternative estimates of the average treatment effect under different assumptions. For further queries exploring mediation and moderation of causal effects by various nodes in each DAG we used the four model specifications which do not assume exclusive priors as we considered the inclusive priors assumption to be more appropriate on reflection, as discussed in the results section. These queries are described in detail in Part 5 of the Results section below, along with their results. For each query, we compared estimates between prior, posterior and data-only specifications to explore the role of expert priors in our analysis.

3. Data

Two forms of data were collected for each partner trial, that is, priors data and trial data. The priors data in each of the two partner trials was collected from experts that were either project designers, implementers or members from the evaluation teams. The trial data of relevance to nodes in the DAG was identified during the DAG elicitation process as described above. This data was then shared by the data manager of each partner project and transformed into binary or ternary node values by the POInT data manager in consultation with the partner teams.

Description of priors data

Despite attempting to simplify the task, and refining our tool in response to piloting, the UPAVAN prior elicitation exercise was difficult for participants to grasp. Of nine participants, one opted out of the prior elicitation exercise just before distribution of the prior elicitation tool, after having made limited contributions to the DAG elicitation process. This participant judged at that point that their knowledge of intervention theory was not sufficient to make a useful contribution. Two participants began but did not complete the exercise, two required a lot of help, and four proceeded without much help. This represents quite a high level of attrition and suggests that this prior elicitation tool may only be suitable for more technical and/or more motivated participants. Although the participants who required a lot of help were the participants with the least training in statistics and causal reasoning, the participants who abandoned the effort entirely were among the more technically-trained. These participants did begin the exercise but abandoned it incomplete, suggesting that they found the exercise too demanding on their time.

The final UPAVAN priors data contained a total of 580 point estimates for 208 parameters from six participants, each associated with a level of certainty allowing a distribution to be fitted around the point estimate.

The prior elicitation tool for the DIG partnership was produced using the same process as for the UPAVAN partnership. As the DAG was much smaller, there were less contingency tables

for participants to complete. Five of the six participants responded promptly and filled out the tool successfully within the first week. The sixth participant had not responded after one week and was chased but did not respond. However, this good response rate masks some evident confusion about the tool from participants.

The final DIG priors data contained a total of 144 point estimates for 24 parameters from six participants, each associated with a level of certainty allowing a distribution to be fitted around the point estimate.

Description of trial data

The UPAVAN trial data had fourteen (14) variables considered for data collection as described in Table 7, namely: PRODIVERSITY, YIELD, SALES, LAND, WATER, STSC, MARGIN, PURCHASES, VIDEO, PRODDECISION, DIETDECISION, DIETDIVERSITY, VISIT and AGRI. A total of 2,097 observations were collected for the UPAVAN trial data. Missing data was observed and multiple data imputation by chained equations (MICE) was implemented to ensure complete data as described in Part 4 of the Empirical Methods subsection of the Methodology section. Table 7 is a summary of the measures adopted for nodes in the UPAVAN DAG.

Table 7: Description of Selected Measures Adopted for Nodes in the UPAVAN DAG

Node label	CausalQueries shortening	Causal category	Node description	Measure source	Measure description	Binary / Ternary coarsening	Notes
Supportive household	HOUSEHOLD	Moderator	Household is supportive of womens' empowerment	Unmeasured	n/a	n/a	Unmeasured
Allocated to AGRI	AGRI	Allocation	Household allocated to the AGRI arm	Women Endline	Cluster level allocation measure	Allocated or not	
Marginalised group	MARGIN	Moderator	Household member of a marginalised group in their community	Males Endline	6.8 "Do you belong to any of the following [castes]?" 6.9 "Which of the following tribes do you belong to?"	HH is a member of a group that is in the minority in that cluster, or not.	
Available land for cultivation	LAND	Moderator	Household owns more land than a kitchen garden	Males Endline	"In total, how much land" (acres)	Greater than 0.5 acres or not	
Regular videos in groups	VIDEO	Intervention activities	Woman regularly attends group video sessions (at least every other month)	Women Endline	Attended one meeting in last 6 months	Yes or no	
Regular home visits	VISIT	Intervention activities	HH is visited regularly (at least every other month)	Women Endline	"Has anyone from VARRAT visited your household since the start of the year (1st January 2019)?"	Yes or no	
Non-monsoon water	WATER	Moderator	Household has access to a water supply that would operate outside the monsoon season	Males Endline	How much land irrigated?	Any irrigated or not	

Node label	CausalQueries shortening	Causal category	Node description	Measure source	Measure description	Binary / Ternary coarsening	Notes
			such as a well				
Influence food production decisions	PRODDECISION	Mediator	Woman in household influences most food production decisions	Women Endline	“How much input or influence did you have in making decisions about food production, including any decisions about income from food production? Input means gave an opinion, was asked an opinion, or was able to influence an action. Probe: Did your spouse or others in the household consult / ask with you when making decisions about the activity or what to do with the income from that activity?” also 15.6.1 and .7.1 about cash cropping and livestock farming	Lump the three (food, crop, livestock) together. Data is "no input" "input in some" "input in most or all". Code as decision maker if 'most or all' for at least one of those three questions.	
Influence dietary decisions	DIETDECISION	Mediator	Woman in household influences most decisions over her own nutrition	Women Endline	“How much input or influence did you have in making decisions about minor household expenditures? Minor food expenditures might include foods for daily consumption or other household needs. Input means gave an opinion, was asked an opinion, or was able to influence an action.	Code as decision maker if 'input in most or all'	Q 15.9 is only about purchasing decisions but is best measure available

Node label	CausalQueries shortening	Causal category	Node description	Measure source	Measure description	Binary / Ternary coarsening	Notes
ST/SC	STSC	Moderator	Household member of a scheduled tribe or caste	Males Endline	6.8 "Do you belong to any of the following [castes]?" 6.9 "Which of the following tribes do you belong to?"	HH is member of a scheduled caste or scheduled tribe, or not	
High produce yield	YIELD	Intermediate outcome	Household produces a high yield (giving them the ability to more often eat the crops grown)	Males Endline	yield and costs questions	Household breaks even, producing yield worth more than costs, or not.	This measure will have a lot of errors resulting from imperfect recall. However, it is the best measure available
Produce sales	SALES	Intermediate outcome	Household sells crops (giving them the ability to buy nutritious foods)	Males Endline	sales questions	Any sales or not	
Produce diversity	PRODDIVERSITY	Intermediate outcome	Household produces a diverse set of produce over the year (giving them the ability to eat more of their own)	Males Endline	production questions	Three levels: create groups of foods and produce a variable. 0 = grains; 1 = '0' plus any other pulses, vegetables, fruits, nuts or seeds	

Node label	CausalQueries shortening	Causal category	Node description	Measure source	Measure description	Binary / Ternary coarsening	Notes
						OR animal-derived foods; 2 = '0' plus both other categories	
Food purchases	PURCHASES	Intermediate outcome	Household buys diverse foods	Males Endline	purchases	Three levels: create groups of foods and produce a variable. 0 = grains; 1 = '0' plus any other pulses, vegetables, fruits, nuts or seeds OR animal-derived foods; 2 = '0' plus both other categories	
Maternal dietary diversity	DIETDIVERSITY	Primary outcome	Woman achieves sufficient maternal dietary diversity (by eating at least 5 out of 10 food groups each 24 hours)	Women Endline	consumption in last 24 hours	Equal to or more than 5 of 10 food groups consumed, or not.	

The DIG trial data had five (5) variables considered for data collection as described in Table 8, namely: ALLOC, ASSET, CLASSIC, DIINDIV and ACTIVE. A total of 1,108 observations were collected for the DIG trial data. Missing data was imputed using MICE to ensure completeness of data following the same process as for the UPAVAN data. Table 8 is a summary of the measures adopted for nodes in the DIG DAG.

Table 8: Description of Selected Measures Adopted for Nodes in the DIG DAG

Node label	CausalQueries term	Causal category	Node description	Measure source	Measure description	Binary / Ternary coarsening	Notes
Allocated to treatment	ALLOC	Allocation	Primary beneficiary for the household is in a cluster allocated to treatment	HH endline survey	allocation to treatment	treatment = 1, control =0	
Asset transfer received	ASSET	Intervention component	Household receives asset transfer	HH endline survey	Has anyone in your HH ever received [support] from DIG (BRAC, HI or NUWODU?) Type of support: 5 Poultry 6 Sheep 7 Goat 8 Pigs/piglets 9 Cow/bull 10 Goat 11 Crop seeds 13 Business capital 1 = Yes 0 = No→ Next	At least one 'yes' to any of these questions = 1, otherwise = 0	

Node label	CausalQueries term	Causal category	Node description	Measure source	Measure description	Binary / Ternary coarsening	Notes
Other classic graduation interventions received	CLASSIC	Intervention component	Primary beneficiary receives at least one other classic graduation intervention e.g. training, consumption support, health subsidy	HH endline survey	Has any HH member attended [training]? Type of training: 01 Agricultural training or extension services, such as training in new agricultural practices or in farmer field school groups 02 Vocational skills training, such as training in tailoring, carpentry, mechanic, brick-making, etc. 03 Business and financial skills training 04 Village Savings & Loan Association (VSLA) or other savings training 05 Leadership or management training 06 Community-based organisation development, including group formation, capacity building, constitution-writing, etc. 1 = Yes 0 = NoNext training item.	At least one 'yes' to any of these questions = 1, otherwise = 0	

Node label	CausalQueries term	Causal category	Node description	Measure source	Measure description	Binary / Ternary coarsening	Notes
					AND Has anyone in your HH ever received [support] from DIG (BRAC, HI or NUWODU?) Type of support: 14 - Consumption support 16 - Health subsidy 18 - Individual-level support/coaching for income generation 1 = Yes 0 = No→ Next		

Node label	CausalQueries term	Causal category	Node description	Measure source	Measure description	Binary / Ternary coarsening	Notes
Individual-level disability-inclusive interventions received	DIINDIV	Intervention component	At least one individual-level disability-inclusive intervention received e.g. assistive device, physiotherapy	HH endline survey	Has anyone in your HH ever received [support] from DIG (BRAC, HI or NUWODU?) Type of support: 15 Assistive devices No number Physio therapies for PWDs 17 Home or place of work modification 1 = Yes 0 = No→ Next	At least one 'yes' to any of these questions = 1, otherwise = 0	
Cluster-level disability-inclusive interventions received	DICLUSTER	Intervention component	Household is in a cluster which received at least one cluster-level disability-inclusive intervention e.g. training and logistical support for paralegals, sensitivity training for village leaders	HI/NUWODU administrative data - unavailable in our timeframe	n/a	n/a	Unavailable in our timeframe

Node label	CausalQueries term	Causal category	Node description	Measure source	Measure description	Binary / Ternary coarsening	Notes
Actively engaged in economic activity	ACTIVE	Outcome	Primary beneficiary for the household is actively engaged in economic activity	HH endline survey	Section 7: Other household income-generating activities 7.2 In the last week, did [NAME] do more than one income generating activity? 7.3 In the last week, did [NAME] do any work for payment...? 7.4 Does [NAME] have any such paid work ...? 7.5 In the last week did [NAME] run a business of any size, for ...? 7.6 Does [NAME] run any business that [NAME] will ...? 7.7 In the last week did [NAME] help in any kind of business run by another ...? 7.8 In the last week was [NAME] an apprentice? 7.9 Does [NAME] have an apprenticeship that [NAME] will ...? 7.10 In the last week did [NAME] work on the HH farm or with... ? 7.11 Will [NAME] definitely work on the HH farm ...? 7.12 In the last four weeks was [NAME] looking ...?	At least one 'yes' to any of these questions = 1, otherwise = 0	

Node label	CausalQueries term	Causal category	Node description	Measure source	Measure description	Binary / Ternary coarsening	Notes
					7.13 In the last four weeks was [NAME] trying to ...?		

Read-only copies of the workbooks with priors data for UPAVAN and DIG are linked to in the relevant results section.

4. Results

In this section we describe the main findings from our project. In our CEDIL working papers we provide code chunks for each analysis step completed in software. Here, we omit the code and are more descriptive, though we refer to the commands used where this is helpful to understanding.

Part 1: Partnerships

In developing and testing the POInT approach we intended to partner with three partner evaluations. The first of these partnerships was with the UPAVAN trial. The second partnership, as discussed in the Methods section, was with the DIG intervention. We also pursued a third partnership with the AMETHIST trial as also discussed in Empirical Methods subsection - Table 3.

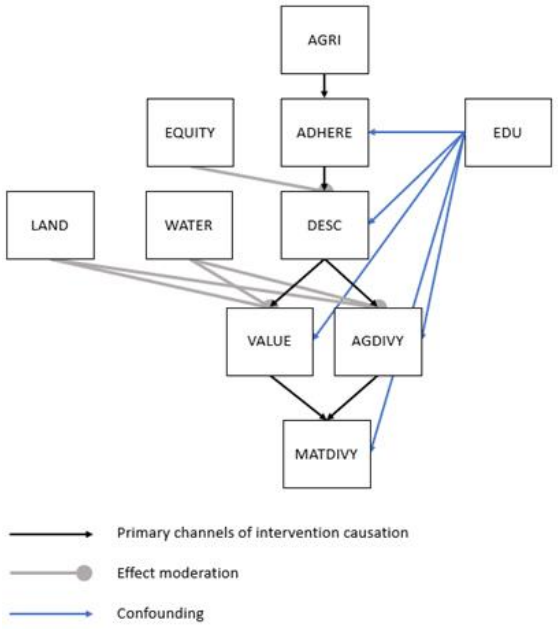
Our partnership with the AMETHIST Team only managed to yield an unfinalised draft DAG, after following all the necessary DAG elicitation processes as proposed earlier in the Methods section. Thus, we were only able to partially implement step one (adaptation of the partner team's theory of change into a formal causal model) of the six steps of the proposed POInT approach. Progress was stalled mainly because our partner team was heavily engaged in the process of collecting, collating and writing up the AMETHIST project evaluation report. However, we drew some valuable insights in our engagement with the AMETHIST project partnership.

All of our proposed partnerships involved no data collection on our part other than data on our colleagues' recollections of their initial understandings of intervention causation (their 'priors' about causation in the intervention). We had initially hoped to augment this approach with data collection on the priors' of recipients of the intervention, but COVID-related delays and the impossibility of securing a no-cost extension to our project meant that we were forced to limit our scope exclusively to secondary re-analysis as described in the Methods section.

Part 2: DAGs

UPAVAN DAG elicitation

Before meeting with expert participants from the UPAVAN team we developed a candidate DAG based on our analysis of the UPAVAN trial protocol and papers, with particular attention to the theory of change diagram developed for the UPAVAN protocol(Kadiyala et al., 2018). This candidate DAG was able to serve as a teaching aid and as a starting point for the elicitation process that followed. This candidate DAG and the definitions of the nodes in that DAG is presented in Figure 9, below.



Node	Type of causal entity	Description of theoretical entity
AGRI	Allocation	Household allocated to the AGRI arm.
ADHERE	Intervention activities	Woman in household sees any AGRI videos.
DESC	Mediator	Woman in household involved in or making a majority of important agricultural decisions.
VALUE	Intermediate outcome	Household achieves higher-than-average value of agricultural production over the year.
AGDIVY	Intermediate outcome	Household achieves adequate diversity of agricultural production by producing at least 5 out of 10 food groups over the year.
MATDIVY	Primary outcome	Woman achieves sufficient maternal dietary diversity by eating at least 5 out of 10 food groups each 24 hours.
EQUITY	Moderator	Woman in household is at least as empowered as man.
LAND	Moderator	Household owns more land than a kitchen garden
WATER	Moderator	Household has access to a water supply that would operate outside the monsoon season such as a well.
EDU	Confounder	Woman in household has completed primary school.

Figure 9: Initial candidate UPAVAN DAG and node definitions

We engaged in two cycles of DAG workshop, email comments, and refinement by the POInT team of the candidate DAG before arriving at a candidate analysis DAG that was agreed by all members of the UPAVAN team. The DAG elicitation process was a significant time investment for members of the UPAVAN team with two 1.5-hour workshops conducted online as well as engagement with an email discussion.

Over the course of the two workshops and the email discussion, the group of relevant experts for the UPAVAN DAG and prior elicitation process was refined to nine people. Four of these people were local implementation partners and five were academics working on different aspects of the trial. At the end of the second DAG workshop, a candidate analysis DAG was agreed, which is presented below as Figure 10. We present this DAG and pick up the process of refining the DAG at this point in order not to over-burden this report with every stage of the process.

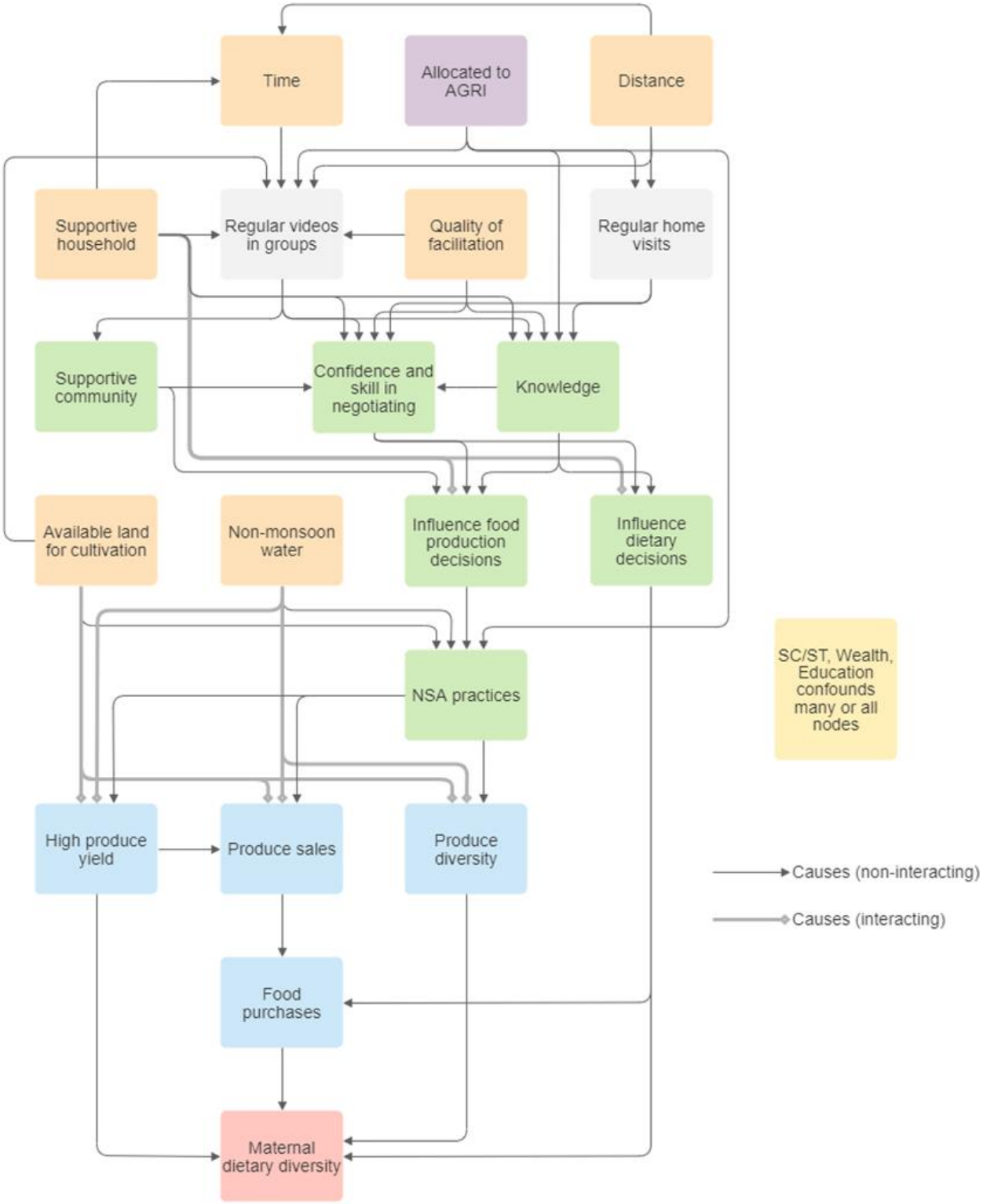


Figure 10: Candidate UPAVAN analysis DAG post second-workshop

This DAG reflected the team's understanding of their theory of change for the intervention fairly well, though the contextual features or confounders most relevant to causal effects in the intervention had not yet been specified as nodes, leaving a 'floating' yellow node to the right of the DAG. A table of definitions for nodes was also produced during the DAG workshop process which is presented as Table 9, below.

Table 9: Node definitions for UPAVAN after second DAG-development workshop

Node	Type of causal entity	Description of theoretical entity
Allocated to AGRI	Allocation	Household allocated to the AGRI arm
Regular videos in groups	Intervention activities	Woman regularly attends group video sessions (at least every other month)
Regular home visits	Intervention activities	HH is visited regularly (at least every other month)
Supportive community	Mediator + Moderator	Woman is embedded in a supportive community – **defined by effect or can we unpack?**
Confidence and skill in negotiating	Mediator	Woman has confidence sufficient to speak up – **defined by effect or can we unpack?**
Knowledge	Mediator	Woman has adequate knowledge of her circumstances and nutrition-sensitive agricultural practices – **can we define a level required for causal effect?**
Influence food production decisions	Mediator	Woman in household influences most agricultural decisions
NSA practices	Mediator	The household uses at least some of the cultivation methods taught in videos
Influence dietary decisions	Mediator	Woman in household influences most decisions over her own nutrition
High produce yield	Intermediate outcome	Household produces a high yield (giving them the ability to more often eat crops grown)
Produce sales	Intermediate outcome	Household sells crops (giving them the ability to buy nutritious foods)
Food purchases	Intermediate outcome	Household buys nutritious foods
Produce diversity	Intermediate outcome	Household achieves adequate diversity of agricultural production by producing at least 5 out 10 food groups over the year (giving them ability to eat more their own)
Maternal dietary diversity	Primary outcome	Woman achieves sufficient maternal dietary diversity (by eating at least 5 out of 10 food groups each 24 hours)
Time	Moderator	Woman has enough free time to travel and attend groups at least every other month
Distance	Moderator	Woman is within walking distance of groups
Quality of facilitation	Moderator	Sessions are facilitated to a good standard
Supportive household	Moderator	Woman in household supported by others – **defined its effect, or can we unpack?**
Available land for cultivation	Moderator	Household owns more land than a kitchen garden
Non-monsoon water	Moderator	Household has access to a water supply that would operate outside the monsoon season such as well.
SC/ST/Wealth/Education	Confounders	Multidimensional, to be decomposed and specified more precisely for analysis DAG

Following the second DAG development workshop more work was done to clarify the meaning of nodes in this table before arriving at a final version. This was achieved via email

discussion and brief one-on-one meetings between members of the POInT and UPAVAN teams. In parallel, the candidate analysis DAG was refined so as to be suitable for analysis using the POInT approach.

The first step in refining the DAG was to remove nodes that had not been measured quantitatively. Although the POInT approach permits the integration of different quantitative data sources, CausalQueries is a fundamentally quantitative tool and we have not been able to make good use of the qualitative data collected by either of the partner trials in this iteration of our work on the POInT approach. We discuss the possibilities and limitations of an approach to including qualitative trial data in a Bayesian analysis framework in the Conclusions section of this paper. Figure 11 shows the UPAVAN candidate analysis DAG with nodes for which no quantitative data was available greyed-out.

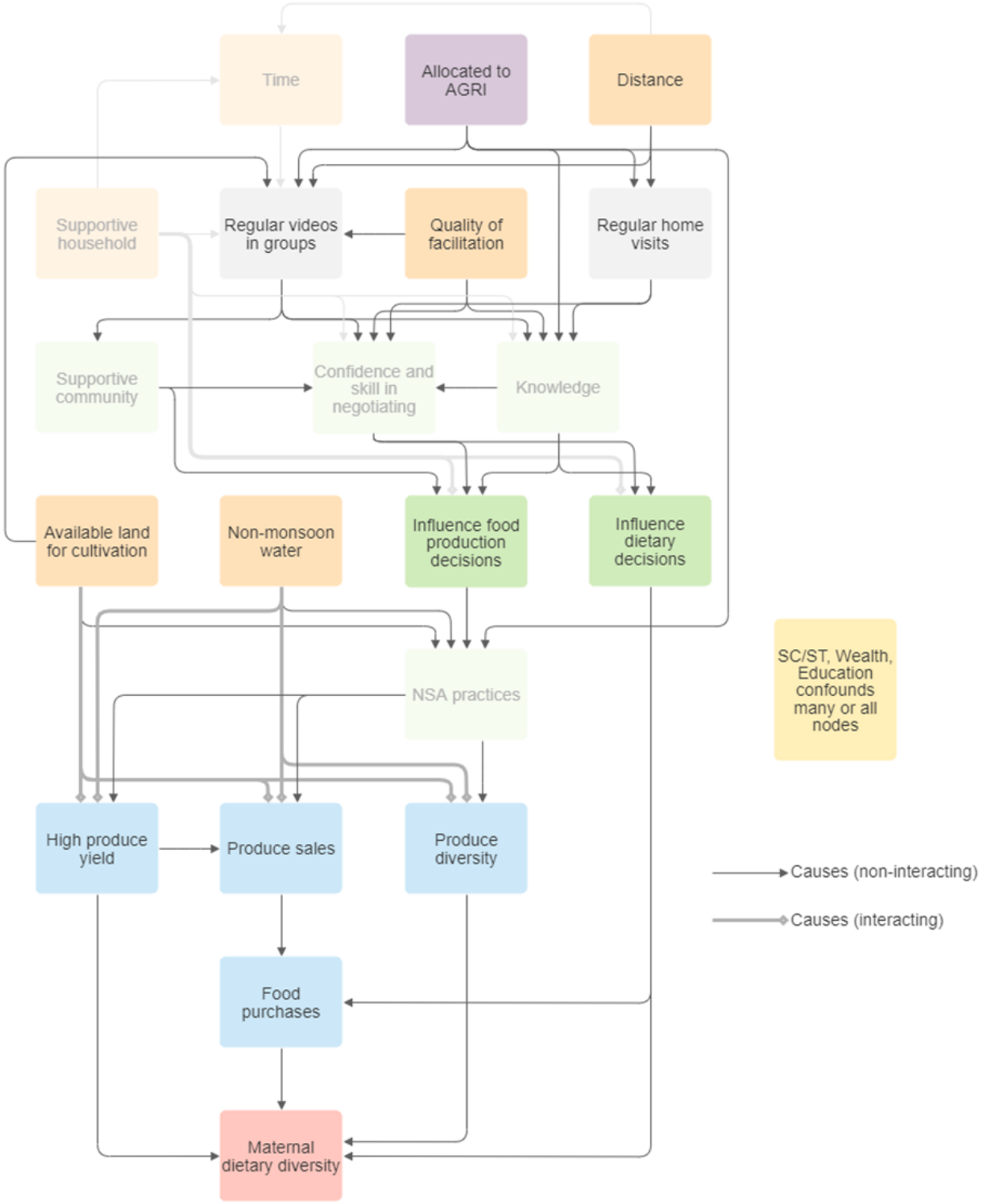


Figure 11: Candidate UPAVAN analysis DAG with unquantified nodes removed

Despite having restricted our analysis to a subset of the nodes on the post-second meeting candidate UPAVAN analysis DAG, we wanted to ensure that the reduced DAG was compatible with the full DAG. This meant we had to preserve the channels of causation reflected by edges in the full DAG in the reduced DAG. For example, in Figure 11, above, we can see that there is no direct edge from 'Regular videos in groups' to 'Influence food production decisions.' However, there are three edges via nodes that must be removed and so the reduced DAG

must preserve an edge between 'Regular videos…' and 'Influence food…' The dotted, two-headed edge in Figure 12, below, represents the confounding effect between the nodes it connects to the removed node 'Supportive household.'

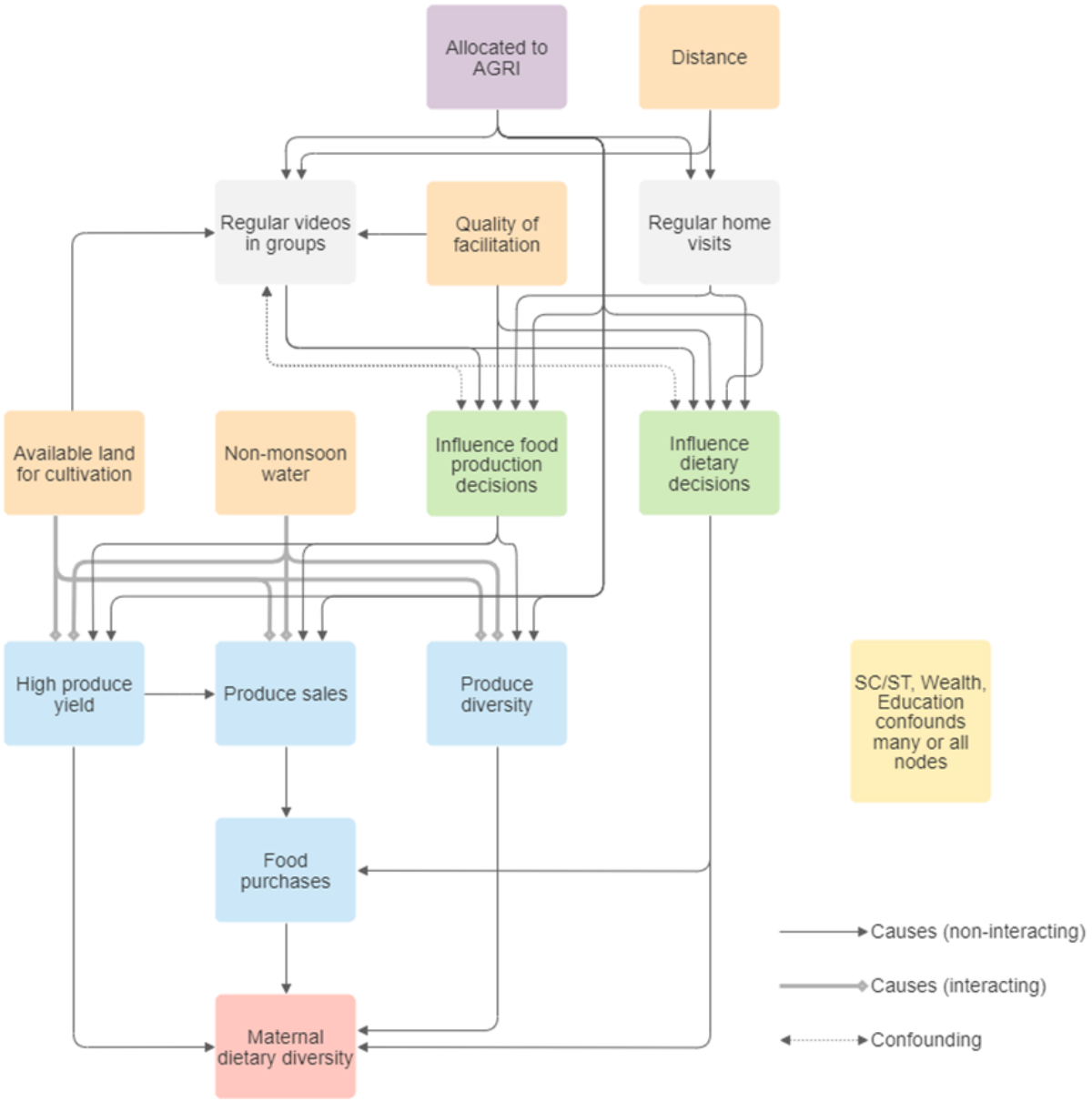


Figure 12: Candidate UPAVAN analysis DAG with causation via removed nodes reflected

The candidate DAG above remained uncomfortably complex for a first application of the POInT approach despite the fact that contextual factors had still not been elaborated. We therefore decided that if at all possible, the DAG should be simplified further by pruning any nodes that were judged least causally significant by our UPAVAN partners. In discussion with them, we decided to prune 'Quality of facilitation' and 'Distance.' In the first instance, this was because the quality of the data available was so low as to render that node effectively unmeasured. In the second case, the intervention had been designed so as to attempt to eliminate the causal effect of distance by setting up a regular meeting for every group of approximately 20 households, meaning all households in the sample should be close to a meeting location.

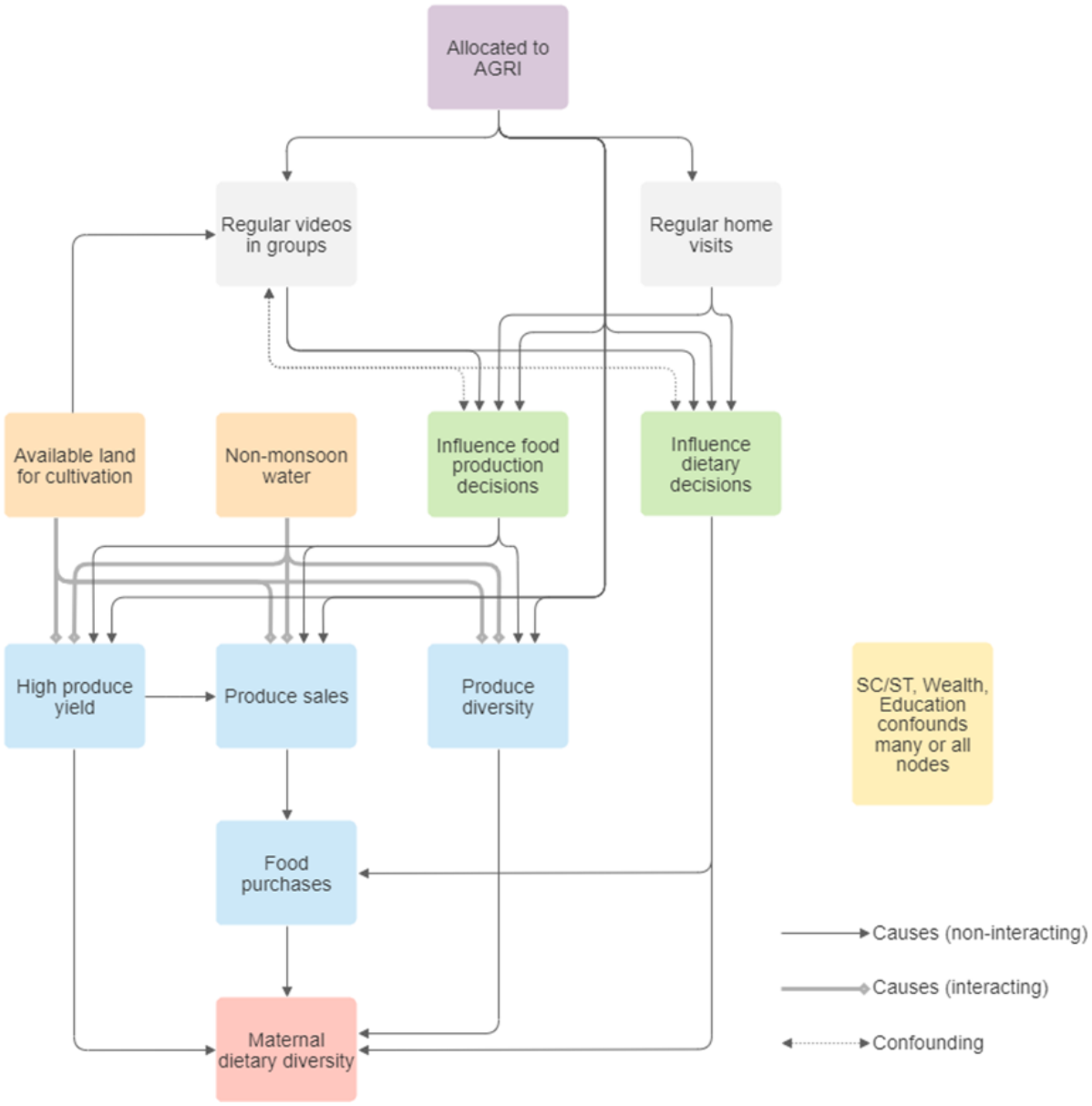


Figure 13: Candidate UPAVAN analysis DAG with two nodes pruned

Once these changes had been agreed with the UPAVAN team by consensus, we were able to work with them over email to add two key features of context to the DAG and to rule out adding other potential contextual features on the grounds that they were not expected to be significant by the UPAVAN team and their inclusion would make the model too complex for the POInT approach. Specifically, we were aware that five parent nodes per child node was a hard limit on what was computable using CausalQueries and so any further parent nodes for several nodes in the DAG would make it uncomputable. Including two key features of context led to the following candidate analysis DAG presented in Figure 14.

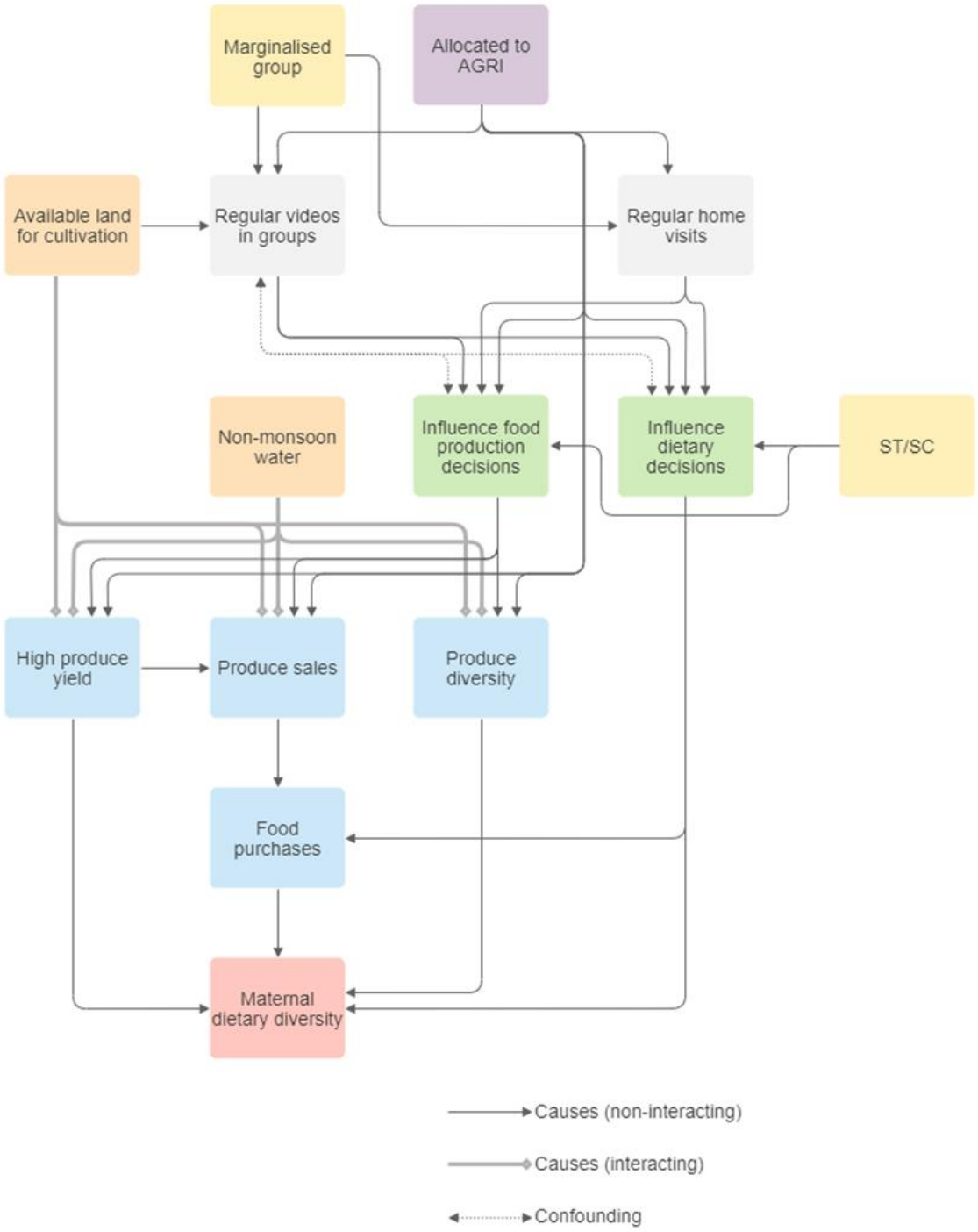


Figure 14: Candidate UPAVAN analysis DAG with confounders added

The full analysis DAG used for the prior elicitation process is equivalent to Figure 14, above, but includes the confounding effect of 'Supportive household' by explicitly representing that unmeasured node, as in Figure 15, below. We used that presentation for the prior elicitation process in order to be able to elicit priors on the confounding effect of that unmeasured node.

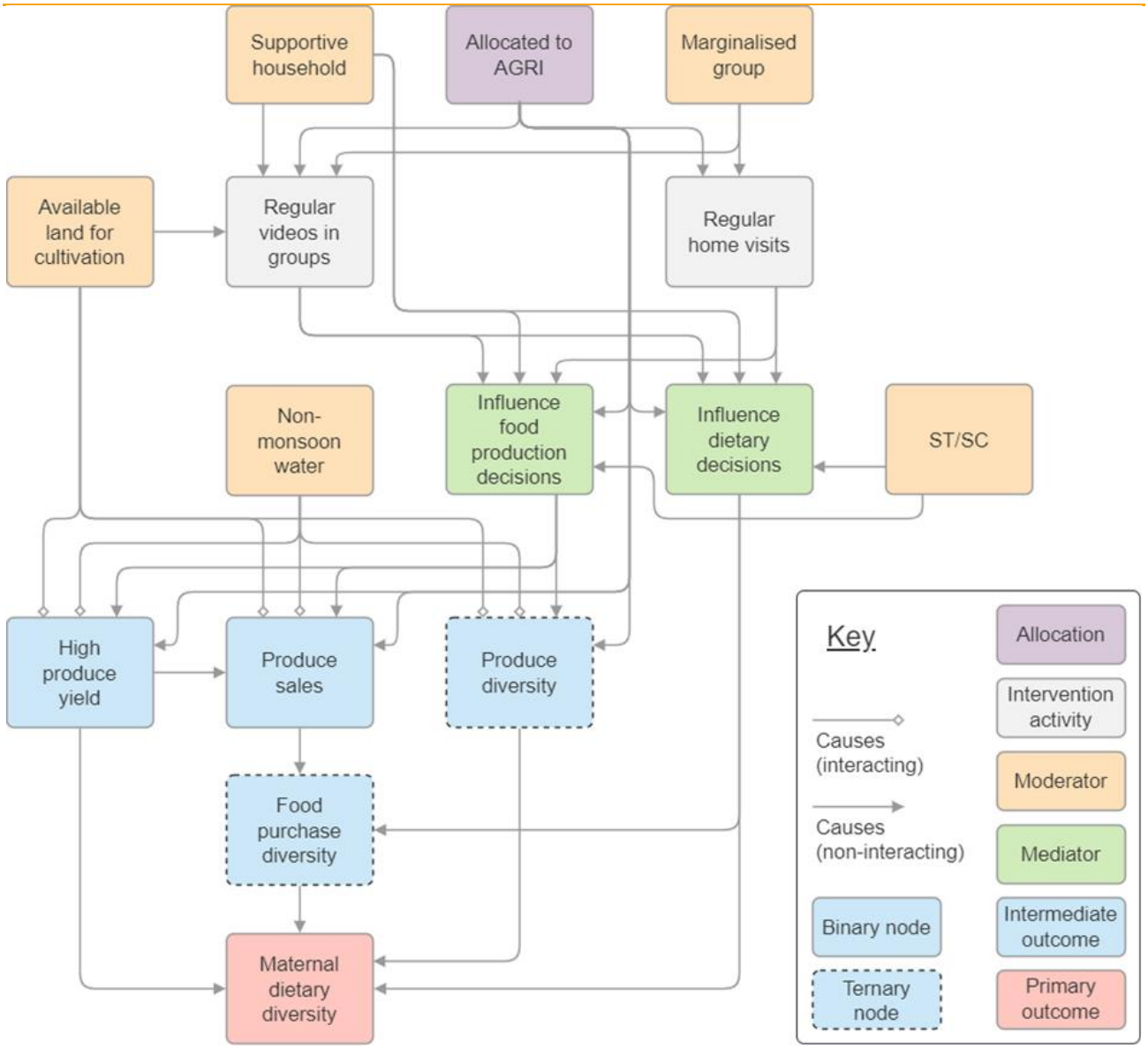


Figure 15: Final UPAVAN analysis DAG

In addition to deciding the structure of the DAG, the DAG elicitation process led to clear definitions in binary or ternary terms of all nodes in the DAG. These definitions are reproduced in Table 7, below. Most nodes could be defined in binary terms and so were. However, the UPAVAN team expressed that limiting the definitions of two nodes to binary would have done unacceptable violence to their theory. For these nodes, 'Produce diversity' and 'Food purchase diversity' a definition was given in ternary (three-valued) terms. This is reflected in the extended DAG notation of Figure 15, above, and in Table 7 in the Data section above.

As well as the structure of the DAG and the definitions of nodes, the DAG elicitation process had to lead to a shared understanding of the assumptions made in the causal model created from the DAG. The UPAVAN team understood and accepted the assumption of monotonicity of effects at all nodes. Interactions were theorised to be required between two moderators (land and water), but otherwise were not required by the experts' theory. This is reflected in

the extended DAG notation of Figure 15, above, where 'interacting' causation is distinguished from 'non-interacting' causation.

DIG DAG elicitation

Before meeting with expert participants from the DIG team we developed a candidate DAG based on our analysis of the DIG trial protocol with particular attention to the two different theory of change diagrams that had been produced for the intervention funding proposal and for the evaluation protocol. These diagrams were both highly ambiguous, meaning that creating an unambiguous DAG on the basis of the diagrams would be challenging and involve a lot of interpretation. However, the supporting text in the evaluation protocol was a little clearer and we were able to produce a candidate DAG to use as an aid for teaching DAG principles and as a starting point for discussion with DIG experts. This candidate DAG is presented as Figure 16, below:

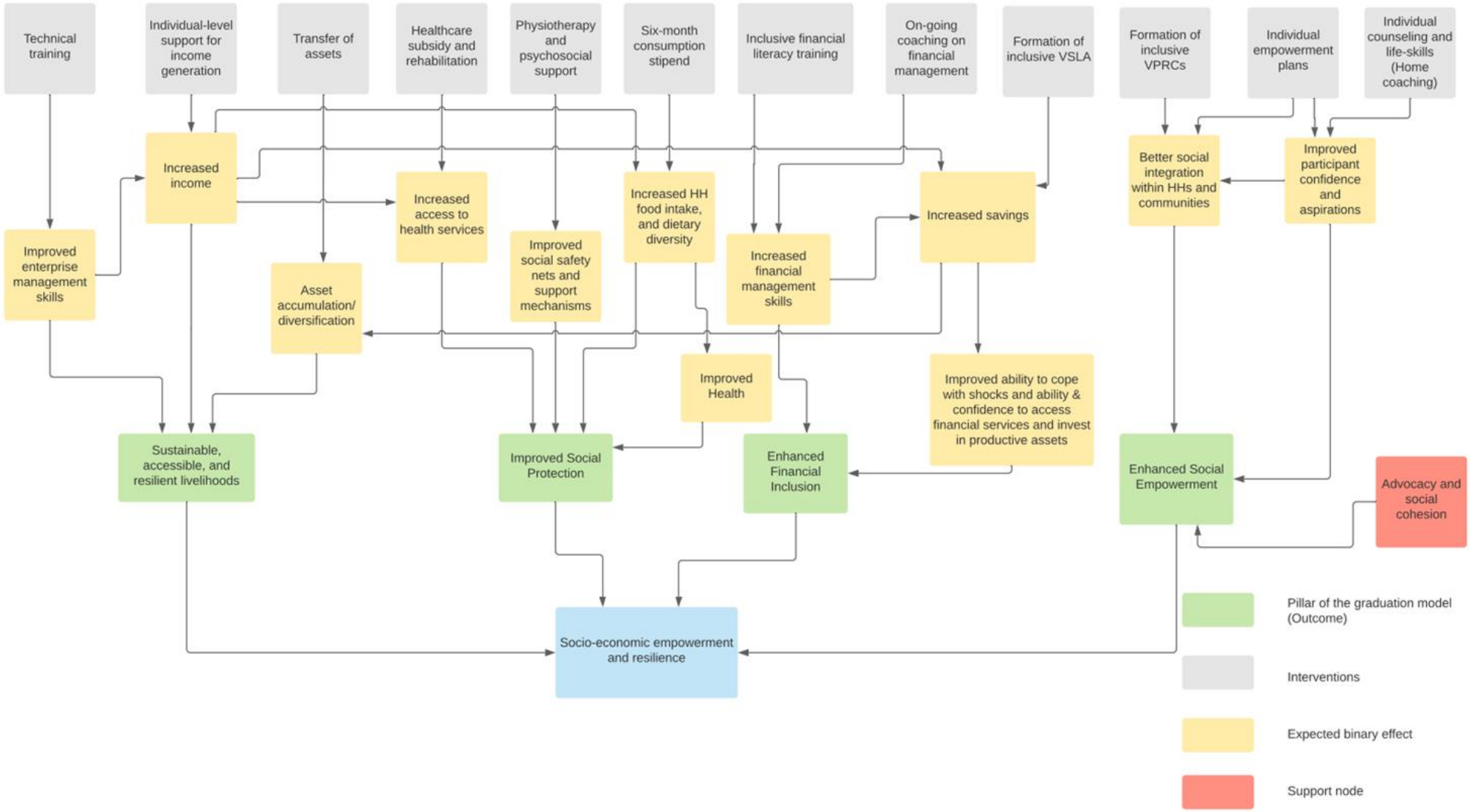


Figure 16: Initial candidate DIG DAG

Unlike for the UPAVAN partnership, the DIG evaluation was ongoing when our partnership began and it was therefore possible to engage more junior implementation team members who were located closer to the mechanisms of the intervention. We therefore operated the first cycle of DAG elicitation workshops in Uganda with members of local BRAC branches who were implementing the intervention. These workshops produced three candidate DAGs which are presented below as Figure 17, Figure 18 and Figure 19 respectively.

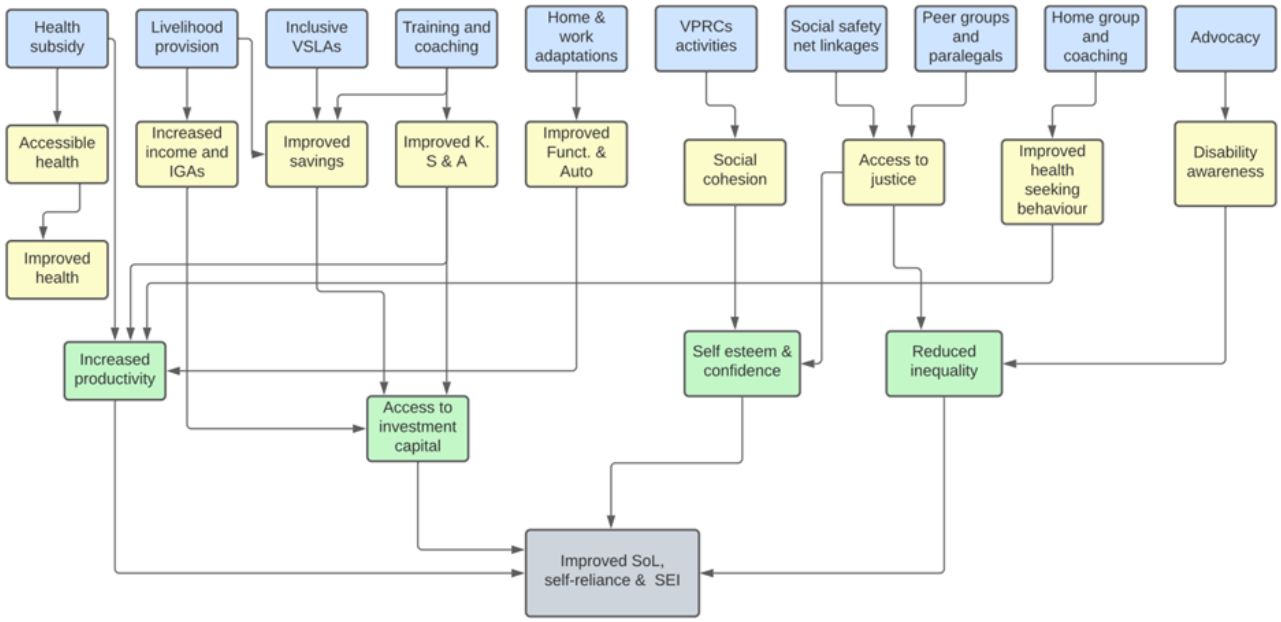


Figure 17: Candidate DIG DAG Uganda Group 1

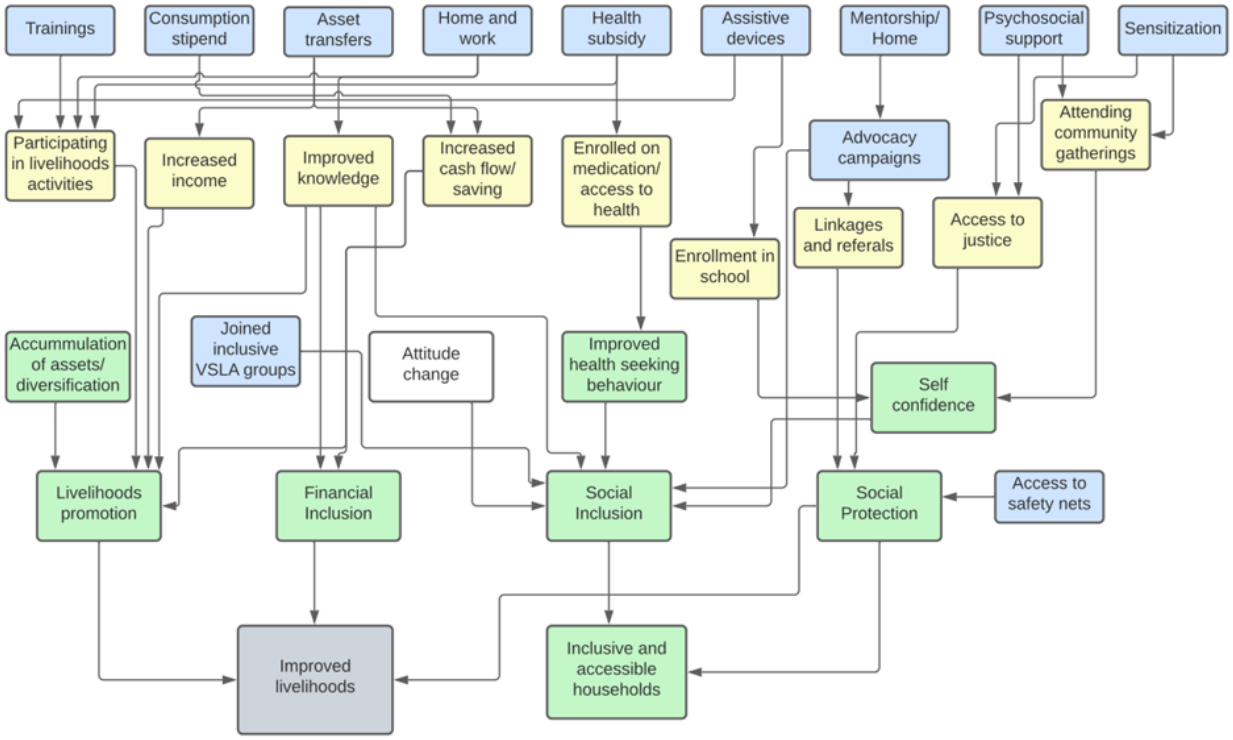


Figure 18: Candidate DIG DAG Uganda Group 2

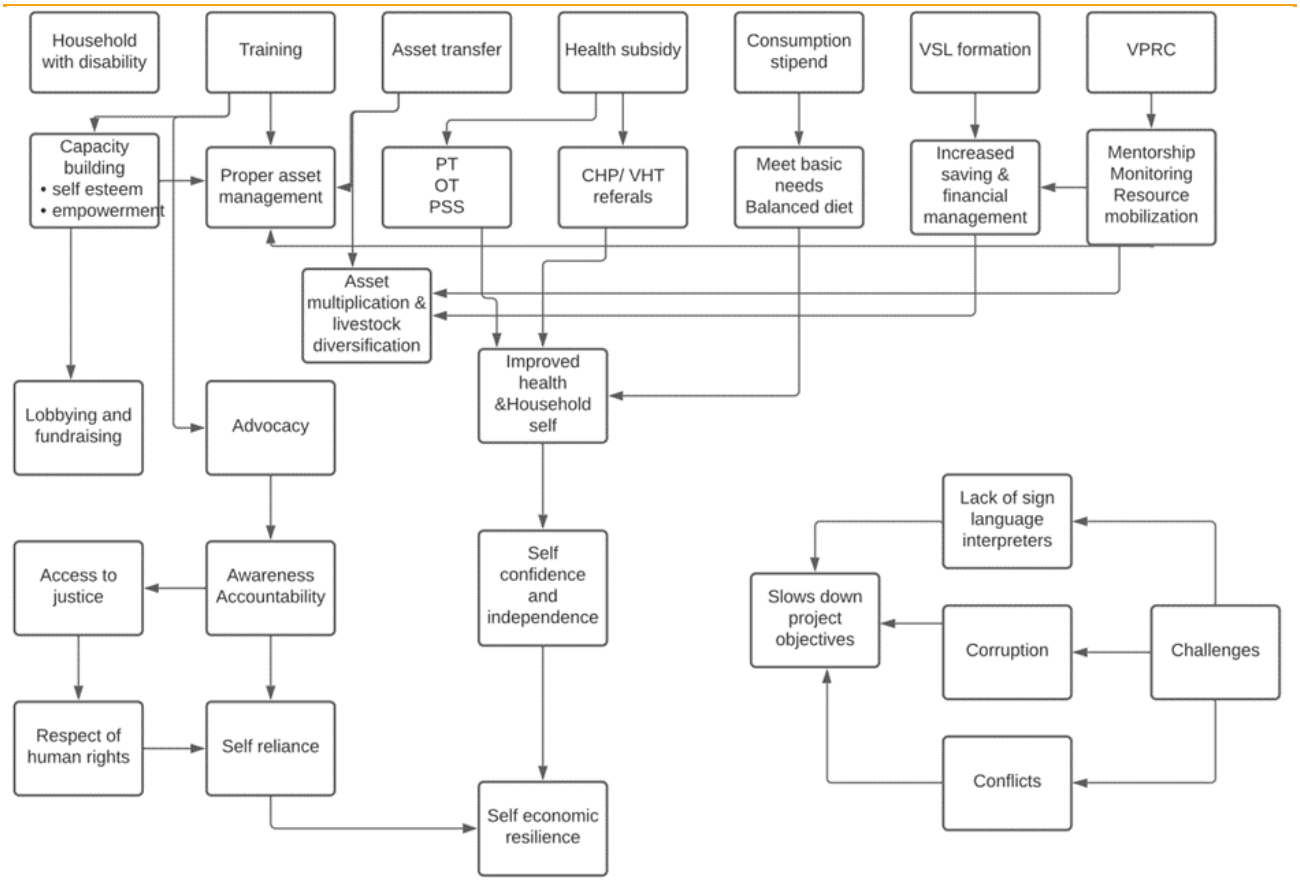


Figure 19: Candidate DIG DAG Uganda Group 3

The candidate DAGs which emerged from the three Ugandan group workshops were unfinished. They did not obey all of the rules of DAGs, with some unconnected nodes and some nodes defined so as to encompass more than one theoretical entity. It was also not clear that all of the edges in the DAG represented causal relationships rather than relationships of precedence in time or conceptual connection. This was due in part to the very complex nature of the DIG intervention meaning that time constraints were more binding than had been anticipated on the basis of the completed UPAVAN process.

Another cause of difficulty with the DIG intervention was the way in which the composite outcome of the intervention was conceptualised and presented in program documents. Looking at Figure 17, above, we can see that the overall outcome is defined as 'socio-economic empowerment and resilience.' Four nodes are depicted as causing this outcome, for example 'sustainable, accessible, and resilient livelihoods.' Depicting this as a causal relationship is a confusion. Rather, the four green-shaded nodes identified as 'pillars of the graduation model' are linked to the composite outcome conceptually rather than causally.

It was not possible to include the local implementation team members who took part in the Ugandan workshops in the full DAG and prior elicitation processes due to the difficulty of remaining in communication with those team members. Rather, once the three workshops had taken place, we reverted to the same selection strategy as for UPAVAN, and conducted the rest of the process with senior members of the implementation teams from BRAC, HI and NUWODU as well as with academics working on the evaluation. We conducted two cycles of

DAG elicitation workshops and email followup with this group using the DAGs issued from the Ugandan workshops as the stimulus for discussion. Over the course of the two workshops and the email discussion, the group of relevant experts for the DIG DAG and prior elicitation process was refined to six people. At the end of the first DAG workshop with this group, a candidate analysis DAG was agreed, which is presented below as Figure 20.

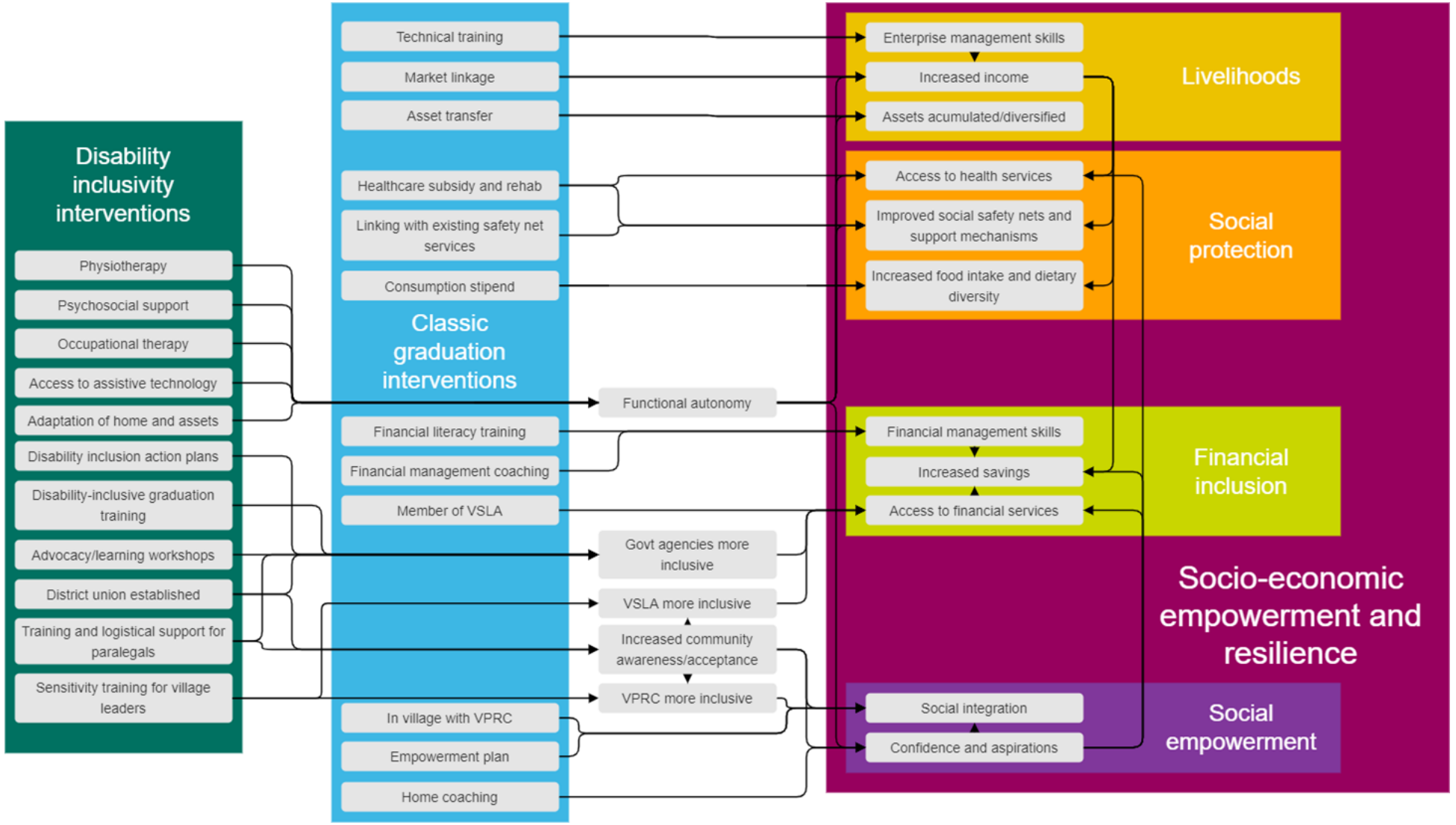


Figure 20: Candidate DIG analysis DAG post-first workshop

This DAG specified every intervention component identified by the documentation, by all three Ugandan workshops and began to explore the intermediate outcomes and mediators between these intervention components and the complete list of outcomes of interest for the intervention. The confusion between causal and conceptual relationships of the earlier DAGs was resolved by only using edges to depict causal relationships and by grouping nodes into conceptual groupings using shaded rectangles.

The DAG is composed of very many nodes and edges; too many for the full POInT approach to be practicable. In particular, specifying every relevant confounding factor and intermediate outcome would have been extremely demanding on the partner team members' time. Some reduction in scope was therefore necessary for an application of the POInT approach. In the second workshop a reduction in scope was agreed, to focus on the 'Increased income' outcome. Restricting the scope of the analysis to one key outcome reduced the total number of nodes to a level that might be manageable. However, the simple DAG of intervention components to outcome has too many parents for the 'Income' node, making the number of nodal types explode and meaning that the model is not calculable in CausalQueries.

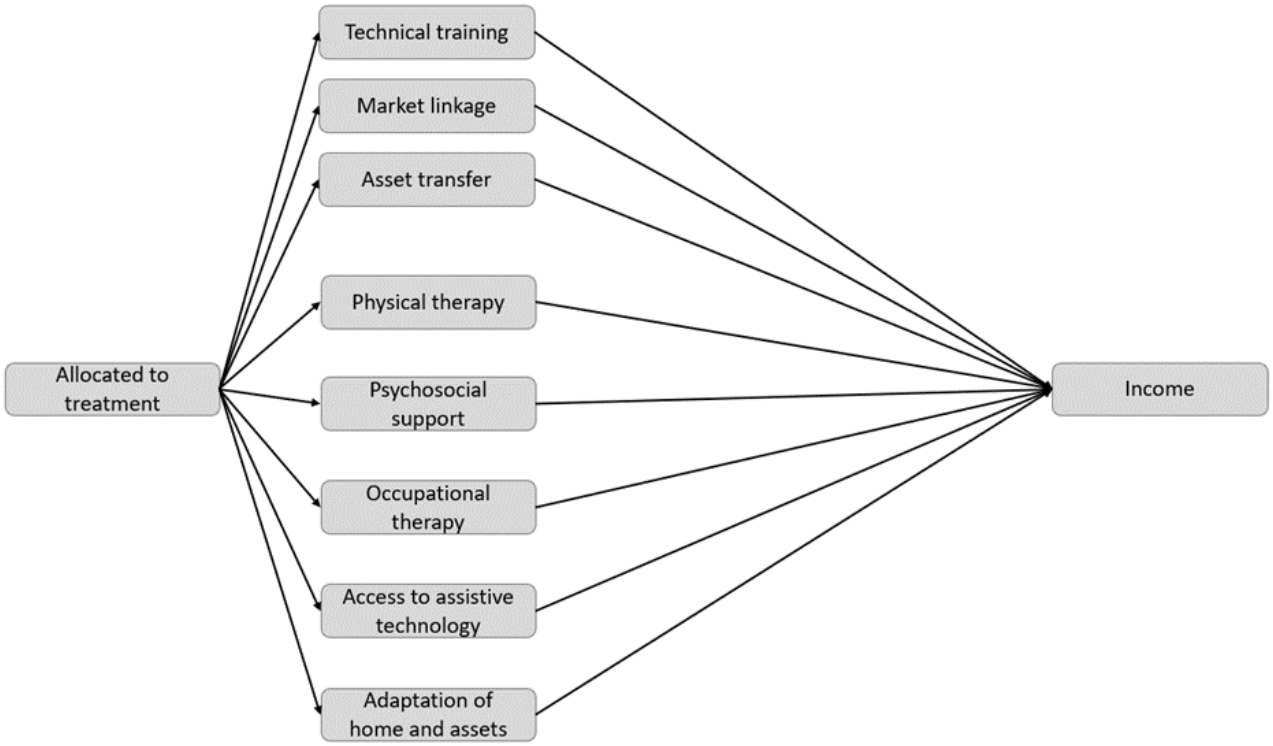


Figure 21: DIG intervention components causing changes in income

It was possible that including more mediators and intermediate outcomes would allow us to reduce the number of parents of 'Income.' Unfortunately, plausible intermediate outcomes and mediators are shared between all of the disability-specific interventions (the bottom five intervention nodes on Figure 22. For example, consider 'Depression'. This is an important hypothesised channel of intervention causation. However, it cannot reduce the number of parent nodes of 'Income' because all of the disability-specific interventions also have effects on income that are not mediated by depression.

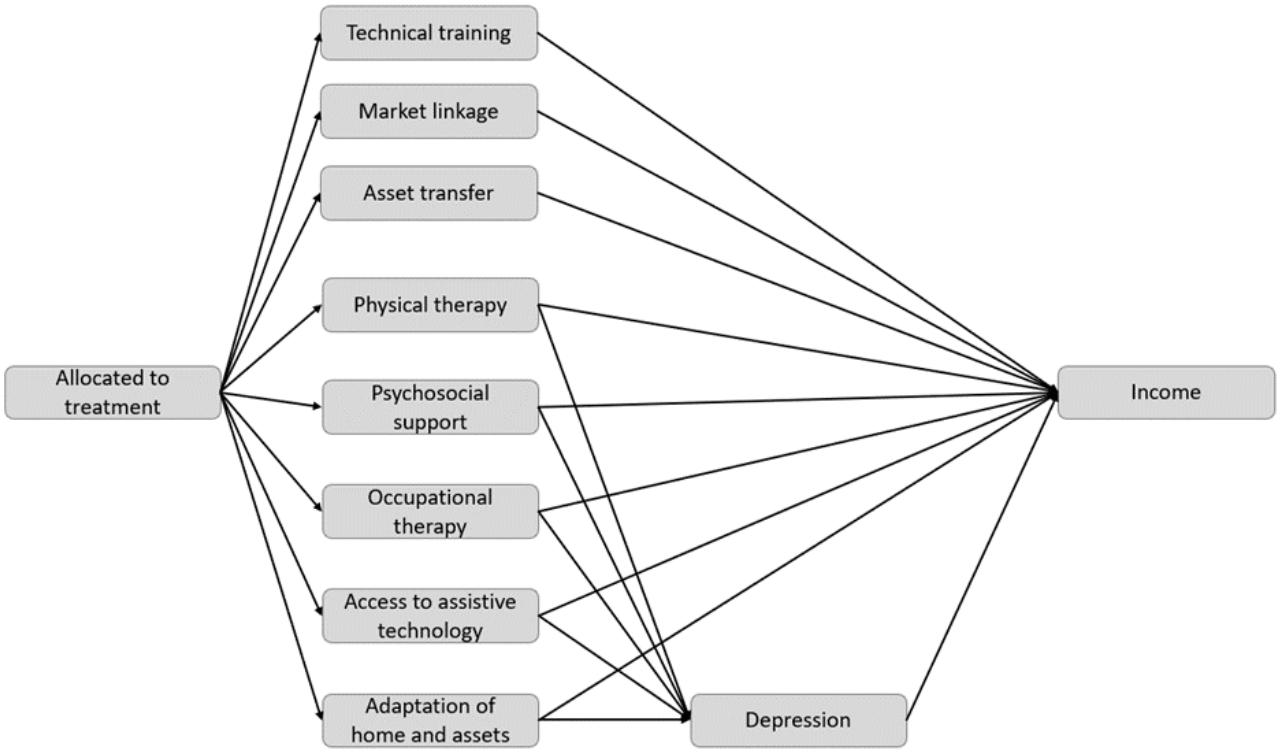


Figure 22: Candidate DIG analysis DAG with depression added

Interviews with members of the DIG team were not able to uncover a set of intermediate outcomes and mediators that could produce a DAG structure that did not contain uncomputably high levels of interaction, suggesting that the theory behind the intervention was irreducibly interactive, with many channels of causation from many intervention components to outcomes of interest. Therefore, we were not able to build a full DAG of intervention components, intermediate outcomes, mediators and moderators even for a reduced focus on one outcome of the DIG intervention. Rather, we were forced to consider what higher-order questions we might be able to ask about the aggregate causal effects of combinations of intervention components in the DIG intervention.

Further interviews with experts on the DIG intervention revealed that the questions that they were most interested in the POInT approach addressing centred on recipients of the intervention being 'actively engaged in economic activity' rather than reaching a certain income level. They considered it unlikely that at the time of the endline survey recipients with disabilities would have been able to achieve significantly increased income levels, but considered whether or not they were actively engaged in economic activity would be a powerful indicator of success in moving recipients with disabilities onto the 'pathway of graduation' from extreme poverty.

Furthermore, it became clear that due to budget cuts and problems of implementation due to the COVID-19 pandemic, the only data available to us within our timeframe would be the baseline surveys and a limited endline survey that had sampled all and only recipients and control individuals with disabilities, rather than also sampling the roughly 80% of eligible individuals who did not have disabilities. Therefore, our analysis DAG would have to focus

exclusively on eligible individuals with disabilities. That discussion took us to the candidate analysis DAG in Figure 23 below.

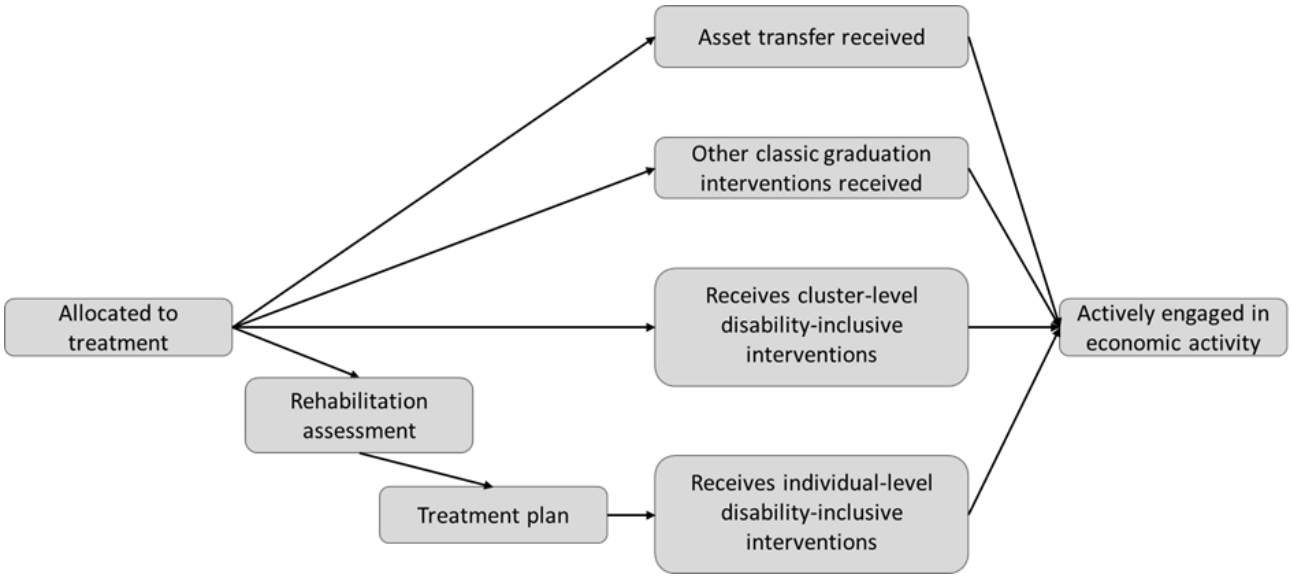


Figure 23: Candidate DIG analysis DAG for 'Actively engaged in economic activity' outcome

It further became clear that it was not going to be possible to link process data on rehab assessments and treatment plans with endline survey data so those nodes could not be assessed. Changing orientation and tidying up leads to the following final analysis DAG for the POInT approach to the DIG intervention presented in Figure 24.

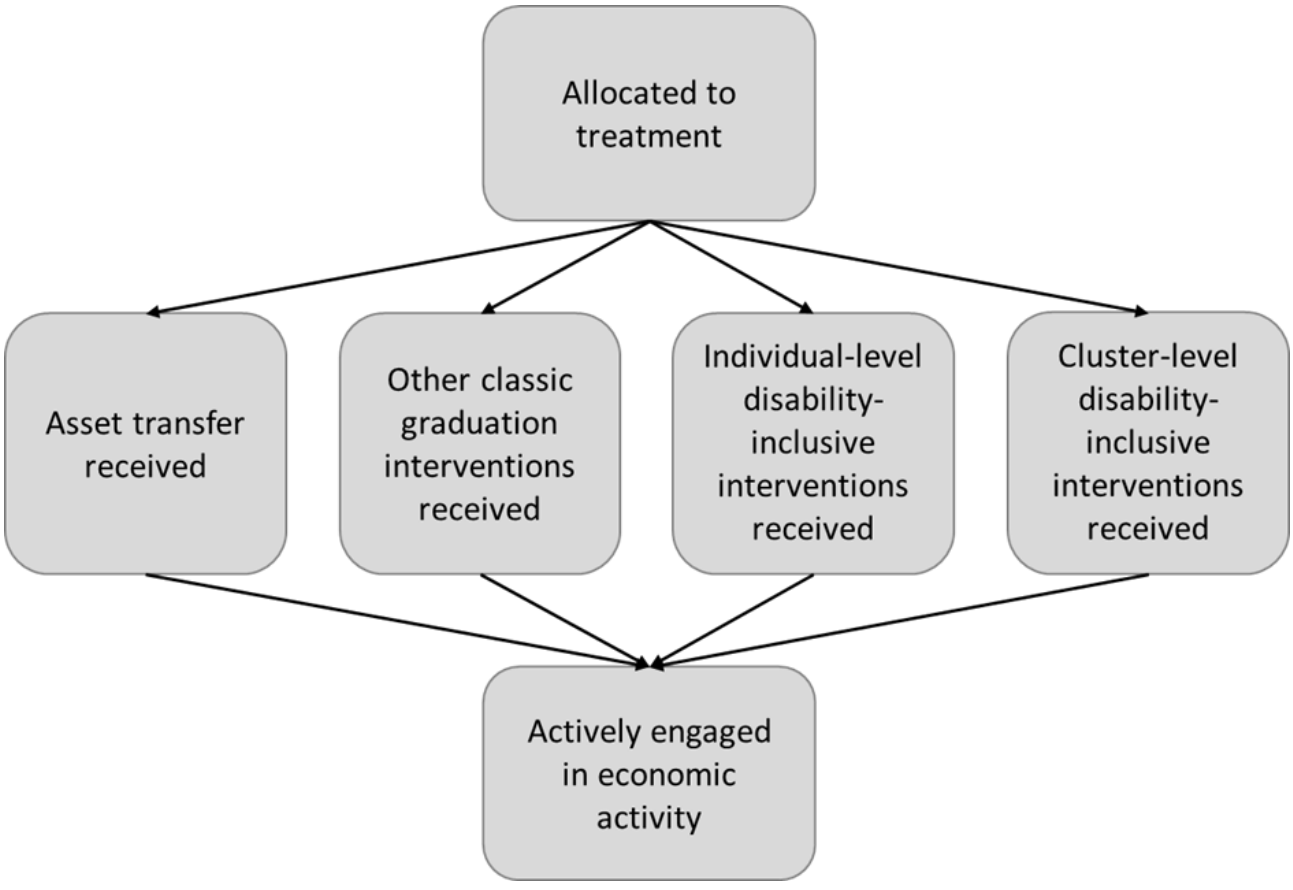


Figure 24: Final DIG analysis DAG

As discussed above, of necessity this analysis DAG is a much reduced version of the full ToC DAG. It is designed only to investigate a very limited and specific question, namely: 'for DIG recipients and control individuals with disabilities, what are the interactions between different intervention components of that person being actively engaged in economic activity?'

As for the UPAVAN intervention, the DAG-elicitation process for DIG also led to a set of definitions for nodes and of measures for nodes as presented in Table 8 in the Data section, above.

Part 3: Priors

Once the DAGs had been agreed with the partner teams, the next step in the process was to elicit the partner teams' prior beliefs about the details of the causal relationships represented in the DAG, as described in the methods section.

UPAVAN prior elicitation

The first step in eliciting experts' priors was to develop the contingency tables for all endogenous nodes in the DAG as well as a table of any exogenous nodes. These tables were created and formatted as described in the methods section to create the UPAVAN prior elicitation tool. A link to a read-only public copy of the tool is provided here.

Collecting priors for endogenous nodes

The UPAVAN DAG contained nine endogenous nodes, requiring the creation of nine contingency tables such as that depicted in Figure 25, below. The questions asked were specified at the level of the unit of the analysis, individual women. However, many nodes were defined at the household level, requiring questions to be framed as 'for 100 women from 100 households with the following characteristics…' as shown on the next page.

Node meaning reminder

Node	Description of theoretical entity	Possible values
Supportive household	Household is supportive of woman's empowerment	TRUE, FALSE
Allocated to AGRI	Household allocated to the AGRI arm	TRUE, FALSE
Marginalised group	Household members of a marginalised group in their community	TRUE, FALSE
Available land for cultivation	Household owns more land than a kitchen garden	TRUE, FALSE
Regular videos in groups	Woman in household regularly attends video screenings	TRUE, FALSE

Collecting your priors

For 100 women from 100 households with the following combinations of values for the parent nodes in the table below, how many of those women would you expect to regularly attend video screenings (for how many would 'Regular videos in groups' = TRUE)? For each answer, respond with the value you think most likely to be correct. Please also first select a level of certainty for this collection of answers, remembering that 'totally uncertain' would mean that you think all values are equally likely.

Certainty

Supportive household	Allocated to AGRI	Marginalised group	Available land for cultivation	Number 'Regular videos in groups' = TRUE
TRUE	TRUE	TRUE	TRUE	
TRUE	TRUE	TRUE	FALSE	
TRUE	TRUE	FALSE	TRUE	
TRUE	TRUE	FALSE	FALSE	
TRUE	FALSE	TRUE	TRUE	
TRUE	FALSE	TRUE	FALSE	
TRUE	FALSE	FALSE	TRUE	
TRUE	FALSE	FALSE	FALSE	
FALSE	TRUE	TRUE	TRUE	
FALSE	TRUE	TRUE	FALSE	
FALSE	TRUE	FALSE	TRUE	
FALSE	TRUE	FALSE	FALSE	
FALSE	FALSE	TRUE	TRUE	
FALSE	FALSE	TRUE	FALSE	
FALSE	FALSE	FALSE	TRUE	
FALSE	FALSE	FALSE	FALSE	

Figure 25: Contingency table, explanation and question used for eliciting priors over causal effects for one endogenous UPAVAN node

In addition to asking for the proportion of women for whom the child node would take the value 'true,' Figure 25 asks for the participant's level of certainty about their answers for this contingency table as described in the methods section.

Collecting priors for exogenous nodes

The UPAVAN DAG contained many exogenous nodes representing contextual factors of relevance to the effect of the intervention. In addition to collecting priors on causal effects in the model we collected priors on the level of exogenous nodes to create a prior model that was calibrated to the intervention context. To do this, we created a table of exogenous nodes and created a question about the proportion of households which would take the value 'TRUE' for each node as in Figure 26, below:

Node meaning reminder

Node	Description of theoretical entity	Possible values
Supportive household	Household is supportive of woman's empowerment	TRUE, FALSE
Marginalised group	Household members of a marginalised group in their community	TRUE, FALSE
Available land for cultivation	Household owns more land than a kitchen garden	TRUE, FALSE
Non-monsoon water	Household has access to a water supply that would operate outside the monsoon season such as a well	TRUE, FALSE
ST/SC	Household members of a scheduled tribe or caste	TRUE, FALSE

Collecting your priors

This section is a little different.

We want to know about your prior beliefs about the context of the evaluation. For that reason, below we ask you to estimate the proportion of households for whom each of the contextual nodes would hold true. So, thinking about 100 women from 100 households, for how many would you expect the household to be supportive of the woman's empowerment (for how many would 'Supportive household' = TRUE), for how many would you expect the household to be members of a marginalised group in their community (for how many would 'Marginalised group' = TRUE), etc.? Please also select a level of certainty for each of your answers.

Node	Number Node = TRUE	Certainty
Supportive household		
Marginalised group		
Available land for cultivation		
Non-monsoon water		
ST/SC		

Figure 26: Table of exogenous nodes, explanation and question used for eliciting priors over the values of exogenous nodes

Distribution of participants over the DAG

Our process allowed participants to provide answers quickly. Piloting using some of the suggested contingency tables and questions for UPAVAN within our team suggested that answering 56 sub-questions about 56 distributions (two to four full contingency tables for UPAVAN) was possible in just under 30 minutes for most participants. The UPAVAN participant exercise contained 208 sub-questions. Therefore, it was not possible to ask

participants to answer the full set of sub-questions in a reasonable time. We estimated that 45 minutes was the maximum length of session for which participants should be expected to remain engaged, giving a very rough sub-question 'budget' per participant of 90. In distributing participants over the full prior elicitation exercise we placed two constraints on ourselves. First, we wished to minimise confusion about the DAG by only asking participants about consecutive child-nodes in the DAG, following the flow of causation from 'top' (allocation) to 'bottom' (the final outcome). Second, given that it seemed that three groups of participants each answering roughly 90 sub-questions would be sufficient to cover the whole DAG with some overlap, we wished to use this overlap to ask different groups about some of the same nodes. This meant that the 9 participants for the UPAVAN project could be divided into three groups with each group answering sub-questions as follows - Table 10:

Table 10: Distribution of participants over the UPAVAN DAG for prior elicitation

Contingency tables	Number of sub-questions	Group 1	Group 2	Group 3
All exogenous nodes	6	✓	✓	✓
Child node 1	16	✓		
Child node 2	4	✓		
Child node 3	32	✓		
Child node 4	32	✓	✓	
Child node 5	16		✓	
Child node 6	32		✓	✓
Child node 7	32			✓
Child node 8	8			✓
Child node 9	36			✓
Total subquestions	208	90	86	114

Although this distribution of sub-questions is not equal, it is the most even distribution possible given the number of sub-questions for each node and the restriction of answering only on sets of sequential child nodes.

Response rate and participant experience

Despite attempting to simplify the task, and refining our tool in response to piloting, the UPAVAN prior elicitation exercise was difficult for participants to grasp. Of nine participants, one opted out of the prior elicitation exercise just before distribution of the prior elicitation tool, after having made limited contributions to the DAG elicitation process. This participant

judged at that point that their knowledge of intervention theory was not sufficient to make a useful contribution. Two participants began but did not complete the exercise, two required a lot of help, and four proceeded without much help. This represents quite a high level of attrition and suggests that this prior elicitation tool may only be suitable for more technical and/or more motivated participants. Although the participants who required a lot of help were the participants with the least training in statistics and causal reasoning, the participants who abandoned the effort entirely were among the more technically-trained. These participants did begin the exercise but abandoned it incomplete, suggesting that they found the exercise too demanding on their time. Most participants reflected that the exercise was difficult and confusing, despite our efforts to make it as clear as possible.

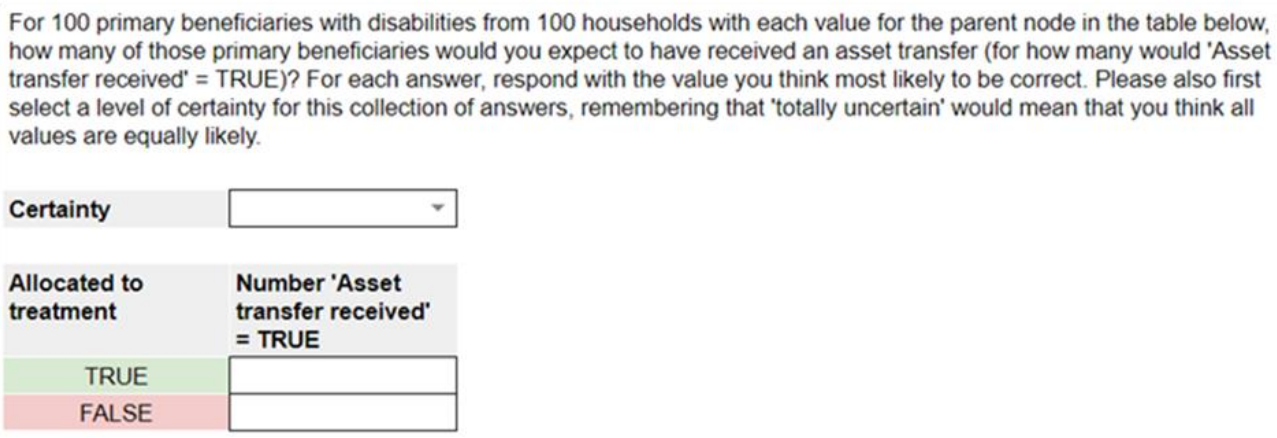
DIG prior elicitation

The prior elicitation tool for the DIG partnership was produced using the same process as for the UPAVAN partnership. As the DAG was much smaller, there were less contingency tables for participants to complete. All the contingency tables were for the five endogenous nodes in the DAG. It was not necessary to collect priors on the single exogenous node in the DAG as this was allocation to the intervention, whose probability of taking the values 1 or 0 is known with certainty as it is within the control of the intervention designers. A link to a read-only public copy of the tool is provided here.

Response rate and participant experience

Five of the six participants responded promptly and filled out the tool successfully within the first week. The sixth participant had not responded after one week and was chased but did not respond. However, this good response rate masks some evident confusion about the tool from participants.

All participants were contacted in the first week of prior elicitation to offer feedback on their answers and to give them a chance to edit those answers, in accordance with our analysis protocol. This was because we had noticed an anomalous response pattern that suggested that all the responding participants had not understood the first four contingency tables correctly. This confusion resulted from the fact that the first four contingency tables were about a child node that only had one parent and took the format below - Figure 27.


For 100 primary beneficiaries with disabilities from 100 households with each value for the parent node in the table below, how many of those primary beneficiaries would you expect to have received an asset transfer (for how many would 'Asset transfer received' = TRUE)? For each answer, respond with the value you think most likely to be correct. Please also first select a level of certainty for this collection of answers, remembering that 'totally uncertain' would mean that you think all values are equally likely.

Certainty	

Allocated to treatment	Number 'Asset transfer received' = TRUE
TRUE	
FALSE	

Figure 27: Format of first four contingency tables for DIG prior elicitation

At the time of the email to participants all four of the responding participants' answers to the first four of the first four questions suggest that they were dividing a pool of 100 households such that the two answers add up to 100. It is possible that these two answers add up to 100, but it is not necessary. Rather, the question is asking, "for 100 households who ARE allocated to treatment, how many will receive the asset transfer?" and "for 100 (different) households that ARE NOT allocated to treatment, how many will receive the asset transfer?" These questions are independent, rather than both referring to the same pool of 100 households. So, possible answers include "90, 10" but also "10, 0" and "70, 70". This information was provided to participants and they were given the opportunity to change their answers if they wished.

One participant responded to the exercise but uniformly entered point estimates of 100 for every answer with certainty levels of 'Very certain.' This participant was given the opportunity to have a quick call with one of our team to clarify the purpose of the tool and give them some feedback on how we interpreted their results. On the call we walked through the training exercise, step-by-step with this participant. They were then confident they understood the exercise well and sent in new responses which were included in our analysis.

Assessing violations of model assumptions

UPAVAN prior assessment

Checking for homogeneity of experts priors

Before combining the priors of different experts for the same partner intervention, the homogeneity of experts' priors was assessed. To assess the level of homogeneity of priors of different experts on the same child nodes, proportions of 1's for each person in each child node were plotted using a composite graph with size of dot representing the level of certainty with which the expert answered as translated to a number of observations that would merit a belief with that level of certainty. The results for UPAVAN priors are presented in Figure 28.

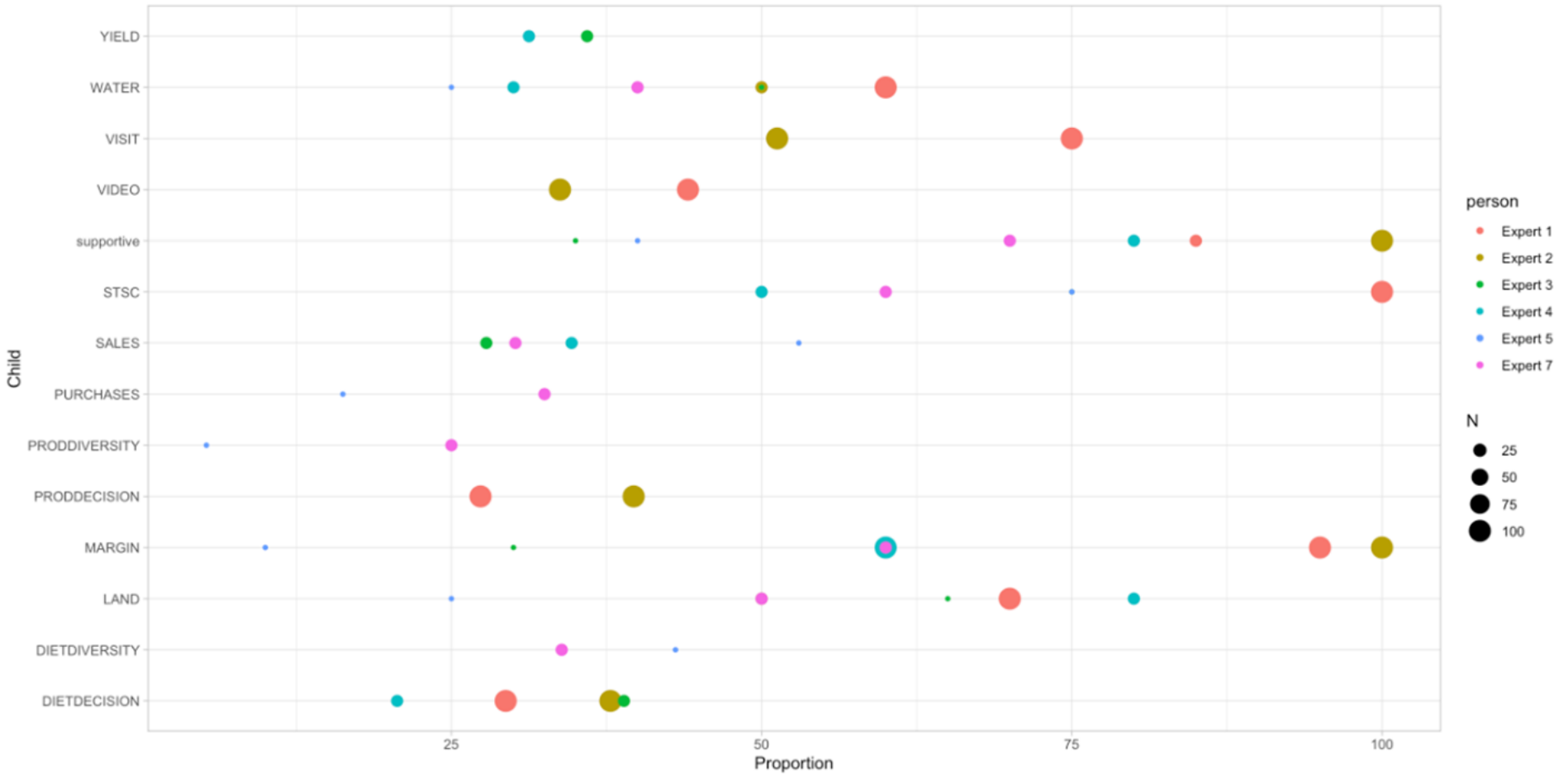


Figure 28: Proportions of 1s for each expert for each node for UPAVAN

The priors reported by the different experts were evidently different for the different child nodes and the certainty levels differed too. Nevertheless, different experts' answers are correlated with each other.

Checking for monotonicity

The consistency of the aggregated priors with the assumption of monotonicity of effects was checked using a three-step approach as described in the methods section. Results for the UPAVAN team's priors are presented in Table 11. In the majority of cases, the monotonicity assumption was upheld. Only the effect of STSC on dietary decisions, and the effects of the AGRI intervention and available water on sales were dependent on the values of other variables in the model, i.e. there was evidence of non-monotonicity.

Table 11: Consistency of priors with monotonicity assumption for endogenous nodes in the UPAVAN analysis DAG

Child	Parent	Monotonic
DIETDECISION	AGRI	TRUE
	VIDEO	TRUE
	VISIT	TRUE
	STSC	FALSE
DIETDIVERSITY	YIELD	TRUE
	PRODDIVERSITY	TRUE
	PURCHASES	TRUE
	DIETDECISION	TRUE
PRODDECISION	AGRI	TRUE
	VIDEO	TRUE
	VISIT	TRUE
	STSC	TRUE
PRODDIVERSITY	AGRI	TRUE
	LAND	TRUE
	WATER	TRUE
	PRODDECISION	TRUE

Child	Parent	Monotonic
PURCHASES	SALES	TRUE
	DIETDECISION	TRUE
SALES	AGRI	FALSE
	LAND	TRUE
	WATER	FALSE
	PRODDECISION	TRUE
	YIELD	TRUE
VIDEO	AGRI	TRUE
	MARGIN	TRUE
	LAND	TRUE
VISIT	AGRI	TRUE
	MARGIN	TRUE
YIELD	AGRI	TRUE
	LAND	TRUE
	WATER	TRUE
	PRODDECISION	TRUE

These findings contrast with experts' explicit priors on monotonicity collected during the DAG elicitation process. During that exercise experts agreed that all causation in the DAG could be assumed to be monotonic.

Checking for interactions

The existence of interaction effects for the causation of each endogenous node in the UPAVAN DAG was checked as described in the methods section. These results were compared to the team's priors about interaction effects as stated during the DAG elicitation process. Results for the UPAVAN team's priors are presented in Table 12.

Table 12: Likelihood Ratio test (Full vs Reduced) UPAVAN

	p-value
DIETDECISION	0.0000031
DIETDIVERSITY	0.8984347
PRODDECISION	0.0384034
PRODDIVERSITY	0.9849163
PURCHASES	0.6180769
SALES	0.0713830
VIDEO	0.0049022
VISIT	0.1199301
YIELD	0.9993057

The interaction effects in the logistic regression models for maternal dietary diversity, produce dietary diversity, purchases, produce sales, regular home visits and high produce yield nodes are not significant. However, interactions were significant for influences dietary decisions, influences food production decisions and regular videos. This indicates that in the priors elicited from the participants there is evidence that they are assuming some interaction between parents of these nodes. These findings contrast with participants' explicit priors about interactions collected during the DAG elicitation process, as can be seen by comparing with Figure 15.

For the purposes of updating and describing the UPAVAN prior model, the experts' explicit priors as collected during the DAG elicitation process are treated as primary. Monotonicity assumptions are made throughout the model and interactions are permitted only at the locations that experts explicitly stated they were needed. The conflict between explicit and implicit priors will be explored in the work of the next paper by relaxing assumptions in the DAG and testing for interactions in the prior model, the posterior model updated on implicit priors and trial data, and a model with flat priors updated only on trial data.

DIG prior assessment

Checking for homogeneity of experts priors

To test for homogeneity (internal consistency) of priors of different experts on the same child nodes, proportions of 1's for each person in each child node were plotted using a composite graph for DIG. The results are presented in Figure 29.

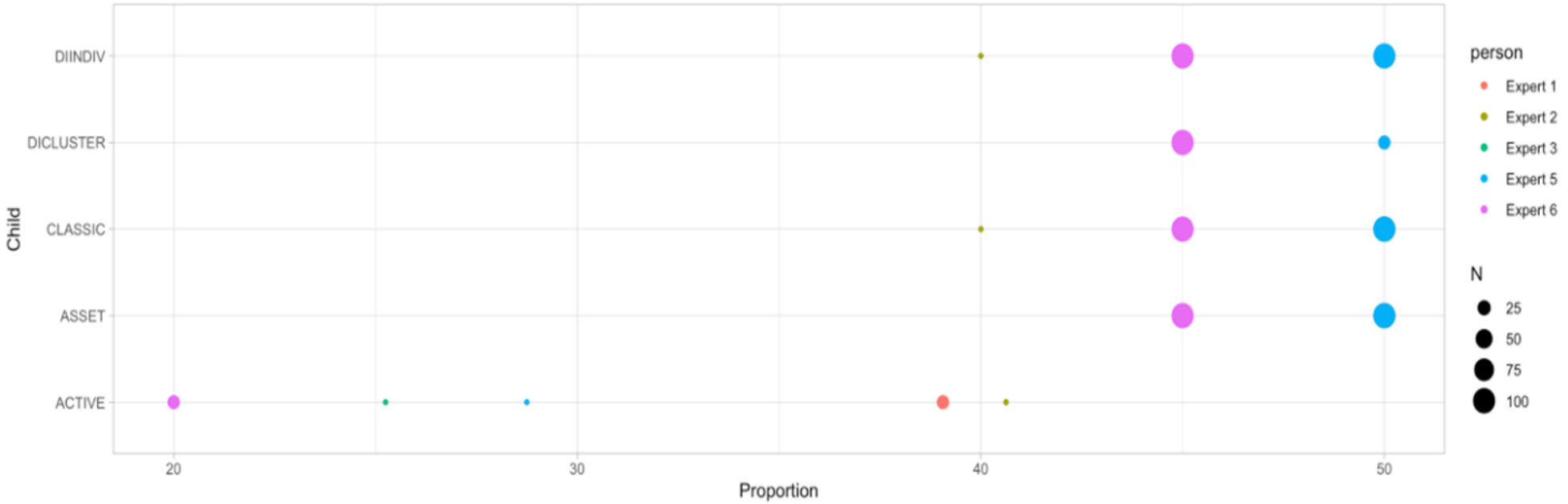


Figure 29: Proportions of 1s for each expert for each node for DIG

The priors reported by the different experts were evidently different for the different child nodes and the certainty levels differed also. However, as with the UPAVAN team, experts' priors were correlated, suggesting some level of agreement. Priors were least similar on causal effects for the final node in the DAG, though these priors were also least certain.

Checking for monotonicity

The consistency of the aggregated priors with the assumption of monotonicity of effects was checked using a three-step approach as described in the methods section. Results for the DIG team's priors are presented in Table 13. In this case there was no evidence of violation of the monotonicity assumption.

Table 13: Consistency of priors with monotonicity assumption for endogenous nodes in the DIG analysis DAG

Child	Parent	Monotonic
ACTIVE	ASSET	TRUE
ACTIVE	CLASSIC	TRUE
ACTIVE	DIINDIV	TRUE
ACTIVE	DICLUSTER	TRUE

Checking for interactions

The existence of interaction effects for the causation of each endogenous node in the DIG DAG was also checked as described in the methods section. These results were compared to the team's priors about interaction effects as stated during the DAG elicitation process. Results for the DIG team's priors are presented in Table 14. Surprisingly, the interaction effects in the logistic regression models were not significant and therefore, a model without interactions would suffice for the DIG priors. This is despite the fact that the DIG analysis DAG's focus was suggested to investigate the interactive effects that are key to the explicit theory of change behind the DIG intervention. It will be very informative to see whether the data from the trial agrees with the explicit rationale for the intervention or with experts' implicit priors as collected by the prior elicitation exercise.

Table 14: Likelihood Ratio test (Full vs Reduced) DIG

	p-value
ACTIVE	0.5833363

Updating and describing the prior models

Both the DIG and UPAVAN models were successfully created in CausalQueries as naive models with the correct structure and restrictions, but flat, uninformative priors and updated on the priors data generated by the prior elicitation process. This process was completed using the augmented CausalQueries code designed to update independent submodels and then stitch them back together into a more complex model that was described in the methods section.

Updating the models was completed using a consumer-grade laptop in a matter of minutes, and did not require recourse to a server, as we had feared it might. This was in part due to the improvements we made to the underlying computation behind CausalQueries, moving several functions from R to C++ and implementing an approach to computing that allowed more processes to be completed in parallel to speed up performance.

We updated the naive models on two different sets of priors data to create two different prior models. The different sets of priors data were aggregated in two different ways as described in the methods section. 'Prior type 1' on the following charts refers to 'exclusive priors' aggregated on the assumption that experts' beliefs are independent of one another, based on entirely different observations for each expert. 'Prior type 2' refers to 'inclusive priors' aggregated on the assumption that experts' beliefs are based on somewhat overlapping samples from one shared set of observations and therefore are not independent of each other.

Figure 30 shows the distribution of nodal types in the UPAVAN model that has been updated by the elicited priors. The peak of the distribution shows approximately where the mean or median of the proportion of participants expected to be of a particular nodal type. The spread of the distributions indicates the uncertainty about this value. As expected, prior type 1 led to more confident model predictions, with higher peaks to the distributions of priors aggregated using this first method. However, the two methods led to similar point estimates for most nodes. The same is true for the DIG model displayed in Figure 31, further down.

Figure 30: Posteriors for nodal types in the full UPAVAN prior model

Note that the x-axis does not extend from 0 to 1 but from 0 to 0.6. Prior types 1 and 2 are shown superimposed.

To download the full-size image click here.

Figure 31 shows the distribution of nodal types for the DIG model. The top-left graph shows, for example, that the experts expected that around 25% of participants would remain inactive regardless of being exposed to any or all aspects of the intervention.

Figure 31: Posteriors for nodal types in the full DIG prior model

Note that the x-axis extends from 0 to 1 in this figure. Prior types 1 and 2 are shown superimposed. To download the full-size image click here.

On the top right in Figure 31, we have the distribution for the nodal type for the ACTIVE child '0000111100001111', which the experts were relatively certain was a very uncommon type. What does this mean? The following tables can be used to help interpret the nomenclature. The first four 0s indicate that ACTIVE will take the value of 0 whenever DICLUSTER and DIINDIV are both zero, i.e. both of the disability-specific components - Table 15.

Table 15: ACTIVE will take the value of 0 whenever DICLUSTER and DIINDIV are both zero

Value that ACTIVE takes:	ASSET	CLASSIC	DICLUSTER	DIINDIV
[*]***************	0	0	0	0
*[*]**************	1	0	0	0
[*]************	0	1	0	0
[*]**********	1	1	0	0

The following four 1s indicate that ACTIVE will be 1 when DICLUSTER is 1, i.e. the cluster-level disability intervention components, regardless of the other node values - Table 16.

Table 16: ACTIVE will take the value 1 when DICLUSTER is 1

Value that ACTIVE takes:	ASSET	CLASSIC	DICLUSTER	DIINDIV
****[*]***********	0	0	1	0
*****[*]**********	1	0	1	0
******[*]*********	0	1	1	0
*******[*]********	1	1	1	0

The following four 0s indicate that ACTIVE will be 0 when the DICLUSTER component is absent but the DIINDIV is present - Table 17.

Table 17: ACTIVE will be 0 when the DICLUSTER component is absent but the DIINDIV is present

Value that ACTIVE takes:	ASSET	CLASSIC	DICLUSTER	DIINDIV
********[*]*******	0	0	0	1
*********[*]******	1	0	0	1
**********[*]*****	0	1	0	1
***********[*]****	1	1	0	1

And finally, the final four 1s indicate that ACTIVE will take the value of 1 when the DICLUSTER node is 1 as well as the DIINDIV node, regardless of the other values - Table 18.

Table 18: ACTIVE will take the value of 1 when the DICLUSTER node is 1 as well as the DIINDIV node, regardless of the other values

Value that ACTIVE takes:	ASSET	CLASSIC	DICLUSTER	DIINDIV
**********[*]***	0	0	1	1
**********[*]**	1	0	1	1
**********[*]*	0	1	1	1
**********[*]	1	1	1	1

In summary, for someone with this nodal type for ACTIVE they respond to the DICLUSTER node only, and not to the DIINDIV node, even in the presence of the ASSET and CLASSIC components of the intervention. This is counterintuitive, as is shown in the confident small proportion expected in the priors.

Part 4: Updating on trial data

Once the models have been elicited, encoded and updated on data representing the elicited priors, the next step in the POInT approach is to update the models also on trial data. This section describes the process of preparing the trial data for updating and then updating the models using CausalQueries.

Coarsening the ternary nodes to binary

As discussed in the Empirical Methods subsection of the Methods section, the work reported in Annex A to extend CausalQueries to handle non-binary data was not complete in time for the analysis of the partner trials. Therefore, we were forced to coarsen non-binary measures to binary for this analysis. This was done so as to minimise information loss, by coarsening so as to combine the least-observed outside value (“0” or “2”) with the central value (“1”), as in Table 19.

Table 19: Coarsening Ternary Data Generated for Ternary Nodes to Binary Data

	PRODDIVERSITY	PURCHASES
1	261	1
2	532	181
3	1304	1915

For both variables, the least common outer value is “0” and so this is combined with “1” when coarsening to binary. This coarsening does not change the headline correlation between allocation to the AGRI arm and maternal dietary diversity (correlation 0.029).

Imputation of missing data

We could not have any missing data in the data used to update the models, as discussed in the Empirical Methods subsection of the Methodology section. This is because we were forced by the nature of the priors data to use the ‘update_stitch’ command to independently update all of the conditionally independent sub-DAGs in the model before stitching them together.

As discussed in the same subsection as above, the imputation method chosen to impute any missing values in the observed data was multiple data imputation by chained equations (MICE). The three fundamental steps of this method were performed, namely: classification of missing data, data imputation, and data accuracy diagnostics of imputed data.

Classification of missing data

Variables were identified, characterised and the level of missingness determined. Key determinants for the characterisation process were: variable name, data structure, number of missing observations, percentage of missing cases of the variable, variable type and proposed imputation predictive model. Variables were classified as either numeric, binary, unordered categorical or ordered categorical variables. Proposed imputation methods included the predictive mean matching, normal, logistic regression, polytomous logistic regression and proportional odds depending on the classification and data structure of the respective variable. Summaries of data classification for both UPAVAN and DIG trial data are presented in Table 20 and Table 21 respectively.

Table 20: Classification of missing data for UPAVAN trial

No.	Variable Name	Description of Variable	Data Structure	Missing Observations	% of Missing Observations	Variable Class	Proposed Imputation Method
1	PRODDIVERSITY	Produce Diversity	0, 1, 2	258	12.3	Numeric	Predictive mean matching
2	YIELD	High Produce Yield	0, 1	1	0.01	Binary	logistic regression
3	SALES	Produce Sales	0, 1	1	0.01	Binary	logistic regression
4	LAND	Available Land For Cultivation	0, 1	8	0.4	Binary	logistic regression
5	WATER	Non-monsoon Water	0, 1	266	12.7	Binary	logistic regression
6	STSC	Schedule Tribe/Caste	0, 1	1	0.01	Binary	logistic regression
7	MARGIN	Marginalised Group	0, 1	1	0.01	Binary	logistic regression
8	PURCHASES	Food Purchase Diversity	0. 1, 2	1052	50.2	Numeric	Predictive mean matching
9	VIDEO	Regular Videos in Group	0, 1	0	0	Binary	logistic regression
10	PRODDECISION	Influence Food Production Decisions	0, 1	287	13.7	Binary	logistic regression
11	DIETDECISION	Influence Dietary Decisions	0, 1	96	4.6	Binary	logistic regression
12	DIETDIVERSITY	Maternal Dietary Diversity	0, 1	0	0	Binary	logistic regression
13	VISIT	Regular Home Visits	0, 1	29	1.4	Binary	logistic regression
14	AGRI	Allocated to AGRI Arm	0, 1	0	0	Binary	logistic regression

Table 21: Classification of missing data for DIG trial

No.	Variable Name	Description of Variable	Data Structure	Missing Observations	% of Missing Observations	Variable Class	Proposed Imputation Method
1	ALLOC	Allocated to treatment	0, 1	0	0	Binary	Not applicable
2	ASSET	Asset transfer received	0, 1	0	0	Binary	Not applicable
3	CLASSIC	Other classic graduation interventions received	0, 1	0	0	Binary	Not applicable
4	DIINDIV	Individual-level disability-inclusive interventions received	0, 1	0	0	Binary	Not applicable
5	ACTIVE	Actively engaged in economic activity	0, 1	8	0.72	Numeric	Predictive mean matching

The UPAVAN data had 2,097 cases. The variable PURCHASES had the highest proportion of missing cases - 50.2%, followed by PRODDECISION - 13.7%, WATER - 12.7%, PRODDIVERSITY - 12.3% and DIETDECISION - 4.6%. The levels of missingness in the observed data is presented in Figure 32 as a histogram of missing data and the pattern representing combinations of missing/observed values and variables.

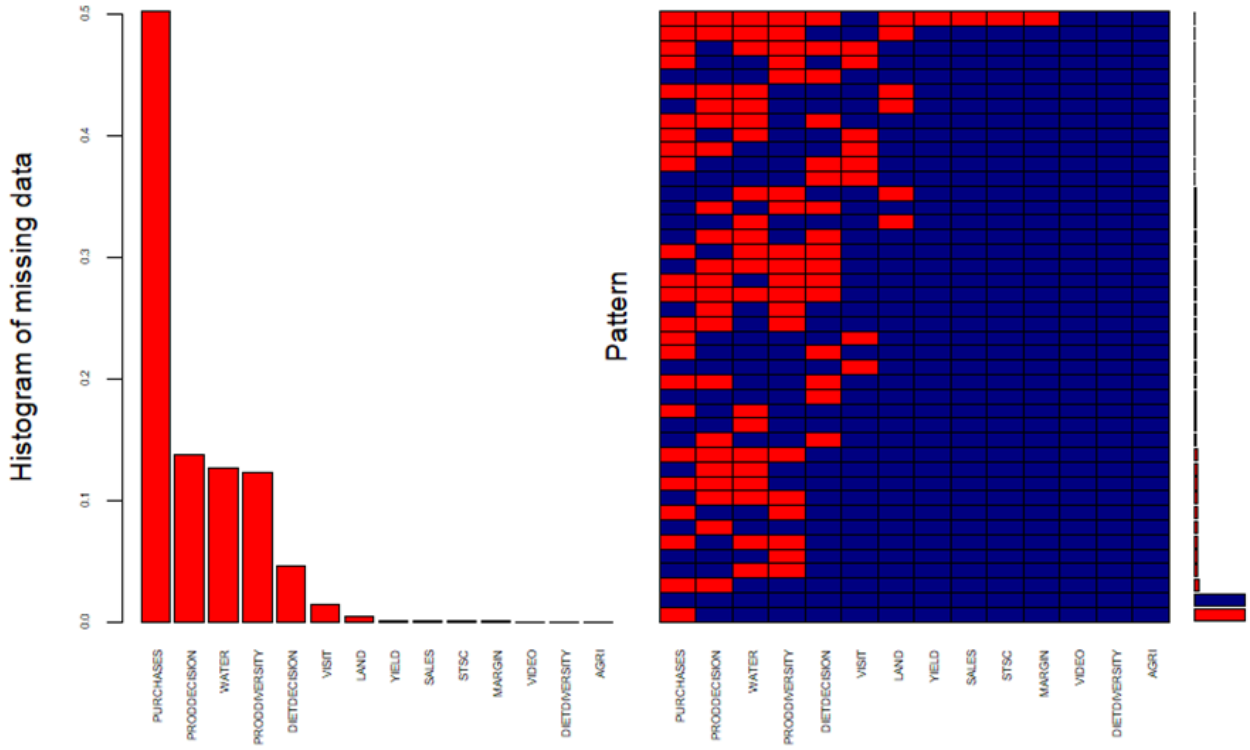


Figure 32: Histogram of missing data for UPAVAN trial

The DIG trial data had 1,108 observations. The variable ACTIVE was the only one that had missing cases - 0.72%. Notably, the DICLUSTER was completely missing as it was not measured in time to be included in the analysis. The levels of missingness in the observed data is presented in Figure 33 as a histogram of missing data and the pattern representing combinations of missing/observed values and variables.

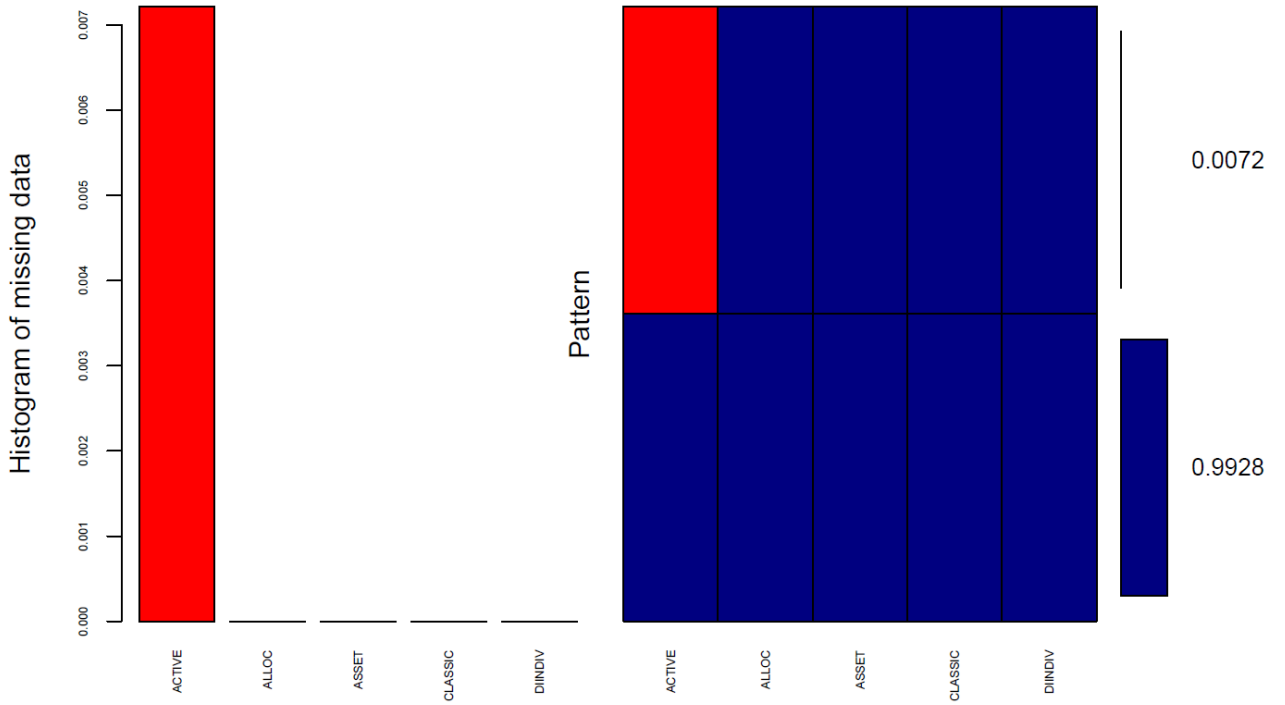


Figure 33: Histogram of missing data for DIG trial

Data Imputation

Five imputation iterations were applied using different predictive models for each variable. When all the data accuracy checks were performed and satisfied, a file of a complete dataset was retained and used for updating the model.

Checking for Data Accuracy

In general, a good imputed value is a value that could have been observed had it not been missing. Two powerful graphical functions were used to make diagnostic plots for checking the imputations namely: density plots and strip plots for both the observed and imputed data. Below, we present both of these checks for the UPAVAN and DIG datasets.

Density plots in Figure 34 and Figure 35 for both UPAVAN and DIG respectively, resemble a good fit for all the variables considered with missing data proportions of more than 0.35%. The density plots, as expected for the observed (blue) and imputed (red) data have a similar structure. The matching shape is a good indicator that the imputed values are indeed plausible values. Huge differences in the densities between the observed and the imputed values would have suggested a problem that needs to be further checked. Such cases would resemble an important scenario of having “pregnant fathers” in imputed variables.

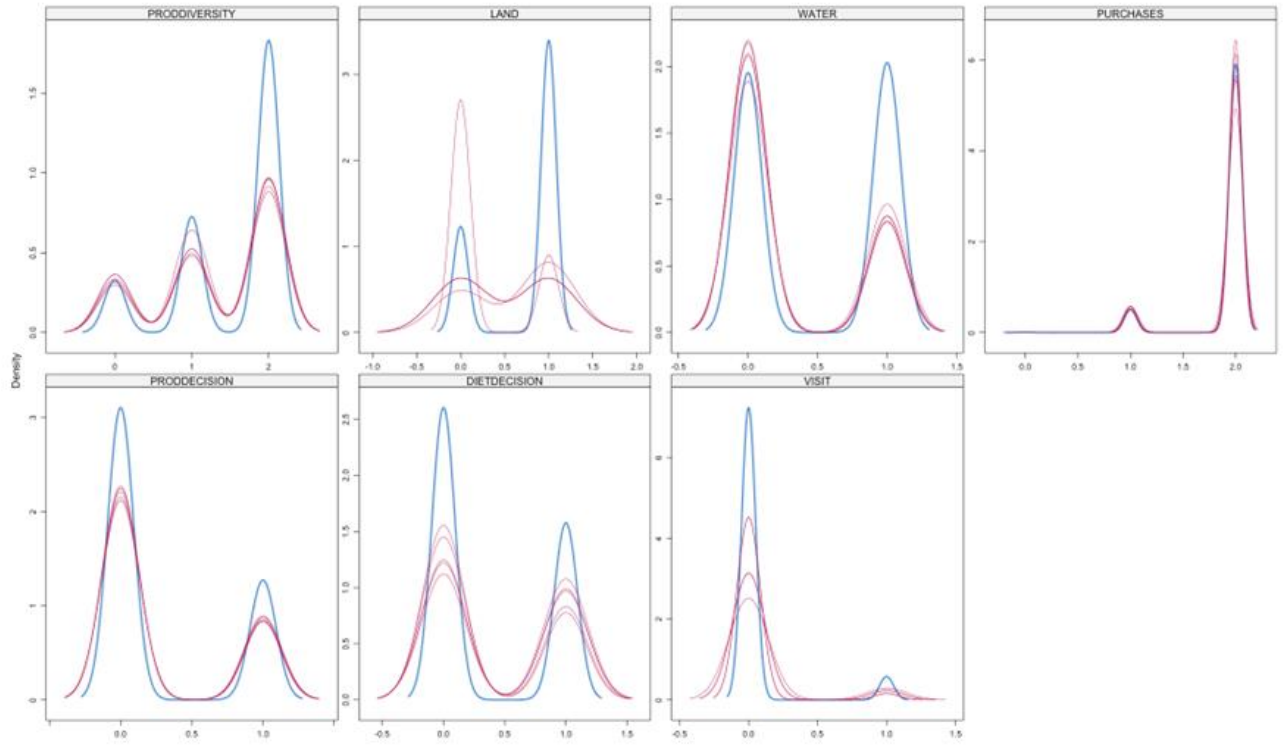


Figure 34: Density plots of imputed data overlaid on observed data for UPAVAN trial

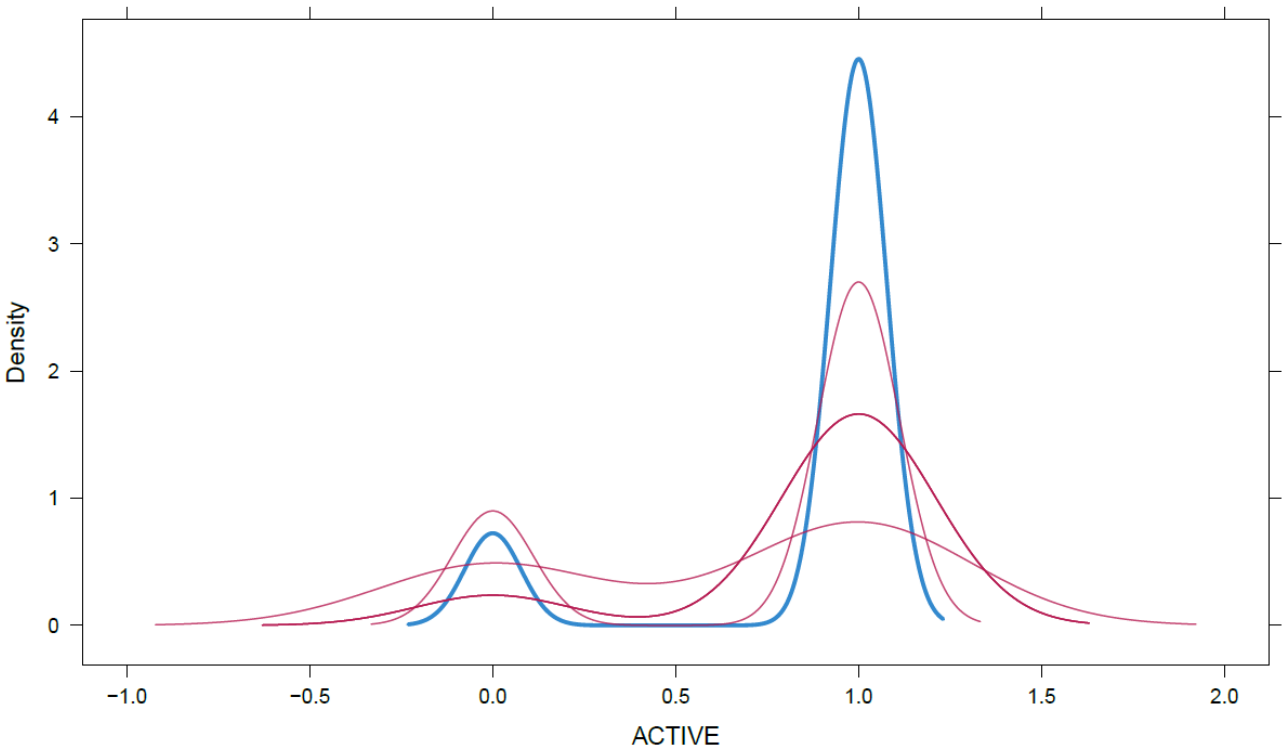


Figure 35: Density plots of imputed data overlaid on observed data for DIG trial

Strip plots present distributions of each of the variables as individual points of values for each imputation iteration. The data points of observed data (blue) and imputed data (red) in Figure 36 - UPAVAN and Figure 37 -DIG, below, are clustered together as an indicator for a good fit. It

is key to observe that the imputed values (red) assume values within the same range of the observed data (blue). It is expected that, where values are not imputed, the plot of the individual points in each variable will only be blue in colour. Examples of such variables included VIDEO, DIETDIVERSITY and AGRI for the UPAVAN trial data - Figure 36 and ALLOC, ASSET, CLASSIC and DIINDIV for the DIG trial data - Figure 37.

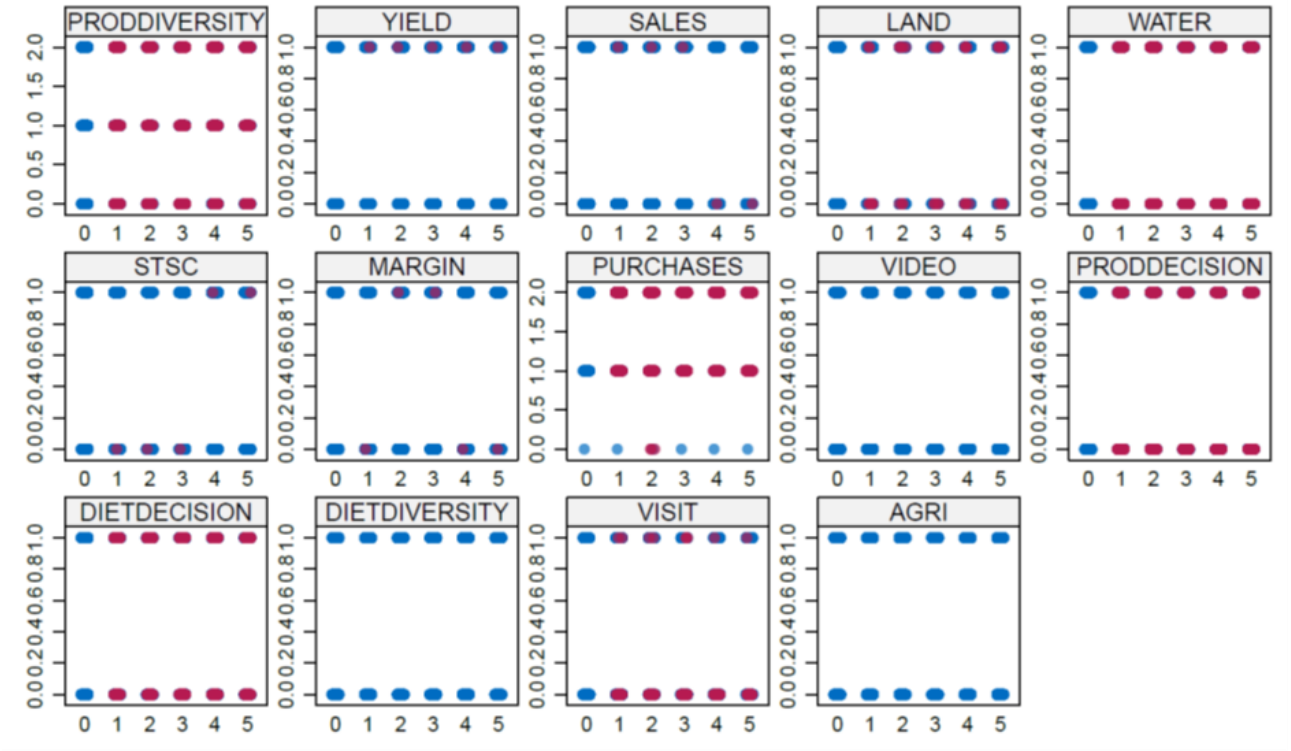


Figure 36: Strip plots of imputed and observed data for UPAVAN trial data

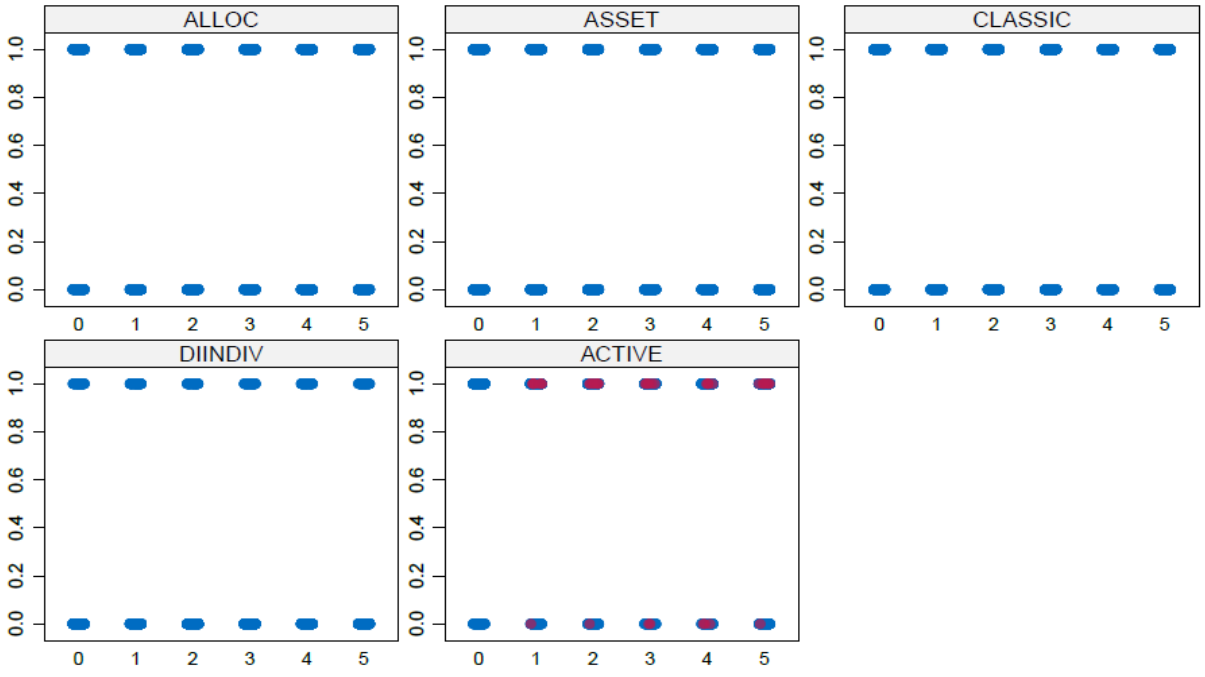


Figure 37: Strip plots of imputed and observed data for DIG trial data

Updating the model on the data

To update the whole model we use our stitching function to update all of the conditionally independent sub-DAGs in the whole DAG and then stitch those models back together, preserving the posteriors. Before doing this, it is worth taking some time to show what is happening 'under the hood' for the sub-DAGs. We do this in the following subsections using the more complex UPAVAN model to illustrate the process that was completed. Throughout, we illustrate by updating only on the trial data post-imputation.

Sub-DAGs

The process of splitting the whole DAG into sub-DAGs necessitates selecting child nodes in the DAG and including all their parents and all and only the edges between parent and child in a sub-model based on that child. For example, the first child node in the DAG is "Regular videos in groups", coded as VIDEO. The sub-DAG based on this node must include all parents of the node, excluding "Supportive household" (which was missing / not collected), and so looks like this - Figure 38:

Figure 38: sub-DAG ending in "Regular videos in groups"

The next child node in the full DAG is "Regular home visits," coded as VISIT. Therefore, the next sub-DAG to be created includes that node and all of the nodes in the full DAG with edges ending at VISIT as presented in Figure 39.

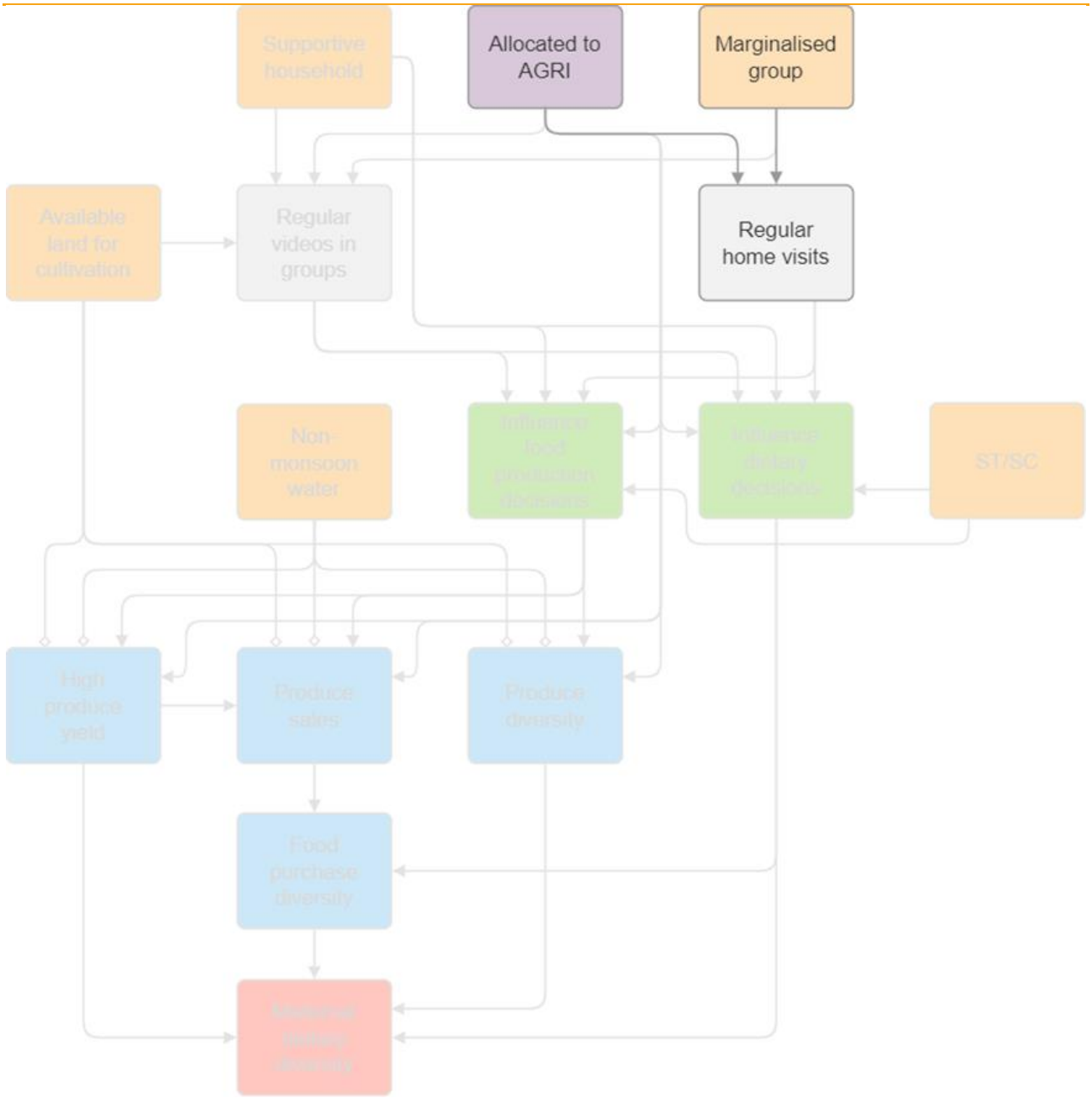


Figure 39: sub-DAG ending in "Regular home visits"

This process is repeated for every child node in the DAG to create the full set of conditionally-independent sub-DAGs, ending with the sub-DAG for the final outcome, "Maternal dietary diversity," coded as DIETDIVERSITY - Figure 40.

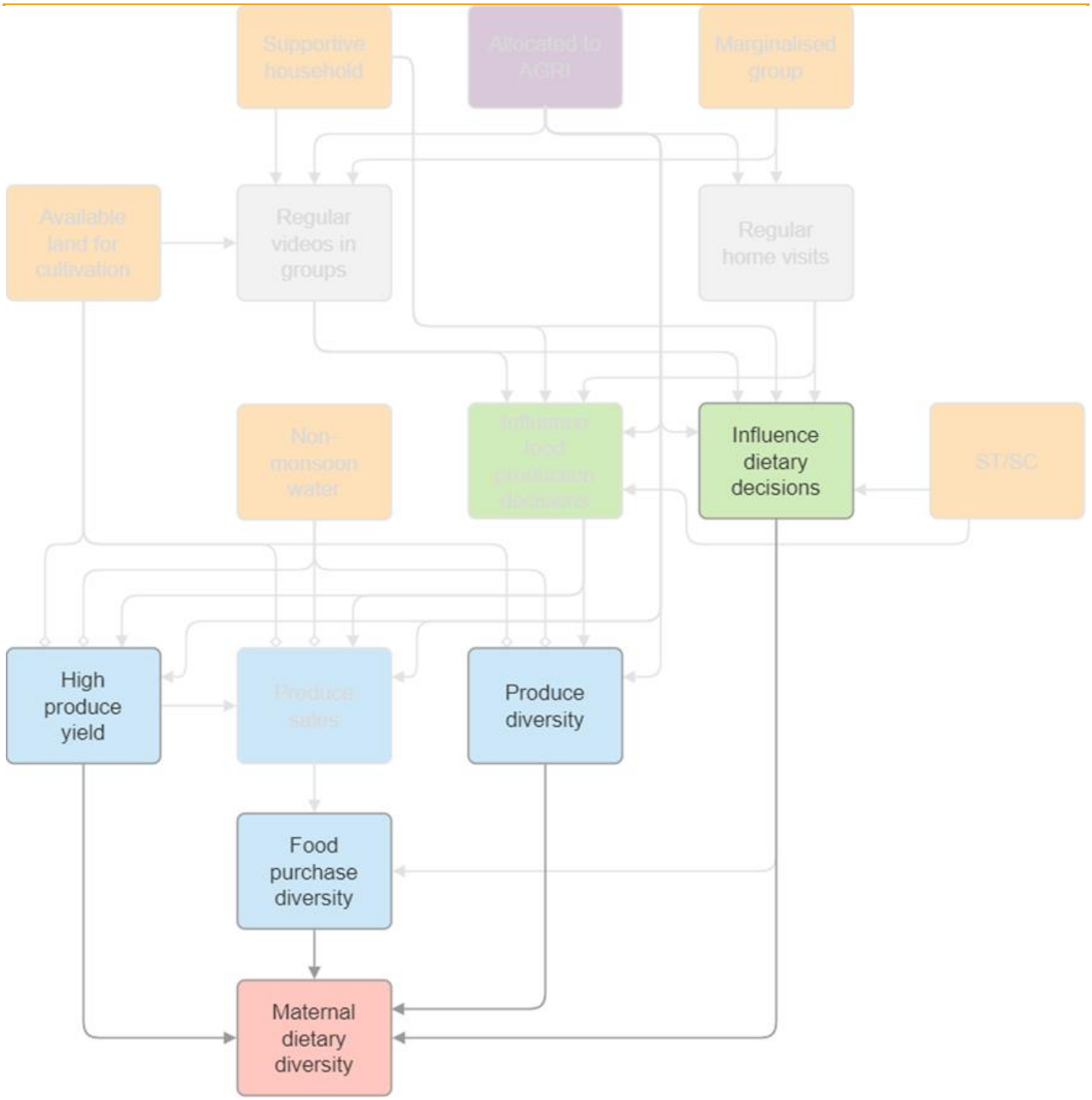


Figure 40: sub-DAG ending in "Maternal dietary diversity"

In CausalQueries, using the split function, the process produces the following sub-DAGs in Table 22:

Table 22: List of sub-DAGs as Modelled in CausalQueries

Child	Parents
DIETDECISION	AGRI, STSC, VIDEO, VISIT
DIETDIVERSITY	DIETDECISION, PRODDIVERSITY, PURCHASES, YIELD
PRODDECISION	AGRI, STSC, VIDEO, VISIT
PRODDIVERSITY	AGRI, LAND, PRODDECISION, WATER
PURCHASES	DIETDECISION, SALES
SALES	AGRI, LAND, PRODDECISION, WATER, YIELD
VIDEO	AGRI, LAND, MARGIN
VISIT	AGRI, MARGIN
YIELD	AGRI, LAND, PRODDECISION, WATER

Diagrammatically, these are presented as in Figure 41:

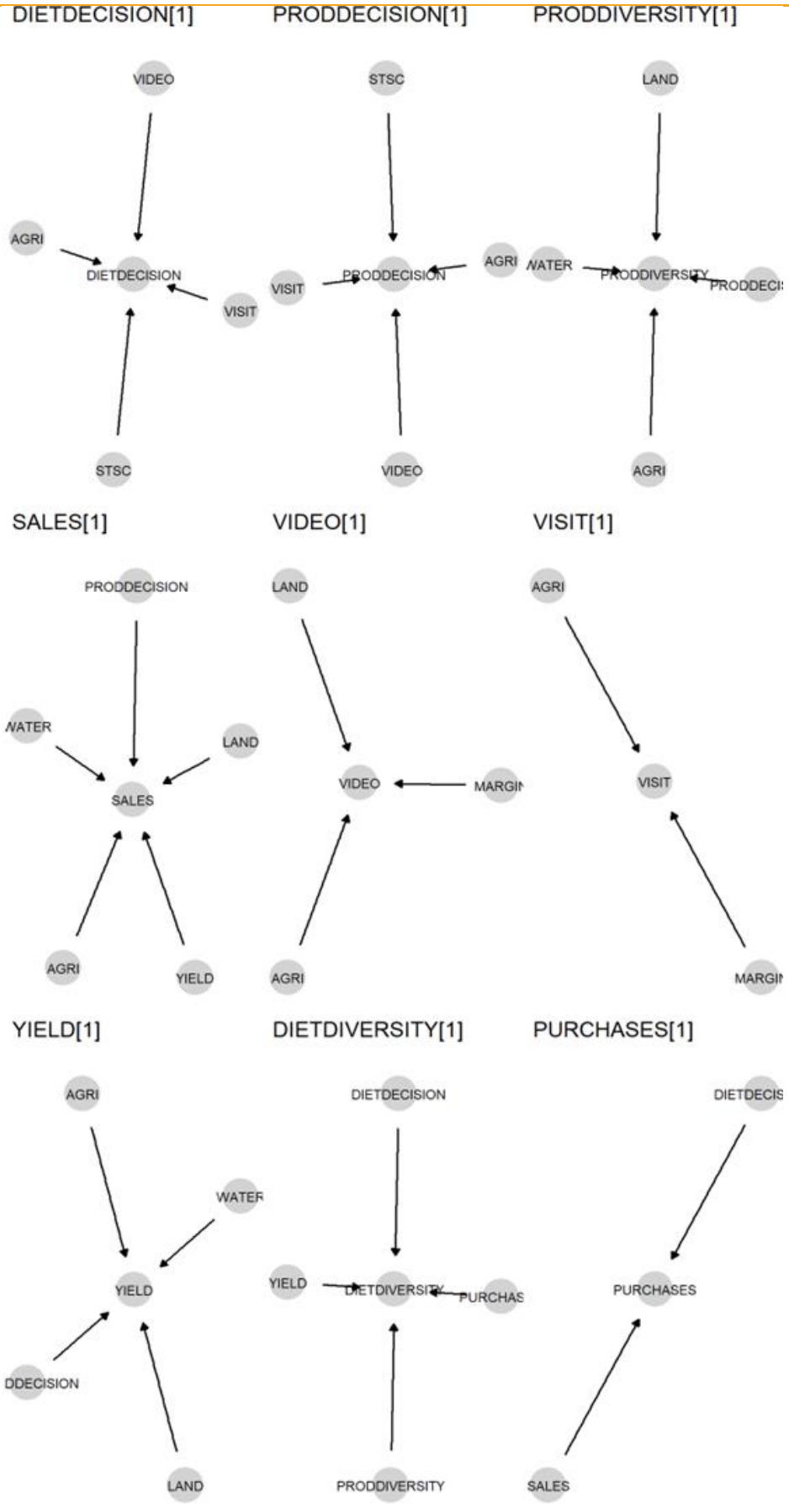


Figure 41: Complete set of sub-DAGs plotted

Updating one sub-model and querying the posteriors

Each of the sub-DAGs formed from the overall DAG was updated using the imputed data. To illustrate, we present the second sub-DAG in the UPAVAN model as an example - Figure 42. This is a simple, two-parent, one-child DAG with monotonicity and non-interaction restrictions. We select the model from the split DAG, and then update_model with the imputed data.

Figure 42: sub-DAG ending in VISIT plotted

We can check that this process has been completed as we would expect at the sub-DAG level before we look at the results of updating the complete model.

Table 23 shows how the model parameters – the distributions of the nodal types – have been updated by the data. Since interpreting the nodal types requires reference to the ordering of the parent nodes, an interpretation table is also shown in Table 24.

Table 23: Prior and Posterior Probabilities Parent Nodes of a sub-DAG

AGRI.0	0.50	0.40
AGRI.1	0.50	0.60
MARGIN.0	0.50	0.59
MARGIN.1	0.50	0.41
VISIT.0000	0.25	0.86
VISIT.0101	0.25	0.14
VISIT.0011	0.25	0.00
VISIT.1111	0.25	0.00

Table 24: Interpretation of the Ordering of Parent Nodes in a sub-DAG

node	position	display	interpretation
VISIT	1	VISIT[*]***	VISIT | AGRI = 0 & MARGIN = 0
VISIT	2	VISIT*[*]**	VISIT | AGRI = 1 & MARGIN = 0
VISIT	3	VISIT**[*]*	VISIT | AGRI = 0 & MARGIN = 1
VISIT	4	VISIT***[*]	VISIT | AGRI = 1 & MARGIN = 1

As expected, the proportions for the treatment (AGRI) are not changed since the proportion in each arm in the data is approximately equal. For being in a marginalised group (MARGIN), there is a slight update, since the proportion in that group in the data was closer to 40%. But neither of these nodal types represent causal processes since these are both exogenous nodes. The nodal types for being visited by an outreach worker (VISIT) represent the responses to the values of AGRI and MARGIN. VISIT.0000 are the nodal types where VISIT takes a value of 0 regardless of the values of AGRI or MARGIN; most (86%) of the women in the sample are of this type. The remainder are of the type 0101, which means that they take the value of 1 when AGRI = 1 and when MARGIN = 0 or 1, but otherwise take a value of zero; these are the people for whom AGRI has a causal effect.

To further check that updating has been completed as expected for the sub-DAG, we can plot the posterior distributions for nodal types for each node in the sub-DAG. These are presented as Figure 43, below.

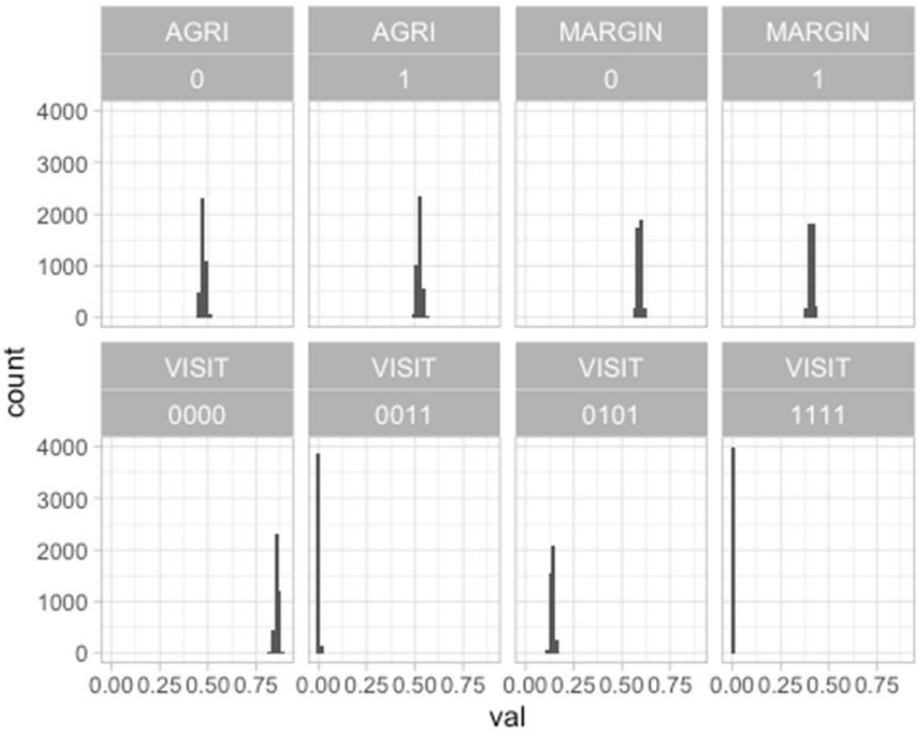


Figure 43: Posterior distributions of nodal types for VISIT sub-DAG

To interpret this plot, we first have to understand what is being plotted in each subplot. These plots are probability distributions that capture the distribution of the probability that each nodal type obtains for units to whom the model applies. The concept of nodal types is introduced in Part 2 of the Theory subsection of the Methodology section but is also illustrated below.

To interpret the plot above, we also have to understand the titles of the sub-plots that make up the whole. These titles are the name of the node concerned and then a list of digits that is a description of a nodal type, as described in Part 2 of the Theory subsection of the Methodology section. For AGRI and MARGIN, there are only two plots as these nodes are binary nodes that have no parents (they are 'exogenous') and therefore only have two nodal types; the nodal type that is taking the value '0', and the nodal type that is taking the value '1.' For endogenous nodes interpreting the nodal type label is more complicated. We can refer to Table 24 to help us interpret the labels of nodal types for VISIT. Looking at that table, we can see that for each nodal type label, the first digit refers to the value taken when both parent nodes take the value '0' and the last digit refers to the value taken when both parent nodes take the value '1.' The second and third digits refer to the value taken when AGRI = '1' and MARGIN = '0' and vice-versa. Because the UPVAN model is constrained by restrictions that say that causation at this node must be monotonic (only positive changes in the outcome result from positive changes in the inputs) and non-interacting (causation is only additive so, for example, it is not the case that the outcome will only take the value '1' if both causes are '1') there are only four nodal types for VISIT. The first nodal type is the nodal type in which VISIT takes the value '0' no matter what (sometimes called 'cursed'), the last is the nodal type in which VISIT takes the value '1' no matter what (sometimes called 'destined'), and the second and third nodal types are those in which the value of VISIT is causally determined by either MARGIN, or AGRI, respectively.

Inspecting the plots, we can see that AGRI takes the value '1' 50% of the time with high confidence, whereas MARGIN takes the value '1' more like 40% of the time, with slightly lower confidence. We can see from the fourth, rightmost VISIT subplot that VISIT is expected never to be 'destined,' with very high confidence. From the second subplot we can see that it is expected almost never to respond positively to MARGIN, with very high confidence. From the first and third subplots we can see that it is expected to be 'cursed' roughly 86% of the time and to respond positively to VISIT roughly 14% of the time, with fairly high confidence. These numbers all correspond to counts of the data and estimates generated with linear regression, confirming the successful updating of the DAG.

Updating the complete DAG

The complete UPAVAN and DIG models were updated using the procedure above on each sub-DAG, with the results stitched together, preserving the nodal types present in the full model and discarding the nodal types of sub-DAGs that are not present in the full model (i.e. the nodal type 0 or 1 for nodes that are exogenous in a sub-DAG but endogenous in the full

model.) This was completed using just prior data, just trial data, or a mixture of the two as described for all six model specifications in Table 25, duplicated below.

Table 25: The six model specifications

Model name	Updated on priors dataset	Updated on trial data
Naive model	None	No
Prior model 1	Priors 1 (Exclusive)	No
Prior model 2	Priors 2 (Inclusive)	No
Data only model	None	Yes
Posterior model 1	Priors 1	Yes
Posterior model 2	Priors 2	Yes

The process was completed on a consumer-grade laptop in a matter of minutes, and did not require recourse to a server, as we had feared it would. This was in part due to improvements that were made to the underlying computation behind CausalQueries.

Describing the UPAVAN model

Once updated we can describe the distributions of the posteriors for each nodal type for each node in each partner trial DAG. The distributions of posteriors for each nodal type for each node in the UPAVAN DAG are presented in Figure 44. The annotations are explained on the following page.

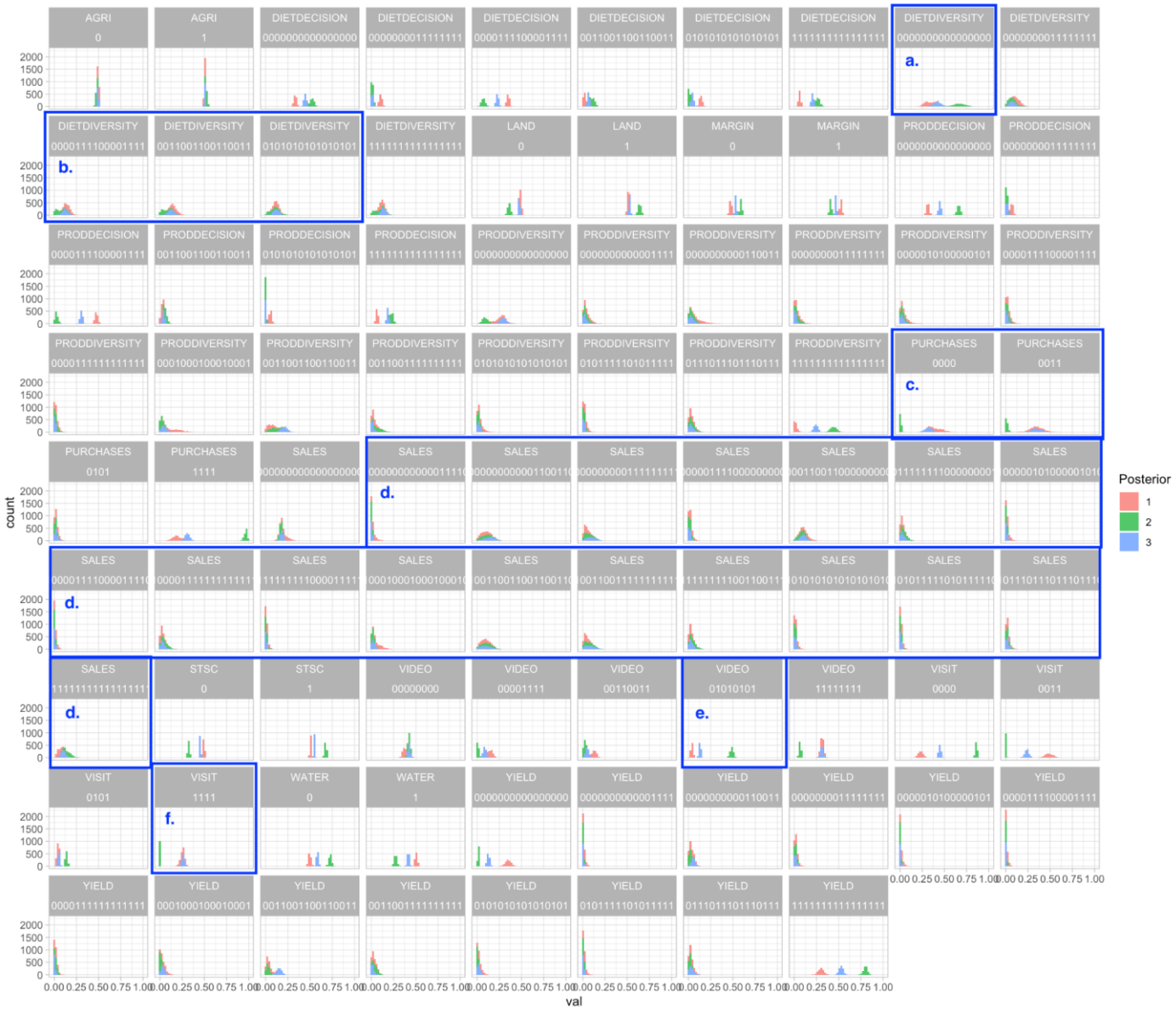


Figure 44: Distributions of posteriors for each nodal type for each node in the UPAVAN DAG

Distributions 1, 2 and 3 are for model specifications Priors 2 (only priors data, aggregated assuming inclusive priors), Data Only, and Posteriors 2 (data plus inclusive priors), respectively.

To download the full-size image, click here.

Figure 44 has been annotated with labelled boxes to draw the reader's attention to particular subplots. These annotations are explained in this list:

a) This causal type is labelled with a string of 0s. This string represents the fact that the nodal type whose distribution is plotted here concerns units who take the value '0' regardless of the values of the parent nodes of DIETDIVERSITY. This nodal type is sometimes referred to as 'cursed' for obvious reasons. The distributions plotted show that on the priors only model specification the proportion of 'cursed' participants is underestimated with respect to DIETDIVERSITY, relative to the data (or overestimated by this data relative to the priors.)

b) Priors slightly over-estimated the proportion who have their DIETDIVERSITY causally affected, relative to the data. This is consistent with there being a higher number of 'cursed' participants in the data than priors expected.

c) The data do not support their being many cursed or casually affected participants with respect to PURCHASES, and instead expects most of them to be positive for PURCHASES regardless ('destined'). This was one of the two ternary variables that was coarsened to be binary which may contribute to increased measurement error in this case.

d) The correspondence between priors and data were good for SALES. This was a relatively simple concept that was operationalised similarly in the prior elicitation process and in the data. It was defined as 'any sales or not.' This may have led to a very low level of conceptual error and measurement error, leading to a strong correspondence between data and priors.

e) The effect of AGRI on VIDEO was strong in the data, but not in the priors. This is curious, as it is surprising that our expert participants should be pessimistic about allocation to the intervention leading to one of the primary intervention components. This is probably a result of measurement error. The node was defined and understood by participants as 'woman attends regular video screenings.' The only measure available in the data was whether a woman recalled having attended at least one screening in the past 6 months. This bad measure may have resulted in many false positives in the data relative to the more exigent definition of the node. Additionally, it may have resulted from not reflecting on the causal significance of this process and the importance of collecting good process data, reinforcing the possible utility of the DAG-development process for future interventions.

f) In the priors, there was a surprising proportion of participants who were expected to be 'destined' to be visited by project workers regardless of the values of the parent nodes including allocation to the intervention. This discrepancy has no obvious explanation to us.

Describing the DIG model

As discussed above, we use the same update_stitch function as for the UPAVAN model due to the fact that update_model cannot update on the priors data. Once the models are updated

we can describe the distributions of the posteriors for each nodal type for each node in the DAG. The distribution of posteriors for each nodal type for each node in the DIG DAG is presented in Figure 45. The annotations are explained on the following page.

Figure 45: Distributions of posteriors for each nodal type for each node in the DIG DAG
Distributions 1, 2 and 3 are for model specifications Priors 2 (only priors data, aggregated assuming inclusive priors), Data Only, and Posteriors 2 (data plus inclusive priors), respectively.
To download the full-size image click here.

a) This subplot concerns the nodal type for units that are not economically active except in the case where only the individual-level disability inclusivity components were active. The priors only model overestimates the probability of this nodal type relative to the data by a large margin. This is a pattern repeated for the three subplots to the left of the highlighted plot and for the five plots on the row below, albeit to a less marked extent. These subplots are all of the various nodal types that reflect some kind of causal effect of some combination of the intervention components, applying to units who are neither 'destined' nor 'cursed'. The fact that the priors predict probability distributions that are further to the right is reflective of the fact that experts' priors were for much larger causal effects than those observed in the data.

b) This subplot concerns the nodal type of those who are 'destined' to be ACTIVE regardless of the values of the nodes representing intervention components. The very large difference between priors and posteriors here reflects the fact that experts enormously underestimated the probability of a unit having a 'destined' nodal type relative to the data. This appears from this plot to be the main explanation for the experts' overestimation of causal effects.

c) In the data, CLASSIC could include intervention elements that were not delivered by DIG. This was due to imperfect measurement and results in a high probability of false positives for this node. This may explain why there are more participants destined to receive the CLASSIC intervention components than were expected by the experts.

Part 5: Querying the models

Once the causal models had been updated on trial data and/or on expert priors according to the six model specifications, these models could be queried. The results could be interpreted to learn about causation in the trial and compared to learn about the relative roles of priors and data in the analysis under different sets of assumptions.

Querying the DIG model

We begin by reporting queries on the DIG model along with their results and our interpretation of those results.

Average Treatment Effect

We ran a query to estimate an average treatment effect under all six model specifications described in Table 26. This query as expressed in CausalQueries syntax is "ACTIVE[ALLOC = 1] - ACTIVE[ALLOC = 0]". This query asks, for causal types in the model, is the value of ACTIVE for that causal type greater when conditioning on ALLOC = 1 (allocated to treatment) than when conditioning on ALLOC = 0 (allocated to control,) or not. This query was evaluated by the query_model function by examining all 594 causal types in the DIG model and seeing whether the statement evaluates to 1 or 0, then returning a weighted average of these evaluations. In this weighted average, particular causal types are weighted by the expectation in the posterior model that they will obtain for units to whom the model applies. A summary of results is presented in Table 26.

Table 26: Estimates of average treatment effect under all six model specifications

Model	Query	Given	Using	mean	sd	cred.low	cred.high
Naive	ATE	•	priors	0.27	0.12	0.07	0.55
Priors 1	ATE	•	posteriors	0.57	0.04	0.49	0.64
Priors 2	ATE	•	posteriors	0.55	0.06	0.43	0.67
Data only	ATE	•	posteriors	0.05	0.01	0.03	0.07
Priors 1 + data	ATE	•	posteriors	0.22	0.02	0.19	0.25
Priors 2 + data	ATE	•	posteriors	0.13	0.02	0.10	0.16

The average treatment effects for the priors data only (Priors 1 and Priors 2) are high compared to the other model specifications (0.57 and 0.55, respectively compared to 0.05 in the data only model). This implies that the experts expected that the intervention was going to have effects some 10x greater than are observed only in the data. The composite model specifications combining data and priors result in middling estimates. This is as expected, given that the level of certainty of the experts' combined priors resulted in a combined *N* for the priors 'data' that was similar to the *N* of the trial data available. All estimates are fairly precise as a result of the fact that the estimate is identified in the model due to the random allocation of units to treatment and control.

The exclusive priors represented by Priors 1 are much more resistant to updating by data, resulting in a much higher estimated ATE in the Priors 1 + Data specification (Posteriors 1). This reflects the increased confidence in the Priors 1 model that results from the assumption that experts' priors are arrived at entirely independently of each other. Although the assumption that experts' priors are inclusive (that they result from different draws from entirely overlapping data) is also an oversimplification, it is a less heroic assumption in light of the fact that UPAVAN and DIG participants had both lived through the intervention at the time that their 'priors' were collected. Although we asked experts to cast their minds back to the start of the programme and to reflect their expectations at the start of the project, this is undoubtedly an inadequate compensation for the fact that experts' priors were not really 'prior.' Adopting the assumption of inclusive priors as our preferred method of aggregating priors goes some way to reducing the potential bias resulting from this limitation.

We do not report the estimates from the Naïve model specification for future queries. It is interesting that 'flat' priors result in estimates of causal effects that follow from the structure

of the DAG, but the particular effects are not as interesting as those informed by priors, data, or both.

Attribution and mediation

A distinctive feature of CausalQueries is that we can query the model on counterfactual conditions to return estimates of attribution. For the DIG model, we are interested in the probability that a causal effect can be attributed to the intervention. We can do this using a query that asks about the probability of a causal effect, just like for the ATE above. Above, the query statement is written as "ACTIVE[ALLOC = 1] - ACTIVE[ALLOC = 0]". In the code below we have written it as "ACTIVE[ALLOC = 1] > ACTIVE[ALLOC = 0]". For binary nodes, these statements are equivalent. Although we use an equivalent query statement, we change what we are asking by introducing a given argument: "ALLOC==1 & ACTIVE==1". This means we are asking only about cases in which in fact a unit was allocated to treatment and had a value of '1' for the main effect. We are asking for only these cases, how would the statement have evaluated, had the value of allocation been 0. That is, for units to whom the model applies, what is the probability that a value of 1 for the effect node would have been zero, had the unit not been allocated to treatment.

We can extend this attribution query to investigate mediation by adding further clauses to the given argument. We do this in the code below by adding the two possible values of a mediator as givens to two queries, in addition to stipulating that allocation and effect = 1. That is, we use two different sets of givens: "ALLOC==1 & ACTIVE==1 & ASSET==0" and "ALLOC==1 & ACTIVE==1 & ASSET==1". By evaluating the same query statement in the presence of these two different sets of givens we can see how our estimation of the probability that attribution is justified is affected by learning about the value of the mediator. A summary of results is presented in Table 27.

Table 27: Mediation query results about attribution, by each of the nodes marked 'mediator' as well as the three intermediate outcomes

Model	Query	Given	mean	sd	cred.low	cred.high
Priors 2	Attribution	ALLOC==1 & ACTIVE==1	0.87	0.04	0.77	0.93
Priors 2	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==0	0.75	0.09	0.55	0.89
Priors 2	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==1	0.87	0.04	0.78	0.94
Priors 2	Attribution	ALLOC==1 & ACTIVE==1 & CLASSIC==0	0.77	0.08	0.59	0.90
Priors 2	Attribution	ALLOC==1 & ACTIVE==1 & CLASSIC==1	0.87	0.04	0.78	0.94
Priors 2	Attribution	ALLOC==1 & ACTIVE==1 & DIINDIV==0	0.83	0.06	0.69	0.93
Priors 2	Attribution	ALLOC==1 & ACTIVE==1 & DIINDIV==1	0.87	0.04	0.78	0.94
Data only	Attribution	ALLOC==1 & ACTIVE==1	0.06	0.01	0.04	0.08
Data only	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==0	0.03	0.01	0.02	0.05
Data only	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==1	0.06	0.01	0.04	0.09
Data only	Attribution	ALLOC==1 & ACTIVE==1 & CLASSIC==0	0.04	0.02	0.01	0.08
Data only	Attribution	ALLOC==1 & ACTIVE==1 & CLASSIC==1	0.06	0.01	0.04	0.09
Data only	Attribution	ALLOC==1 & ACTIVE==1 & DIINDIV==0	0.05	0.01	0.03	0.08
Data only	Attribution	ALLOC==1 & ACTIVE==1 & DIINDIV==1	0.08	0.02	0.05	0.11
Priors 2 + data	Attribution	ALLOC==1 & ACTIVE==1	0.16	0.02	0.12	0.20

Model	Query	Given	mean	sd	cred.low	cred.high
Priors 2 + data	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==0	0.07	0.02	0.04	0.11
Priors 2 + data	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==1	0.18	0.02	0.13	0.22
Priors 2 + data	Attribution	ALLOC==1 & ACTIVE==1 & CLASSIC==0	0.04	0.02	0.01	0.08
Priors 2 + data	Attribution	ALLOC==1 & ACTIVE==1 & CLASSIC==1	0.19	0.02	0.15	0.23
Priors 2 + data	Attribution	ALLOC==1 & ACTIVE==1 & DIINDIV==0	0.16	0.02	0.12	0.19
Priors 2 + data	Attribution	ALLOC==1 & ACTIVE==1 & DIINDIV==1	0.17	0.02	0.13	0.21

There is a lot to unpack in this large table. The same pattern of large effects suggested by priors, small effects resultant from the data, and a compromise estimate arising from a composite model specification is reflected here as for the ATE query. Like for the ATE, there is no suggestion that prior estimates followed a very different form from estimates resulting from data, rather the difference is mostly one of magnitude.

To unpack the results of the queries, consider one model specification, that of Priors 2 + Data. Under this model specification the base attribution query returns a mean of 0.16 with a standard deviation of 0.02. This is a precise estimate resulting from the fact that this estimate is identified in the model. This is because we have randomization of allocation to treatment and control for the DIG intervention. Unpacking the mean of 0.16, this means that the central estimate for the probability that the presence of the desired outcome was due to the intervention, for those units who were allocated to treatment, is 0.16. For all mediators, this probability is increased by knowledge that the mediator = 1 and is reduced by knowledge that the mediator = 0. However, the size of this effect on our confidence of attribution is very different between ASSET and CLASSIC on the one hand, and DIINDIV on the other. Knowing the value of DIINDIV changes our confidence in attribution of the effect to the intervention by very little (0.00-0.01). On the other hand, knowing that ASSET = 0 decreases our confidence in attribution from the base case of 0.16 by more than half to 0.07, whereas knowing that ASSET = 1 increases it to 0.18. The power of knowledge about CLASSIC is higher, with a reduction in confidence by 75% to 0.04 when we know that its value is 0 and an increase to 0.19 when we know its value is 1, reflecting a strong estimate of the mediating power of the classic graduation intervention components, even stronger than the mediating effect of the classic asset transfer component.

Alternative mediation query

A more complex alternative mediation query in Table 28, asks about the presence of positive indirect effects of allocation on the outcome for each mediator. This query is written in CausalQueries syntax as: "ACTIVE[ALLOC=1, M=M[ALLOC=1]] > ACTIVE[ALLOC=1, M=M[ALLOC=0]]" where "M" = the mediator in question. By evaluating this query against a probability-weighted average of the causal types in the model, we are asking: if you change the value of the mediator from the value it would take when allocation = 0, to the value it would take if allocation = 1, holding allocation constant at 1 (all other effects held constant), what's the probability of a positive effect?

Concentrating on the Priors 2 + Data model specification, we can see that this alternative mediation query gives results that are consistent with the previous form of query. There is a higher positive indirect effect through ASSET (0.07) and CLASSIC (0.08) as compared to DIINDIV (0.01). However, these coefficients are smaller than those under the attribution style of query as we are not conditioning on a positive treatment effect, but rather investigating all causal types of relevance to all units.

Table 28: An Alternative Mediation Query

Model	Query	Given	mean	sd	cred.low	cred.high
Priors 2	Positive indirect effects via ASSET	•	0.30	0.06	0.18	0.43
Priors 2	Positive indirect effects via CLASSIC	•	0.27	0.06	0.15	0.40
Priors 2	Positive indirect effects via DIINDIV	•	0.16	0.05	0.07	0.28
Data only	Positive indirect effects via ASSET	•	0.02	0.01	0.01	0.05
Data only	Positive indirect effects via CLASSIC	•	0.02	0.01	0.01	0.04
Data only	Positive indirect effects via DIINDIV	•	0.00	0.00	0.00	0.01
Priors 2 + data	Positive indirect effects via ASSET	•	0.07	0.02	0.04	0.11
Priors 2 + data	Positive indirect effects via CLASSIC	•	0.08	0.01	0.06	0.10
Priors 2 + data	Positive indirect effects via DIINDIV	•	0.01	0.00	0.00	0.01

Necessity

For the partner team, the main point of the DIG model we elicited is the investigation of interaction between the three different sets of intervention components that we have described as mediators above. One aspect of interaction can be captured by investigating how necessary a given parent node is. To do this, we used an attribution query, as above, but compared results using alternative exhaustive sets of givens to see how our estimates of attribution are influenced by knowledge that all mediators take the value 1 except for one, which takes the value 0. So, for example, we compare the case where our givens are "ALLOC==1 & ACTIVE==1 & ASSET==1 & CLASSIC==1 & DIINDIV==1" with the case where our givens are "ALLOC==1 & ACTIVE==1 & ASSET==0 & CLASSIC==1 & DIINDIV==1". This lets us compare the effect on our estimation of attribution of knowledge that the asset transfer component is the only intervention component which is missing with knowledge that all intervention components are present. That is, we are able to investigate how necessary the presence of the asset transfer is to our confidence in the attribution of a causal effect of the intervention. A summary of results is presented in Table 29.

Focusing again on the results for the Priors 2 + Data model specification: one is half as likely to attribute a positive outcome to the effect of the intervention when one knows that ASSET = 0 (from 0.21 to 0.10), and one is 1/4 as likely when knowing that CLASSIC = 0 (from 0.21 to 0.06), whereas there is almost no change from learning about DIINDIV (from 0.21 to 0.20). This suggests that none of the sets of intervention components are strictly necessary for all units, but that CLASSIC and ASSET are strongly predictive of a treatment effect.

Table 29: Investigating how necessary a given parent node is

Model	Query	Given	mean	sd	cred.low	cred.high
Priors 2	Attribution	ALLOC==1 & ACTIVE==1	0.87	0.04	0.77	0.93
Priors 2	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==1 & CLASSIC==1 & DIINDIV==1	0.87	0.04	0.79	0.94
Priors 2	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==0 & CLASSIC==1 & DIINDIV==1	0.77	0.08	0.58	0.90
Priors 2	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==1 & CLASSIC==0 & DIINDIV==1	0.78	0.08	0.61	0.91
Priors 2	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==1 & CLASSIC==1 & DIINDIV==0	0.84	0.06	0.71	0.94
Data only	Attribution	ALLOC==1 & ACTIVE==1	0.06	0.01	0.04	0.08
Data only	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==1 & CLASSIC==1 & DIINDIV==1	0.09	0.02	0.05	0.12
Data only	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==0 & CLASSIC==1 & DIINDIV==1	0.06	0.02	0.03	0.10
Data only	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==1 & CLASSIC==0 & DIINDIV==1	0.07	0.02	0.03	0.12
Data only	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==1 & CLASSIC==1 & DIINDIV==0	0.07	0.02	0.04	0.10
Priors 2 + data	Attribution	ALLOC==1 & ACTIVE==1	0.16	0.02	0.12	0.20
Priors 2 + data	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==1 & CLASSIC==1 & DIINDIV==1	0.21	0.02	0.17	0.26
Priors 2 + data	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==0 & CLASSIC==1 & DIINDIV==1	0.10	0.02	0.05	0.14
Priors 2 + data	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==1 & CLASSIC==0 & DIINDIV==1	0.06	0.02	0.02	0.11
Priors 2 + data	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==1 & CLASSIC==1 & DIINDIV==0	0.20	0.02	0.15	0.24

Sufficiency

As above, we use a sort of attribution analysis to investigate the sufficiency of each intervention component. This time, although we continue to set the values of allocation and main effect to 1, our base case for comparison sets the value of the mediators to 0. We then compare the results of the attribution query for this base case with the case in which each mediator's value in turn is set to 1 in the givens statement. This allows us to investigate how sufficient each intervention component is, in the absence of the others, in increasing our confidence in the attribution of causal effects - Table 30.

Focusing again on the results for the Priors 2 + Data model specification, we can see that knowledge that all of the sets of intervention components = 0 reduces our confidence in the attribution of an effect to the intervention for a treated unit to 0. This makes sense, as there are no direct paths in the model from treatment to outcome, only paths mediated by the sets of intervention components. Knowing that CLASSIC =1 when other sets of components = 0 gives an 8% confidence in attribution of a positive outcome to the intervention, whereas that number for ASSET is 4% and for DIINDIV is 2%. It is interesting to compare this picture to the picture resulting from the necessity queries. In that case, knowledge that CLASSIC = 0 when the other components = 1 reduces confidence in a positive outcome being attributable to the intervention by 10 percentage points from the base case of 16% to 6%. However, this doesn't mean that CLASSIC on its own is sufficient for confidence in the attribution of a positive effect. On its own, we are still only half as confident as in the case with no knowledge of mediators that a positive outcome is attributable to the intervention, for treated units.

Table 30: Investigating how sufficient a given parent node is

Model	Query	Given	mean	sd	cred.low	cred.high
Priors 2	Attribution	ALLOC==1 & ACTIVE==1	0.87	0.04	0.77	0.93
Priors 2	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==0 & CLASSIC==0 & DIINDIV==0	0.00	0.00	0.00	0.00
Priors 2	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==1 & CLASSIC==0 & DIINDIV==0	0.68	0.13	0.39	0.88
Priors 2	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==0 & CLASSIC==1 & DIINDIV==0	0.63	0.14	0.31	0.87
Priors 2	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==0 & CLASSIC==0 & DIINDIV==1	0.54	0.14	0.24	0.78
Data only	Attribution	ALLOC==1 & ACTIVE==1	0.06	0.01	0.04	0.08
Data only	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==0 & CLASSIC==0 & DIINDIV==0	0.00	0.00	0.00	0.00
Data only	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==1 & CLASSIC==0 & DIINDIV==0	0.04	0.02	0.01	0.09
Data only	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==0 & CLASSIC==1 & DIINDIV==0	0.04	0.01	0.02	0.06
Data only	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==0 & CLASSIC==0 & DIINDIV==1	0.04	0.02	0.01	0.08
Priors 2 + data	Attribution	ALLOC==1 & ACTIVE==1	0.16	0.02	0.12	0.20

Model	Query	Given	mean	sd	cred.low	cred.high
Priors 2 + data	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==0 & CLASSIC==0 & DIINDIV==0	0.00	0.00	0.00	0.00
Priors 2 + data	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==1 & CLASSIC==0 & DIINDIV==0	0.04	0.02	0.01	0.08
Priors 2 + data	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==0 & CLASSIC==1 & DIINDIV==0	0.08	0.02	0.04	0.13
Priors 2 + data	Attribution	ALLOC==1 & ACTIVE==1 & ASSET==0 & CLASSIC==0 & DIINDIV==1	0.02	0.01	0.00	0.04

Querying the UPAVAN model

As for DIG, we intended to query different specifications of the UPAVAN model using similar queries to those used above. We hoped to generate estimates of the average treatment effect under all of our different model specifications from Table N. We also hoped to conduct mediation analysis in the same way as for the DIG model, above: first, by asking about how learning about the value of the mediator affects our beliefs about the probability that positive effects of treatment on the outcome were due to the treatment; and second, by asking about the probability of positive indirect effects flowing through each mediator.

In addition, the UPAVAN model contains moderators whose effects can be investigated with queries. We intended to do this using the same form of attribution queries used to investigate mediation, asking what is learned about the probability of the attribution of positive effects to the intervention when we know the value of the moderator. For example, we would compare the result of the query DIETDIVERSITY[AGRI = 1] - DIETDIVERSITY[AGRI = 0] under the conditions ``AGRI==1 & DIETDIVERSITY==1", "AGRI==1 & DIETDIVERSITY==1 & LAND==0", and "AGRI==1 & DIETDIVERSITY==1 & LAND==1". This tells us how learning about the value of a moderator effects our estimates of the probability that positive changes in the outcome can be attributed to changes in allocation.

Further, the presence of both mediators and moderators means that we might want to learn about the interactions between mediation and moderation. We intended to investigate this by using queries asking about moderated mediation. Maybe the pathway through which treatment effects operate depends on these exogenous conditions under which households are operating? We could ask the model to find out. That is, we intended to run the mediation queries asking about positive indirect effects conditioning on values of the moderators. For example, we could use the positive indirect effects query "DIETDIVERSITY[AGRI=1, PRODDECISION=PRODDECISION[AGRI=1]] > DIETDIVERSITY[AGRI=1, PRODDECISION=PRODDECISION[AGRI=0]]") and condition on the givens "LAND==0", "LAND==1". This would tell us how learning about the effects of the moderator affects our estimates of the probability of positive indirect effects via the mediator.

These queries could not be run on the full UPAVAN model in time for this version of this working paper because it was not possible to overcome the computational challenges of querying large models in time. We hoped to complete this work much earlier in the project, soon after the point at which we became aware of the problem. However, this work has proved much more complex than expected and it has not been possible to complete it in time. Addressing this problem has required the hiring of a developer who has written a large amount of novel code with useful applications well beyond the scope of this project. We believe that it will soon be complete and will allow us to query the full UPAVAN model and to update this working paper. We discuss this software development work more fully in the Conclusion section and in Annex A, section A2 that follows.

Querying UPAVAN sub-DAGs

Despite not being able to query the full DAG, we are able to construct sub-DAGs that express elements of the relationships in the full model, to update these simple models on data, and to query them. This exploits the fact that the formulation of the theory of change as a DAG allows for principled simplification of the DAG so that the causal structure is respected but in a summarised form(Pearl & Mackenzie, 2018). Once these sub-DAGs have been created in CausalQueries they can be updated on the trial data following the same process as that described for the full DAG (using the update_stitch command) or by using the simpler update_model command from unmodified CausalQueries. These models cannot be updated on the priors data in this way, however. This is because, as discussed, the priors data is generated in such a way that there is no rectangular dataset of priors data on which an arbitrary sub-DAG can be updated. Rather, there is a collection of rectangular datasets of priors data, one for each child node in the full DAG featuring data for the child given different values of its parents in the full DAG. The full model can be updated on these datasets using the update_stitch command, but trying to do so for an arbitrary sub-DAG (rather than a sub-DAG consisting of all and only a child node from the full DAG and its parents) would lead to erroneous results.

Despite not currently being able to make use of the priors data for queries relating to arbitrary sub-DAGs, we can update a naïve specification of an arbitrary sub-DAG on data only and query that model. This gives results that are informative about relationships in the data conditional on the theory of change being roughly correct, albeit leaving out much of the information collected from experts. In future, we hope to be able to translate the results of our prior elicitation strategy into the alpha parameters for Diriclet distributions that CausalQueries requires when inputting priors without updating the model on pseudodata. This would enable us to use the prior information for arbitrary subDAGs of the full model and would reduce the computational load of the approach by substituting updating on priors 'data' for data entry of alpha values. The work we have done towards achieving this is reported in Annex A, Section A3. This work being incomplete, below we report the results of queries on sub-DAGs updated only on data.

Average treatment effect

For the simplest example, consider a two-node DAG beginning with "Allocated to AGRI" and ending with "Maternal dietary diversity." This is all that is needed to estimate the average treatment effect. This simple, two-node DAG can be plotted in CausalQueries, updated on the trial data, and queried using the same form of query as above. The results are presented in Table 31.

Table 31: Average Treatment Effect in The Simplified Causal Structure

Query	Given	Using	mean	sd	conf.low	conf.high
ATE	•	priors	0.00	0.32	-0.63	0.66
ATE	•	posteriors	0.03	0.02	-0.01	0.07

As we suspected, given the small and somewhat uncertain finding of the original trial using frequentist statistics, updating on the data from flat priors justifies belief in a small effect with a fairly narrow credibility interval (much narrower than the estimate from flat priors) albeit one that includes a very small negative effect.

Mediation through three intermediate outcomes

For a more complex example of a reduced DAG, we can explore mediation by three secondary outcomes. To do this we reduce the effects of allocation to the AGRI arm on the final outcome, into effects mediated by changes in the three intermediate outcomes that immediately precede it: "High produce yield," "Food purchase diversity," and "Produce diversity," as well as a direct path (which flows through "Influence dietary decisions" on the full DAG) - Figure 46.

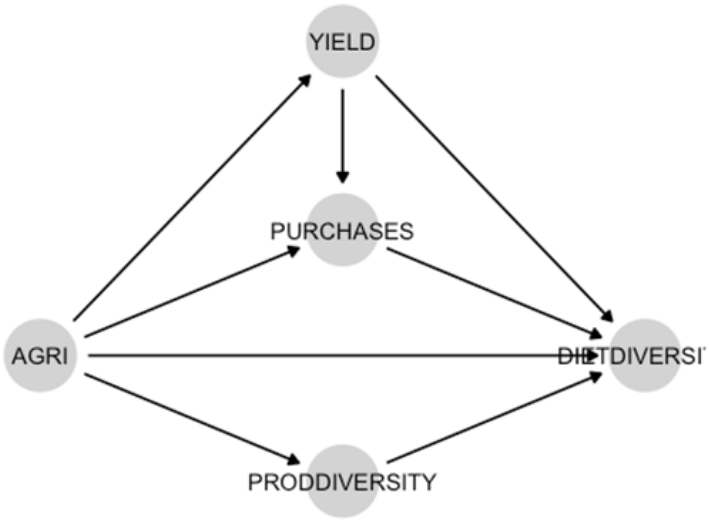


Figure 46: Mediation by three intermediate outcomes sub-DAG

This model can be specified and updated with the imputed data. Model querying results are summarised in Table 32.

All the queries reported in Table 32 are fundamentally about attribution. The queries ask: 'What is the probability that sufficient maternal dietary diversity (DIETDIVERSITY) was caused by allocation to the AGRI arm (AGRI)?'. The givens for the first query in the first line in the table

restrict the query to those cases in which AGRI = 1 and DIETDIVERSITY = 1. In addition, subsequent queries in the table are restricted to those cases in which the intermediate outcomes take a certain value.

By comparing the means of estimates of probability in Table 32, we can compare what changes in belief are justified by knowledge that each of the intermediate outcome nodes takes a certain value, compared to the base case. Looking down the columns for mean, and low and high confidence bounds, we can see that there are some small changes from case to case, but that the only significant change in mean is in the case where it is known that YIELD = 0. Here, the mean of the estimate is almost double the base case, albeit at double the standard deviation. This suggests that a positive effect of allocation to the AGRI arm on the primary outcome, in cases in which allocation and a positive effect have both been observed, is twice as likely in cases where the mother's household has failed to achieve a high crop yield. This is true both by comparison to the case in which it is known that the household did not achieve a high crop yield, and to the case in which yield is unknown. Knowledge about the other intermediate outcomes does not justify any change in belief compared to not knowing about them.

Table 32: Causal Queries Results on Dietary Diversity Through Three Intermediate Outcomes

Query	Given	Using	Case.estimand	mean	sd	conf.low	conf.high
COE	AGRI==1 & DIETDIVERSITY==1	posteriors	FALSE	0.06	0.03	0.02	0.12
COE	AGRI==1 & DIETDIVERSITY==1 & YIELD==0	posteriors	FALSE	0.11	0.06	0.03	0.24
COE	AGRI==1 & DIETDIVERSITY==1 & YIELD==1	posteriors	FALSE	0.06	0.03	0.02	0.12
COE	AGRI==1 & DIETDIVERSITY==1 & PURCHASES==0	posteriors	FALSE	0.06	0.03	0.02	0.12
COE	AGRI==1 & DIETDIVERSITY==1 & PURCHASES==1	posteriors	FALSE	0.06	0.03	0.02	0.12
COE	AGRI==1 & DIETDIVERSITY==1 & PRODDIVERSITY==0	posteriors	FALSE	0.05	0.03	0.01	0.12
COE	AGRI==1 & DIETDIVERSITY==1 & PRODDIVERSITY==1	posteriors	FALSE	0.06	0.03	0.03	0.12

Given YIELD==0 seems to be the only channel of mediation in this reduced model that is significant, we can expand our list of queries to investigate, for cases in which we know the household to have failed to achieve a high crop yield, whether knowing about the level of the other intermediate outcomes justifies any change in belief about attribution of a causal effect of AGRI.

Examining these queries and comparing their results in Table 33, we can see that justified belief in the mediating effect of a low yield is not significantly affected by knowledge of the values of the other mediators in the model.

Table 33: Causal Queries Results on Effects of Low Yields on Dietary Diversity

Query	Given	Using	Case.estimand	mean	sd	conf.low	conf.high
COE	AGRI==1 & DIETDIVERSITY==1 & YIELD==0	posteriors	FALSE	0.11	0.06	0.03	0.24
COE	AGRI==1 & DIETDIVERSITY==1 & YIELD==0 & PURCHASES==0	posteriors	FALSE	0.12	0.06	0.03	0.26
COE	AGRI==1 & DIETDIVERSITY==1 & YIELD==0 & PURCHASES==1	posteriors	FALSE	0.11	0.05	0.03	0.24
COE	AGRI==1 & DIETDIVERSITY==1 & YIELD==0 & PRODDIVERSITY==0	posteriors	FALSE	0.11	0.07	0.01	0.27
COE	AGRI==1 & DIETDIVERSITY==1 & YIELD==0 & PRODDIVERSITY==1	posteriors	FALSE	0.12	0.05	0.04	0.23

Effect modification of water

Effect modification by contextual variables can also be explored in CausalQueries. Since querying the complete DAG is not currently possible, we can construct another summary model that respects the underlying structure of the complete DAG, albeit radically simplifying it. With this much-reduced model we can explore how the effect of the AGRI intervention on maternal dietary diversity is modified by the availability of a non-monsoon water supply - Figure 47.

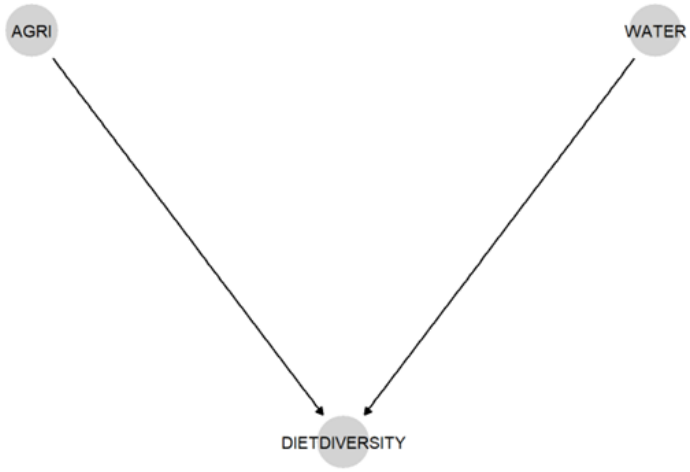


Figure 47: Moderation by non-monsoon water supply sub-DAG

In the full model, there are some chains of effects between AGRI and DIETDIVERSITY which are independent of WATER, and some chains via nodes whose effect is modified by WATER. This simple model does not restrict interactions to reflect this possibility of interactions in the effects on DIETDIVERSITY of AGRI and WATER, though it does impose a monotonicity restriction as agreed with the UPAVAN team. We can update the model and run queries, setting the value of WATER - Table 34.

Table 34: Causal Queries Results on Moderation Effects of Allocation to AGRI on Dietary Diversity

Query	Given	Using	Case.estimand	mean	sd	conf.low	conf.high
COE	AGRI==1 & DIETDIVERSITY==1	posteriors	FALSE	0.11	0.04	0.04	0.19
COE	AGRI==1 & DIETDIVERSITY==1 & WATER==0	posteriors	FALSE	0.14	0.05	0.04	0.25
COE	AGRI==1 & DIETDIVERSITY==1 & WATER==1	posteriors	FALSE	0.09	0.04	0.02	0.18

These queries will be familiar from the mediation analysis above. We can see that, for cases where AGRI = 1 and DIETDIVERSITY = 1, the probability that sufficient maternal dietary diversity has been caused by allocation to the AGRI arm is 0.11 with a standard deviation of 0.04. The question answered by the next two queries is: 'What change in belief about this probability is justified by knowledge of the value of WATER?' What the queries tell us is that learning that WATER = 0 justifies belief in an increased probability of allocation causing

sufficient dietary diversity, to 0.14, albeit with a standard deviation of 0.06. By contrast, for cases where WATER = 1, we should be less confident that allocation to the AGRI arm caused sufficient dietary diversity, with an updated probability of 0.09, standard deviation 0.04.

5. Conclusions

Through the work we have done to advance the POInT approach to integration of process and outcome data with theory, we have successfully applied the method to two real cases and thus built a foundation for future researchers to build on. All of the tools we have developed have been documented and are freely available for other researchers to use. In this section we reflect on the project as a whole and report the major strengths and limitations of our approach as well as implications and recommendations for policy and practice.

Strengths

The work we have completed on this project and reported here has revealed the promise of the POInT approach. In addition, it has advanced the methods and tools required to apply any structured causal modelling approach to the evaluation of real interventions. In this subsection we reflect on the most important aspects of that development work and on the reasons to be excited about the approach that it facilitates.

From proof-of-concept to real theories

Previous examples of a structured causal modelling approach in the literature have been limited to small models and are best characterised as proof-of-concept work. Through our application of the POInT approach to the UPAVAN case we show that it is possible to elicit a theory of change expressed as a DAG from a trial team. We have shown how that model can be iteratively developed so as to arrive at an analysis DAG that can be encoded in CausalQueries as a Bayesian causal model. This causal model represents an underlying counterfactual logic that is not inherent in some other methods, such as Structural Equation Modelling. The code and approach were refined such that updating was not only possible but trivially quick on a conventional laptop; this exceeded our expectation and has raised our confidence that the POInT process can be integrated into evaluators' workflows.

The causal models of our partner teams had to be constrained for computational and conceptual reasons, though we found that the required constraints (monotonicity, binary or ternary nodes, and limited interactions) were broadly acceptable to those partner teams. We

were able to implement the constraints computationally. The models were updated using expert beliefs and trial data by splitting larger models into smaller DAGs that were stitched together. The posterior values of the updated models could be interrogated. Although we uncovered an issue with running queries on the complete UPAVAN model, simplified DAGs could be queried and the full DIG model could be queried.

We have shown that it is possible to update large and complex models with data usually associated with the 'impact' as well as the 'process' components of an evaluation. This was possible for the UPAVAN and DIG models with restrictions on the kinds of nodal types but was otherwise non-parametric and flexible. Initial work with the AMETHIST partner team suggested that this approach would also work for their theory of change diagram, though this application could not be completed due to the sensitivities involved with attempting to produce a complementary analysis that required data sharing before the point at which the primary trial results would be published.

Work with the DIG partner team revealed the impossibility of using the CausalQueries-powered POInT approach for the shape of DAG implied by that partner team's full theory, a useful finding that is covered in the Limitations subsection. For this implementation we were able to use our approach to investigate a subquestion of relevance to the partner evaluation, investigating a useful part rather than the totality of their theory of change. This model was encoded in software, updated with expert beliefs and trial data, and queried for estimates of interest.

Explicit modelling and updating of expert beliefs

In this project, not only have we translated real theories of change into Bayesian causal models, but we have fitted those models to expert beliefs about the nature and strength of causal relations in the model. Eliciting those prior beliefs required us to develop a novel prior elicitation strategy that allowed us to elicit an order of magnitude more parameters than was previously possible. To create a prior model in CausalQueries based on those beliefs we had to develop an approach to interpreting prior beliefs as the data which, if observed, would justify those beliefs, and then update the model on that data. These innovations are discussed in more detail in Parts 3 of both the Empirical Methods subsection and the Results section.

Including these prior beliefs when estimating effects of interest reveals many interesting findings. For example, looking at the plot of posterior distributions for the UPAVAN case and especially at the different estimates for queries for the DIG case, we can see that experts consistently estimated much larger causal effects than those that arise from updating the model on data alone. This finding is consistent with others in the literature, for example (Iacovone et al., 2023) find large estimates for causal effects based on priors elicited and aggregated from different sorts of stakeholders despite a null effect of the intervention in question. Collecting and reporting such priors more often might allow researchers and policymakers to begin to quantify any biases in the prior estimation of effects and to predict the sorts of interventions which are more or less susceptible to these biases.

Another interesting finding from our prior elicitation was the discrepancy between the restrictions that participants were explicitly happy to place on the causal model in advance and the violations of those restrictions implied by their implicit priors as collected through prior elicitation. This suggests that even where causal models are elicited from experts and they are asked some technical questions about the features of those models, there is a role for prior elicitation through the estimation of results to triangulate such statements with implicit beliefs.

The power of queries

Examining the results of queries presented in Section 3.7 shows the potential of using the analysis of causal models for evaluation of development interventions. Most notably, being able to ask directly about attribution is interesting. Evaluators will want to know about average treatment effects, which the language of queries can deliver, but they may also want to know about attribution: how confident can we be, in the case of a positive outcome on a treated unit, that this positive outcome is attributable to the intervention. This is a different question from the ATE as reflected in the numbers for the DIG application. The ATE for the Priors 2 + Data model specification is 0.13 whereas the attribution coefficient is 0.16.

Further, the attribution query is a useful building block for asking more questions about mediation, moderation or even moderated mediation. We can ask, in all these cases, what more is learned when we learn the value of the moderator or mediator, and how does this affect our confidence in attribution.

The alternative mediation query presented above makes use of another interesting feature of CausalQueries syntax: the ability to consider effects under counterfactual conditions. The counterfactual mediation query allows us to ask of a model, for a given mediator, if you change the value of the mediator from the value it would take when allocation = 0, to the value it would take if allocation = 1, holding allocation constant at 1 (all other effects held constant), what's the probability of a positive effect?

Another interesting form of query was explored for the UPAVAN intervention, though we were unable to operationalise it owing to the computational constraints discussed in sections 3.8 and 4.1: moderated mediation. The UPAVAN model contains moderators whose effects can be investigated with queries. We intended to do this using the same form of attribution queries used to investigate mediation, asking what is learned about the probability of the attribution of positive effects to the intervention when we know the value of the moderator.

This approach shares a strong conceptual similarity with the Realist concept of context, mechanism, outcome (CMO) configurations (Pawson and Tilley, 1997). Assessing the probability of mediation through a particular channel is one way of assessing whether a certain mechanism is active, and an exogenous moderator node is one expression of the way in which context can condition causal responses. At least for some CMO configurations, moderated mediation queries may be a way of assessing whether they are significant in a given setting. This could allow us to estimate the probability of causal effects via different

channels in different contextual configurations, increasing our ability to build a middle-range theory of the sorts of settings in which a given type of intervention might be expected to be cost-effective (Pawson, 2000).

Limitations of the project

Our work has also uncovered some limitations of the POInT approach. Some of these limitations apply to our analysis specifically, while others will apply to all analysis using a similar method. We report both types in the following subsections.

Computational limits on the querying of large models

Throughout this project we have been continuously developing the software CausalQueries in order to make it fit for use with larger models that reflect the real theories underlying complex evaluations. Thanks to this work we have successfully encoded, plotted, updated and queried the DIG model and we have encoded, plotted, and updated the full UPAVAN model. As discussed above, we have only been able to query sub-DAGs reflecting simplified elements of the UPAVAN model rather than the full model. This is a result of computational limits that we have not yet overcome. However, we are very close to doing so and hope to have implemented the software development required as part of the CausalQueriesTools package that we have been developing and documenting so that anyone will be able to use CausalQueries for models as large as the UPAVAN model. The reason for the computational problem, the steps taken to address it and the solution we are currently finalising are all presented in Annex A, Section A2.

Some models are too wide

In contrast with the UPAVAN case, the DIG experience has revealed that there are some structures of intervention theory that cannot be analysed using the POInT approach. Theories which imply that very many different entities or processes will collectively cause changes in an outcome may not be suitable for investigation using the POInT approach given current conceptual and computational limits. As discussed above for the DIG case, this is because the POInT approach requires modelling precise effects on nodal types, and the number of nodal types for a given endogenous node quickly becomes very large, to an uncomputable level, if causation theorised at that node is too complex. In our experience, this limit for a binary node is found when there are more than five binary parents that interact. For such cases, regression may be a more appropriate tool, e.g. logistic models with maximum-likelihood estimation, even though this restricts analysis to aggregate causal effects and does not allow for the granular investigation of the balance of different types of effects like using a Bayesian causal model does.

Some theories require higher-valued nodes

As we have discussed, we have worked hard to extend CausalQueries beyond the strictly binary nodes for which it was designed, to accommodate ternary and higher-valued nodes. This work has been completed and is presented in Annex A, Section A1. However, we were not

able to complete this work in time to treat the ternary nodes in the UPAVAN model without coarsening them to binary. This is a limitation of the work presented here, though we intend not to coarsen those nodes to binary in future work which will be presented as academic papers and blog posts.

Incorporating multinomial nodes increases the number of nodal types in a model exponentially, just as increasing the number of parent nodes of any child does. This effect is more pronounced the more values the new node can take. Therefore, including ternary and quaternary nodes in any future application of this method will push on computational limits if there are too many such nodes. Incorporating quinary or higher-valued nodes may not be possible without significant breakthroughs in method and tool development. This means that the method as it currently stands is not suitable for analysing theories which include irreducibly many-valued nodes, and especially not theories which require the analysis of an irreducibly continuous variable. That said, in working with the three partner teams involved with this project, we have found confirmatory evidence for our prior expectation that most causal thinking does not in fact rely on many-valued concepts at the conceptual level. Using frequentist statistics we are used to dealing with continuous *measures* such as 'household consumption'; however, our theories are often about non-continuous *concepts* such as 'meeting basic needs.' We are confident that the scope of application of our methods is wider than it may seem to those who are used to working with continuous measures as a result of working with current statistical tools.

That said, we have only been able to complete the DAG elicitation process with three teams and it is possible that many or even most theories rely on many-valued causal concepts. It is also possible that the teams with whom we worked were particularly keen to accommodate the restrictions of our approach, knowing that it was experimental. Only further work to translate theories of change into causal models will reveal how widely acceptable the limitations of our approach are.

Limited ability to compare our results with traditional approaches

As part of this project, we had hoped to compare our results with those arising from the traditional approaches employed by the partner teams. However, for each partner trial this was frustrated by a different challenge to employing our full approach:

- We were unable to query the full UPAVAN model due to computational limits.
- We were unable to model the full DIG ToC due to the conceptual incompatibility of our approach with such a 'wide,' irreducibly interactive, theory. In addition, the DIG evaluation is ongoing.
- We were unable to progress beyond the DAG-elicitation stage with the AMETHIST team due to sensitivities around data sharing that took a long time to manifest themselves.

We hope in future work to complete our full approach for both the UPAVAN and AMETHIST partner teams and to present this work in journal articles and blogs. For the UPAVAN

partnership this will be possible very soon, as the work reported in Annex A, Section A2 is nearly complete. For AMETHIST, we are continuing our discussions with the team and we plan to reignite our partnership after their traditional analysis is published and thus sensitivities about data sharing have been overcome. Once the DIG evaluation is complete, we should be able to compare the results of our analysis of interaction between intervention components with the insights published as part of the traditional evaluation, but that evaluation is not yet complete.

For this report, the conditions have not yet been met for us to compare our results with the traditional evaluation results of the partner trials. We are therefore constrained in our ability to compare our approach with a traditional analysis. We have done this to some extent in reporting the queries on UPAVAN sub-DAGs in Part 5 of the Results Section, but it will only be in future work that these comparisons can be made well.

Divergence of priors and data

As discussed above, for many of the queries of interest, and as shown in the plots of distributions of nodal types, the expert priors elicited from the two partner teams often diverged markedly from the estimates generated on the trial data alone. In many cases the jointly updated model specification, on priors and data, generates very different estimates than the data only model. This is true even under the assumption of inclusive priors which gives the aggregated priors less weight overall than the assumption of exclusive priors. This is an interesting finding of applying the POInT process to the two partner trials, which we have discussed above. There are many possible explanations for this observation and these are worth exploring briefly. Some reflect limitations of our approach, while others reflect facts about the world. The weight given to each of these possible explanations in each case will determine what each divergence between priors and data tells us about expert beliefs, the causal processes under study, and our method.

Measurement error

Some of the difference between priors and data will be due to measurement error. Both in the case of expert priors and in the case of trial data, we have captured imperfect measures of latent objects. For example, when deciding how to measure the CLASSIC node in the DIG model, we were faced with a choice between probable false negatives and probable false positives depending on how we defined the node measure in terms of the trail measures. This is because there was a specific survey question in the endline that aimed to capture most of the classic graduation intervention components delivered by the DIG team but which didn't include all of the sorts of training that the intervention intended to deliver. Defining the node measure in terms of this survey question would in expectation cause false negatives for that node. On the other hand, including responses to another survey question asking about training but which did not limit the scope of the question only to training delivered by DIG would risk false positives. We believed that the latter choice was more defensible, but this may well be the cause of a divergence between experts' prior beliefs about this node and the picture of nodal types concerning this node that emerges from a data only model.

Similarly, some amount of measurement error is inevitable when capturing the prior beliefs of experts. In our second working paper from this project we discussed a confusion DIG experts had when filling in the prior elicitation arising from the fact that the first four contingency tables only had two possible parent node values. This caused participants mistakenly to allocate shares of 100 recipients to each row of the table such that the two responses summed to 100 recipients when this was not necessary; the two shares from 100 are independent of each other.

Conceptual error

A related source of divergence between priors and data may be related to the way in which node definitions were understood by experts compared to how they were operationalized for measurement. This is not quite the same as measurement error, rather it is a difference in understanding and/or communication between expert participants and POInT team members.

Selection bias

We selected our expert participants on the basis that they were the designers, implementers and evaluators of the partner intervention who were knowledgeable about that intervention. However, this is not the full population of experts with relevant knowledge, and our sample of experts is not random within that full population. This may introduce selection bias to expert priors. For example, our sample was restricted for both partner teams to those senior implementation partners who were permanently employed for the full duration of the project, and easiest to contact. We might expect junior implementation staff to have systematically different knowledge from their seniors and the fact that we could not include these people's priors is a weakness of our approach. Similarly, there will be many experts, academic and otherwise, from outside the partner teams with valuable and perhaps systematically different knowledge. Sampling from the full universe of these experts is not realistic (let alone randomly sampling,) however it might be possible for future projects such as ours to widen their sample of experts to increase precision and reduce potential for bias.

Most tellingly, potential program recipients and other stakeholders who are most deeply knowledgeable of local context do not feature in our samples of experts. Including these priors was outside of the scope of our project but could be a fruitful way of increasing precision and reducing bias.

Another form of selection bias may emerge from the DAG elicitation process. Our method for developing the DAG from the theory of change was discursive and open. The limitation of this is that there might have been divergent views that were difficult to air in the workshops. The UPAVAN team taking part in the workshops was diverse in the roles represented. We are aware that in any team there is the potential for power dynamics and personalities to shape the output of a shared process in ways that do not represent all views equally. We chaired these workshops with an awareness of this potential and will explore this potential limitation in future iterations of the DAG-development process with other evaluations.

It is worth noting that to the extent that selection bias in expert priors is a problem with our analysis, there is no reason to think that this is less true of the informal processes that are currently standard practice when integrating insights from process and impact evaluation and predicting effects in novel settings. Our process is transparent in this regard and can allow someone seeking to interpret results to assess the extent to which our results are reasonable. For example, in Part 3 of this section, dealing with the prior elicitation process, we were able to describe the within-sample variation for partner experts. This is useful information when considering the likely extent of variation between samples, albeit not sufficient on its own to underpin a belief on that matter.

Underestimation of inertia

A general pattern in examining the distributions of nodal types plotted in section 3.6 is that for both partner interventions the experts' priors underestimate the probability of units taking a nodal type of 'destined' or 'cursed,' relative to the pattern in the data. This is equivalent to underestimating inertia, over-estimating the probability of change. This could itself be due to a particular bias in our selection of experts; we selected experts in the design, implementation and evaluation of interventions designed to bring about changes. It is quite possible that a wider selection of experts, especially stakeholders in the local context would have corrected this bias by emphasising the difficulty of change.

Implementation problems

Implementation problems are a plausible explanation for a negative gap between expert priors and the results suggested by data. For both interventions, but especially the DIG intervention, some amount of the divergence between priors and data may be explained by such implementation problems. The DIG intervention was severely affected by the COVID-19 pandemic directly hampering implementation as well as leading to a diverting of funding away from some aspects of the programme. In addition, African swine fever killed many of the pigs that were intended as useful productive assets for recipient families.

Formalising the integration of qualitative data

This project has shown how to combine process and outcome data, using theory to produce integrated inferences that have been arrived at formally, transparently and reproducibly. However, we have limited ourselves to quantitative process data. Much of the data collected by modern process evaluations is qualitative. We have been forced to leave integrating this data into inferences as a problem for future research, as its integration poses complex problems.

To see why, we must be clear what we mean by 'qualitative.' Sometimes, 'qualitative' is used to refer to within-case data that is nonetheless expressed quantitatively. The 'qualitative' data in a typical process evaluation is not of this form – it is generally interview and focus group transcripts, as it was for the UPAVAN trial. A subset of this data can be reduced to quantitative data about the values of nodes in a DAG such as the UPAVAN DAG. ("In my household, we had a non-monsoon water supply, and we were able to grow diverse crops.") We can work with this data easily, reducing it to quantitative, within-case data. If we think this data is 'better'

than survey data – perhaps observed with less measurement error – then we could reflect this in a latent variable specification of the model where measurement error was modelled. However, some of the qualitative data in interview and focus group transcripts is not about one unit's experiences and is therefore not reducible to a set of node values; it's about the causal process. ("No household could grow diverse crops without a non-monsoon water supply.") This 'causal' qualitative data is both harder to interpret and potentially of higher probative value.

We examined the UPAVAN qualitative data to look for this sort of 'causal' qualitative data. Although some instances can be found in the UPAVAN transcripts, these are sparse. This is likely the case in most process evaluation transcripts, which are designed to elicit information about causation within the specific household rather than in the population.[10] Even where data about causation in the population is present, it is challenging to interpret using the POInT approach. We attempted to find ways of translating 'causal' qualitative data into quantitative values than can be interpreted by CausalQueries. The most promising method we have tried is to use the same conjoint questions as for prior elicitation. For a given respondent (one per HH interview or one of several per focus group discussion) we can impersonate the respondent and answer some subset of those conjoint questions (some number of rows in the contingency tables,) only those that we are justified in answering on the basis of the beliefs about causation expressed in the transcript. However, only the strongest statements about necessity and sufficiency can be interpreted well. ("Without a non-monsoon water supply, no household could benefit from the videos.") Often, statements are too ambiguous to justify any particular answer to rows in a contingency table, especially where multiple parent nodes affect a child, and the statement is only about a subset of the parents. ("It's difficult to benefit from the videos without a non-monsoon water supply.")

The most promising approach to using qualitative data with a POInT-style method is to set out from the start to differentiate 'causal' data from within-case node value data, collecting both. For the within-case node value data, traditional structured interviewing is appropriate. For the 'causal' data, we need a new data collection instrument. Using the same sort of contingency tables that we used to elicit expert beliefs to collect data is a promising way forward. This approach would treat respondents in the population as a different sort of 'expert.' However, the prior-elicitation exercise was challenging for our partner team experts. Those questions need to be made much more user-friendly to work for populations of interest for development interventions.

We can imagine a user-interface that is much more approachable than that which we used for prior elicitation and is used by enumerators working with potential recipients of an intervention and other local stakeholders. For instance, we could:

[10] Some process evaluations include qualitative enquiry about norms, which deliberately looks for information about causation at a population-level, or at least within 'communities of belief.' The data arising from this sort of process *is* likely to include a lot of 'causal' qualitative information.

- Focus more clearly on a group of fictitious units (households, families, businesses) and allocate them as shares of the group to outcomes
- Ask for graphical rather than numerical input (e.g. use sliders, shares of pies etc.)
- Take the conjoints one question at a time rather than as a whole contingency table (randomising order?)
- Elicit data on only a subset of causal relationships of interest

Such an approach is an exciting possibility for future research which might be able to make good use of local knowledge about causation in the context. This could be beneficial at the design and piloting stage, before an intervention design is finalised, allowing expert local knowledge to maximise the expected benefit of the intervention across relevant stakeholders.

Areas for future research

This project has required us to develop new methods and tools, laying a foundation for future research. One of our important audiences is the researchers we hope will build on this foundation. Here we list a few ideas for areas of research that might most fruitfully extend what we have started.

DAGs

- A project that focussed exclusively on developing DAGs from ToCs at an early stage in programme development over many programme teams could more systematically catalogue the advantages and challenges of this approach for such teams.

Prior elicitation

- The prior elicitation approach that we have developed allows for the elicitation of an order of magnitude more parameters than was previously possible. However, the approach should be subjected to testing on a wide variety of theories and with a wide variety of teams.
    - It would be fruitful, for example, to explore whether there are advantages (perhaps in reducing sources of bias) and disadvantages (almost certainly in terms of slower response time) to serving conjoint questions individually rather than serving respondents entire contingency tables at once.
    - Further, it could be explored whether serving conjoint questions or contingency tables in a random order has advantages and disadvantages compared to serving them in the order that they appear as causation flows through the DAG.
- The specific tool that we used for prior elicitation could be improved upon, ideally using the R package Shiny, taking just a CausalQueries model as an input and creating a user interface for prior elicitation automatically.
- As discussed above, the prior elicitation tool would have to be refined in order to be used with the recipient populations of development interventions and other local stakeholders, but we believe that doing this work could develop a useful new tool for eliciting the deep expertise of such stakeholders about causal processes in context.

CausalQueries

- We hope in future work to compare the results of querying causal models using CausalQueries with traditional frequentist statistical analysis for the DIG and UPAVAN interventions. However, much more research could be conducted to compare CausalQueries to traditional analysis methods both in theory and for real applications.
- We discuss work to extend CausalQueries to non-binary nodes Annex A, Section A1. This work could be further tested on more real data and for higher-valued nodes.

The POInT approach

- Applying the full POInT approach or an approach inspired by it to an evaluation of a project right from the earliest stages is another exciting avenue for future research. In particular, this could shed light on how useful, or not the DAG-elicitation process and query-framing could be for teams still in the development stage, which we were not able to investigate empirically.

Implications and recommendations for practice

From its inception, the POInT project was about bridging the gap between promising methodological innovations and the practical realities of evaluating real-world programmes. Although we have had to focus on methods development and have only studied a small number of applications, we nonetheless hope that there are some lessons for practitioners that can be drawn from our experience.

DAGs could lead to better design and better data collection

We have learned that developing explicit DAGs in participatory workshops with implementation designers, evaluators and implementers is possible and fruitful. We did not find that this took a lot of time, and it was possible for all participants to converge on an agreed DAG. The explicit causal process represented in a DAG helped to reveal assumptions, such as dependencies on contextual factors. Better awareness of these assumptions could lead to improved intervention design or targeting. By showing that the model can be updated with data from throughout the DAG we have demonstrated the potential to take a more holistic view of an intervention evaluation.

In the literature, translating theories of change into DAGs is argued to lead to various advantages. We have found support for some of these claims. The following subsections discuss in more detail.

Forcing a granular approach to causation identifies more relevant 'third variables'

For all of our partner teams, the DAG referred to causal processes more directly, less ambiguously, and at a finer level of detail than the theory of change that teams had been operating under. For example, for DIG, the original project ToC had a complex composite outcome referred to as 'socio-economic empowerment, and resilience.' Working through the DAG-elicitation process revealed that a key outcome for people with disabilities was being

economically active. For UPAVAN, it was understood that the marginalisation of women might undermine their ability to benefit from the programme. Working through the DAG-elicitation process revealed that, although this was true in general, the effect would be offset to some extent by the fact that women who were members of a scheduled tribe or caste were expected to be *more* likely to participate in decisions about their diet and about agricultural practices, a key mediating node.

Examples like these lend support to the idea that moving from a ToC or logic model to a DAG forces a move to more granular theory which is a better guide to evaluation. This facilitates thinking about crucial causal links (mediators) and the most important features of context (instruments, moderators and confounders.) For example, for UPAVAN, moving from the ToC to the DAG allowed the team to reflect their learning from the ongoing process evaluation that having a supportive household environment was a key moderator for the effectiveness of the intervention.

A useful guide to design and data collection

Uncovering all of the relevant 'third variables' as above can provide a guide to design and to data collection. We have learned that developing an explicit DAG from the theory of change can inform more efficient and complete data collection. For example, the DAG development process revealed that having a supportive household environment was likely important in the underlying causal process assumed for the UPAVAN intervention, but data on this variable was not collected during the evaluation. It is possible that had the DAG development process been completed prior to endline data being collected then data on this variable would have been collected. In general, the process of creating a DAG and defining measures for nodes, if completed early enough in the partner evaluation, will allow evaluation teams to collect data on factors that they now realise to be valuable but on which they may not have been intending to collect data. This might, for example, be a mediator, moderator or confounder which had not been specified during the creation of the original, somewhat ambiguous theory of change but was revealed by the process of creating an unambiguous DAG.

Conversely, we found that the vast majority of data collected in partner trial questionnaires were not related to nodes on the DAG. Since these data were not expected to be significantly causally related to the intervention (as mediators, moderators, or secondary outcomes) it raises the possibility that they could be removed without impacting the quality of the evaluation. This would reduce the burden on respondents to surveys. Reducing the time taken to collect data could also lead to more efficient deployment of field workers, for example by interviewing more participants. These findings suggest a role for DAG-elicitation as a way of helping teams specify more efficient data collection that could improve the power and cost-effectiveness of trials.

Being explicit about the role of prior belief

In addition to generating a DAG at an early stage, programme design and evaluation teams may benefit from engaging in a prior-elicitation exercise, ideally well before programme design is finalised. This could reveal an expectation about headline effects that implies a need

for design changes, for example that the programme's effects are not expected to be large enough to justify its costs, or that they might be too small to detect with the planned sample size, or that they are expected to be minimal or negative for a significant subgroup.

Other expectations might have more subtle implications for design and evaluation. For example, our experience with the DIG prior elicitation exercise revealed that a programme that had been designed as essentially interactive, with many components that are all said to be necessary was not expected by our group of experts to really have *interactive* effects. Rather, effects were expected to be additive, and there may have been some justification for designing the intervention as a cluster-randomised trial that was capable of learning more about the relative value of some intervention components compared to others.

Similarly, at the end of the project, it is valuable to be explicit about the expectations going in. For example, if it is suspected that a programme has faced implementation challenges such as one component not being delivered, it is helpful for transparency and credibility to have a prior model showing a strong expectation of the interactive rather than additive effects of that component. Similarly, if an unexpectedly large or small effect size emerges from the data, it is beneficial to have a principled way of deciding the extent to which this surprising effect should be a basis on which to design future projects or to scale up the intervention. If programme designers really were not sure how large the effect would be, this is a good argument for taking seriously the unexpectedly large or small finding. However, if designers were very sure, on the basis of their existing theory, that the effect should be as they expected it to be, then this is a principled reason for discounting the effect emerging from the data and more seriously investigating whether it might be due to measurement error, confounding, or some other cause.

Plotting posterior distributions

Despite an inability to query the full UPAVAN model due to computational limitations, much can be learned about causation in the intervention and about the different implications of different model specifications from the plots of posterior distributions found in Section 3.6. In particular, the overrepresentation of 'destined' or 'cursed' cases in the data for many nodes is a useful finding easily gleaned from such plots. Even more complex findings about the effects of particular combinations of parent node values can be gleaned from these plots with some practice in deciphering the labels in 1s and 0s applied to nodal types. These labels are a major weakness of such plots, being very time-consuming to decipher without a lot of practice. In addition, it is unclear how this labelling could be improved. Fundamentally, querying is a much more user-friendly and precise way of extracting data from posterior distributions of nodal types. That said, the posterior plots provided in section 3.6 are a useful 'eyeballing' tool that reveal the strongest effects and will reveal many data anomalies or modelling errors. As such, they should be part of a workflow using CausalQueries for causal modelling.

Implications and recommendations for policy

The implications of our project for policymakers follow from those for evaluation practitioners to a large extent. We believe that the considerations in the previous subsection imply that policymakers should make space for or even require some level of adoption of this approach from programme design and evaluation teams.

Create DAGs at an early stage

Most importantly, we think there would be advantages to encouraging ToCs to be expressed as DAGs with single-concept, granular nodes at an early project stage. We believe this would lead to:

- Better design choices
    - Identifying potential instrumental variables
    - Identifying the most causally significant third variables for matching or for stratification before randomisation
    - Identifying colliders that should not be controlled for
- More effective and efficient data collection
    - Making sure to collect data on key causal nodes
    - Not collecting unnecessary data on nodes that are not expected to be causally significant

We have also discussed some less concrete advantages of the DAG process that we observed working with our partner teams. Our partner teams found the DAG process to be beneficial as it facilitated and necessitated interesting, important and sometimes difficult conversations. This was for two main reasons:

- DAGs are unambiguous so teams are forced to have hard conversations
- DAGs are a simple graphical object that can bring different disciplines and specialisms together

It is possible that an organisation that adopted a standardised DAG notation for theories of change could build up an accretion of DAGs that could allow that organisation or a whole field of study to iteratively test key causal structures that appeared across multiple DAGs. These might be DAGs for similar sorts of interventions in different contexts or particular nodes that often feature connected to each other in the same ways even across different sorts of interventions. These structures might put some flesh on the bones of the idea of 'middle-range theory' which is much talked about in evaluation but it is unclear to most how it might be operationalised.

There is a danger to this strategy, which is that the strongest assumptions encoded in the DAG are those concerning absence of nodes or edges. If a DAG is adopted at an early stage and is not subject to review, then the following analysis may be undermined by being blind to an important factor. An open-ended qualitative enquiry in the scoping stage should help to create the best initial analysis DAG possible. However, even a DAG informed by this sort of

process should be open to challenge and change during the implementation process. Qualitative work during implementation is an ideal way of mapping unexpected mechanisms, mediators and moderators that may be active in the context and which should be included in a final analysis. Even if such factors are not identified early enough for quantitative data to be collected about them, their mere presence on the DAG would still aid in creating a more accurate posterior model than one from which they are absent.

Require explicit, quantitative priors

We have demonstrated a method for eliciting explicit, quantitative priors over a full DAG. We think that this process has many benefits for evaluation practitioners, as discussed above. We believe that it is in the interest of programme and evaluation commissioners to require programme and evaluation designers to engage in this exercise for several reasons. It may not always be possible or desirable (for larger models) to elicit these prior beliefs over the entire theory of change. However, it is always possible at a minimum to elicit priors over headline estimates of causal effects. We think requiring programme teams to elicit explicit, quantitative priors of this sort has several advantages for policymakers, including:

- Avoid commissioning programmes that are expected to have null or negative effects as a whole or for a significant subgroup, or whose effects are expected not to be cost-efficient or to be too small to detect at the proposed sample size.
- Better assess 'equipoise' for randomisation. (Can withholding the intervention be justified if we expect large positive effects? Can delivering the intervention be justified if we expect many null or negative effects (for some subgroups)?)
- Test areas of practitioner certainty that may never have been explicitly updated on data
- Focus learning on areas of uncertainty through design and data collection
- Iteratively improve on the tendency to systematically overestimate effects, and identify the conditions under which that overestimation is most and least pronounced.

Combine DAGs and prior elicitation with the most appropriate statistical estimation strategy

It is clear from our work that this approach is not suited to answering all questions over all theories. Some theories, like the full DIG ToC involve too many interacting components for our approach. Others will involve higher-valued or truly continuous nodes that CausalQueries cannot accommodate. For answering questions about these theories and others, it is more appropriate to use an alternative statistical estimation strategy such as regression analysis or structural equation modelling. That said, even in these cases we believe that the rest of our approach, involving unambiguous causal models and explicit, quantitative prior beliefs is appropriate. A very recent paper by Iacovone et al. (2023) at the World Bank combines Bayesian priors with linear regression analysis of intention-to-treat effects, for example. We support that approach where appropriate.

We believe our findings reveal the promise of a structural causal modelling approach like CausalQueries, which permits a fine-grained, granular examination of causation in smaller models with low-valued nodes. We have shown that the queries that can be answered by this approach allow for much more complex and specific questions to be answered than can be answered with traditional approaches that are limited to analysis causation in the aggregate.

Engaging with the language of queries at an early stage

Despite the project having the intention to fully query the whole UPAVAN intervention model, it was not successful in answering queries that depending on full models due to computational challenges. However, one of the reflections of our team and of the partner teams after going through the process of framing queries and asking them of the model is that the language of queries forces evaluators to be precise. This is somewhat analogous to a benefit arising from the DAG development process: that DAGs are unambiguous about causal relationships and so evaluators are forced to be, too. The language of queries, if engaged in at an early stage of a research process, could be a useful way of fostering conversations between evaluators and putting flesh on the bones of a 'questions-driven' approach to evaluation. Moreover with a sharper identification of those queries that go average treatment effects, such as mediation queries or attribution queries, model elicitation could focus on simpler models that are complex enough to address these queries, but not more complex.

Bibliography

Aarons, G. A., Fettes, D. L., Sommerfeld, D. H., & Palinkas, L. A. (2012). Mixed methods for implementation research: Application to evidence-based practice implementation and staff turnover in community-based organizations providing child welfare services. *Child Maltreatment*, *17*(1), 67–79.

Angrist, J. D., & Krueger, A. B. (2001). Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives*, *15*(4), 69–85. https://doi.org/10.1257/jep.15.4.69

Arif, S., & MacNeil, M. A. (2022). Applying the structural causal model framework for observational causal inference in ecology. *Ecological Monographs*. https://doi.org/10.1002/ecm.1554

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Pariente, W., Shapiro, J., Thuysbaert, B., & Udry, C. (2015). A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*, *348*(6236), 1260799–1260799. https://doi.org/10.1126/science.1260799

Bennett, A., & Checkel, J. T. (2014). *Process tracing: From metaphor to analytic tool*. Cambridge University Press.

Breuer, E., Lee, L., De Silva, M., & Lund, C. (2015). Using theory of change to design and evaluate public health interventions: A systematic review. *Implementation Science*, *11*(1), 1–17.

Brito, C., & Pearl, J. (2012). *Generalized Instrumental Variables* (arXiv:1301.0560). arXiv. https://doi.org/10.48550/arXiv.1301.0560

Cartwright, N. (2019). Commentary: Why mixed methods are necessary for evaluating any policy. In M. Nagatsu & A. Ruzzene (Eds.), *Contemporary Philosophy and Social Science: An Interdisciplinary Dialogue* (pp. 173–184). Bloomsbury.

Creswell, J. W., Klassen, A. C., Plano Clark, V. L., & Smith, K. C. (2011). Best practices for mixed methods research in the health sciences. *Bethesda (Maryland): National Institutes of Health*, *2013*, 541–545.

Davey, C. H., Bonell, C., Cartwright, N., Hargreaves, J., Hassan, S., & Humphreys, M. (2017). Gaps in evaluation methods for addressing challenging contexts in development. *CEDIL Pre-Inception Paper*.

Davey, C., Hargreaves, J., Hassan, S., Cartwright, N., Humphreys, M., Masset, E., Prost, A., Gough, D., Oliver, S., & Bonell, C. (2018). Designing evaluations to provide evidence to inform action in new settings. *CEDIL Inception Paper*, *2*.

Davies, R. (2018). Representing theories of change: Technical challenges with evaluation consequences. *Journal of Development Effectiveness*, *10*(4), 438–461. https://doi.org/10.1080/19439342.2018.1526202

Dawid, A. P. (2010). Beware of the DAG! *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008*, 59–86. https://proceedings.mlr.press/v6/dawid10a.html

De Silva, M. J., Breuer, E., Lee, L., Asher, L., Chowdhary, N., Lund, C., & Patel, V. (2014). Theory of change: A theory-driven approach to enhance the Medical Research Council's framework for complex interventions. *Trials*, *15*(1), 1–13.

Errington, R. D., Ashby, D., Gore, S. M., Abrams, K. R., Myint, S., Bonnett, D. E., Blake, S. W., & Saxton, T. E. (1991). High energy neutron treatment for pelvic cancers: Study stopped because of increased mortality. *British Medical Journal*, *302*(6784), 1045–1051.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, *58*(1), 21–29.

Freedman, L. S., & Spiegelhalter, D. (1983). The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *32*(1–2), 153–160.

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. J. (2010). *Impact Evaluation in Practice*. The World Bank. https://doi.org/10.1596/978-0-8213-8541-8

Giddens, A. (1984). *The constitution of society: Outline of the theory of structuration*. Univ of California Press.

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 37–48.

Hughes, M. D. (1991). Practical reporting of Bayesian analyses of clinical trials. *Drug Information Journal*, *25*(3), 381–393.

Iacovone, L., Mckenzie, D., & Meager, R. (2023). *Bayesian Impact Evaluation with Informative Priors: An Application to a Colombian Management and Export Improvement Program* (Text/HTML WPS 1074; Impact Evaluation Series). World Bank. http://documents.worldbank.org/curated/en/099807301092338354/IDU08080a50008e3c047f40a8de0e4e6b95c2d2c

Jimenez, E., Waddington, H., Goel, N., Prost, A., Pullin, A., White, H., Lahiri, S., & Narain, A. (2018). Mixing and matching: Using qualitative methods to improve quantitative impact evaluations (IEs) and systematic reviews (SRs) of development outcomes. *Journal of Development Effectiveness*, *10*(4), 400–421.

Johnson, S. R., Tomlinson, G. A., Hawker, G. A., Granton, J. T., & Feldman, B. M. (2010). Methods to elicit beliefs for Bayesian priors: A systematic review. *Journal of Clinical Epidemiology*, *63*(4), 355–369. https://doi.org/10.1016/j.jclinepi.2009.06.003

Kadiyala, S., Harris-Fry, H., Pradhan, R., Mohanty, S., Padhan, S., Rath, S., James, P., Fivian, E., Koniz-Booher, P., Nair, N., Haghparast-Bidgoli, H., Mishra, N. K., Rath, S., Beaumont, E., Danton, H., Krishnan, S., Parida, M., O'Hearn, M., Kumar, A., … Allen, E. (2021). Effect of nutrition-sensitive agriculture interventions with participatory videos and women's group meetings on maternal and child nutritional outcomes in rural Odisha, India (UPAVAN trial): A four-arm, observer-blind, cluster-randomised controlled trial. *The Lancet Planetary Health*, *5*(5), e263–e276. https://doi.org/10.1016/S2542-5196(21)00001-2

Kadiyala, S., Prost, A., Harris-Fry, H., O'Hearn, M., Pradhan, R., Pradhan, S., Mishra, N. K., Rath, S., Nair, N., Rath, S., Tripathy, P., Krishnan, S., Koniz-Booher, P., Danton, H., Elbourne, D., Sturgess, J., Beaumont, E., Haghparast-Bidgoli, H., Skordis-Worrall, J., … Allen, E. (2018). Upscaling Participatory Action and Videos for Agriculture and Nutrition (UPAVAN) trial comparing three variants of a nutrition-sensitive agricultural extension intervention to improve maternal and child nutritional outcomes in rural Odisha, India: Study protocol for a cluster randomised controlled trial. *Trials*, *19*(1), 176. https://doi.org/10.1186/s13063-018-2521-y

Leviton, L. C. (2017). Generalizing about public health interventions: A mixed-methods approach to external validity. *Annual Review of Public Health*, *38*, 371–391.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., & Imbens, G. (2014). Promoting transparency in social science research. *Science*, *343*(6166), 30–31.

Moore, G. F., Audrey, S., Barker, M., Bond, L., Bonell, C., Hardeman, W., Moore, L., O'Cathain, A., Tinati, T., Wight, D., & Baird, J. (2015). Process evaluation of complex interventions: Medical Research Council guidance. *BMJ*, *350*(mar19 6), h1258–h1258. https://doi.org/10.1136/bmj.h1258

Pawson, R. (2000). Middle-range realism. *European Journal of Sociology / Archives Européennes de Sociologie*, *41*(2), 283–325. https://doi.org/10.1017/S0003975600007050

Pearl, J., & Bareinboim, E. (2014). External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science*, *29*(4), 579–595. https://doi.org/10.1214/14-STS486

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Polit, D. F., & Beck, C. T. (2010). Generalization in quantitative and qualitative research: Myths and strategies. *International Journal of Nursing Studies*, *47*(11), 1451–1458.

Prost, A., Harris-Fry, H., Mohanty, S., Parida, M., Krishnan, S., Fivian, E., Rath, S., Nair, N., Mishra, N. K., Padhan, S., Pradhan, R., Sahu, S., Skordis, J., Danton, H., Koniz-Booher, P., Beaumont, E., James, P., Allen, E., Elbourne, D., & Kadiyala, S. (2022). Understanding the effects of nutrition-sensitive agriculture interventions with participatory videos and women's group meetings on maternal and child nutrition in rural Odisha, India: A mixed-methods process evaluation. *Maternal & Child Nutrition*, e13398. https://doi.org/10.1111/mcn.13398

Rohrer, J. M. (2018). Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Spirtes, P., Glymour, C. N., Scheines, R., & Heckerman, D. (2000). *Causation, prediction, and search*. MIT press.

Tennant, P. W., Harrison, W. J., Murray, E. J., Arnold, K. F., Berrie, L., Fox, M. P., Gadd, S. C., Keeble, C., Ranker, L. R., Textor, J., Tomova, G. D., Gilthorpe, M. S., & Ellison, G. T. (2019). *Use of directed acyclic graphs (DAGs) in applied health research: Review and recommendations* (p. 2019.12.20.19015511). medRxiv. https://doi.org/10.1101/2019.12.20.19015511

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*(3), 219–242.

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*(12), 1049–1064.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–67.

Weiss, C. H. (1972). *Evaluation Research: Methods for Assessing Program Effectiveness*. Prentice-Hall.

Williams, T. C., Bach, C. C., Matthiesen, N. B., Henriksen, T. B., & Gagliardi, L. (2018). Directed acyclic graphs: A tool for causal studies in paediatrics. *Pediatric Research*, *84*(4), Article 4. https://doi.org/10.1038/s41390-018-0071-3

Annex A: Attachments

A1: Causal inference from causal models with non-binary data

Abstract

In our UPAVAN application, we have a set of ternary nodes, nodes taking on three possible values. Less attention has been paid to causal queries over non binary nodes like this. We describe a strategy for Bayesian estimation of counterfactual causal quantities with multinomial data. The strategy involves representing causal relations between multinomial nodes on suitably constructed twin models defined over binary nodes. Updating and querying can then be done using existing tools. We illustrate the approach using data on government trust and attitudes to Covid-19 policies.

The full paper can be found at this link: ***Causal Inference from causal models with non-binary data***

A2: Computational improvements to CausalQueries with benchmarks

Abstract

We encountered some computational challenges to applying the POInT approach using CausalQueries. For example, the querying of large models was one of the computational challenges encountered. Constructing a double matrix of causal type distributions with the standard 4000 draws consumes 64000 Gigabytes (64 Terabytes) of memory (8 bytes per double x 4000 x 2x10^9). These estimates represent lower bounds on memory usage ignoring pointers, metadata, attributes and memory consumed by operations on the data structures. It is clear that querying even simplified causal models of a complexity that is sufficiently representative of practitioners' beliefs about theories of change with out of the box CausalQueries methods far outstrips RAM and disk space of both conventional consumer hardware and even some high-performance computers. Much of our work in adapting CausalQueries for application to evaluations in the POINT project has been to overcome such computational challenges. Several computational changes were made to mitigate these challenges.

A summary paper of the computational changes made with benchmarks for performance improvements can be found at this link: ***Computational improvements to CausalQueries with benchmarks***

A3: Alternative Strategies for Using Prior Elicitation

Abstract

This paper presents alternative strategies for using prior elicitation that can be applied in different contexts, It proposes that we can infer Dirichlet (α) parameters from our measures of conditional probabilities and our measure of uncertainty either by:

1. solving explicitly whenever restrictions imply that we have as many parameters as data points (e.g. when there is one parent, or two parents monotonicity, and no complementarity).
2. assuming independence of potential outcomes. This is easy to apply and produces a result consistent with the data and "central" among feasible alternatives. It is not obvious yet how to apply when there are restrictions.
3. using CausalQueries: we can treat our problem as a problem about learning about the subject's *alphas* given data they provide. It turns out that the problem is formally similar to learning about λ.

Approach 3 seems in general most flexible and should get the same result as approach 1 whenever approach 1 is possible; and results similar to 2 but with an assumption of flat priors (for us, with respect to subject's priors) rather than an assumption of independence.

The full memo can be found at this link: ***Parameters from priors***

www.CEDILprogramme.org