Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn

Trang Quynh Nguyen, Ian Schmid, Elizabeth A. Stuart

Context. Mediation analysis is increasingly popular. It has become more accessible, with the availability of various computing tools. It is encouraged, or even required, by some research funding agencies.

Most mediation analyses use methods based on a definition of the *indirect* effect as the product of two coefficients from two relatively simple regression models (Baron & Kenny, 1986). This *model-dependent* effect definition has a causal interpretation -- as a path-specific effect of the exposure on the outcome -- only in the special case where linear models are used and the two models correctly reflect the true causal models. The recent incorporation of *causal inference* in mediation analysis has led to *model-free* effect definitions with causal interpretation, which generalize the traditional definition and remove parametric assumptions. Based on these effect definitions, the corresponding identifying conditions are then clarified, and estimation methods developed.

A challenge for a substantive researcher interested in using causal mediation analysis is that the relevant methodological literature is fast-growing and complex, which may be confusing if the researcher is unfamiliar with causal inference or unfamiliar with mediation. There are not one, but several types, of causal effects that have been defined for the mediation setting. The researcher needs to choose which of these to target, and judge whether they are identified given the study design and data, before attempting estimation.

Purpose. The goal of this and a sibling paper is to help ease the understanding and adoption of causal mediation analysis. This specific essay aims to convince the researcher of the causal inference approach, and to help the researcher select the target causal effect(s) that best match their research question.

Part 1: The need for explicit causal thinking in mediation analysis. After explaining the above-mentioned key difference between the traditional and causal inference approach, we argue for the need for explicit causal thinking in mediation analysis. We put forth two main arguments. First, unlike analysis in a non-mediation setting (which may target conditional associations rather than causal effects), mediation analysis is unavoidably about causal effects. Its conceptualization often involves drawing arrows that represent the influences of variables on one another, and its results are generally interpreted by the research consumer in causal terms. Second, causal effects are much less intuitively clear in the mediation setting compared to the non-mediation setting, hence it is beneficial to adopt a formal framework for reasoning about them. Specifically, the notion of the effect of an exposure on an outcome suggests comparing the outcome under exposure and under nonexposure, both observable conditions. But path-specific effects in the mediation setting cannot be mapped to observable conditions, and thus are more abstract. It is therefore beneficial to adopt a formal framework for logical reasoning about such effects.

Part 2: Defining the target causal effect(s) based on the research question. The bulk of the paper explains in as-plain-as-possible language existing effect types, paying special attention to motivating these effects with different types of research questions, and using concrete examples for illustration.

Starting with the total effect, we introduce the potential outcome framework, defining the total effect (i) for the individual, as the difference between the two potential outcomes under exposure and nonexposure; and (ii) for the inference population, as the difference between the means of these potential outcomes. A toy example of a college readiness intervention makes concrete things concrete. We also introduce the causal directed acyclic graph (DAG), a tool that captures causal relationships as they are conceptualized, which we use throughout the paper to visually represent effect definitions.

Turning to effects that are more of interest in the mediation setting, the paper differentiates two perspectives (or purposes of analysis): the *explanatory* perspective (aiming to explain the total effect) and the *interventional* perspective (asking questions about hypothetical interventions on the exposure and mediator, or hypothetically modified exposures). These correspond to two classes of causal effects.

Using the same toy example and zooming in on an individual, we present the *natural (in)direct effects* (Robins & Greenland, 1992; Pearl, 2001), which *explain* (or more precisely, decompose) the total effect. As the total effect contrasts two exposure conditions, decomposition involves imagining an in-between condition – where exposure is set to one condition, but mediator is set to the value it would take under the other exposure condition. We provide two heuristics to aid understanding of these effects, using the metaphors of information flows and double exposure. Addressing the fact that there are two different decompositions of the total effect, we explain what this means, and discuss how the researcher might choose between the two or choose both – via being more precise about the research question.

Next, we introduce *interventional effects*, a class of effects that corresponds to *what if* questions about hypothetical interventions on the exposure and/or mediator – setting them to specific values or distributions. This class is relevant to a range of research interests. One relevant theme in intervention research is imagination of modifications to the current intervention (the college prep program), for example retaining only the direct or an indirect effect mechanism. Another theme is consideration of a new intervention that may impact on the mediator, with examples from disparities research treating bullying experience as mediator. Another application concerns the effectiveness of an existing intervention in a new context where the mediator is externally affected. This class of effects contains as special cases many known effects, including *interventional (in)direct effects* (Didelez, Dawid, Geneletti, 2006; VanderWeele, Vansteelandt & Robins, 2014), *controlled direct effects* (and also the total effect), but also offers many more possibilities. Noting that the (in)direct effects named in the literature are restrictive given the range of real-world research questions, we argue that when approach a mediation analysis with a *what if* question, the researcher should leverage this general class. It allows the researcher to flexibly define target causal effect(s) to best match the research question.

Closing remarks. The paper ends with remarks that orient the researcher to the literature on the next steps: identification and estimation of the effects they choose to target.